Towards Elucidating Regulatory Structure of Metabolic Networks for Dynamic Modeling

Justin Y. Lee*. Carlos Orosco**. Britney Nguyen***
Mark P. Styczynski****

*Georgia Institute of Technology, Atlanta, GA 30332
USA (e-mail: justlee@gatech.edu).
** Georgia Institute of Technology, Atlanta, GA 30332
USA (e-mail: corosco3@gatech.edu)
*** Georgia Institute of Technology, Atlanta, GA 30332
USA, (e-mail: bnguyen82@gatech.edu)
****Georgia Institute of Technology, Atlanta, GA 30332
USA (Tel: 404-894-2825; e-mail: mark.styczynski@chbe.gatech.edu).

Abstract: The ability to understand and manipulate metabolism is of great value in the chemical industry, as it opens the door to engineering organisms to make valuable small molecule chemicals and intermediates. However, even simple organisms like bacteria and yeast have extremely complex metabolic networks, consisting of typically well-characterized stoichiometric relationships and often poorly-characterized regulatory relationships. We have recently developed a framework for constraint-based dynamic modeling of metabolic networks, but one of the outstanding challenges in applying this framework is the need for better ways to infer the regulatory network structure in cases where only stoichiometry, not regulatory structure, is known. We will discuss the applications of machine learning relevant to developing a predictive understanding of cellular metabolism, including the use of data from systems-scale measurement of small molecules (known as metabolomics) coupled with inferred or explicitly measured metabolic flux distributions to characterize these unknown relationships. By training on a few simple models, we are able to substantially prune the large search space of candidate regulatory interactions, yielding improved identification of the true interactions from that search space. Taken together this approach is promising for future modeling and engineering of these complex biochemical systems.

Keywords: Biosystems, Modeling and identification, Bioinformatics, Metabolic engineering, Microbial technology

1. INTRODUCTION

In metabolic systems, many reaction rates are directly regulated by the small molecule biochemical components of metabolism (metabolites). While inclusion of these regulatory interactions is critical for creating accurate metabolic models, knowledge of where in the metabolic network these regulatory interactions occur is often unknown or incomplete for systems that are not well-studied. Despite potentially containing information that could help identify these interactions, metabolomics data (the systems-scale measurement of metabolites in biological systems) have been sparingly used with computational methods to discover regulatory interactions. These data are increasingly more commonly being acquired as part of systems biology experiments, typically through techniques like gas or liquid chromatography coupled to mass spectrometry.

We have recently developed an approach called linear kinetics dynamic flux balance analysis (LK-DFBA) that takes a widely-used genome-scale metabolic modelling approach, flux balance analysis (FBA), and tries to expand it to allow for dynamic metabolic models (Dromms *et al*., 2020). While many different versions of FBA exist that exploit different types of datasets to allow for improved model performance, most of the published FBA implementations include an assumption of metabolic steady state. However, metabolic processes are typically not at steady state, making these models necessarily limited in their potential accuracy and predictivity. Dynamic FBA approaches have been developed, but they typically entail some additional characteristic (e.g., an accompanying differential equation) that make them difficult to implement at a truly genomic scale in a computationally feasible fashion. LK-DFBA attempts to address this shortcoming by linking together FBA models at different time points, tracking metabolite concentrations via pooling fluxes, and expressing reaction and regulation kinetics via linear approximations to typically nonlinear equations (e.g., Michaelis-Menten kinetics), allowing for dynamic metabolic modelling with the potential for genome-scale applicability. However, this approach requires knowing the regulatory structure of the metabolic network, which may not be available in the literature for non-model organisms or non-central metabolic pathways. The identification of these regulatory structures from metabolomics data (the same type of data that is used in LK-DFBA and not fully exploited in FBA in general) has not received significant attention in the literature, yet will be a central problem moving forward in the development of dynamic genome-scale models of metabolism.

Here, we present our efforts to use metabolite profiling and flux data to infer potential regulation of specific reaction fluxes in a model by specific metabolite concentrations. We use machine learning approaches to triage potential interactions from among the many possibilities.

