Bioinformatics, 2023, **39(6)**, btad367
https://doi.org/10.1093/bioinformatics/btad367
Advance access publication 5 June 2023
Applications Note

OXFORD

Systems biology

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

Matthew L. Jenior[1,†], Emma M. Glass [1,†], Jason A. Papin [1,2,3,*]

[1]Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, United States
[2]Department of Medicine, Division of Infectious Diseases & International Health, University of Virginia, Charlottesville, Virginia, United States
[3]Department of Biochemistry & Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States

*Corresponding author. Department of Biomedical Engineering, University of Virginia, Box 800759, Charlottesville, VA 22908, USA. E-mail: papin@virginia.edu (J.A.P.)
[†]Equal contribution.

Associate Editor: Janet Kelso

Abstract

Motivation: 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 (i) do not readily integrate with some of the widely-used suites of packaged methods available for network analysis, (ii) lack effective network curation tools, (iii) are not sufficiently user-friendly, and (iv) often produce low-quality draft reconstructions.

Results: Here, we present Reconstructor, a user-friendly, COBRApy-compatible tool that produces high-quality draft reconstructions with reaction and metabolite naming conventions that are consistent with the ModelSEED biochemistry database and includes a gap-filling technique based on the principles of parsimony. Reconstructor can generate SBML GENREs from three input types: annotated protein .fasta sequences (Type 1 input), a BLASTp output (Type 2), or an existing SBML GENRE that can be further gap-filled (Type 3). While Reconstructor can be used to create GENREs of any species, we demonstrate the utility of Reconstructor with bacterial reconstructions. We demonstrate how Reconstructor readily generates high-quality GENRES that capture strain, species, and higher taxonomic differences in functional metabolism of bacteria and are useful for further biological discovery.

Availability and implementation: The Reconstructor Python package is freely available for download. Complete installation and usage instructions and benchmarking data are available at http://github.com/emmamglass/reconstructor.

1 Introduction

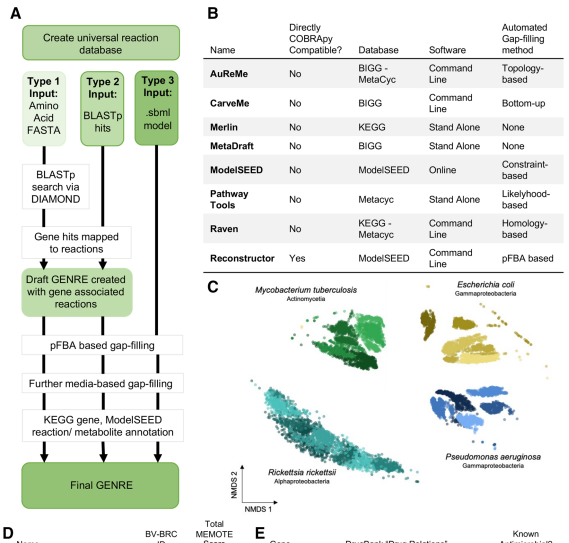
Genome-scale metabolic network reconstructions (GENREs) are valuable tools for understanding the link between the genotype and phenotype of an organism. GENREs can enable greater understanding of the effects of genetic and environmental perturbation on cellular function and can help identify novel drug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

The generation 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), CarveMe (Machado *et al.* 2018), and among others (Dias *et al.* 2015, Aite *et al.* 2018; Olivier 2018, Wang *et al.* 2018, Karp et al. 2021) (Fig. 1B). However, the reconstructions generated by these tools typically require additional compatibility modules for integration with COBRApy (Ebrahim *et al.* 2013), and subsequent manual or automated curation (King *et al.* 2015, Moretti *et al.* 2016, Mundy *et al.* 2017, Camborda *et al.* 2022, Saadat *et al.* 2022, Schneider *et al.* 2022; see Supplementary).

