Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries

Shu Pan[1,2] and Jennifer L Reed[1,2]

With rapid improvements in next-generation sequencing technologies, our knowledge about metabolism of many organisms is rapidly increasing. However, gaps in metabolic networks exist due to incomplete knowledge (e.g., missing reactions, unknown pathways, unannotated and misannotated genes, promiscuous enzymes, and underground metabolic pathways). In this review, we discuss recent advances in gap-filling algorithms based on genome-scale metabolic models and the importance of both high-throughput experiments and detailed biochemical characterization, which work in concert with *in silico* methods, to allow a more accurate and comprehensive understanding of metabolism.

Addresses
[1] Department of Chemical and Biological Engineering, University of Wisconsin-Madison, 1415 Engineering Dr, Madison, WI 53706, United States
[2] Great Lakes Bioenergy Research Center, Madison, WI 53706, United States

Corresponding author: Reed, Jennifer L (reed@engr.wisc.edu)

Current Opinion in Biotechnology 2018, **51**:103–108

This review comes from a themed issue on **Systems biology**

Edited by **Nathan Price** and **Eran Segal**

https://doi.org/10.1016/j.copbio.2017.12.012

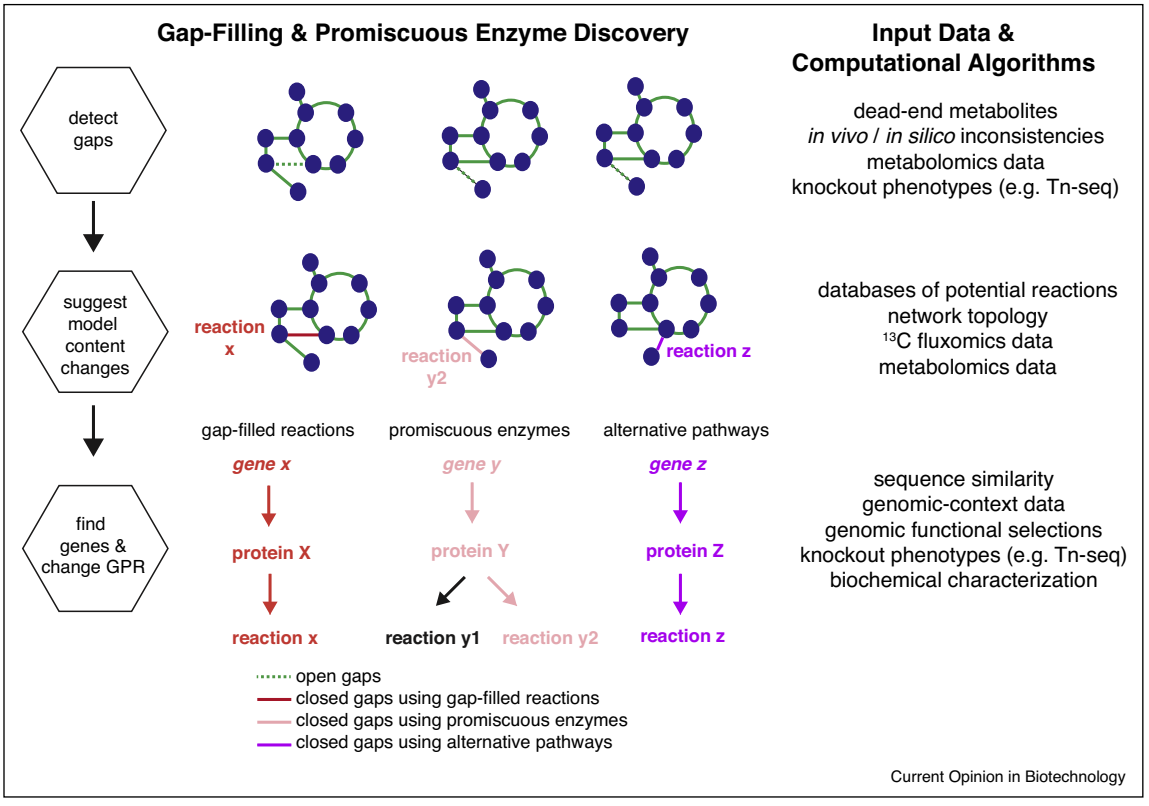
Introduction

A genome-scale metabolic model is a mathematical representation of the metabolic capabilities of an organism, which is inferred primarily from genome annotations. Such models have shown great utility in predicting biological capabilities, metabolic engineering, and systems medicine [1–4]. A draft model is often generated automatically by software platforms, which use genome annotations of a specific organism and connect genes to metabolic reactions using reference databases [5,6]. A draft model has to be further refined and evaluated in multiple steps to ensure its quality [5,6]. This refinement and evaluation process includes gap-filling, which improves the network connectivity by modifying content of the metabolic model. Gap-filling analyses can lead to discoveries of missing reactions, unknown pathways, unannotated and misannotated genes, as well as promiscuous enzymes and underground metabolic pathways. Classic gap-filling algorithms have been reviewed previously by Orth and Palsson [7]. These algorithms generally include three steps: detecting gaps, suggesting model content changes (i.e., add/remove reactions, change biomass compositions, or change reaction reversibility), and identifying genes responsible for the gap-filled reactions (Figure 1). In the first step, gap-filling algorithms identify dead-end metabolites (metabolites which cannot be consumed or produced in the network), and/or inconsistencies between model predictions and experimental data (e.g., growth phenotypes). They then solve for a set of reactions from metabolic databases of potential reactions that if added to the metabolic model 'activate' dead-end metabolites or resolve the inconsistencies. In the third step, some gap-filling algorithms discover genes that could be responsible for these reactions, which can be further tested biochemically or genetically. A simple gap-filling example is illustrated in Figure 2. Here, we first review recent gap-filling methods, which are more efficient or operate under different assumptions. Then we discuss how advances in experimental techniques have significantly advanced gap-filling methods by identifying model-data inconsistencies. Finally, we describe recent studies that have used gap-filling analyses to discover the promiscuous functions of enzymes.

Advances in gap-detection and reaction-addition algorithms

