OXFORD

Systems biology

DEMETER: efficient simultaneous curation of genome-scale reconstructions guided by experimental data and refined gene annotations

Almut Heinken [1,2], **Stefanía Magnúsdóttir**[3], **Ronan M. T. Fleming** [1,4] and **Ines Thiele** [1,2,5,6,]*

[1]School of Medicine, National University of Galway, H91 TK33 Galway, Ireland, [2]Ryan Institute, National University of Galway, H91 TK33 Galway, Ireland, [3]Center for Molecular Medicine, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands, [4]Leiden Academic Centre for Drug Research, Leiden University, 2333 CC Leiden, The Netherlands, [5]Division of Microbiology, National University of Galway, H91 TK33 Galway, Ireland and [6]APC Microbiome Ireland, University College Cork, T12 K8AF Cork, Ireland

*To whom correspondence should be addressed.
Associate Editor: Olga Vitek

Abstract

Motivation: Manual curation of genome-scale reconstructions is laborious, yet existing automated curation tools do not typically take species-specific experimental and curated genomic data into account.

Results: We developed Data-drivEn METabolic nEtwork Refinement (DEMETER), a Constraint-Based Reconstruction and Analysis (COBRA) Toolbox extension, which enables the efficient, simultaneous refinement of thousands of draft genome-scale reconstructions, while ensuring adherence to the quality standards in the field, agreement with available experimental data and refinement of pathways based on manually refined genome annotations.

Availability and implementation: DEMETER and tutorials are freely available at https://github.com/opencobra.

Contact: ines.thiele@nuigalway.ie

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction

The Constraint-Based Reconstruction and Analysis (COBRA) approach relies on genome-scale metabolic reconstructions that have been curated based on genomic, biochemical and physiological data, a laborious process consisting of 96 steps (Thiele and Palsson, 2010). On the other hand, existing automated reconstruction pipelines, such as ModelSEED (Henry *et al.*, 2010), provide limited support for curation based on organism-specific experimental and genomic data.

Here, we present Data-drivEn METabolic nEtwork Refinement (DEMETER), a reconstruction pipeline that enables the efficient and simultaneous refinement of thousands of draft genome-scale reconstructions. DEMETER specializes in reconstructing human-associated microbes, which previously enabled the reconstruction of 773 gut microbes, AGORA (Magnusdottir *et al.*, 2017), as well its expansion, AGORA2, accounting for 7206 human microbial strains (Heinken *et al.*, 2020). The refinement of draft reconstructions in DEMETER is guided by a wealth of experimental data, such as carbon sources, fermentation pathways, and growth requirements, for over 1000 species, as well as by strain-specific comparative genomic analyses. Hence, DEMETER ensures the resulting refined reconstructions capture known traits of the target organisms.

2. Features

The DEMETER pipeline consists of three main steps: (i) data collection and integration, (ii) draft reconstruction refinement, testing and debugging, and (iii) computation of model properties (Fig. 1).

