Title

Nicholas L. Larus-Stone

Queens' College

UNIVERSITY OF CAMBRIDGE

A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for the degree of Master of Philosophy in Advanced Computer Science

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: nl363@cl.cam.ac.uk

May 25, 2018

Declaration

I Nicholas L. Larus-Stone of Queens' College, being a candidate for the M.Phil in Advanced Computer Science, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Total word count: 0

Signed:

Date:

Abstract

This is the abstract. Write a summary of the whole thing. Make sure it fits in one page.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Despite more than a century of active biology research, running biological experiments is still time consuming and difficult. Therefore, having a good computational model of the biological system of interest is incredibly valuable. What makes a computational model good? First of all, and most importantly, a good model needs to accurately describe how the biological system acts in the real world. If the model accurately represents the biological system, then experiments can be run in silico instead of in vivo, speeding up the pace of research. Secondly, a good model should be consistent–meaning its results vary only due to intrinsic biological sources of variation. Though biological experiments can have wide variance due to external conditions, we want our model to vary due to the underlying biology. Exact reproducibility is very difficult in laboratory environments, but a computer model can and should eliminate the extrinsic sources of variance that one encounters in a laboratory setting. Finally, the best models not only describe the system, but also facilitate deeper insights into the underlying biology. A good model should help researchers improve their understanding of the system that is modeled and provide a starting point for novel insights.

Biological systems are incredibly complex and can be modeled at many different levels. For example, there has been a recent rise in using multi-omic data–data that includes genomic, transcriptomic, metabolomic measurements–to model the behavior of an organism. This dissertation will focus on using metabolic information to build a good computational model for a specific type of biological system called a cell-free system. In particular, we use a variational autoencoder (VAE) on top of a metabolic model to determine which reactions are most important for cell-free systems. The goal is to provide a model for biologists to better understand cell-free systems. This will allow them to anticipate how different starting conditions will affect their experiments before actually running the experiment.

Metabolic Models Metabolism involves all of the chemical reactions that occur within an organism in order to sustain its life and growth. A huge number of applications in biology require examining the metabolism of an organism. In particular, metabolic engineering has the potential to create advances in medicine and energy [1]. Thus, metabolic models have gained popularity over the past few decades as an important tool to understanding what is going on inside of a cell. These models are used to predict how the phenotype of the organism will respond to various environmental inputs.

Most metabolic models involve mathematical representations of the processes occurring within the cell. The different types of metabolic models are usually collected under the heading of constraint-based reconstruction and analysis (COBRA) models [2]. In particular, Flux Balance Analysis (FBA), is one of the most popular metabolic modeling techniques and has been used in applications from optimizing metabolic pathways [3] to investigating gene networks [4]. FBA is applied to genome scale metabolic models (GEMs) to solve for a specific outcome of interest. As progress has been made in the sequencing and understanding of the genomes of various organisms, these GEMs have become commonplace for many organisms. However, these GEMs describe living organisms, not cell-free systems.

Cell-Free Systems Cell free protein systems (CFPS) are a simplified version of cells that expose the crucial internal transcription and translation machinery, allowing the system to continue produce proteins without being alive. CFPS have existed since the early 1960's as a way to express proteins that are otherwise difficult to express in a living cell []. Recent work involving the removal of certain genes have stabilized amino acid synthesis and improved the yield of these protein expression systems [5]. This has led to a resurgence in the field of using CFPS in both academic and commercial contexts.

CFPS are ideal in two respects for this dissertation. First of all, running experiments in CFPS is much faster than typical in vivo methods because there is no need to grow cells. Due to the time limitations of the dissertation schedule, this meant that I was able to generate relevant data within a few months instead of the years that a cell-based system could take. Additionally, CFPS contain a strict subset of the reactions that occur in a cell-based system. Therefore, we hypothesized that it would be easier to model a CFPS than a full E. coli system.

There have been a few attempts to model cell-free systems, which are explored in Section 2.2. However, very few of them have used metabolic models to examine cell-free systems. Since CFPS are not living organisms, their functionality is entirely based on energy constraints. So, it makes sense to examine these systems from the perspective of metabolism. Existing metabolic models for CFPS are hand-constructed instead of using the GEMs that already exist for the original organism.

Contributions This dissertation focuses on a number of novel aspects with regards to metabolic modeling and cell-free systems.

- This dissertation provides an automated reduction system that tailors GEMs specifically for cell-free systems. That means the system can be used for any cell-free system which is derived from an organism that has a fully specified GEM, of which there are many dozens.

- Additionally, it is the first use to our knowledge of applying deep learning or autoencoders to FBA models. While PCA has been applied in the past, it lacks the generalizability of an autoencoder.

- Finally, this dissertation provides an end-to-end system to accelerate the design-build-test (DBT) cycle of a lab. Given biological data on

