Title

Reconstructor: A COBRApy compatible tool for automated genome-scale metabolic network reconstruction with parsimonious flux-based gap-filling

Authors

Matthew L Jenior+[1]

Emma M Glass+[1]

Jason A Papin*[1,2,3]

+ denotes co-first authorship

* denotes corresponding author

Affiliations

1. Department of Biomedical Engineering
2. Department of Medicine, Division of Infectious Diseases & International Health
3. Department of Biochemistry & Molecular Genetics

University of Virginia, Charlottesville, VA, 22902 USA

Abstract

Summary

Genome-scale metabolic network reconstructions (GENREs) are valuable for understanding cellular metabolism *in silico*. Several tools exist for automatic GENRE generation. However, these tools frequently (1) do not readily integrate with some of the widely-used suites of packaged methods available for network analysis, (2) lack effective network curation tools, and (3) are not sufficiently user-friendly. Here, we present Reconstructor, a user-friendly COBRApy compatible tool with ModelSEED namespace compatibility and a pFBA-based gap-filling technique. We demonstrate how Reconstructor readily generates high-quality GENRES that are useful for further biological discovery.

Availability and Implementation

The Reconstructor package is freely available for download via pip in the command line (pip install reconstructor). Usage instructions and benchmarking data are available at http://github.com/emmamglass/reconstructor.

Contact

Jason Papin: papin@virginia.edu

Introduction

Genome-scale metabolic network reconstructions (GENREs) are valuable tools for understanding the link between the genotype and phenotype of an organism. GENREs enable greater understanding of the effects of genetic and environmental perturbation on cellular function and can help to identify novel drug targets, among many other applications (Haggart *et al.*, 2011; Gu *et al.*, 2019; Kim *et al.*, 2012).

The synthesis of GENREs can be an incredibly laborious and complex process, requiring the integration of data from multiple sources (Thiele and Palsson, 2010). The creation of a GENRE begins with the annotated genome sequence to predict reactions to include in the draft GENRE, and then further model curation steps are performed to gap-fill missing reactions. While GENREs can be generated and curated manually, methods for the automated creation of GENREs have emerged (Mendoza *et al.*, 2019).

Several platforms exist for automated GENRE creation, including ModelSEED (Seaver *et al.*, 2021) and CarveMe (Machado *et al.*, 2018), among others (Chevallier *et al.*, 2018; Dias *et al.*, 2015; Olivier, 2018; Karp *et al.*; Wang *et al.*, 2018) (Figure 1B). However, additional compatibility modules are necessary to use CarveMe- and ModelSEED-created GENREs with the COBRApy analysis toolbox (Moretti *et al.*, 2016; Mundy *et al.*, 2017), due to the web-based nature of ModelSEED and the use of the BiGG reaction database by CarveMe. Additionally, conventional automated GENRE creators use gene-protein-reaction (GPR) associations for reaction scoring and subsequent gap-filling, which could lead to the inclusion of reactions in the final model that are unnecessary. GPR associations cannot be solely relied upon for gap-filling because their formulation relies on sufficient genome annotation data and experimentally obtained information, which can be lacking for certain organisms (Mendoza *et al.*, 2019).

Here, we introduce Reconstructor, an automated GENRE creation tool that creates COBRApy-compatible reconstructions in the ModelSEED namespace. Additionally, we include a gap-filling technique based on parsimonious flux balance analysis (pFBA), a more biologically tractable gap-filing technique than other techniques based exclusively on GPR mapping.

Results

Universal reaction database construction

A universal database of metabolic reactions was created based on all available reactions and metabolites in the ModelSEED database. Reconstructor GENREs are in the ModelSEED namespace, but are also directly compatible with COBRApy, without needing the use of additional compatibility modules. All ModelSEED reactions and metabolites were added to the universal database via reaction and metabolite dictionaries, missing exchange reactions were identified and corrected, and biomass function was updated. It is important to note that the gram-positive and gram-negative biomass functions were differentially defined. The universal reaction database was further curated to remove mass-imbalanced reactions and reactions with no reactants. Intracellular sink reactions were added. The result was a universal database that contains a reaction collection from which the genome-informed model can select reactions for gap-filling. The user can also curate their own universal database to use with Reconstructor by altering the ModelSEED reaction and metabolite dictionaries. The ability to curate readily this provided database or to make use of any other user-provided universal database in the same name-space is a key feature of Reconstructor.

