Using deep learning to predict gene essentiality in bacteria

Ian Lo and Carolin Schulte

Project description

The aim of this project is the development of a deep learning model that accurately predicts gene essentiality based on gene- and protein-level information that are readily available for most bacterial species. Identifying essential genes in bacteria is of major importance for various research areas: with regard to basic research questions, elucidating gene essentiality has the potential to expand our understanding of fundamental concepts underlying the functioning of a cell, improving genome annotations, and determining context-specific metabolic pathway usage[1]. More applied research endeavors have focused on leveraging knowledge of gene essentiality for creating synthetic cells with a minimal number of genes[2] or designing drugs targeted against various pathogens[1].

Determining gene essentiality on the genome-scale was initially a very labor-intensive process whereby single gene deletion mutants had to be created experimentally[3]. Recent advances in experimental methods, such as transposon-insertion sequencing approaches, have greatly enhanced our ability to identify essential genes in a high-throughput manner. However, these methods still require substantial experimental effort and are limited to genetically tractable organisms that can be cultured in laboratory conditions[4]. Due to these drawbacks and the rapidly increasing availability of genome sequences, computational methods are now commonly used for predicting gene essentiality. While homology- and metabolic modeling-based approaches have been successfully applied to this end, these methods either rely on the existence of conserved orthologs or are limited to genes involved in metabolic processes[5], which substantially limits their scope. In contrast, machine learning models trained on data for model organisms have the potential to identify universal features of essential genes and accurately predict essential genes in unstudied species. Several studies have explored machine learning methods such as Logistic Regression, Decision Tree or Random Forest[6], however Deep Learning methods have so far not been extensively applied[7].

This project will use supervised machine learning methods, in particular neural networks, to predict gene essentiality in bacteria. Models will be trained on data for bacterial species for which gene essentiality has been experimentally determined[8]. A major part of this project will be appropriate featurization of available data: in addition to nucleotide and amino acid sequences, more complex features such as codon usage, homology to known essential genes and gene expression levels will be evaluated as possible inputs, since use of these features has been shown to substantially improve the performance of machine learning models[9]. For each genome, non-essential genes will typically outnumber essential genes, which will probably necessitate the application of under- or oversampling techniques. Central to our approach will be a focus on using only those information that are readily available for a broad range of non-model species, and we will aim to achieve accurate predictions for species that are genetically distant from those used in the training set.

Data availability

Genome and protein sequences will be obtained from the NCBI RefSeq database[10] and gene expression data will be retrieved from the Gene Expression Omnibus (GEO)[11]**.** Since experimentally obtained gene essentiality data are known to be imperfect, information from several databases , such as the Fitness Browser[8] and the Database of Essential Genes (DEG)[12], will be combined to obtain reliable essentiality labels. In particular, the Fitness Browser contains gene essentiality data for 46 bacterial strains. Additional information will be obtained using suitable computational tools (e.g. OrthoFinder[13] to identify orthologous genes) and databases (e.g. the STRING database[14] for protein-protein interactions) as needed.

Suggested supervisor: Prof. Ernest Fraenkel

References

1. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet* **19**, 34–49 (2018).
2. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52–56 (2010).
3. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
4. Cain, A. K. *et al.* A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* **21**, 526–540 (2020).
5. Aromolaran, O., Aromolaran, D., Isewon, I. & Oyelade, J. Machine learning approach to gene essentiality prediction: a review. *Briefings in Bioinformatics* **22**, bbab128 (2021).
6. Dong, C. *et al.* Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment. *Briefings in Bioinformatics* **21**, 171–181 (2020).
7. Hasan, M. A. & Lonardi, S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. *BMC Bioinformatics* **21**, 367 (2020).
8. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
9. Fu, C. *et al.* Leveraging machine learning essentiality predictions and chemogenomic interactions to identify antifungal targets. *Nat Commun* **12**, 6497 (2021).
10. RefSeq: NCBI Reference Sequence Database. https://www.ncbi.nlm.nih.gov/refseq/.
11. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2013).
12. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research* **42**, D574–D580 (2014).
13. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
14. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612 (2021).