COMPSCI 602
Project Report 2

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

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

1 System, task, environment, and implementation

The system here consists of nonlinear simple ML models outlined in [2](Random Forest, XGBoost, Multilayer Perceptron), and ensembles of these models. 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 that of a dataset that contains details about prokaryote operons for E.Coli, Vibrio, and Cyano[4]. The dataset will be constructed using the following repositories: `https://github.com/TattaBio/DGEB`
`https://github.com/sandialabs/OperonSEQer`
`https://github.com/BioinformaticsLabAtMUN/OpDetect`

2 Phenomena

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 be measuring this primarily through the area under the ROC 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).

3 Variables

The experiments will primarily be ablations of the features in the dataset. The dataset will be constructed by obtaining RNA-seq data for the protein-coding gene pairs in each of the three organisms. There will be 6 replicates for each organism. The data will also contain the intergenic distances, strand concordance, orientation pattern, GC-content difference, functional category match, as well as promoter and terminator motifs. I will be ablating each of these features to identify the relative importance to each model, and observing how the performance is affected.

4 Frontier

The current state of the art tools for Operon-prediction make it evident that models assuming linear relationships between the features and the classification output are insufficient, and non-linear tree-based models [2] or neural-network based models [1] are better suited to the task. There is currently no literature describing why this is. Interpretability and ablation studies on this baseline should fill this gap. Domain-informed features such as intergenic distance, GC content, and conserved promoter or terminator motifs [3] provide foundational, biologically interpretable signals that anchor operon prediction in established genomic rules. However, RNA-seq data [1, 2] captures the dynamic,

environment-dependent transcriptional landscape, including regulatory effects and condition-specific operon boundaries, thus offering a complementary and real-time window into gene expression that static sequence features alone cannot provide. Combining both approaches should enable more accurate and context-aware identification of operons, and also enable us to understand the relative importance of the features to the models, but such work has not been done till date.

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