Deep Learning Based Cyanobacterial Promoter Strength Prediction

Yun-young Choi, Jong-hwan Yoon, Ah-Ram Kim*

School of Life Science, Handong Global University, Pohang 37554, South Korea

Abstract

Cyanobacteria are a phylum of autotrophic gram-negative bacteria that can obtain biological energy through oxygen photosynthesis. Through various genetic engineering, cyanobacteria with special functions can be created. When performing genetic engineering, it is essential to have a promoter that can express the desired gene. In this study, we try to predict the gene expression level as a strength of a promoter when introducing a heterologous protein. Since promoters have different characteristics of each species and their consensus sequences are also different, we specifically used *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942 among various cyanobacteria. Using computational analysis, we found the difference between *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942 promoters. Additionally, we estimated the promoter strength through mRNA expression level using dRNA-seq data, and predicted the strength of new promoter sequences using a deep learning-based CNN model. This study will provide the foundation for more in-depth analysis in the future by building a model pipeline for promoter analysis.

Introduction

Promoter is a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript. Prokaryotes lack many enhancers or transcription factors, so promoters have an absolute impact on gene expression. In addition, bacterial promoters have simpler consensus sequences compared to eukaryotic promoters. They typically consists of two main regions: TATAAT at -10 region, and TTGACG at -35 region. With various facts about bacterial promoters, we can use deep learning methods to create a new promoter sequence and predict its strength.

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Especially, convolutional neural network (CNN) models were developed and used to predict the strength of a promoter based on its nucleotide sequence. In order to train CNN model, train data and its labels are needed. Promoter sequence was used as train data, which were assumed upstream -100 bp from the TSS (transcription start site) of each gene. Also, number of normalized reads in dRNA-seq was used as the promoter strength. If the prepared promoter sequences pass through the feature extraction and regression layer in the CNN model, the unique pattern from the promoter sequences can be extracted. Then it would be able to predict the promoter strength accordingly.