Here, we introduce Reconstructor, an automated GENRE creation tool that generates COBRApy-compatible reconstructions in the ModelSEED namespace. Additionally, we include a two-step gap-filling technique based on parsimonious flux balance analysis (pFBA) (Lewis *et al.* 2010), a more biologically tractable gap-filling technique than other techniques based exclusively on gene–protein-reaction mapping. pFBA is motivated by the possible notion that high metabolic flux has high enzyme turn-over and that the synthesis of enzymes is energetically costly. Consequently, the model will minimize overall flux (and thus costly high enzyme turn-over), but still maximize for a given objective function (Lewis *et al.* 2010). Using a pFBA approach to gap-filling ensures that we account for all reactions with genetic evidence while generating a reconstruction that is consistent with these principles of

Received: September 19, 2022. **Revised:** May 22, 2023. **Editorial Decision:** May 31, 2023. **Accepted:** June 2, 2023



B

Name	Directly 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

D

Name	BV-BRC ID	Total MEMOTE Score
Porphyromonas gingivalis	837.83	85%
Bacillus amyloliquefaciens	1390.556	84%
Citrobacter braakii	57706.84	83%
Acinetobacter bereziniae	106648.24	84%
Tropheryma whipplei	218496.4	85%
Proteus mirabilis	529507.6	83%
Clostridiodes difficile	699034.5	84%
Campylobacter jejuni	1349827.3	84%
Helicobacter pylori	1382925.3	85%
Escherichia coli	2848143.3	83%

E

Gene	DrugBank "Drug Relations"	Known Antimicrobial?
pabB	Formic acid	**Yes**[1,2]
pabC	D-[3-hydroxy-2-methyl-5-phosphonooxymethyl-pyridin-4-ylmethyl]-N,O-cycloseryamide	No
serC	Triethylene glycol	**Yes**[3]
aroC	Flavin mononucleotide, Cobalt hexamine ion	No, No
fabI	Soneclosan, Triclosan	No, **Yes**[4]
panC	Tris-Hydroxymethyl-Methyl-Ammonium, 2,4-Dihydroxy-3,3-Dimethyl-Butyrate, alpha,beta-Methyleneadenosine 5'-triphosphate, Pantoyl Adenylate, beta-Alanine	**Yes**[5], No, No, No, No
nadC	Quinolinic acid	**Yes**[6]
gmk	Formic acid	**Yes**[1,2]
pyrE	5-O-phosphono-alpha-D-ribofuranosyl diphosphate, Orotic acid	No, No
pyrB	Alanosine, Sparfosic acid	**Yes**[7], No
pyrD	Formic acid, Orotic acid, Flavin mononucleotide	**Yes**[1,2], No, No

Figure 1. Reconstructor overview. (A) Flowchart detailing the functionality of the Reconstructor tool. (B) Comparison of other widely used GENRE construction tools including Reconstructor, adapted from Mendoza *et al.* (2019). (C) Flux sampling was performed on 20 bacterial reconstructions generated with Reconstructor (five strains for each of four species). Each dot represents a sampled flux distribution. Five hundred sampled flux distributions were captured for each reconstruction. Samples were dimensionally reduced using non-metric multidimensional scaling and plotted on a 2D plane for visualization. (D) GENREs were created via Reconstructor for each of the 10 bacterial species listed, genome sequences were downloaded from the BV-BRC (Davis *et al.* 2020), BV-BRC IDs for each species are listed. (E) Essential genes for a generated *Pseudomonas aeruginosa* (strain NCGM2.S1) reconstruction, existing drugs that target those essential genes according to DrugBank, and whether the identified drugs are known antimicrobials. [1]Ricke *et al.* (2020), [2]Alzahrani *et al.* (2022), [3]Chrószcz *et al.* (2022), [4]Kondratenko *et al.* (2020), [5]Narui *et al.* (2009), [6]Murthy *et al.* (1945), and [7]Yazdankhah *et al.* (2006).

parsimony. We also demonstrate how metabolic network models derived from Reconstructor can be used to generate experimentally testable hypotheses.