Input data formats and draft GENRE scaffold extraction

Reconstructor automates the build of a GENRE from three different types of user defined input. Type 1 requires inputs of an annotated genome sequence in the form of an amino acid FASTA file and the gram status of the bacterial species. Type 2 requires an input of BLASTp hits and the gram status of the bacterial species, bypassing the BLASTp search step. Type 3 requires an existing GENRE in sbml format, and further gap-filling is performed based on pFBA gap-filling (as described further below). Additionally, the user can define their own media conditions for a given GENRE by providing metabolite names present in their defined media condition.

The GENRE creation process is described below from the starting point of a Type 1 input. The amino acid FASTA is aligned to the KEGG database by performing a BLASTp search with the DIAMOND sequence aligner tool (Buchfink *et al.*, 2014). Then, the gene hits are processed and translated into reactions. These reactions and associated gene names are used to create a draft GENRE based solely on gene associated reactions. Additionally, reactions are added to the draft GENRE based on defined media conditions.

pFBA-based approach to gap-filling draft GENREs

Several gap-filling methods exist (Pan and Reed, 2018), many of which use parsimony as a guiding principle in which a minimum number of reactions are added to satisfy criteria like growth in defined media (Prigent *et al.*, 2017; King *et al.*, 2018; Zimmermann *et al.*, 2021). In Reconstructor, the draft GENRE is gap-filled by adding reactions to the draft GENRE from the universal reaction database through a pFBA approach. The pFBA gap-filler first modifies the universal reaction database by removing any reactions that are present already in the gene associated reaction draft GENRE. Then, draft GENRE reactions are added to the universal reaction database. The optimal objective flux for the draft GENRE is calculated and used to constrain further optimization to a fraction of this level of flux. Each reaction is assigned a weight in the new objective function with non-gene associated, universal reactions at a maximum weight of 1. The sum of the reaction and linear coefficient combinations for each forward and reverse reaction pair is used as the new objective function. This linear coefficient assignment forces flux through gene-associated reactions and minimizes the inclusion of non-gene associated reactions. New reactions are added to the draft GENRE if the absolute value of the solution flux is greater than 1×10^{-6}.

Secondary Gap-filling, component annotation, and GENRE output

Further gap-filling is performed based on defined media conditions. The GENRE is then annotated with KEGG genes, ModelSEED metabolites/reactions, and reactions implicated in biomass are defined. Finally, exchange reactions are corrected, and basic model checks are completed to report the number of genes, reactions, and metabolites in the draft and final GENREs, how many reactions were gap-filled, and the final objective flux. Finally, the model is saved to sbml format, the current community standard (Hucka *et al*., 2003).

COBRApy compatibility

Current widely-used GENRE creation tools, ModelSEED and CarveMe, both require additional modules to be used in conjunction with COBRApy (Moretti *et al.*, 2016; Mundy *et al.*, 2017). GENREs created with Reconstructor are directly compatible with COBRApy; they can be directly imported into python after creation and do not require additional compatibility modules to take advantage of the powerful COBRApy analysis toolbox. Reconstructor's direct COBRApy compatibility allows users to streamline automated GENRE analysis pipelines, potentially accelerating GENRE-based discovery and hypothesis generation.