2. APPROACH

Here, we have started to develop a machine learning classification framework that uses stoichiometric information about the biological system and metabolite concentration and flux data to determine where metabolite-dependent regulatory interactions likely occur.

2.1 Model systems

We used two defined model systems as the basis for generating gold standard *in silico* data, as shown in Figures 1 and 2. Each are greatly simplified representations of metabolic models, but they incorporate key aspects of true models.

The Small Regulation Model is comparatively small (six reactions) and simple (one branch point) in its metabolic topology. However, its regulatory topology presents additional potential richness, including negative feedback regulation and cross-talk between divergent pathways. This serves as a useful, extremely small, well-controlled synthetic network with multiple biologically relevant features.

The Big Regulation Model is not large in an absolute sense and is still much smaller than real genome-scale metabolic models, but it adds significant complexity to the Small Regulation Model. While it only contains ten fluxes, it has three branch points and extremely complex regulation, including positive and negative feedback regulation and pathway cross-talk. Its increased size provides a combinatorial increase in potential regulatory interactions compared to the Small Regulation Model, which makes the active fraction of potential regulatory interactions in the model more representative of what one would expect to find in a true biological or biochemical system. It also allows us to get a sense of how the machine learning challenge will scale with problems of increasing size.

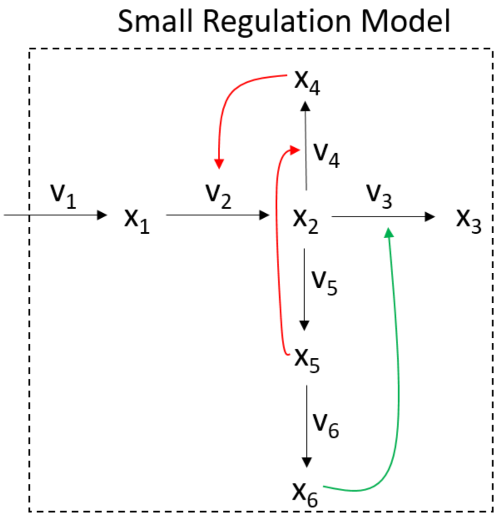

Figure 1: Small model network. There are 20 possible regulatory interactions in this network, but only three actual regulatory interactions: two inhibitory (red arrows) and one activating (green arrow).

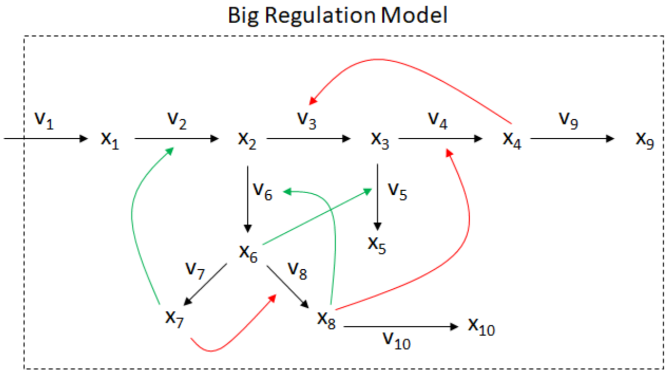

Figure 2: Larger model network. There are 72 possible regulatory interactions in this network, but only six actual regulatory interactions: three each of inhibitory (red arrows) and activating (green arrows). This network is still an order of magnitude smaller than real biological systems, but serves as a challenge of additional complexity compared to the Small Regulation Model.

2.2 Gold standard data generation

Biochemical Systems Theory kinetic equations based on generalized mass action kinetics were used for the generation of concentration and flux profiles. 10 random initial conditions were used to create 10 different datasets for each model. Noise was added to the data by drawing a random value from $N_{i,k} \sim (y_i(t_k), CoV \cdot y_i(t_k))$, where $y_i(t_k)$ is the value of species (metabolite or flux) i at timepoint k, and CoV is the coefficient of variance. A CoV of 0.05 was used in this work. Noisy data were smoothed using a previously described impulse function (Dromms and Styczynski, 2015).