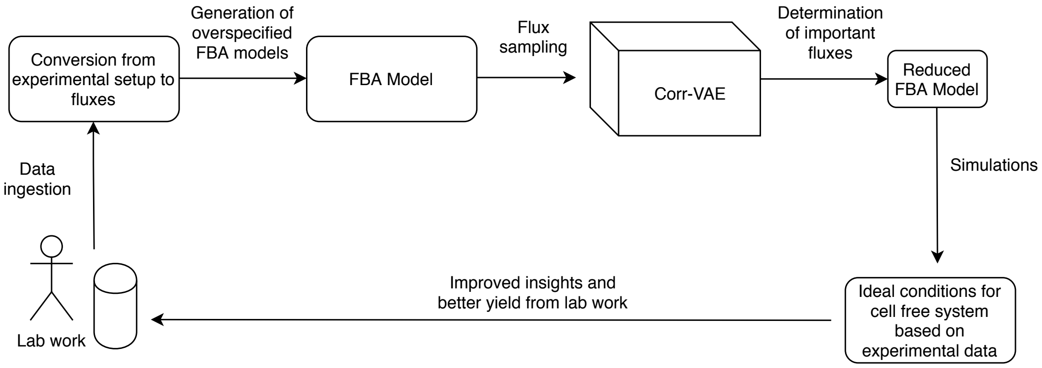
Conversion from experimental setup to fluxes
Generation of overspecified FBA models
FBA Model
Flux sampling
Corr-VAE
Determination of important fluxes
Reduced FBA Model
Simulations
Ideal conditions for cell free system based on experimental data
Improved insights and better yield from lab work
Data ingestion
Lab work

Figure 1.1: A figure showing the overall pipeline for the system to generate cell-free metabolic models.

one end, the computational pipeline then produces a model that allows the prediction of data for subsequent experiments.

Paper outline The structure of this dissertation is as follows. The next chapter provides background and related work. We focus on describing the current state of the art in metabolic modeling, specifically focusing on FBA and the creation of a GEM for E. coli. Next, we focus on efforts to model cell-free systems, describing the few examples of FBA applied to cell-free systems. Then, we describe the relevant literature in autoencoders, in particular VAEs. Finally, we describe previous applications of machine learning to FBA.

Chapter 3 describes the system we have designed and built to model cell-free systems. The entire pipeline is shown in fig 1. We begin by describing the biological experiments and the type of data that was generated. Next, we describe how to ingest the experimental setup and convert it into various FBA models. Then, we explain how we flux sampled from the FBA models to generate a larger dataset for our deep learning algorithms. That dataset is then ran through a VAE, which generates a latent representation. That latent representation is used to reduce the FBA models. Finally, we describe how we the reduced FBA model to improve future experimental yield.

Chapter 4 describes the results of the experiments. This includes results showing the improvment of my technique over naive solutions as well as novel biological insights that were unearthed due to my system.

Finally, we conclude with a chapter of discussion and directions of future work in this area.

Chapter 2

Related Work

This chapter introduces related work in the relevant fields of computer science and biological modeling. The key insight is that very few techniques have focused on modeling cell-free systems and no one has yet applied deep learning techniques to metabolic models.

2.1 FBA and GEMs

The earliest work on metabolic modeling comes from the 1980s with the use of stoichiometric equations to predict the yield of specific fermentation products [6]. This work was soon expanded to other systems and the use of linear programming (LP) to solve for the optimal result was proposed [7]. Constraint based analysis was soon applied to the description of the metabolic system in E. coli [8]. Improvements on this initial model have been stepwise in the years following. For instance, a large step involved the addition of catabolic pathways to the model [9].

With the growth in genome sequencing, however, came the ability to describe metabolic systems on the basis of their underlying genome.

2.2 Modeling Cell-Free systems

2012 cell free model Murray kinetic models Varner ssFBA

2.3 VAEs

Kingma 2013

2.4 Applying ML to FBA

PCA + FBA Bayesian FBA Network models on FBA

Chapter 3

Design and Implementation

This chapter may be called something else. . . but in general the idea is that you have one (or a few) "meat" chapters which describe the work you did in technical detail.

3.1 Data gathering

3.1.1 Cell-free systems

3.1.2 Experimental setup

See Appendix A for more details

3.2 Data ingestion and incorporation

3.3 FBA and flux sampling

3.3.1 FBA

3.3.2 Flux sampling

3.4 VAEs

3.5 Correlated VAEs

3.6 FBA model reduction

[10]

Chapter 4

Results

4.1 Comparison of models

4.2 Emphasis on where FBA-VAE does particularly well

4.3 Noise addition

4.4 Transfer learning

4.5 Biological insights

Chapter 5

Summary and Conclusions

As you might imagine: summarizes the dissertation, and draws any conclusions. Depending on the length of your work, and how well you write, you may not need a summary here.

You will generally want to draw some conclusions, and point to potential future work.

5.1 Improved data gathering

5.2 RL system for reduction

5.3 Improved starting model

5.4 Improved VAE

Appendix A

Experimental Setup

Bibliography

[1] Jay D Keasling. Synthetic biology and the development of tools for metabolic engineering. *Metabolic engineering*, 14(3):189–195, 2012.

[2] Jan Schellenberger, Richard Que, Ronan MT Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, et al. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2. 0. *Nature protocols*, 6(9):1290, 2011.

[3] E_ Almaas, B Kovacs, T Vicsek, ZN Oltvai, and A-L Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839, 2004.

[4] Tomer Shlomi, Yariv Eisenberg, Roded Sharan, and Eytan Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3(1):101, 2007.

[5] Kara A Calhoun and James R Swartz. Total amino acid stabilization during cell-free protein synthesis reactions. *Journal of biotechnology*, 123(2):193–203, 2006.

[6] Eleftherios Terry Papoutsakis. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 26(2):174–187, 1984.

[7] David A Fell and J Rankin Small. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochemical Journal*, 238(3):781–786, 1986.

[8] RA Majewski and MM Domach. Simple constrained-optimization view of acetate overflow in e. coli. *Biotechnology and bioengineering*, 35(7):732–738, 1990.

[9] Amit Varma and Bernhard O Palsson. Metabolic capabilities of escherichia coli: I. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology*, 165(4):477–502, 1993.

[10] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology*, 7(1):74, 2013.