Proceedings of the 14th International Symposium on Process Systems Engineering – PSE 2021+
June 19-23, 2022, Kyoto, Japan

Knowledge-matching based computational framework for genome-scale metabolic model refinement

Kiumars Badr, Q. Peter He and Jin Wang*

Department of Chemical Engineering, Auburn University, Auburn, AL, 36849, USA
**Corresponding Author's E-mail: wang@auburn.edu*

Abstract

Genome-scale metabolic models (GEMs) are mathematically structured knowledge base reconstructed from annotated genome of different organisms. With the advancement of next-generation sequencing technology, many organisms have had their genomes sequenced. However, obtaining a high-quality GEM is highly time-consuming, even with the introduction of several genome-scale reconstruction tools that offer automated draft network generation and gap filling. It has been recognized that the iterative process of manual curation and refinement is the limiting step of GEM development, and how to expedite the GEM refinement is still an open question. As cellular metabolism is a complex system with very high degree of freedom and redundancy, the principles and techniques developed in process systems engineering can be adapted to expedite GEM refinement. In this work we present a knowledge-matching based computation framework for GEM refinement, and demonstrate the effectiveness of the proposed solution using the refinement of a GEM for *Clostridium tyrobutyricum* .

Keywords: Genome-scale metabolic model; manual curation; system identification; multivariate analysis; visualization.

1. Introduction

A genome-scale metabolic model (GEM) is a mathematically structured knowledge base that is reconstructed from annotated genome of an organism (King et al., 2016)). A GEM contains a list of biochemical reactions, metabolites and (annotated) genes involved in the cellular metabolism for a specific organism, as well as a set of biophysical constraints (e.g., nutrient uptake and substrate availabilities, etc.). A high-quality GEM can be used to conduct simulations to answer various questions about the capabilities of the organism, serve as a framework to integrate and interpret omics data collected through experiments, and guide the design of mutant for metabolic engineering, etc.

The GEM reconstruction process generally consists of the following steps (Thiele and Palsson, 2010): (1) a draft network is reconstructed based on the annotation of a genome and the prediction of candidate metabolic functions; (2) the draft reconstruction is refined or curated by the user in an iterative manner through an exhaustive review of each reaction, metabolite and gene in the network; (3) the reconstruction is transformed into a mathematical structure, with an objective function and a set of constraints to account for different culture conditions. With the mathematical representation, the resulting GEM can be evaluated by reproducing the experimental data. If the GEM predictions do not match the experimental data, the manual refinement process will be repeated till the quality of the GEM is satisfactory.

It has been recognized that the iterative process of manual curation and refinement is the limiting step of GEM development (Mendoza et al., 2019)). To expedite the development process, several genome-scale reconstruction tools that offer automated draft network generation and gap filling have been reported. However, a systematic assessment of seven GEM reconstruction tools (Mendoza et al., 2019)) concluded that none of the tools performed well in all of the evaluated categories, and there was a relatively large discrepancy between the automatic reconstruction and the high-quality manual curation. Therefore, it is no surprise that the currently validated high-quality GEMs collected in BiGG Models are all manually curated (King et al., 2016)).

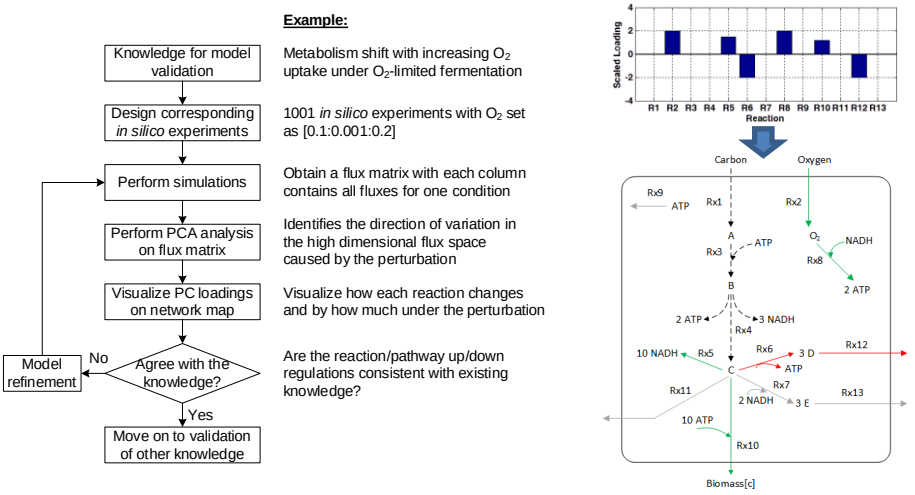
For GEM refinement, one big challenge is to quickly identify the root cause of an erroneous model prediction. Because of the complex interconnectivity in the GEM, many times seemingly unrelated reactions located far away from the "problematic" reactions (i.e., reactions that are not carried out in the expected way) play a key role in correcting the model behavior. Directly comparing the model prediction with experimental data usually yields limited information on the "hidden" relations between the erroneous model prediction and its corresponding root cause reactions. Currently, GEM refinement relies heavily on the modeler's knowledge and capability to sort out clues from various simulation results, and is labor intensive and time consuming.