Results

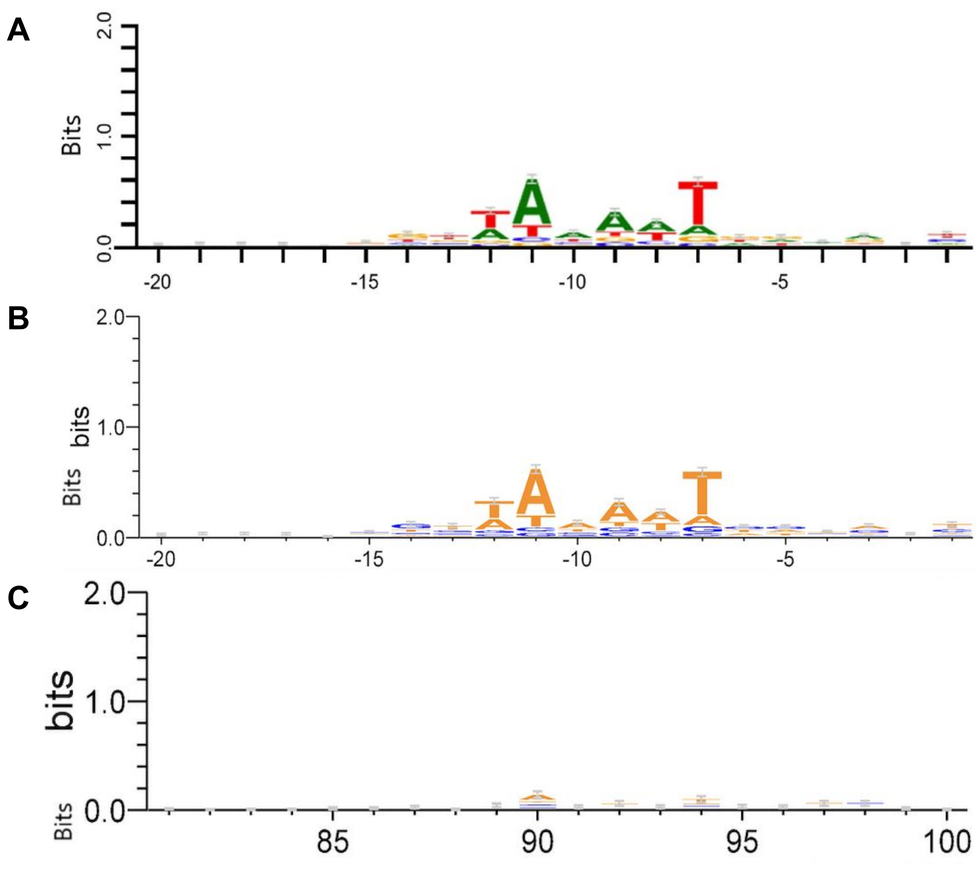

Figure 1. Position Weight Matrix(PWM) results of *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942

(A) *Synechocystis sp.* PCC6803 promoter alignment from reference article (Euijin Seo et al (2023)) (B) Reproduced result of *Synechocystis sp.* PCC6803 promoter alignment (C) *Synechococcus elongatus* PCC7942 promoter alignment.

Through figure A and B, the visualization tool was validated, and it can be seen that TANNNT is a conserved region in the PCC6803 promoter. However, we can clearly see that the promoter upstream -20 bp of PCC7942 is weakly conserved than that of PCC6803.

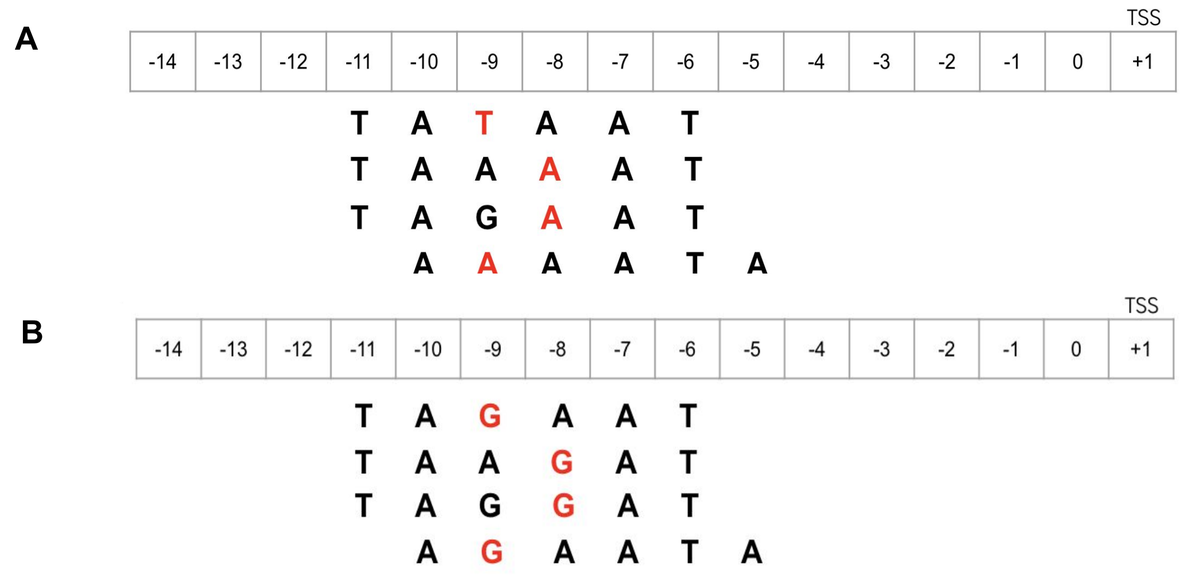

Figure 2. Comparison of conserved sequences of *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942

(A) Conserved sequences of *Synechocystis sp.* PCC6803. (B) Conserved sequences of *Synechococcus elongatus* PCC7942.

Figure 2 is the result of visualizing the top four 6-mers with the highest frequency in each of PCC6803 and PCC7942. In the case of PCC6803, it can be seen that T and A are the main components of the promoter sequence. On the other hand, in the case of PCC7942, T and A are also dominated, but also G is characteristically distributed unlike PCC6803. The nucleotide marked as red is the part where the substitution occurred with G when comparing the promoter sequences of PCC6803 and PCC7942.

