Learning genomic synteny contributions to gene expression levels using transformers

Lucas Waldburger & Emaad Khwaja
Professor Jennifer Listgarten
Spring 2022

Transcriptional neighborhoods have been shown to contribute an effect on gene expression level [1]. Convolutional neural networks (CNN) were recently applied to predict gene expression from a promoter sequence [2]. However, the gene expression activity at the proximal ends of a given gene is believed to be more predictive of gene expression than the promoter sequence.

In this proposal, we replace the Gradient Boosted Regression Trees (GBRT) from Brooks et al. with a transformer to learn the contributions of transcriptional neighborhoods to gene expression level [3]. We expect that a transformer can better integrate long range interactions than GBRT, CNN, or other methods. Furthermore, gene expression occurs along a linear 1D axis, e.g. the DNA transcript, however genes both upstream and downstream of the current translational position can have regulatory effects which alter the relative abundance of protein produced. As such, we intend to utilize a bi-directional transformer architecture, similar to BERT [4].

First, we will reproduce the GBRT results from the paper then test a transformer architecture. Second, we will apply similar evaluation metrics such as mean-squared error loss on the predicted expression level of the genes of interest. Third, we intend to do thorough visualization of the activation maps of the attention heads to contribute towards explainability of the prediction and interdependence of gene expression from the entire transcript. Lastly, we also intend to take this one step further by reverse engineering the model to generate transcripts which maximize expression of particular genes within pre-selected boundary conditions.

In theory, complementary wet-lab experiments could be performed to validate these sequences. This work provides a step towards predictive modeling of complex biological systems for genome design.

References

[1] Aaron N. Brooks, Amanda L. Hughes, Sandra Clauder-Münster, Leslie A. Mitchell, Jef D. Boeke, and Lars M. Steinmetz. Transcriptional neighborhoods regulate transcript isoform lengths and expression levels. *Science*, 375(6584):1000–1005, 2022.

[2] Eeshit Dhaval Vaishnav, Carl G. de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A. Thompson, Joshua Z. Levin, Francisco A. Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901):455–463, March 2022.

[3] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.