**Machine learning on microbial phenotypes to classify gene functions**

*Abstract*— High-throughput studies of microbial phenotypes hold the potential in elucidating the functions of genes. Here we combined 2 high-throughput phenotype datasets with 5 categories of annotations as labels to classify genes with distinct functions. Preliminary results using complete phenotype data yielded mediocre performance, possibly resulted from incomplete annotations and/or non-separable nature of functional associations of genome-wide studies. However, selecting small numbers of mutually exclusive classes significantly improves the performance, indicating that the power of high-throughput phenotyping can be coupled with machine learning to associate genes that are functionally connected.

Keywords—microbial phenotypes, functions, supervised machine learning

# Introduction

Phenotypes play important roles in understanding functions of genes, leading to better understanding of disease models and thus contribute to new drug discoveries. Among model organisms that can be easily manipulated to test hundreds of thousands of phenotypes in parallel, E. coli serves as one of the best. We here would like to combine 2 different datasets [1] [2] that contain all phenotype data under mutation of almost every single gene of E. coli as the features (486 features for 3525 genes) and use 5 sets of gene annotations [3-5] as the target variables to classify genes of same functions. These annotations are thought to be highly accurate, yet incomplete and not mutually exclusive. In addition, there are many annotations that don’t have enough number for machine learning methods to train. Therefore, for each annotation sets, we picked 3 gold standard class and used several well-established supervised learning techniques to demonstrate the power of high-throughput microbial phenotypes to explain gene functions.

# Methods

A close up of a green screen

Description automatically generatedA screenshot of a cell phone

Description automatically generatedSubset of the phenotype data based on various annotation labels are selected. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector machine and Convoluted Neuro network were used do perform supervised learning.

A screenshot of a cell phone

Description automatically generated

Figure 1. (upper left) no. of annotated samples for each annotation set (upper right) A table showing the number of samples drawn for each class within each annotation sets as independent labels. 3 classes were selected for each class. (bottom) Unsupervised learning workflow.

# Results

For each selected category of annotations, 6 supervised learning techniques were applied. Maximum performance for each annotation ranges from 73% to 100% for both accuracy and precision (Figure 2). Overall, the phenotype data used here seem to be most effective in learning genes that constitute the same protein complex, and worst in genes that are co-regulated. This result is reasonable since the deletion of any subunit in a protein complex is very likely to cause the same malfunction and downstream phenotypes, while co-regulated genes might perform utterly different functions from different operons.

Looking into each supervised learning method (Figure 2), we observed that Logistic regression stands out as the simplest yet possibly strongest method, while Decision tree fell behind all other methods. It is surprising that although phenotypes as features shouldn’t be independent of each other, more complicated models such as SVM or CNN don’t significantly surpass the performance of logistic regression. Further experimentation with larger sample size might help answer this question.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

Figure 2. (Upper) Accuracy and (Lower) precision calculated using various supervised machine learning methods when using different annotations (labels)

A screenshot of a cell phone

Description automatically generated

Table 1. Best hyperparameters for each supervised learning method for each annotation (labels). For Logistic regression and Decision tree there were no hyperparameter tuning.

# conclusions

In this study, we have performed several supervised learning methods with a combined microbial phenotype dataset from 2 high-throughput studies. For every kind of annotations as labels, we get decent accuracy and precision. The results guarantee the utility of high-throughput, indirect phenotype measurement in explaining functions of genes.

# discission

Small subsets of data picked by biochemical knowledge guarantee enough samples which are mutually exclusive under distinct labels, therefore resulted in descent performance to learn functions via phenotypes. However, the impact of using machine learning on complete phenotype data with better labeling is yet to be done. It is worth noting that the 5 annotation sets selected here are mostly curated from experimental results from very large number of publications. It is spectacular that different biochemical or molecular biological experiments that yield the corpus of these annotations can be used as high-quality labels to examine phenotype data.

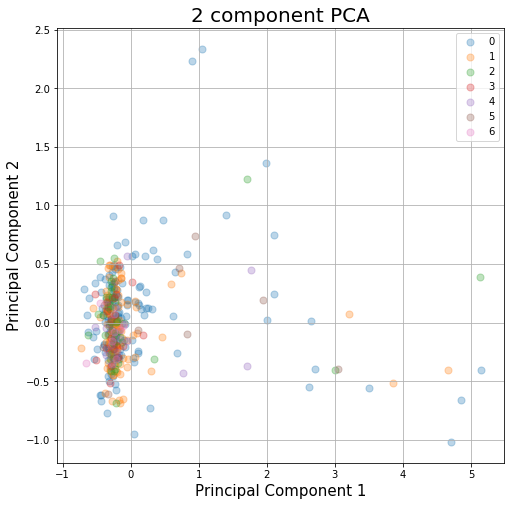
We have tried to reduce the dimensions of annotations by hierarchical clustering and separate genes into distinct categories. However, we obtained almost no separation by PCA, t-SNE or self-organizing map (Figure.3, Figure 4), and poor performance on all the supervised learning methods tested (~40% accuracy). Hopefully, with more annotations become available in the future, the genome-wide phenotypic data we have shown here can be much better exploited.