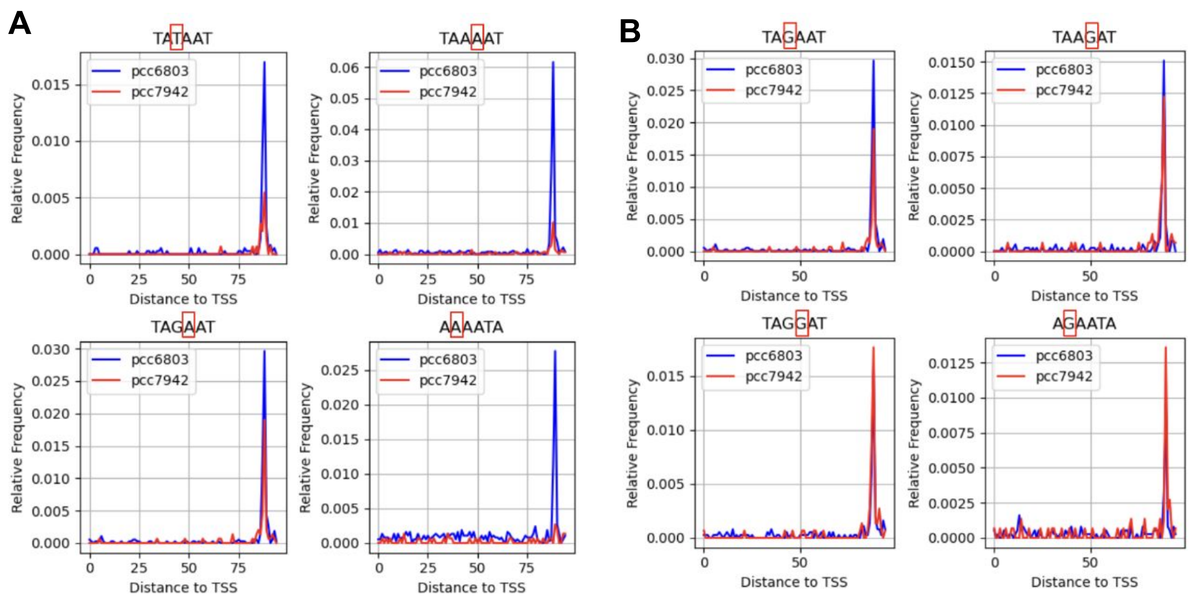

Figure 3. Comparison of 6-mer frequency analysis of *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942

(A) Relative frequency of *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942 with original 6-mers from PCC6803 (B) Relative frequency of *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942 with adjusted 6-mers from PCC7942.

When comparing the relative frequency with original 6-mers, it can be seen that PCC7942 has a much smaller frequency than PCC6803. On the other hand, when comparing the relative frequency with adjusted 6-mers, it can be seen that PCC7942 has a significantly higher frequency. This means that G, which frequently occurs in PCC7942, can affect the consensus sequence. Therefore, based on this result, we can conclude that PCC6803 and PCC7942 will obtain better model performance when the consensus sequence is differently applied to each organisms.

Conclusion

Through this study, the difference between *Synechocystis sp.* PCC6803 and *Synechococcus elongatus* PCC7942 promoters could be clearly understood and applied differently. In addition, by comparing the results of the reference paper and the reproduced results, the code used to create the results could be validated . Based on the established data from this study, future works would be improving the performance of CNN models to predict the expression level of promoters. These studies will serve as a good prediction tool before expressing new proteins to Cyanobacteria and can also be proven experimentally.

References

1. Seo, E., Choi, Y. N., Shin, Y. R., Kim, D., & Lee, J. W. (2023). Design of synthetic promoters for cyanobacteria with generative deep-learning model. Nucleic Acids Research, 51(13), 7071-7082.
2. Vijayan, V., Jain, I. H., & O'Shea, E. K. (2011). A high resolution map of a cyanobacterial transcriptome. Genome biology, 12, 1-18.

*The code used to create the results of this study is uploaded in this link: https://github.com/foryourjoy/bio-ai-capstone