2 Results

2.1 Universal reaction database construction

A universal database of metabolic reactions was adapted from the ModelSEED database, removing all reactions that are unbalanced (see Supplementary). From this modified ModelSEED database, we generated reaction and metabolite dictionaries, missing exchange reactions were identified and corrected, and the biomass function was updated. The result was a universal database that contains a reaction collection from which the genome-informed model can select reactions for inclusion and 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 readily curate this existing universal database or to make use of any other user-provided universal database in the same name-space is a key feature of Reconstructor.

2.2 Input data formats and draft GENRE scaffold extraction

Reconstructor automates the build of a GENRE from three different types of user-defined inputs. Type 1 requires inputs of an annotated genome sequence in the form of an amino acid FASTA file. It is important to note that the user must annotate the genome beforehand, as genome DNA files are currently not supported as inputs in Reconstructor. Type 2 requires an input of BLASTp hits, bypassing the BLASTp search step. Type 3 requires an existing GENRE in SBML format in the same namespace and with the same construction (e.g. definition of intracellular/extracellular compartments) as Reconstructor network reconstructions, and further pFBA gap-filling is performed (as described in Supplementary). Additionally, for input Types 1 and 2, the user can define their own media conditions for a given GENRE by providing metabolite names in their defined media condition (further discussed in Supplementary).

The GENRE creation process is described below from the starting point of a Type 1 input. The amino acid FASTA file is aligned to the KEGG database by performing a BLASTp search with the DIAMOND sequence aligner tool (Buchfink *et al.* 2015). Then, the KEGG gene hits are processed and translated into ModelSEED 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 media conditions.

2.3 Parsimonious flux balance analysis-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, we introduce a two-step gap-filling process based on (i) parsimonious flux principles and (ii) user-defined media conditions. Our gap-filling technique works by minimizing the flux through an optimal reaction set (all gene-associated reactions and a set of non-gene associated reactions that minimizes flux), rather than minimizing the number of reactions added to the network (see Supplementary). After the optimal reaction set is chosen, reactions are added to the GENRE.

2.4 Component annotation and final GENRE output

The final gap-filled GENRE is then annotated with KEGG (Kanehisa *et al.* 2016) gene IDs, ModelSEED metabolites, and reaction names. Finally, basic model statistics are reported including the number of genes, reactions, and metabolites in the draft and final GENREs, how many reactions were the result of gap-filling, and the final biomass objective flux so the user can ensure the gap-filling process was successful. Finally, the model is saved in SBML format, the current community standard (Hucka *et al.* 2003).

2.5 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). Reconstructor GENREs are directly compatible with COBRApy; they can be generated via command line and easily imported into Python, or generated directly in a Python script using the reconstruct() function to easily and

immediately take advantage of the powerful COBRApy analysis toolbox. Reconstructor's direct COBRApy compatibility allows users to streamline GENRE analysis pipelines, potentially accelerating GENRE-based discovery and hypothesis generation.

2.6 Reconstructor utility

We demonstrate the utility of Reconstructor GENREs by addressing three key aspects: (i) quality of reconstructions for a range of bacteria with different levels of literature investigation, (ii) ability of GENREs to capture strain-level differences, and (iii) ability to quickly generate testable biological hypotheses.

While Reconstructor could be used for any annotated amino acid .fasta file, we demonstrate here the utility of Reconstructor with bacterial reconstructions. We generated a total of 10 GENREs representing unique bacterial strains for analysis and benchmarking through the metabolic model testing suite (MEMOTE) (Lieven *et al.* 2020). We selected a diverse set of bacterial species to ensure we can generate high-quality reconstructions for both well studied/annotated species like *Clostridium difficile* and lesser-known species like *Tropheryma whipplei*. MEMOTE scores and SBML reconstructions for each of the 10 species (Fig. 1C) are available at http://github.com/emmamglass/reconstructor. The overall MEMOTE scores for the reconstructions ranged from 83% to 85% (Fig. 1D). MEMOTE score comparisons between similar ModelSEED and CarveMe reconstructions are discussed in Supplementary.

