Title: Exploring Transcription Factor Binding Sequence and Orientation Patterns

Author(s): Timothy James Becker, Dashzeveg Bayarsaihan

Affiliation(s): Connecticut College, University of Connecticut Health

Abstract: Transcription Factors (TF) are important DNA binding proteins that act as master transcriptional regulators in humans. They play a pivotal role in developmental processes where multiple genes are regulated simultaneously such as in palate formation. TFs perform this function through a primary DNA binding domain that attaches to regulatory features like enhancers or promoters along with a secondary protein-to-protein domain forming complexes that facilitate or repress transcriptional activity. The rules that govern TF binding combinations are not widely known other than limited *in vivo* studies that have focused on combinations of two or three TFs. Using TF binding motifs from the JASPAR database, we investigate the latent information content in TF sequences across genes, regulatory features and intergenic background regions of arbitrary size. Our novel framework generates positional TF sequences using histone marks (ChIP-seq bed data) and RefSeq annotations to train an encoder-only transformer language model that uses genomic features (and or cell types) as classes. We formulate specialized balancing, metrics and losses to maximize accuracy across user specified search labels thereby enabling exploration of the sequence patterns. Using these pretrained-models, the user can then purify data (filtering out incorrectly predicted unseen data) to then visualize and cluster (by TF sequence-similarity metrics) the results.

Keywords: Human Regulatory Genomics, Transcription Factor Binding Sequences, Machine Learning, Latent Pattern Discovery

Author Profiles (s):

https://scholar.google.com/citations?user=6hv4-c4AAAAJ

https://scholar.google.com/citations?user=19rhQLwAAAAJ&hl=en&oi=ao