FitnessMap: Investigating Co-Essential Gene Pairs in E. coli Using Tn-Seq Data and Structural Modeling

The FitnessMap project aimed to uncover functional interactions between genes in *Escherichia coli* by integrating transposon sequencing (Tn-Seq) fitness data with structural modeling of protein-protein interactions. Bacterial cells rely on networks of proteins working together to perform critical functions, and by analyzing patterns of gene essentiality across many conditions, it is possible to infer functional relationships between genes. This project leveraged the Fitness Browser, a rich functional genomics resource containing gene fitness profiles derived from thousands of experimental screens across diverse bacterial species.

At the core of the project was the use of gene fitness profiles—vectors that represent how the deletion of each gene affects bacterial growth across many conditions. These vectors were assembled in a fitness matrix in which rows represent genes and columns represent experimental conditions. With 3790 genes and 132 conditions, this matrix enabled the identification of co-essential genes—those with similar patterns of fitness across conditions. After conducting Exploratory Data Analysis (EDA), we found that many columns correlated with each other, which would bias a comparison of genes across experiments. To account for the correlations between experiments, we employed Generalized Least Squares (GLS) to measure the similarity between gene fitness vectors. GLS allowed for more robust detection of gene-gene relationships by whitening the data to correct for interdependencies between conditions and computing regression coefficients that quantify how much the fitness of one gene predicts the fitness of another.

Gene pairs with high GLS coefficients and low p-values were prioritized as candidates for further investigation. These pairs were interpreted as potentially functionally linked, either through direct protein-protein interaction or shared roles in cellular pathways. All pairs were filtered to cross a specified p-value threshold. Subsequently, the 50 pairs with the highest magnitude GLS coefficients were obtained, as well as several pairs with the lowest magnitude GLS coefficients to serve as controls. Protein sequences were retrieved from the Fitness Browser and Uniprot. Interestingly, in the 50 pairs with the highest GLS coefficients, 8 genes had unknown protein sequences. 8 of these corresponded to probable RNAs, including sRNAs and a tyrosine tRNA.

Protein-coding gene pairs were inputted into AlphaFold-Multimer v3, a powerful deep learning model that predicts the structure of protein complexes. Due to memory constraints on Google Colab, some structures were also predicted on the AlphaFold server, which utilized AlphaFold 3 (https://alphafoldserver.com/). Although the use of the two different models was not ideal, both are effective predictors of protein-protein interactions. The output structures were evaluated using inter-chain Predicted Aligned Error (PAE) scores, which provide a residue-residue estimate of alignment uncertainty between chains. By calculating the minimum PAE of interaction—defined as the lowest average cross-interface error—the project ranked gene pairs by structural confidence. Protein pairs with strong structural contacts and low inter-chain PAE of below <1.5 are considered likely physical interactors. In total, 27 gene pairs

within the top 50 interacting gene pairs were analyzed. These had an average of 17.59 min PAE, and the lowest min PAE value observed was 6.00. This indicated that for most of the 27 pairs, there was unlikely to be a physical interaction. The results of our preliminary analysis of the high GLS coefficient pairs and controls demonstrate that GLS coefficient may not be a good indicator of protein-protein interactions. As p-value may be a superior predictor, a future direction is to perform AlphaFold on the gene pairs with the smallest p-values.

There were a number of genes (20) which appeared more than once within the top 50 gene pairs. One of the genes, yhiM, was present in 9 different pairs. Intiguingly, yhiM and the 9 other genes share biological functions related to phosphate transport and responding to changing phosphate levels. Other genes which correlated together in the top GLS coefficient hits may similarly be involved in shared cellular pathways. Further insights may be obtained by delving deeper into the fitness data itself, such as by comparing scatterplots of fitness values between gene pairs.

For future directions, several extensions of the FitnessMap project could deepen our understanding of co-essential gene relationships and protein-protein interactions in *E. coli*. Expanding the number of AlphaFold predictions to include gene pairs across a broader spectrum of GLS coefficients would allow for more systematic analysis of how structural interaction likelihood varies with GLS coefficient. The analysis could be repeated utilizing p-values instead of GLS coefficients as the primary metric for statistical co-essentiality, as p-value may prove a better predictor. Adopting a consistent and up-to-date AlphaFold modeling pipeline will be important for reproducibility and model quality. Finally, targeted literature reviews of top-interacting gene pairs could reveal functional links that tie back to the fitness profiles.

Overall, the FitnessMap project demonstrated the power of combining statistical modeling with structural biology to map the interaction landscape of a bacterial genome. The approach not only identifies potential physical interactions but may also highlight candidates for further biological investigation. Looking ahead, this framework can be extended to other species in the Fitness Browser to discover conserved gene modules and functional pathways. In conclusion, the FitnessMap project demonstrates how functional genomics and machine learning-driven structural modeling can be combined synergistically to reveal novel insights into microbial gene networks.