Reconstructor generates high quality reconstructions

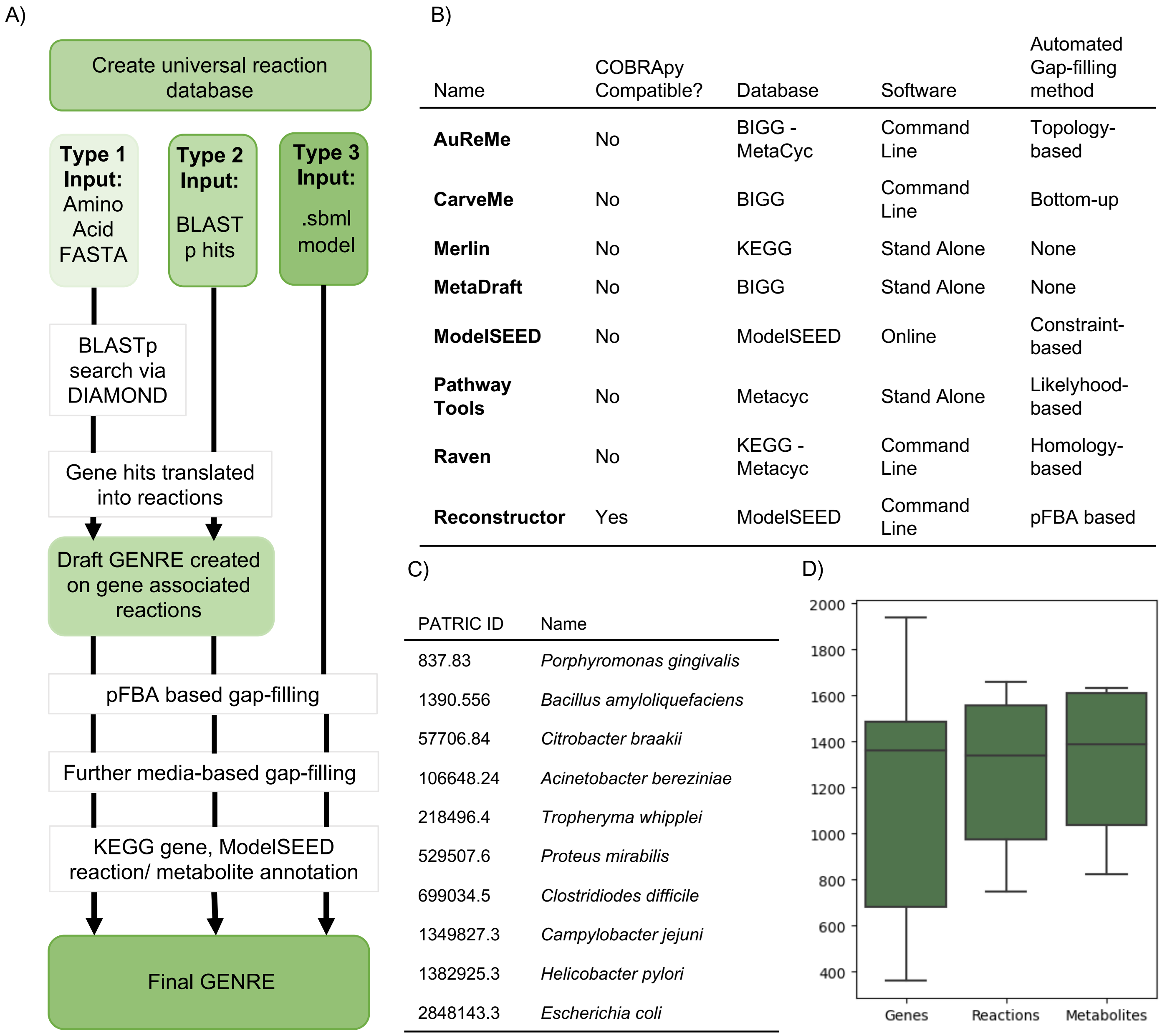
As a demonstration of the utility of Reconstructor, 10 GENREs representing unique bacterial strains were created by Reconstructor for analysis and benchmarking through the metabolic model testing suite (MEMOTE). MEMOTE scores for each of the 10 reconstructions (Figure 1C) are available at http://github.com/emmamglass/reconstructor. Each of the 10 generated GENREs were imported into COBRApy and the number of genes, metabolites, and reactions were determined (Figure 1D). The number of genes, reactions, and metabolites present within Reconstructor-generated models is consistent with GENREs created with other tools (Machado *et al.*, 2018).

Conclusion

Reconstructor automatically creates and curates COBRApy-compatible, genome-scale metabolic network reconstructions in the ModelSEED namespace and uses a pFBA based gap-filling technique (Figure 1A) that is more efficient and consistent with parsimony principles in metabolic modeling than conventional gap-filling techniques (Jenior *et al.,* 2020). Direct COBRApy compatibility enables the user to import GENREs directly into python for further downstream analysis via the robust COBRApy toolbox. Reconstructor generates high-quality GENREs as evidenced through MEMOTE benchmarking, and Reconstructor GENREs are characteristically similar to those created by other tools.