While the benchmarking quality of Reconstructor GENREs is high, we wanted to ensure that Reconstructor creates GENREs that are capable of capturing strain-, species-, and class-level variation in metabolic functionality. To address this question, we further generated reconstructions for five distinct strains of each of four bacterial species, *Pseudomonas aeruginosa, Mycobacterium tuberculosis, Escherichia coli*, and *Rickettsia rickettsii*, for a total of 20 reconstructions. Through flux balance analysis and visualization with nonmetric multidimensional scaling, we show that the Reconstructor network reconstructions are able to capture functional metabolic differences in strain, species, and class, as evidenced by distinct clustering of flux samples (Fig. 1C) (see Supplementary).

Since we determined that Reconstructor GENREs are able to capture differences in metabolic functionality with significant detail, we wanted to demonstrate the utility of Reconstructor GENREs for generating testable biological hypotheses rapidly through integration with COBRApy. We generated a metabolic network reconstruction of a *Pseudomonas aeruginosa* strain, NCGM2.S1, that has not been previously created. Because of Reconstructor's direct COBRApy compatibility, we were able to apply COBRApy tools to run a gene essentiality analysis. We then mapped these essential genes to targets of existing drugs in DrugBank (Wishart *et al.* 2018). We determined that 7 identified drugs are known inhibitors of microbial growth, while 13 other drugs had not been tested previously (Fig. 1E). These untested drugs represent new hypotheses that can readily be tested experimentally (see Supplementary).

3 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 (Fig. 1A) that is more 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, captures class-, species-, and even strain-level differences in functional metabolism and can be used for rapid experimental hypothesis generation.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation (NSF) [1842490], National Institutes of Health [grant numbers T32 GM 145443-1, R01-AI154242, R01-AT010253], and the TransUniversity Microbiome Initiative.

Data availability

The data presented in this article are available at http://github.com/emmamglass/reconstructor and in the online supplementary material.

References

Alzahrani O, Elumalai P, Nada H *et al. Pseudomonas putida*: sensitivity to various antibiotics, genetic diversity, virulence, and role of formic acid to modulate the immune-antioxidant status of the challenged nile tilapia compared to carvacrol oil. *Fishes* 2022;**8**:6.

Buchfink B, Xie C, Huson DH *et al.* Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.

Camborda S, Weder J-N, Töpfer N *et al.* CobraMod: a pathway-centric curation tool for constraint-based metabolic models. *Bioinformatics* 2022;**38**:2654–6.

Aite M, Chevallier M, Frioux C *et al.* Exploration for "À -La-Carte" reconstructions of genome-scale metabolic models. *PLoS Comput Biol* 2018;**14**:e1006146.

Chrószcz MW, Barszczewska-Rybarek IM, Kazek-Kęsik, A *et al.* Novel antibacterial copolymers based on quaternary ammonium urethane-dimethacrylate analogues and triethylene glycol dimethacrylate. *Int J Mol Sci* 2022;**23**:4954.

Davis JJ, Wattam AR, Aziz RK. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res* 2020;**48**:D606–12.

Dias O, Rocha M, Ferreira EC *et al.* Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res* 2015;**43**:3899–910.

Ebrahim A, Lerman JA, Palsson BO *et al.* COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst Biol* 2013;**7**:74.

Gu C, Kim GB, Kim WJ *et al.* Current status and applications of genome-scale metabolic models. *Genome Biol* 2019;**20**:18.

Haggart CR, Bartell JA, Saucerman JJ *et al.* Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Enzymol* 2011;**500**:411–33.

