COMPSCI 602
Project Report 4

Anonymous Author

Project Title: Beyond Black Boxes: Interpretability-Focused ML for Operon Prediction

1 Introduction

The system in this project consists of nonlinear simple ML models outlined in OperonSEQer [1] (Logistic Regression, Gaussian Naive Bayes, Random Forest, XGBoost, Multilayer Perceptron). The task they are being evaluated on is that of operon prediction (if two genes are transcribed together, they form an operon; otherwise they do not). The environment in which the study is performed is the DGEB [5] E.Coli dataset augmented with genomic data from NCBI [4] and biologically motivated features [2]: intergenic distances, strand concordance, orientation pattern, GC-content difference, and functional category match computed from the genomic data, as well as promoter and terminator motifs obtained from RegulonDB [3].

The phenomena I would like to study are the performance of simple ML models on the operon prediction task as well as the relative importance of each feature on the prediction output. I will measure this primarily through the area under the ROC curve (ROC-AUC) on the test set, but also the precision, recall, and F1 score of the models (baseline, fine-tuned, as well as each ablation study instance). The primary methodology will be ablation studies, where individual features or feature groups are removed to assess their contribution to model performance.

2 Research Question 1: Model Learning Mechanisms

Research Question: How do these models learn operon prediction?

Hypotheses:

- **H1.1:** There are features that carry more signal than others, identifiable through ablation studies, which reveal the fundamental mechanisms that models use for operon prediction. Features will show significant variation in their contribution to model performance when removed.
- **H1.2:** Regulatory motif features (promoters and terminators) carry the most signal for operon prediction, as suggested by prior work, and their removal will result in the largest performance degradation compared to other feature groups.

Testing Methods: Both hypotheses will be tested through systematic ablation studies. Each feature or feature group (regulatory motifs, genomic distances, strand information, sequence features, functional categories) will be removed individually, and the resulting change in ROC-AUC will be measured. Features will be ranked by their contribution, with the largest performance drop indicating the highest signal. H1.2 will be specifically evaluated by comparing the performance degradation when regulatory motifs are ablated versus when other feature groups are ablated.

3 Research Question 2: Non-Linear Model Performance

Research Question: Why do non-linear models perform better than linear models?

Hypotheses:

- **H2.1:** There is a fundamental difference in how linear and non-linear models learn from the data, with non-linear models capturing patterns that linear models cannot represent. This difference will be evident in divergent feature importance rankings across model types.
- **H2.2:** Non-linear relationships exist between features that best-performing models exploit, detectable through ablation studies with polynomial and interaction terms added to linear models, or through differential feature importance across model types.

Testing Methods: H2.1 will be tested by comparing feature importance rankings derived from ablation studies across Logistic Regression, Random Forest, XGBoost, and MLP models. Low correlation between rankings will indicate different learning mechanisms. H2.2 will be tested through three approaches: (1) augmenting linear models with polynomial features (e.g., distance^2, GC-content^2) and measuring performance improvement, (2) adding interaction terms (e.g., distance × GC-content) to linear models, and (3) identifying features whose ablation disproportionately affects non-linear models compared to linear models, indicating non-linear feature effects.

4 Research Question 3: Ensemble Model Performance

Research Question: Why do ensemble models perform better than individual models?

Hypotheses:

- **H3.1:** There are differences in feature weightages between Multi-Layer Perceptrons (MLPs) and tree-based models (XGBoost/Random Forest), allowing them to capture complementary information. This will be demonstrated through divergent feature importance rankings from ablation studies.
- **H3.2:** An ensemble of MLPs will outperform XGBoost/Random Forest ensembles on the test set (measured by ROC-AUC improvement of at least 2
- **H3.3:** The computational cost of ensembling MLPs (measured in training time and peak memory usage) will be significantly larger than tree-based ensembles. The performance gain (if any) will be quantified relative to this computational cost to determine if the tradeoff is worthwhile.
- **H3.4:** Heterogeneous ensembling strategies (mixing linear and non-linear models) will outperform homogeneous strategies (all linear or all non-linear models) due to increased diversity in learned representations.

Testing Methods: H3.1 will be tested by performing ablation studies on both MLP ensembles and tree-based ensembles, comparing which feature removals most impact each ensemble type. H3.2 will be evaluated by training ensembles of multiple MLPs (with different random initializations) and comparing their ROC-AUC, precision, recall, and F1 scores to XGBoost/Random Forest ensembles on the test set. H3.3 will be measured by recording training time and peak memory usage during ensemble training and inference using Python's built-in timing functions and memory profiling tools. H3.4 will be tested by comparing test set ROC-AUC of homogeneous ensembles (only Logistic Regression, only Random Forest, or only MLPs) against heterogeneous ensembles (combining Logistic Regression, Random Forest, XGBoost, and MLP).

References

[1] Raga Krishnakumar and Anne M. Ruffing. Operonseqer: A set of machine-learning algorithms with threshold voting for detection of operon pairs using short-read rna-sequencing data. *PLoS Comput Biol*, 18(1):e1009731, 2022.

[2] Gabriel Moreno-Hagelsieb and Julio Collado-Vides. A comparative genomics approach to predict operons in prokaryotes. *Bioinformatics*, 18(suppl 1):S329–S336, 2002.

[3] Alberto Santos-Zavaleta, Socorro Gama-Castro, Irma Contreras-Moreira, Julio Diaz-Peredo, Laura Mendez-Cruz, Hilda Solano-Lira, Martin E. D. Osorio Garcia, Alejandra Ledezma-Tejeida, Alejandra Garcia-Alonso, Alexander N. T. Schaub, Hilda Moreno-Hagelsieb, Peter D. Karp, and

Julio Collado-Vides. Regulondb v12.0: A comprehensive resource of transcriptional regulation in escherichia coli k-12. *Nucleic Acids Res*, 52(D1):D255–D263, 2024.

[4] Eric W. Sayers, John Beck, Edward E. Bolton, J. Rodney Brister, Justin Chan, Roy Connor, Michael Feldgarden, Andrea M. Fine, Kendal Funk, Jennifer Hoffman, Sekar Kannan, Colleen Kelly, William Klimke, Sunghoon Kim, Scott Lathrop, Aron Marchler-Bauer, Tasha D. Murphy, Cameron O'Sullivan, Robert Schmieder, Yuliya Skripchenko, Adam Stine, Francoise Thibaud-Nissen, Jian Wang, Jian Ye, Elizabeth Zellers, Valerie A. Schneider, and Kim D. Pruitt. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Res*, 53(D1):D20–D29, 2025.

[5] Thomas West-Roberts, Anshul Kundaje, and James Zou. Dgeb: A diverse genomic embedding benchmark for functional genomics tasks. *bioRxiv*, 2024.