Some recent algorithms aim to detect and fill gaps more efficiently than earlier gap-filling algorithms [7]. For example, FASTGAPFILL [8] is a scalable algorithm that computes a near minimal set of added reactions for a compartmentalized model. Another method, GLOBALFIT [9], reformulates the mixed integer linear programming problem of gap-filling into a simpler bi-level linear optimization problem. It efficiently identifies the minimal set of network changes needed to correct multiple *in silico* predictions that are inconsistent with *in vivo* observations simultaneously. Meneco [10] and a hybrid metabolic network completion algorithm [11] reformulate the reaction-addition problem using answer set programming, a declarative programming paradigm intended to solve difficult combinatorial search problems. Their usage of answer set programming allows for stoichiometry constraints to be

Figure 1



Steps, input data, and computational algorithms of gap-filling. First, dead-end metabolites, *in silico* and *in vivo* inconsistencies, metabolomics data, and knockout phenotypes allow detection of gaps in metabolic models. Then, the model content (i.e., reactions and biomass compositions) is changed to resolve these inconsistencies. In this step, missing reactions can be added from databases, and network topology analysis can rank these potential reactions. Metabolomics and ^{13}C fluxomics data could also suggest reactions that should be included in the model. Finally, the genes responsible for the filled gaps are identified using sequence similarity, genomic-context data, genomic functional selections, or knockout phenotypes and are verified by biochemical characterization. Similarly, promiscuous enzymes and underground metabolic pathways can also be identified when analyzing the gaps in the models.

violated, potentially resulting in solutions which are less biased by the inaccurate stoichiometry of a model. The hybrid metabolic network completion approach combines answer set programming with linear stoichiometry constraints, and offers a better solution for restoring highly degraded models [11].

Algorithms, such as GAUGE [12•], have been developed to exploit alternative mechanisms for gap-identification. GAUGE exploits flux coupling analysis (FCA) [13] that detects how two reactions depend on each other. Using FCA, GAUGE finds gaps involving genes that are associated with fully dependent reactions but show uncorrelated expression patterns. However, GAUGE can only analyze a subset of a model where gene-protein-reaction (GPR) associations are defined and isozymes or multifunctional genes do not create possibilities for uncorrelated gene expression patterns.

Some algorithms exploit alternative mechanisms for adding reactions. The novel algorithm DEF [14] is based upon filling reactions to 'activate' dead-end metabolites in a manner similar to eukaryotes engulfing mitochondria to find the most efficient pathways for consuming oxygen. Following this quasi-endosymbiosis theory, DEF aims to add reactions that maximize production/consumption of dead-end metabolites in the original model. A DEF solution could contain more reactions that are biologically reasonable compared to a parsimonious gap-filling solution, which often is a minimal set of reactions.

Inherently different from all algorithms mentioned above, pattern-based gap-filling algorithms do not contain an explicit gap-identification or reaction-addition step. In a metabolite pattern and probabilistic method [15], feature propagation Markov models (HMMs) are used to rank potential gap-filled reactions by how closely they are related to the network. In MATBoost and BoostGAPFILL [16,17•], a training incidence matrix, S, with artificial gaps is created by deleting reactions randomly from a network. Then a machine learning technique, matrix factorization, completes the missing entries creating