COMP 561 – Final Project Proposal

Predicting TF Binding with DNA Shape Features

Claire Yang [ID: 260898597]

The main question of this project how DNA shape features such as minor groove width, roll, propeller twist, and helical twist help to predict transcription factor (TF) binding. Our goal is to test if these structure features can separate bound and unbound sites.

Input data: (i)Human genome (hg19); (ii)GM12878 active regulatory regions; (iii)Factorbook TF binding sites and PWMs; (iv)DNA shape data from DNAshapeR or precomputed shape tracks.

Pipeline:

1. Pick one TF with enough ChIP-seq peaks (e.g., CTCF).
2. Extract positive sites (binding peaks) and negative sites (non-binding regions in active chromatin).
3. Generate DNA sequences for all sites (e.g., ±50 bp around motif).
4. Compute one-hot sequence matrix.
5. Get DNA shape values using DNAshapeR or read from shape files.
6. Merge sequence and shape features, split by chromosomes into train/valid/test.

Timeline

Phase 0: Train logistic regression and SVM using DNA shape features only.

Phase 1:

Build a simple CNN for sequence input with One-hot encode DNA sequences.

Concatenate CNN features with shape features → MLP → output probability.

Optional: Try cross-attention fusion between sequence

Phase 2: Evaluation:

Compare AUROC, AUPRC between all models. Test statistical significance with bootstrap. Use SHAP or Integrated Gradients to show which shape features matter most.

Metrics: (i) AUROC, AUPRC, F1 score; (ii) Training time and parameter count; (iii)Feature importance ranking (for interpretability)

Expected: Phase 1 model showing added value from combining sequence and shape

References

- Rohs D et al. (Nature, 2009).https://www.nature.com/articles/nature08473
- Zhou T et al. (PNAS, 2015). https://www.pnas.org/doi/10.1073/pnas.1422023112
- Chiu TP et al. (Bioinformatics, 2016). https://pubmed.ncbi.nlm.nih.gov/26668005/
- Kabir A et al. (NAR Genomics and Bioinformatics, 2024). https://pmc.ncbi.nlm.nih.gov/articles/PMC10827174/