COMPSCI 602
Project Report 5

Chandana Magapu
College of Information and Computer Science
University of Massachusetts
Amherst, MA 01003
hippo@cs.umass.edu

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

1 Research Questions and Hypotheses

Research Question: How do simple ML models learn operon prediction?

Hypotheses:

- **H1.1:** There are some features that carry more signal than others, which reveal the fundamental mechanisms that models use for operon prediction. They 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 [1], and their removal will result in the largest performance degradation compared to other feature groups.

2 Type of Empirical Analysis

Both hypotheses will be tested through experiments; specifically 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 Data collection

The system in this project consists of nonlinear simple ML models outlined in OperonSEQer [2] (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 [3]: 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 (including the -10 and -35 boxes, and Rho-independent terminators) detected using position weight matrices and log-likelihood scoring, following standard practices described in prior studies [3]. I will vary each of biologically motivated features by dropping the columns and comparing the models' ROC-AUC score before and after dropping.

4 Analytic techniques

Along with the ablations/inverse ablations, I will measure changes in model performance with Bar charts and double bar charts as well as confusion matrices.

References

[1] Rezvan Karaji and Lourdes Peña-Castillo. Opdetect: A convolutional and recurrent neural network classifier for precise and sensitive operon detection from rna-seq data. *PLoS One*, 20(8):e0329355, 2025.

[2] 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.

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

[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.