MyCompoundID 2.0: An Evidence-Based Metabolome Library for Metabolite Identification

Bowen Yang[1], Hao Li[1], Shuang Zhao[2], Liang Li[1,2]

1. University of Alberta, 2. The Metabolomics Innovation Centre (TMIC)

Introduction

Metabolites play important roles in biological functions, and the development of analytical techniques such as LC-MS has enabled the high-throughput detection of metabolites. However, one of the major challenges in this field is that **only approximately 30% if the detected features** can be identified due to the absence of corresponding matches in the databases.

To address this issue, MyCompoundID (MCID) database (www.MyCompoundID.org) was developed. It employs the prediction of biochemical reaction products to expand metabolite coverage, and has been widely used, garnering over 610 citations. However, it hasn't been updated for nine years. Hence, recognizing advancements in metabolomics in recent years, especially as **new metabolites from different species** have been discovered and **new cheminformatics methods** have been developed, we have updated and introduced MCID 2.0 to improve the accuracy of the predicted products and expand their application scope.

Methods

The construction of MCID 2.0 involves three components:

Initial library creation: The initial library is divided into different sections based on the diversity of the species included, with each section assembled by combining databases via InChIKey (**Fig. 2**).

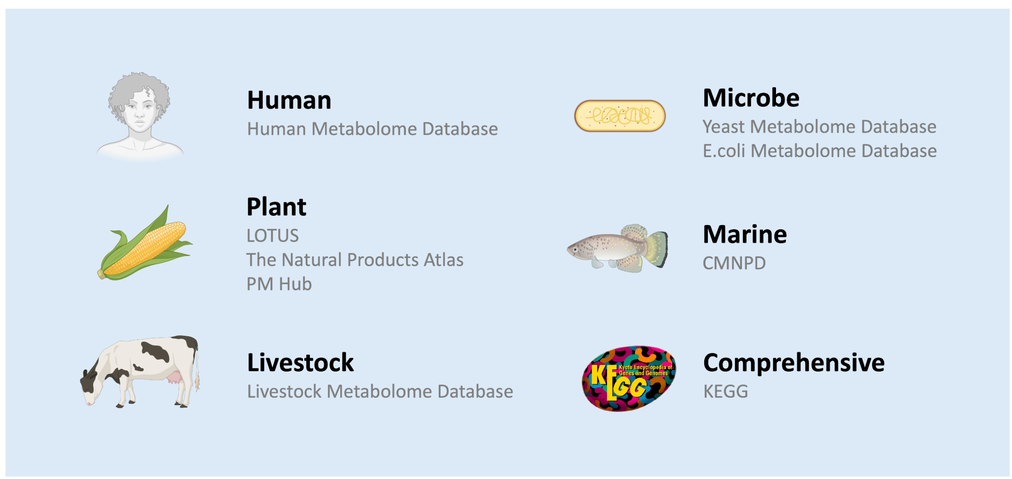

Figure 2. The initial library of MCID 2.0

Products prediction: The products of reactions are predicted using RDKit to build the one- and two-reaction libraries. The reaction rules are represented by SMIRKS, and based on these rules, we developed a comprehensive pipeline to generate the predicted product libraries (**Fig. 3**).

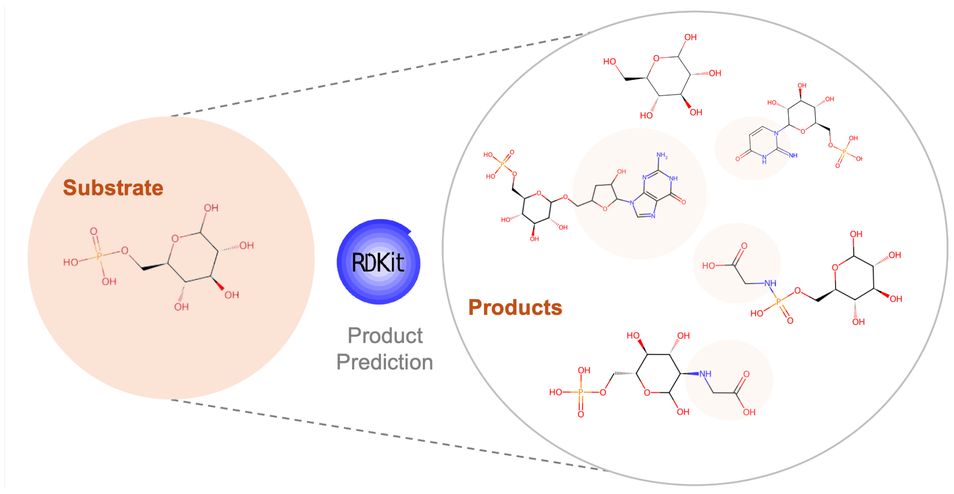

Figure 3. The pipeline of products prediction

Software implementation: Two usage modes for MCID 2.0 are created: a web server and a standalone application. SQLite manages the database behind MCID 2.0, while Python powers the backend functions. The web server version is built with the Flask framework and connects the backend to a React-based frontend. The standalone application's graphical user interface is created using Tkinter (**Fig. 4**).