2.1. Data collection and integration

The minimal prerequisite is the availability of a sequenced genome for the organisms of interest. An essential step is the generation of draft genome-scale reconstructions, e.g., using ModelSEED (Henry *et al.*, 2010) or KBase (Arkin *et al.*, 2018). Where possible, gram status and species-specific experimental data are propagated to the target organisms. Moreover, strain-specific comparative genomic analyses retrieved from PubSEED subsystems (Aziz *et al.*, 2012) can be mapped to DEMETER.
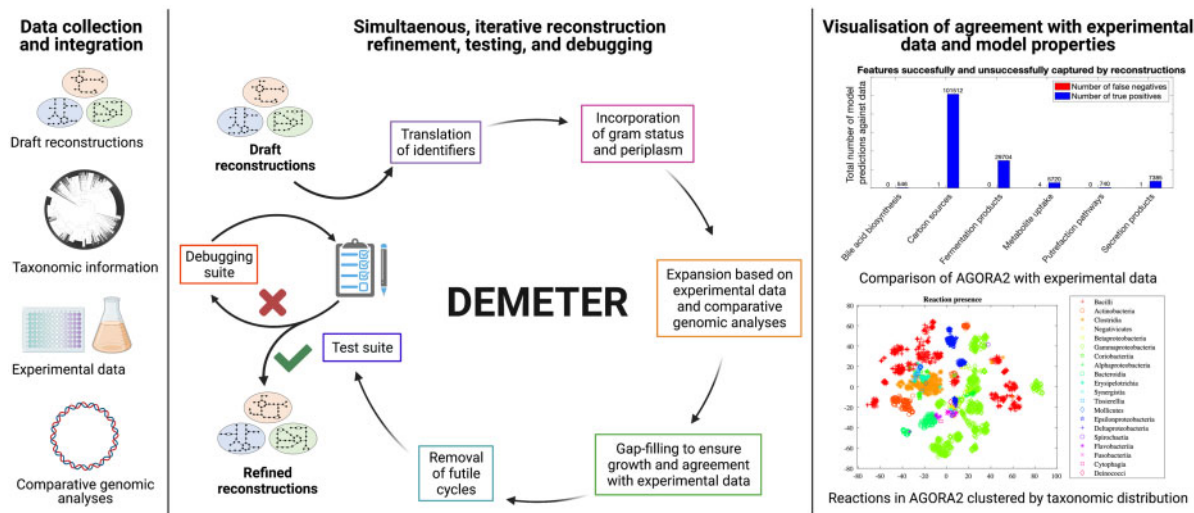
Fig. 1. Overview of the DEMETER workflow consisting of (i) data collection and integration, (ii) simultaneous refinement, testing and debugging of the draft reconstructions, and (iii) visualization of test results and computation of model properties. Created with BioRender.com

2.2. Refinement step

During the refinement step, the draft reconstructions are systematically improved (Fig. 1). Briefly, the following steps are performed:

- Translation from ModelSEED to Virtual Metabolic Human (Noronha *et al.*, 2019) reaction and metabolite nomenclature.
- Curation of the biomass objective function based on gram status and, where appropriate, generation of a periplasmatic compartment.
- Inclusion of species-specific pathways for carbon source utilization, fermentation products, and consumed and secreted metabolites.
- Refinement of pathways and gene-protein-reaction associations based on strain-specific comparative genomic analyses.
- Removal of futile cycles to ensure thermodynamic feasibility.
- Gap-filling to ensure growth and agreement with provided experimental data, including complex and defined media.
- Quality-controlled rebuilding of the resulting refined reconstruction.

2.3. Test and debugging suite

To ensure high quality and predictive potential of the refined reconstructions generated by DEMETER, a test suite is provided that performs systematic quality control and quality assurance (Fig. 1). Any errors are subsequently corrected through a provided automated debugging suite. Some reconstructions may require additional manual inspection.

2.4. Analysis of model properties

To elucidate how metabolic traits are spread across strains, model features including reaction and metabolite content, metabolite uptake and secretion potential, and internal metabolite biosynthesis potential are computed and subsequently visualized. Taxonomically close strains reconstructed by DEMETER are also similar in their reaction content (Fig. 1).

3. Implementation and code availability

DEMETER is written in MATLAB (Mathworks, Inc.) and is freely available at the COBRA Toolbox GitHub https://github.com/open cobra/cobratoolbox (Heirendt *et al.*, 2019). A comprehensive

tutorial in form of a MATLAB live script (Supplementary File S1) is provided at https://github.com/opencobra/COBRA.tutorials.

4. Discussion

Refined reconstructions built through DEMETER adhere to the quality standards in the COBRA field and capture the known metabolic features of the target organisms. Hence, they are suitable for predictive modeling studies, such as the construction and interrogation of personalized microbiome models. Note that while DEMETER was initially developed for the human microbiome, it can be applied to any bacterial or archaeal species.

References

Arkin,A.P. *et al.* (2018) KBase: the United States department of energy systems biology knowledgebase. *Nat. Biotechnol.*, **36**, 566–569.

Aziz,R.K. *et al.* (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One*, **7**, e48053.

Heinken,A. *et al.* (2020) AGORA2: large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities. *bioRxiv*, 2020.2011.2009.375451.

Heirendt,L. *et al.* (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.*, **14**, 639–702.

Henry,C.S. *et al.* (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.

Magnusdottir,S. *et al.* (2017) Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.*, **35**, 81–89.

Noronha,A. *et al.* (2019) The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.*, **47**, D614–D624.

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