As cellular metabolism is a complex system with very high degree of freedom and redundancy, we hypothesize that the principles and techniques developed in process systems engineering can be adapted to address some of the challenges associated with GEM refinement. In this work, we first review the foundation of the proposed solution, i.e., a knowledge-matching based computational framework for GEM analysis, then we present the developed GEM refinement approach, and demonstrate the effectiveness of the proposed solution using the development of a GEM for *C. tyrobutyricum*.

2. Knowledge-matching based framework for GEM analysis

In essence, a high-quality GEM is a comprehensive knowledge base of the organism's cellular metabolism. If the key qualitative knowledge captured by a GEM could be extracted and visualized, then GEM refinement could be expedited through knowledge-matching, i.e., by comparing the extracted knowledge with the available ones. The knowledge captured by a GEM is usually embedded in numerical model predictions under various environmental perturbations (e.g., culture media and conditions) and/or genetic perturbations (e.g., mutant), which are difficult to extract and visualize.

To address this challenge, we have developed a system identification (SID) based computational framework for knowledge-matching based GEM analysis (Damiani et al., 2015). In the SID framework, three main steps are involved for GEM analysis, as shown in Figure 1a. First, a set of *in silico* experiments are designed to cover the whole transition path between two metabolic states, where each simulation represents an incremental change along the path. The experiments represent a one-dimensional perturbation to the cellular metabolism and result in a flux matrix containing a series of cellular metabolic states along the transition path; next, system identification tools (such as principal component analysis or PCA) are applied to extract the knowledge contained in the flux matrix (e.g., how the cells respond to the perturbation); finally, the extracted knowledge is visualized against the metabolic network map and compared with the existing knowledge for GEM analysis. Here we use an illustrate example (Figure 1b) to demonstrate how the SID-based GEM analysis works.



(a) (b)

Fig. 1 (a) Illustrated flow chart of the SID-based GEM analysis; (b) an example showing the loadings (top) and their visualization on a simulated network (bottom).

As shown in the bottom part of the Figure 1b, the illustrative example is a toy metabolic network with 13 reactions, which consumes both carbon and oxygen to produce biomass and three potential by-products (C, D and E). To determine if the model correctly captures the metabolic capabilities of the network, we examine if the knowledge captured by the model on how the network respond to an increased O_2 supply under oxygen-limited condition is correct. This response is well understood and well conserved among different microbes, and we expect to see increased flux through electron transport chain due to the availability of additional O_2 (electron acceptor) and increased biomass growth. To do so, we conducted a set of *in silico* experiments, where the carbon uptake was fixed, while the upper limit of O_2 uptake was gradually increased from 0.1 to 0.2 mmol/gDCW/min. These simulations resulted in a 13×1001 flux matrix where each column contains all fluxes within the network under a given O_2 uptake flux. Next, PCA, which we have applied in a novel closed-loop subspace identification algorithm (Wang and Qin, 2006), is applied to extract the knowledge contained in the flux matrix. Due to the linear network structure and one-dimensional perturbation, one principal component (PC) is expected to capture 100% of variation in the flux matrix provided that the whole transition path is located within the same phenotype. Indeed, our result showed that one PC captured 100% of variation. The PC loading captures the knowledge of how the perturbation propagates through the network, i.e., how each reaction flux in the network is affected by the perturbation. The scaled loading is plotted in Figure 1b (top part) and visualized on the network structure in Figure 1b (bottom part). Figure 1b shows that with increasing oxygen supply, the carbon flux in the network shifts from production of by-product D toward biomass production, together with upregulated electron transport chain which converts reducing power (NADH) to produce ATP. This model response agrees with existing understanding as described before, and confirms the model quality. This example further illustrates that although rooted in numerical simulations, the knowledge-matching offered by the SID-based framework is qualitative in nature, which offers enhanced robustness against systematic error among experimental results reported in different literature.

3. SID-based GEM refinement

To identify the root cause of an erroneous model prediction, we adapt the SID-based framework to identify a few candidates which drastically narrows the search space for the potential root causes. The SID-guided GEM refinement process consists of three steps: (1) conducting a set of *in silico* experiments with forced correct model behaviors by applying additional hard constraints; (2) applying PCA to determine how each reaction is affected by the forced correct behavior; (3) identify the candidate root cause reactions – we hypothesize that the reactions affected the most by the forced model behavior are the potential root cause reactions that contribute to the erroneous model behavior. Once the candidate reactions are identified, they will be examined against existing knowledge to see if any of them is the actual root cause. Below we use the development of a GEM for *C. tyrobutyricum* as an example to illustrate how the SID-guided GEM refinement works.