Hucka M, Finney A, Sauro HM *et al.* The systems biology markup language (SBML): a medium for representation and exchange biochemical network models. *Bioinformatics* 2003;**19**:4. https://doi.org/10.1093/bioinformatics/btg015.

Jenior ML, Moutinho TJ, Dougherty BV *et al.* Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. *Plos Computat Biol* 2020;**16**:4. https://doi.org/10.1371/journal.pcbi.1007099.

Kanehisa M, Sato Y, Kawashima M *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**: D457–62.

Karp PD, Sato Y, Kawashima M *et al.* Pathway tools version 23.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 2021;**22**:109–26.

Kim TY, Sohn SB, Kim YB *et al.* Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr Opin Biotechnol* 2012;**23**:617–23.

King B, Farrah T, Richards MA *et al.* ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions. *Bioinformatics* 2018;**34**:1594–6.

King ZA, Dräger A, Ebrahim A *et al.* Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput Biol* 2015;**11**:e1004321.

Kondratenko YA, Nikonorova AA, Zolotarev AA *et al.* Tris(hydroxymethyl)methyl ammonium salts of biologically active carboxylic acids. Synthesis, properties and crystal structure. *J Mol Struct* 2020;**1207**:12813.

Lewis NE, Hixson KK, Conrad TM *et al.* Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* 2010;**6**:390.

Lieven C, Beber ME, Olivier BG *et al.* MEMOTE for standardized genome-scale metabolic model testing. *Nat Biotechnol* 2020;**38**:272–6.

Machado D, Andrejev S, Tramontano M *et al.* Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 2018;**46**:7542–53.

Mendoza SN, Olivier BG, Molenaar D *et al.* A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol* 2019;**20**:1–20.

Moretti S, Martin O, Van Du Tran T *et al.* MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res* 2016;**44**:D523–6.

Murthy YK, Thiemann JE, Coronelli C *et al.* Pharmacology alanosine, a new antiviral and antitumour agent isolated from a Streptomyces. *Nature* 1966;211:1198–9.

Mundy M, Mendes-Soares H, Chia N *et al.* Mackinac: a bridge between ModelSEED and COBRApy to generate and analyze genome-scale metabolic models. *Bioinformatics* 2017;**33**:2416–8.

Narui K, Noguchi N, Saito A *et al.* Anti-infectious activity of tryptophan metabolites in the L-tryptophan-L-kynurenine pathway. *Biol Pharm Bull* 2009;**32**:41–4.

Olivier BG. 2018. SystemsBioinformatics/cbmpy-metadraft: MetaDraft is now available. https://zenodo.org/record/2398336 (3 October 2022, date last accessed).

Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr Opin Biotechnol* 2018;**51**:103–8.

Prigent S, Frioux C, Dittami SM *et al.* Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. *PLoS Comput Biol* 2017;**13**:e1005276.

Ricke SC, Dittoe DK, Richardson KE *et al.* Formic acid as an antimicrobial for poultry production: a review. *Front Vet Sci* 2020; **7**:563.

Saadat NP, van Aalst M, Ebenhöh O *et al.* Network reconstruction and modelling made reproducible with moped. *Metabolites* 2022;**12**: 275.

Schneider P, Bekiaris PS, von Kamp A *et al.* StrainDesign: a comprehensive python package for computational design of metabolic networks. *Bioinformatics* 2022;**38**:4981–3.

Seaver SMD, Liu F, Zhang Q *et al.* 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* 2021;**49**:D575–88.

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

Wang H, Marcišauskas S, Sánchez BJ *et al.* RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput Biol* 2018;**14**:e1006541.

Wishart DS, Feunang YD, Guo AC *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**: D1074–82.

Yazdankhah SP, Scheie AA, Arne Høiby E *et al.* Triclosan and antimicrobial resistance in bacteria: an overview. *Microb Drug Resist* 2006;**12**:83–90.

Zimmermann J, Kaleta C, Waschina S *et al.* Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* 2021;**22**:35.