Bioinformatics, 38(20), 2022, 4831–4832
https://doi.org/10.1093/bioinformatics/btac596
Advance Access Publication Date: 1 September 2022
Applications Note

Systems biology

MetaboAnnotator: an efficient toolbox to annotate metabolites in genome-scale metabolic reconstructions

Ines Thiele [1,2,3,4,*], German Preciat[5] and Ronan M.T. Fleming [1,5]

[1]School of Medicine, University of Galway, Galway H91TK33, Ireland, [2]Ryan Institute, University of Galway, Galway H91TK33, Ireland, [3]Division of Microbiology, University of Galway, Galway H91TK33, Ireland, [4]APC Microbiome Ireland, University College Cork, Cork T12K8AF, Ireland and [5]Analytical BioSciences Division, Leiden Academic Centre for Drug Research, Leiden University, 2311 EZ, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on March 24, 2022; revised on August 16, 2022; editorial decision on August 25, 2022; accepted on August 29, 2022

Abstract

Motivation: Genome-scale metabolic reconstructions have been assembled for thousands of organisms using a wide range of tools. However, metabolite annotations, required to compare and link metabolites between reconstructions, remain incomplete. Here, we aim to further extend metabolite annotation coverage using various databases and chemoinformatic approaches.

Results: We developed a COBRA toolbox extension, deemed MetaboAnnotator, which facilitates the comprehensive annotation of metabolites with database independent and dependent identifiers, obtains molecular structure files, and calculates metabolite formula and charge at pH 7.2. The resulting metabolite annotations allow for subsequent cross-mapping between reconstructions and mapping of, e.g., metabolomic data.

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

Contact: ines.thiele@universityofgalway.ie

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

1 Introduction

Over the past decade, the systems biology community has observed a paradigm shift, moving from annotating metabolic reconstruction content with only one database-dependent identifier (as one could theoretically retrieve all other ones from the one collected) per metabolite entry to as many identifiers (IDs) as possible. Furthermore, many genome-scale metabolic reconstructions have a significant number of entries without any IDs. At the same time, available tools, e.g., MEMOTE (Lieven *et al.*, 2020), evaluate the presence of IDs in a reconstruction but do not provide provisions to fill missing IDs automatically.

To date, numerous manual and (semi-)automated approaches have been suggested (Haraldsdottir *et al.*, 2014) (Supplementary Table S1); however, their ease of use remains limited. Furthermore, the identifiers collected by these tools remain limited in the extent of databases they capture, mostly focusing on KEGG (Kanehisa *et al.*, 2017), CheBI (Hastings *et al.*, 2013), HMDB (Wishart *et al.*, 2018), and PubChem (Kim *et al.*, 2019). In contrast, to extend the interoperability of metabolic reconstructions, a broader range of database-dependent IDs is desirable. Additionally, many of the available tools do not collect and standardize molecular structure files, required for, e.g., atom mapping. MetaboAnnotator overcomes these challenges.

2 Features

Here, we present MetaboAnnotator, a pipeline for semi-automatic annotation with metabolite identifiers, which is fully compatible with the COBRA toolbox (Heirendt *et al.*, 2019), and implemented in Matlab (Mathworks, Inc.). MetaboAnnotator requires, minimally, a metabolite name, from which it tries to retrieve further database dependent and independent IDs (Fig. 1). It performs searches based on full name matches as well as frequently used synonyms. Ideally, at least one identifier, e.g., InChI String, is provided to increase the confidence in mapping and ensure that more IDs are found. Up to 72 IDs are retrieved from various resources, e.g., BridgeDB (van Iersel *et al.*, 2010), using different search terms (e.g., names, provided IDs) (Supplementary Table S2). Additionally, a molecular structure file is retrieved, which is then used to determine the charged metabolite formula and charge (at pH 7.2) using ChemAxon (http://www.chemaxon.com). Importantly, all retrieved information is captured in a metabolite-centred metabolite structure, listing for each metabolite the retrieved IDs (or their absence) as well as the source providing each ID. The latter is valuable for tracing the provenance of metabolite annotations as well as the propagation of IDs.