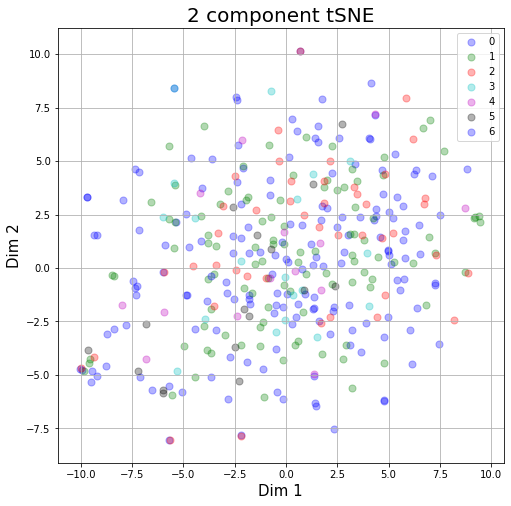
In addition to the 2 phenotype datasets we describe here, there are many others that measure distinct types of phenotypes [6-8], whereas in this study, our phenotypes (features) are simply growth rates measured by number of pixels of colony sizes under different stress/growth conditions. Incorporating new studies as a whole might be interesting future direction that help decipher functions of genes in more detail as well as facilitating more generalized machine learning models to be developed.

A close up of text on a white background

Description automatically generated

Fig. 3 PCA using complete phenotype dataset. There are no obvious functional clusters observed





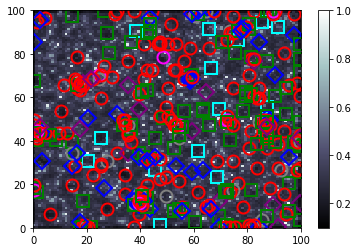


Figure. 4. With all described 5 annotations (labels), we tried to hierarchically cluster them and divide them into mutually exclusive groups, and then remove the number of groups that have less than 9 as the new labels. This resulted in 6 groups ready for supervised learning. However, the best accuracy obtained never go over 50%. As for unsupervised learning methods on this subset of phenotype data, PCA (upper), t-SNE (middle) and self-organizing map (bottom) reveal not-easily separated nature, which is complementary to the supervised learning methods.

A close up of text on a white background

Description automatically generated

Figure 5. Selected Protein complexes in phenotype-defined functional space, generated by Gaussian Mixture Model with Expectation Maximization. Phenotype data naturally separate by labels.

##### acknowledgement

This paper is a side project of Ontology of Microbial Phenotypes (microbialphenotypes.org). I (Peter) really thank the input from Chiou-Jiin Huang and Hao-Yu Miao for the code optimization and performance tuning. I also appreciate the lead from Dr. Bobak Mortazavi and my graduate advisor, Dr. James C. Hu.

##### References

[1] R. J. Nichols *et al.*, "Phenotypic landscape of a bacterial cell," *Cell,* vol. 144, no. 1, pp. 143-56, Jan 07 2011, doi: 10.1016/j.cell.2010.11.052.

[2] M. N. Price *et al.*, "Mutant phenotypes for thousands of bacterial genes of unknown function," *Nature,* vol. 557, no. 7706, pp. 503-509, May 2018, doi: 10.1038/s41586-018-0124-0.

[3] I. M. Keseler *et al.*, "The EcoCyc database: reflecting new knowledge about Escherichia coli K-12," *Nucleic Acids Res,* vol. 45, no. D1, pp. D543-D550, Jan 4 2017, doi: 10.1093/nar/gkw1003.

[4] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Res,* vol. 44, no. D1, pp. D457-62, Jan 4 2016, doi: 10.1093/nar/gkv1070.

[5] S. Gama-Castro *et al.*, "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Res,* vol. 44, no. D1, pp. D133-43, Jan 4 2016, doi: 10.1093/nar/gkv1156.

[6] A. Typas *et al.*, "High-throughput, quantitative analyses of genetic interactions in E. coli," *Nat Methods,* vol. 5, no. 9, pp. 781-7, Sep 2008

[7] T. Fuhrer, M. Zampieri, D. C. Sevin, U. Sauer, and N. Zamboni, "Genomewide landscape of gene-metabolome associations in Escherichia coli," *Mol Syst Biol,* vol. 13, no. 1, p. 907, Jan 16 2017, doi: 10.15252/msb.20167150.

[8] M. Campos, S. K. Govers, I. Irnov, G. S. Dobihal, F. Cornet, and C. Jacobs-Wagner, "Genomewide phenotypic analysis of growth, cell morphogenesis, and cell cycle events in Escherichia coli," *Mol Syst Biol,* vol. 14, no. 6, p. e7573, Jun 25 2018, doi: 10.15252/msb.20177573.