C. tyrobutyricum is a novel and promising industrial chassis strain that plays an important role in carboxylic acids production from lignocellulosic substrates. Since a GEM for *C. tyrobutyricum* is currently not available, we first developed a draft GEM based on a published GEM for *C. beijerinckii*, iCM925 (Milne et al., 2011). Genome analysis revealed that *C. tyrobutyricum* does not use phosphotransbutyrylase nor butyrate kinase for butyric acid production; instead, it uses a CoA transferase to mediate the butyric acid production from butyryl-CoA by reassimilation of acetic acid (Lee et al., 2016). Therefore, the corresponding reaction pathways were deleted from or added to iCM925 to obtain the draft GEM for *C. tyrobutyricum* (iKB917).

Experimental results showed that when cultivated on glucose, *C. tyrobutyricum* excretes butyrate, acetate, hydrogen and carbon dioxide, in addition to cell growth (Lee et al., 2016). However, when the draft GEM was tested by using the experimentally reported glucose uptake rate as the only constraint to predict cell growth and product excretion, the model does not produce butyrate at all, but over-produces acetate, hydrogen and biomass, while under-produces CO_2 (Table 1). Since the draft GEM contains 939 reactions, it is not clear which reaction pathway(s) could be the causes for this erroneous modeling behavior and should be modified. To improve the GEM, we applied the SID-based GEM refinement, as shown in Figure 2.

Table 1. Comparison of measured data with predicted data by the GEM

No.	Condition	Glucose	Acetate	Butyrate	H_2	CO_2	Biomass
1	Experimental measurement	-2.594	0.717	1.895	4.75	4.34	0.052
2	Draft GEM prediction	-2.594	4.223	0	7.96	4.17	0.074
3	Refined GEM prediction	-2.594	0.823	1.834	4.81	4.45	0.053

(1) To force the correct model behavior, i.e., butyrate production, we added a hard constraint to force butyrate production flux gradually increased from 0 to 1.0 mmol/gDCW/min, to produce a flux matrix of 939×1001.

(2) PCA was applied to analyze the flux matrix, and the PC loadings of selected reactions are plotted in Fig. 2 (bottom plot). Clearly, the affected reactions were widespread across multiple pathways. However, the reactions that were affected the most are concentrated in two reaction pathways, i.e., H_2 and H_2O syntheses.

(3) Additional review of literature suggests that due to physiological constraints, the amount of H_2 and CO_2 produced by the cells are proportional to each other (Jo and

Kim, 2016). Since the GEM does not contain any regulatory mechanism, it over produces H_2. To correct this erroneous behavior, we implemented a soft constraint on the ratio of H_2/CO_2 production. Such a soft constraint could provide similar effects as regulatory mechanisms without adding hard constraints, therefore provides more flexibility for the GEM to simulate different phenotypes under different conditions.

(4) The refined GEM (with the added soft constraint) was tested again, which showed significantly improved prediction accuracy (low row of Table 1). The refined GEM not only predicts the excretion of butyrate, but also predicts the excretion of all the byproducts whose fluxes showed excellent agreement with experimental results.

This example highlights the challenge of GEM refinement and effectiveness of the SID-guided refinement. Without the guidance of the SID framework, it can take much longer to figure out what would be the root cause of the erroneous model behavior.

4. Conclusions

GEM has been recognized as a highly effective tool to elucidate the complex cellular metabolisms. It offers a foundation to integrate various omics data, and helps reveal genotype-phenotype relationships, which is fundamental to biology. Despite the recent advancement in automated GEM draft reconstruction and gap filling, manual curation and refinement of GEM remains the limiting step in GEM development. To help address this challenge, we developed a knowledge-matching based computational framework to expedite the GEM refinement. By adapting the principles and techniques in process systems engineering, the proposed SID-guided GEM refinement can quickly identify the "hidden" root cause for the erroneous modeling behavior, therefore significantly expedite the GEM refinement process. The effectiveness of the SID-guided GEM refinement framework is demonstrated through improving a draft GEM of *C. tyrobutyricum*.

References

Damiani, A.L., He, Q.P., Jeffries, T.W., Wang, J., 2015. Comprehensive evaluation of two genome-scale metabolic network models for Scheffersomyces stipitis. Biotechnology and Bioengineering 112, 1250–1262.

Jo, J.H., Kim, W., 2016. Carbon material distribution and flux analysis under varying glucose concentrations in hydrogen-producing Clostridium tyrobutyricum JM1. Journal of Biotechnology 228, 103–111.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., Lewis, N.E., 2016. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research 44, D515.