2.3 Features for machine learning

Features are key characteristics of datasets that help differentiate between true and nonexistent regulatory interactions. Features used for machine learning included the correlation between a controller metabolite and target flux (likely greater in magnitude if the interaction exists), the goodness-of-fit of a surface fitting between multiple controller metabolites and the target flux (likely greater for a true interaction), and whether the (metabolite concentration, flux) ordered pairs represent a function with a single target flux for each concentration (likely untrue for interactions that are not true).

2.3 Machine learning methods

The *in silico* data were divided into a training set and a testing set for 100 bootstrap replicates. A random forest machine learning classifier was trained using correctly labeled interactions (true or nonexistent) from the training set and assessed for performance on the testing set for each replicate.

3. RESULTS

When using clean synthetic data with no noise added (CoV = 0), the machine learning classifier was able to predict true interactions with 100% sensitivity (true positives divided by true positives plus false negatives) and 98.7% specificity (true negatives divided by true negatives plus false positives), supporting the potential utility of both the feature set and the machine learning approach (Figure 3). However, when noise was added to the data, classifier performance degraded. Smoothing of the noisy data (which would normally be done for real experimental data) was necessary in order for the classifier to predict that any interactions were true. After smoothing, the classifier had 33.9% sensitivity and 87.6% specificity across ten different noisy datasets used for the bootstrap analysis (Figure 4).

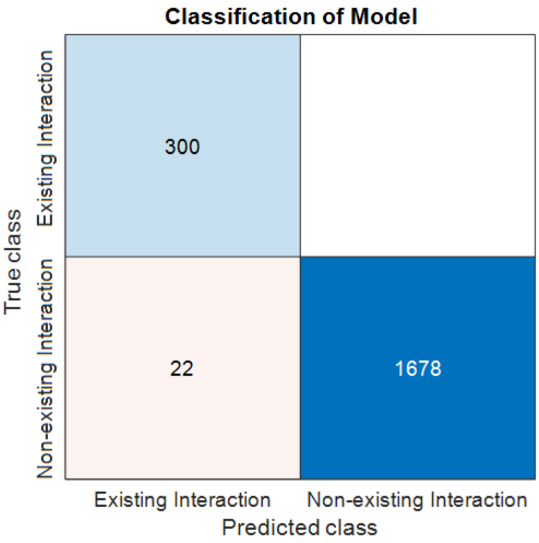

Figure 3: Confusion matrix for in silico data generated without any added noise.

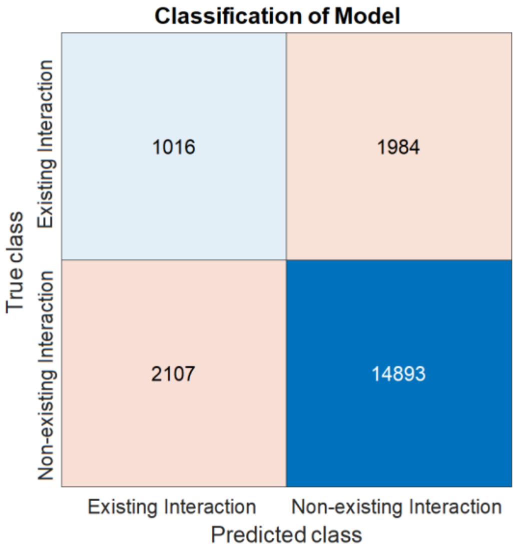

Figure 4: Confusion matrix for in silico data generated with noise based on a hypothetical coefficient of variation of 5%.

4. CONCLUSIONS

Use of realistic metabolite profiling data with even a comparatively small amount of noise that would arise from biological or technical variability presents challenges, but the effectiveness of our approach on clean data suggests the potential for broader utility. Future steps will include the identification or generation of additional features to better enable machine learning classification.

REFERENCES

Dromms, R.A., Lee, J.Y., and Styczynski, M.P. (2020) LK-DFBA: A Linear Programming-based modeling strategy for capturing dynamics and metabolite-dependent regulation in metabolism. *BMC Bioinformatics*, 21, 93.

Dromms, R.A., and Styczynski, M.P. (2015) Improved metabolite profile smoothing for flux estimation. *Molecular Biosystems,* 11 (9), 2394-2405.