Project Description

Representation learning of fluxomic data for therapy

Supervisors: Dr Diego A. Oyarzún and Dr Oisin Mac Aodha

Abstract

This project aims to improve the analysis of cellular metabolism for diagnostics and therapy, using genome-scale metabolic models (GEMs) and self-supervised learning. The interconnected nature of cellular metabolism presents challenges in predicting systemic effects when specific reactions are inhibited through therapeutic interventions. We will develop an end-to-end pipeline for graph representation learning of GEMs, test the utility of low-dimensional embeddings for therapy and diagnostics, and build a user-friendly web platform for model training. The approach will be based on the integration of metabolic graph structure into variational autoencoders trained on large fluxomic data. Applications include classification of GEMs from healthy and disease states, metabolic stratification of tumours, and detection of lethal targets against pathogens.

Introduction

Cell metabolism is a highly interconnected network with thousands of reactions, many of which are attractive targets for therapy in a range of conditions, including diabetes, gout, and rare disorders such as phenylketonuria or Gaucher disease. In cancer treatment, metabolic drug targets have received increasing attention as tumours can display metabolic vulnerabilities that can be exploited for therapy [1]. In infectious diseases, several promising antibiotic treatments target specific reactions to disrupt the metabolic capacity of pathogens. However, the complex connectivity of cellular metabolism makes it challenging to predict the systemic effects caused by inhibition of specific reactions.

Genome-scale metabolic models (GEMs) are a mathematical description of the connectivity of metabolic fluxes in an organism. Such models have been built for over 3,000 microbial species and, recently GEMs have gained increased adoption to model human metabolism. A key development was the release of Recon3D, an all-encompassing model of human metabolism that can be instantiated to specific cell types with 'omics data [2]. This enables the construction of patient-specific models for personalized medicine and leverage computational approaches to analyse the complex interactions within an individual's metabolic network.

From a mathematical perspective, a genome-scale metabolic model is a set of linear constraints that define a *flux cone*. The high-dimensionality of the flux space (>2,000 dimensions) limits our ability to analyse the shape of the cone and its response to drug treatments, loss-of-function mutations, and other perturbations relevant for therapy and diagnostics. In a recent proof-of-concept study, we trialled the use variational autoencoders to compress the dimensionality of the flux cone while preserving structure [3]. These promising results offer a novel route for the analysis of GEMs based on self-supervised training.

Research Challenge

Our general aim is to develop an improved framework for the analysis of genome-scale metabolic models. Our specific objectives are:

1) Build an end-to-end pipeline for graph representation learning of genome-scale metabolic models, using a combination of autoencoders and the underlying graph describing the connectivity of metabolic fluxes.

2) Test the utility of the low-dimensional embedding for a range of downstream supervised and unsupervised relevant tasks for therapy and diagnostics, using publicly available GEMs.
3) Build a web platform for users to train our latent representation on their in-house genome-scale metabolic models.

Data & Methodology

The project will rely on well adopted packages for GEM analysis and simulation (COBRApy, COBREXA). Models will be retrieved from publications and the BiGG database [4]. For model training, we will generate a large corpus *fluxomic* data from specific GEMs. These consist of high-dimensional flux vectors sampled from the flux cone defined by each model; sampling will be performed with Markov Chain Monte Carlo (MCMC) methods using existing packages.

In **Objective 1**, we will modify the variational autoencoder (VAE) in our recent work [3] to include the graph structure of the metabolic network. The expectation is that inclusion of connectivity between different flux features will improve the expressiveness of the learnt embeddings, and therefore produce more accurate low-dimensional representations that can employed for downstream tasks. As a graph backbone, we will utilize the Mass Flow Graphs we have introduced in a previous work for accurate prediction of gene essentiality with graph neural networks [5]. We will explore various model architectures, training strategies and data pre-processing techniques to improve the expressiveness and generalization ability of the embeddings.

In **Objective 2**, we will focus on various relevant tasks, including: a) classification of GEMs obtained from healthy and disease states; b) metabolic stratification of tumour types from their GEMs; c) clustering of cell-type specific GEMs from single-cell transcriptomics; d) detection of lethal targets against microbial and fungal pathogens.

In **Objective 3**, we will build a full stack web application that can readily interact with existing well-adopted GEM tools (RAVEN, COBRApy, COBREXA) to increase adoption by the community. The software will be aimed at end-users and offer a no-code platform for the analysis of GEMs in a low-dimensional embedding space.

RRI/Ethical considerations

Improving personalized therapy and diagnostics raises ethical concerns regarding its potential to deepen global health inequalities. High costs of sequencing and specialized treatments may widen the gap between high- and low-income nations, as affluent countries often pioneer personalized medicine initiatives, leaving poorer nations lagging behind. During the project we will explore the application of our tools to detect metabolic drug targets against pathogens with high prevalence in the developing world (Objective 2d), such as *M. leprae* (leprosy), *T. cruzi* (Chagas disease), and *P. falciparum* (malaria). Genome-scale metabolic models for these pathogens are publicly available.

Expected outcome & Impact

The project will deliver a suite of data, models and software for the analysis of metabolism in disease. Our approach will provide a framework to distil the metabolic space into semantically rich vectors that can be used as a foundation for predictive modelling. This can potentially contribute to the discovery of new therapeutic targets or diagnostic biomarkers across a range of conditions, and contributes to the fast-moving body research at the interface of machine learning and personalized therapies.

References

[1] Z. E. Stine, Z. T. Schug, J. M. Salvino, and C. V. Dang, 'Targeting cancer metabolism in the era of precision oncology', Nat Rev Drug Discov, vol. 21, no. 2, pp. 141–162, Feb. 2022, doi: 10.1038/s41573-021-00339-6.

[2] E. Brunk et al., 'Recon3D enables a three-dimensional view of gene variation in human metabolism', Nat Biotechnol, vol. 36, no. 3, pp. 272–281, Mar. 2018, doi: 10.1038/nbt.4072.

[3] Cain, Samuel, Merzbacher, Charlotte, and Oyarzún, Diego A., 'Low-dimensional representation of genome-scale metabolism', Submitted, 2024.

[4] Z. A. King et al., 'BiGG Models: A platform for integrating, standardizing and sharing genome-scale models', Nucleic Acids Research, vol. 44, no. D1, pp. D515–D522, Jan. 2016, doi: 10.1093/nar/gkv1049.

[5] R. Hasibi, T. Michoel, and D. A. Oyarzún, 'Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality', npj Syst Biol Appl, vol. 10, no. 1, pp. 1–10, Mar. 2024, doi: 10.1038/s41540-024-00348-2.