Funding Information

This work was supported by the National Science Foundation (NSF) [1842490]; National Institutes of Health [grant numbers T32 GM 145443-1, R01-AI154242, R01-AAT010253]; and the TransUniversity Microbiome Initiative.


A)
Create universal reaction database
Type 1 Input: Amino Acid FASTA
Type 2 Input: BLASTp hits
Type 3 Input: .sbml model
BLASTp search via DIAMOND
Gene hits translated into reactions
Draft GENRE created on gene associated reactions
pFBA based gap-filling
Further media-based gap-filling
KEGG gene, ModelSEED reaction/ metabolite annotation
Final GENRE
B)
C)
D)
2000
1800
1600
1400
1200
1000
800
600
400
Genes
Reactions
Metabolites

Name	COBRApy Compatible?	Database	Software	Automated Gap-filling method
AuReMe	No	BIGG - MetaCyc	Command Line	Topology-based
CarveMe	No	BIGG	Command Line	Bottom-up
Merlin	No	KEGG	Stand Alone	None
MetaDraft	No	BIGG	Stand Alone	None
ModelSEED	No	ModelSEED	Online	Constraint-based
Pathway Tools	No	Metacyc	Stand Alone	Likelyhood-based
Raven	No	KEGG - Metacyc	Command Line	Homology-based
Reconstructor	Yes	ModelSEED	Command Line	pFBA based

PATRIC ID	Name
837.83	*Porphyromonas gingivalis*
1390.556	*Bacillus amyloliquefaciens*
57706.84	*Citrobacter braakii*
106648.24	*Acinetobacter bereziniae*
218496.4	*Tropheryma whipplei*
529507.6	*Proteus mirabilis*
699034.5	*Clostridiodes difficile*
1349827.3	*Campylobacter jejuni*
1382925.3	*Helicobacter pylori*
2848143.3	*Escherichia coli*

Figure 1. Reconstructor overview | A) Flowchart detailing the functionality of the Reconstructor tool. B) Comparison of other widely used GENRE construction including Reconstructor, adapted from (Mendoza *et al.*, 2019). C) GENREs were created via Reconstructor for each of the 10 bacterial species listed, genome sequences were downloaded from the Pathosystems Resource Integration Center (PATRIC) (Davis *et al.*, 2020), PATRIC IDs for each species are listed. D) Boxplots of the number of genes, reactions, and metabolites in the 10 GENREs.

References

Buchfink,B. *et al.* (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Davis,J.J. *et al.* (2020) The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Res.*, **48**, D606–D612.

Gu,C. *et al.* (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.*, **20**, 1–18.

Haggart,C.R. *et al.* (2011) Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Enzym.*, **500**, 411–433.

Hucka , M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics.,* **19**, 524-531.

Jenior, M. L., *et al.* (2020) Transcriptome-guided parsimonious flux balance analysis improves predictions with metabolic networks in complex environments. *Plos. Comp. Bio.*, **16**, 1-26.

Kim,T.Y. *et al.* (2012) Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr. Opin. Biotechnol.*, **23**, 617–623.

King,B. *et al.* (2018) ProbAnnoWeb and ProbAnnoPy: Probabilistic annotation and gap-filling of metabolic reconstructions. *Bioinformatics*, **34**, 1594–1596.

Machado,D. *et al.* (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, **46**, 7542–7553.

Mendoza,S.N. *et al.* (2019) A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.*, **20**, 1–20.

Moretti,S. *et al.* (2016) MetaNetX/MNXref - Reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.

Mundy,M. *et al.* (2017) Mackinac: A bridge between ModelSEED and COBRApy to generate and analyze genome-scale metabolic models. *Bioinformatics*, **33**, 2416–2418.

Pan,S. and Reed,J.L. (2018) Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr. Opin. Biotechnol.*, **51**, 103–108.

Prigent,S. *et al.* (2017) Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks.

Seaver,S.M.D. *et al.* (2021) The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.*, **49**, D575–D588.

Thiele,I. and Palsson,B. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Zimmermann,J. *et al.* (2021) Gapseq: Informed Prediction of Bacterial Metabolic Pathways and Reconstruction of Accurate Metabolic Models. *Genome Biol.*, **22**, 1–35.