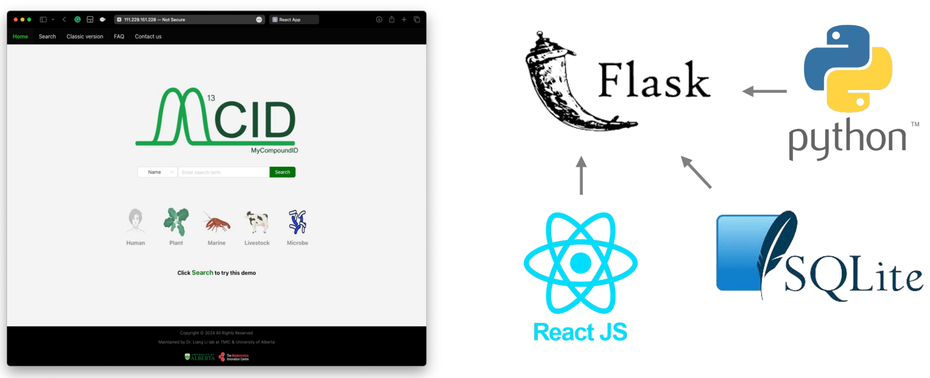

Figure 4. The software architecture of MCID 2.0

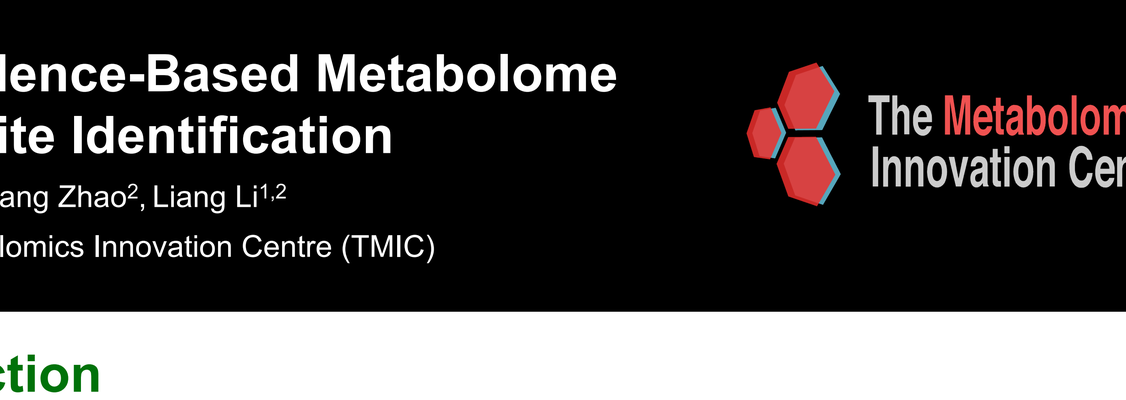

Figure 1. The workflow and limitation of LC-MS based-metabolomics

Novel Aspects

- Expand the coverage of compounds and species to accommodate different research subjects
- Construct a scalable and maintainable rule-based reaction product prediction pipeline
- Provide a more modern, user-friendly interface applicable to various scenarios

Preliminary Results

MCID 2.0, an advanced **cross-platform tool** for metabolite identification, is developed. It features a new evidence-based metabolome database to enhance LC-MS feature identification coverage.

We have divided our initial libraries into **six sections based on the type of organisms** they cover (**Fig. 2**) and predicted the one-step reaction outcomes for each of these initial libraries to construct one-reaction libraries. Subsequently, we used these one-reaction libraries as substrates to predict two-reaction libraries.

To facilitate the prediction of reaction products, we developed a **maintainable and scalable pipeline**, allowing for rapid integration of new chemical transformation rules and their application to new datasets. We also have established rules for **76 common biochemical reactions**, which are currently being applied to the initial libraries.

As shown in **Table 1**, by predicting products from these reactions, the number increases and covers a greater chemical space. This expansion can help explain some LC-MS features that were previously unannotated in the original datasets.

Table 1. The number of compounds in each library

	zero-reaction	one-reaction	two-reaction
Human	45,433	1,920,565	ongoing
Plant	218,893	ongoing	ongoing
Livestock	1,192	36,526	ongoing
Microbe	4,667	263,666	ongoing
Marine	27,233	1,037,557	ongoing
KEGG	15,473	494,411	ongoing

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, Canadian Institutes of Health Research, Canada Foundation for Innovation, Genome Canada and Alberta Innovates.

Reference

1. Huan, T., et al. (2015). Anal. Chem. 87, 20, 10619–10626
2. Kruger, F., et al. (2020). Chem. Inf. Model. 60, 7, 3331–3335
3. Weininger D. (1990). J. Chem. Inf. Comput. Sci. 30, 3, 237–243
4. RDKit: Open-source cheminformatics. http://www.rdkit.org.

Online Browsing