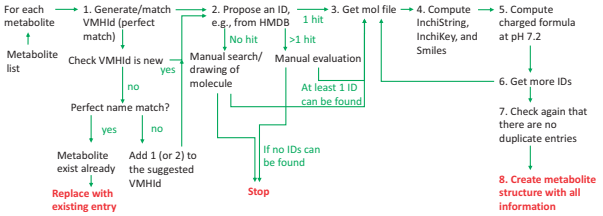

Fig. 1. Overview of the semi-annotation procedure implemented in MetaboAnnotator

2.1 Input data and timing

Three types of input scenarios are envisaged:

1. A *metabolic reconstruction* (‘*model2MetStructure.m*’) that is loaded and its metabolites are converted into a metabolite structure. The metabolite abbreviations will be used as field names for the metabolite structure. If metabolite IDs are provided, they will be used to pre-populate the metabolite structure. For each metabolite, the pipeline will be run (Fig. 1).
2. A *spreadsheet* with metabolites (‘*list2MetaboliteStructure.m*’) is loaded. If abbreviations are provided for the metabolites (in a column named ‘VMH’), these abbreviations are used. Otherwise, new metabolite abbreviations are generated, using the rules defined in Thiele and Palsson (2010) and used as fields in the metabolite structure. The implementation ensures that there are no duplicate abbreviations with the rBioNet (Heinken *et al.*, 2021) and the VMH (Noronha *et al.*, 2019) databases. If IDs are provided, they will be used to populate the metabolite structure.
3. A *cell array* with metabolites (‘*list2MetaboliteStructure.m*’). Again, metabolite abbreviations will be generated if absent from the cell array. Any provided metabolite IDs will be used to populate the metabolite structure.

For each annotated metabolite, the annotation source, type (e.g., automatic), and date will be provided for tractability. Note that any annotations are suggestions, which may require manual curation.

3 Implementation

MetaboAnnotator is written in MATLAB (Mathworks, Inc.) and is freely available at the COBRA Toolbox GitHub https://github.com/opencobra/cobratoolbox (Heirendt *et al.*, 2019). Comprehensive tutorials in form of a MATLAB live script are provided at https://github.com/opencobra/COBRA.tutorials. MetaboAnnotator relies on openBabel (O’Boyle *et al.*, 2011), and if desired, ChemAxon (ChemAxon), for which a free academic license can be obtained.

4 Discussion

MetaboAnnotator comprehensively collects database independent and dependent IDs and thereby, allows to connect the metabolic reconstructions to novel application areas. While MetaboAnnotator largely depends on available resources for finding metabolite IDs, the combination of the various resources results in a comprehensive coverage of metabolite IDs for a genome-scale metabolic reconstruction. By systematically collecting and standardizing molecular structure files, MetaboAnnotator enables the use of chemoinformatic tools in conjunction with metabolic reconstructions. At the same time, any input metabolite is also mapped or translated to the nomenclature of the virtual metabolic human (VMH) (Noronha *et al.*, 2019), thus, providing information whether a metabolite is present in the human metabolic reconstruction (Brunk *et al.*, 2018; Thiele *et al.*, 2020) or in gut microbial reconstructions (Heinken *et al.*, 2020). Disease-relevant resources, such as https://clinicaltrials.gov/, are also mapped, allowing to further broaden the application of genome-scale metabolic reconstructions to biomedical applications. Finally, by enabling to annotate metabolites from scratch (input types 2 and 3), metabolites identified in metabolomic studies can be annotated and mapped onto the VMH database.

Funding

This study was funded by grants from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme [757922 to I.T.]; and by the National Institute on Aging [RF1AG058942 and U19AG063744].

Conflict of Interest: none declared.

References

Brunk,E. *et al.* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, **36**, 272–281.

Haraldsdottir,H.S. *et al.* (2014) Comparative evaluation of open source software for mapping between metabolite identifiers in metabolic network reconstructions: application to recon 2. *J. Cheminform.*, **6**, 2.

Hastings,J. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–463.

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

Heinken,A. *et al.* (2021) DEMETER: efficient simultaneous curation of genome-scale reconstructions guided by experimental data and refined gene annotations. *Bioinformatics (Oxf., Engl.)*, **37**, 3974–3975.

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.

Kanehisa,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Kim,S. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.

Lieven,C. *et al.* (2020) MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.*, **38**, 272–276.

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.

O’Boyle,N.M. *et al.* (2011) Open Babel: An open chemical toolbox. *J. Cheminformatics*, **3**, 33.

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

Thiele,I. *et al.* (2020) Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.*, **16**, e8982.

van Iersel,M.P. *et al.* (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, **11**, 5.

Wishart,D.S. *et al.* (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.