Lee, J., Jang, Y.-S., Han, M.-J., Kim, J.Y., Lee, S.Y., 2016. Deciphering Clostridium tyrobutyricum Metabolism Based on the Whole-Genome Sequence and Proteome Analyses. mBio 7, e00743--16.

Mendoza, S.N., Olivier, B.G., Molenaar, D., Teusink, B., 2019. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology 2019 20:1 20, 1–20.

Milne, C.B., Eddy, J.A., Raju, R., Ardekani, S., Kim, P.-J., Senger, R.S., Jin, Y.-S., Blaschek, H.P., Price, N.D., 2011. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. BMC Syst Biol 5, 130.

Thiele, I., Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 2010 5:1 5, 93–121.

Wang, J., Qin, S.J., 2006. Closed-loop subspace identification using the parity space. Automatica 42, 315–320.

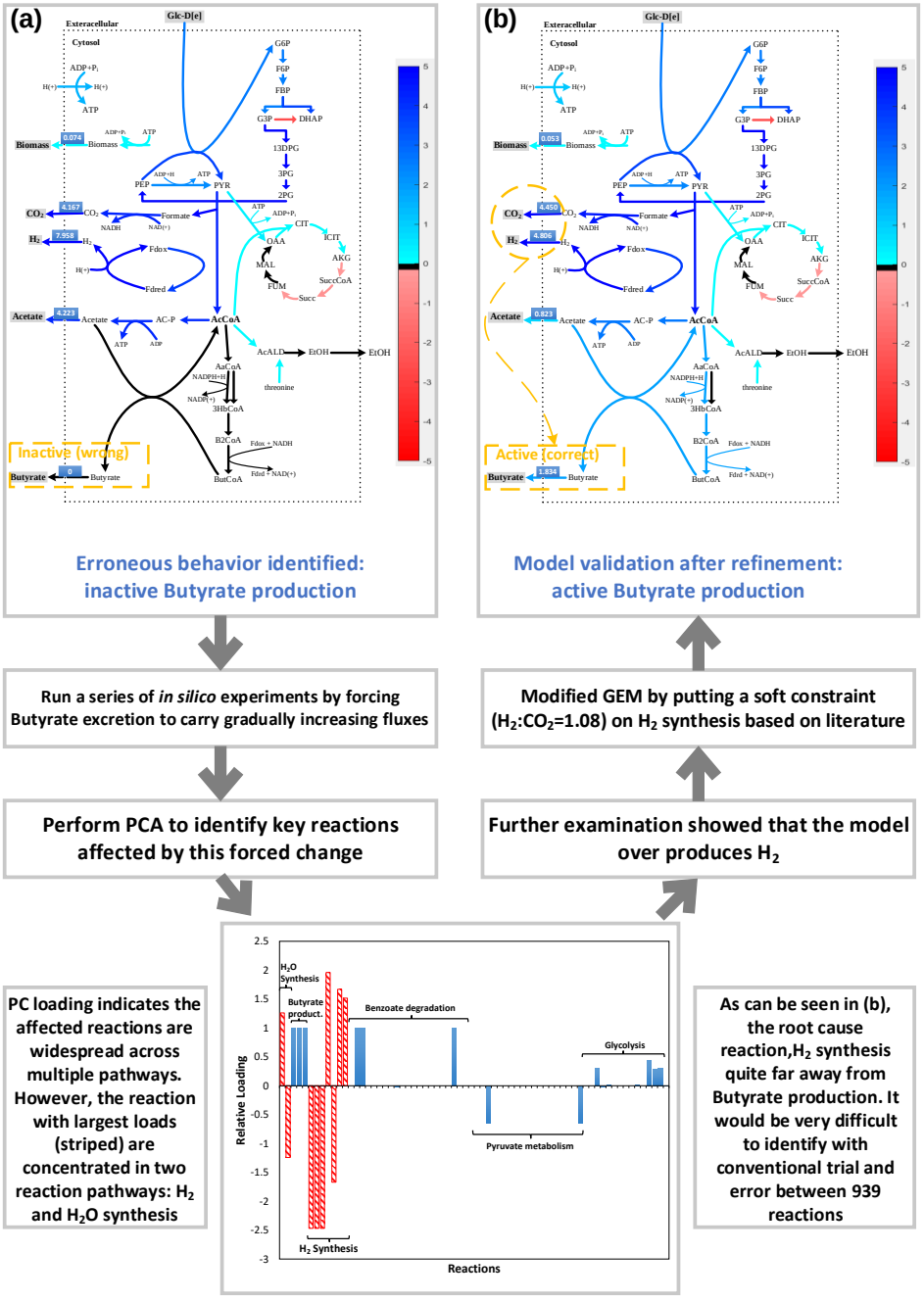


Figure 2. Demonstration of the SID based framework for GEM refinement