Peter I-Fan Wu UIN: 925008954, Hao-Yu Miao UIN: 329007009, Chiou-Jiin Huang UIN: 128003840

CSCE 633 Machine Learning project proposal – Classification of bacterial genes by high-throughput phenotype data

Introduction

Phenotypes play important roles in understanding functions of genes, which leads us to better understand 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. In recent years, there has been vast amount of phenotype data being published, using various experimental approaches. Yet from different studies, the effort to make those phenotype data interoperable to one another hasn't been conducted. In addition, machine learning approaches that query inferences to lead further experimental designs are still in its infancy .We here would like to associate 3 different datasets (Fuhrer et al., 2017, Nichols et al., 2011, Price et al., 2018) that contain almost all phenotype data under mutation of every single gene of *E. coli* as the explanatory variables, and use 5 sets of gene annotations (Gama-Castro et al., 2016, Kanehisa et al., 2016, Keseler et al., 2017) as the response variables to classify genes of similar functions.

Specific aims

Popular unsupervised/supervised learning methods that have been proven useful in general with methods that are useful for biological datasets are to be tested as follows:

i. Unsupervised learning approaches: PCA, t-SNE, Self-organizing Map
ii. Supervised learning approaches: Random forest, SVM, Neuro network

Research strategies

i. Data preparation
    a. clean, normalize and combine the 3 datasets
    b. (For supervised learning) Render a good representation of responses based on the 5 best *E. coli* annotation sets
ii. Unsupervised/ Supervised learning methods listed above will be tested:
    a. For unsupervised learning approaches, the attributes of the clusters/groups of genes will be assessed with the 5 annotation sets.
    b. For supervised learning, AUROC/AUPRC will be reported

Expectations

i. For unsupervised learning: It is expected that some genes that share similar functions or involved in the same biological processes (e.g. Glucose metabolism, DNA synthesis) will be classified in the same group or clustered together. This can be assessed with the 5 annotation sets.

ii. For supervised learning it is expected that some genes of unknown function can be classified with those that share similar functions or involved in the same biological processes.

References

Fuhrer, T., Zampieri, M., Sevin, D. C., Sauer, U., & Zamboni, N. (2017). Genomewide landscape of gene-metabolome associations in Escherichia coli. Mol Syst Biol, 13(1), 907. doi:10.15252/msb.20167150

Peter I-Fan Wu UIN: 925008954, Hao-Yu Miao UIN: 329007009, Chiou-Jiin Huang UIN: 128003840

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res, 44(D1), D133-143. doi:10.1093/nar/gkv1156

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res, 44(D1), D457-462. doi:10.1093/nar/gkv1070

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., . . . Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res, 45(D1), D543-D550. doi:10.1093/nar/gkw1003

Nichols, R. J., Sen, S., Choo, Y. J., Beltrao, P., Zietek, M., Chaba, R., . . . Gross, C. A. (2011). Phenotypic landscape of a bacterial cell. Cell, 144(1), 143-156. doi:10.1016/j.cell.2010.11.052

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., . . . Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. Nature, 557(7706), 503-509. doi:10.1038/s41586-018-0124-0