**metaPathwayMap: A tool to predict metabolic pathway neighborhoods from untargeted metabolomics data**

Gaurav Moghe1\*, Susan R. Strickler2

1 Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA

2 Boyce Thompson Institute, Ithaca, NY, USA

\* Corresponding author ([gdm67@cornell.edu](mailto:gdm67@cornell.edu))

**Abstract**

XX

**Introduction**

Untargeted tandem liquid chromatography mass spectrometry (LC-MS/MS) of plant extracts produces thousands of peaks; unfortunately, <1% of them can be reliably identified by matches to spectral databases. Thus, the significance of global metabolome shifts experienced upon stress or in mutants remains under-studied compared to transcriptomics, where database searches can reliably identify properties such as protein domains, sequence homology/orthologous groups, and Gene Ontology. To circumvent this lack of identifiable peaks, metabolite profiling studies frequently resort to using targeted metabolomics, wherein only compounds of known masses are detected; however, this experimental design ignores global changes occurring in the metabolome. Thus, new strategies to better identify perturbed metabolic pathways from untargeted metabolomics data are needed.

A recent, deep learning based tool named CANOPUS (Dührkop *et al.*, 2020) circumvents the problem of specific compound identification by proposing structural categories to a peak based on its MS/MS profile. CANOPUS, which is part of the SIRIUS package, takes Mascot Generic Format (mgf) files as input, and uses pre-trained deep learning models to predict up to five hierarchical structural categories (Superclass, Class, Subclass, Level5, Most specific class) representing the likely dominant structural motif, as well as “all classifications” representing other motifs or other structural possibilities. These categories are based on the well-established ChemOnt ontology schema (Feldman *et al.*, 2005), which has also been applied to entries in the Chemical Entities of Biological Interest (ChEBI) database (Degtyarenko *et al.*, 2008) using the tool ClassyFire (Djoumbou Feunang *et al.*, 2016). Using this ChemOnt classification of ChEBI entries as a common thread, we developed a novel tool metaPathwayMap that matches compounds in the MetaCyc pathway database to peaks obtained via untargeted metabolomics. The outcome is prediction of the structural and pathway “neighborhood” of the metabolomics peak, which is useful in biologically interpreting metabolomics signals.

**Materials and Methods**

metaPathwayMap uses two extrinsic inputs: (1) pathway models derived from MetaCyc (Hawkins *et al.*, 2021) (here, PlantCyc), and (2) CANOPUS predictions of compound structural classes of LC-MS derived peaks, in tsv format.

PlantCyc flat files provide compound names, pathways and ChEBI IDs for most compounds. Since each compound can theoretically belong to >1 pathways, we first developed a pathway similarity network **(Fig. 1)** based on the similarity of the ChemOnt categories of the pathway’s substrates. Distance (1-Jaccard Coefficient) was calculated between pathway pairs, and a threshold of 0.20 was applied to make pathway clusters. This threshold was found to be better at associating similar metabolic pathways vs. a more relaxed random bootstrap-based 5th percentile threshold of 0.43; however, this option is available to the users to implement if needed. Pre-compiled network clusters in the format required by metaPathwayMap were generated, however, users can also generate files required as input using the provided Python scripts.

metaPathwayMap further computes Jaccard Coefficients between each CANOPUS prediction and each PlantCyc compound, and further filters the predictions using a user-defined threshold (relaxed: 0.6, stringent: 0.7). If Compound A is part of Pathways X,Y,Z, all pathways will be listed as of interest. Furthermore, up to 3 additional pathways that are compositionally similar to X,Y,Z based on the Jaccard Distance are also provided as output for further assessment by the user.

A GUI-based web application implementing the metaPathwayMap pipeline was developed using the Django v4.0.2 web framework and PostgresSQL 14.1. Output files are downloadable, and networks can be visualized in Cytoscape.js v 3.21.0. The web application can be accessed at https://solgenomics.net/pages/solcyc. The application is packaged as a Docker image available at Docker Hub. All code is available at GitHub.

**Results**

We tested the accuracy of metaPathwayMap in predicting the right compound and pathway neighborhoods using **(Set 1)** seventeen high-confidence anthocyanins from sweet potato (Bennett *et al.*, 2021) and **(Set 2)** thirteen mixed compounds from the MassBank database (Horai *et al.*, 2010) and additional resin glycosides and acylsugars (Landis *et al.*, 2021; Kruse *et al.*, 2021). CANOPUS annotated 4/17 Set 1 anthocyanins incorrectly as Flavonoid\_3-O-p-coumaroyl\_glycosides (Most Specific Class), while others were annotated as Anthocyanidin\_3-O-glycosides, with other classes being same. All compounds were assigned to ANTHOCYANIN-SYN pathway type, although the specific pathway models were different **(Supplementary File 1A)**. The Jaccard Coefficients of all top metabolite hits were > 0.7. Of the 13 Set 2 compounds, 12 and 9 received CANOPUS and metaPathwayMap annotations, respectively. The three without metaPathwayMap annotations were completely incorrect, and thus no associated pathway clusters were identified above a Jaccard threshold > 0.7. Of the 9, CANOPUS annotations of four compounds were close but not completely accurate, thus, the correct pathway neighborhood was obtained only among the top five predictions for each compound. The SIRIUS score, used by CANOPUS to prioritize predictions, was not found to be associated with CANOPUS prediction accuracy. For the four partially correct metaPathwayMap predictions, the Jaccard Coefficient for the top five pathway matches was less than 0.7. For the remaining five with correct CANOPUS annotations, metaPathwayMap predicted the correct pathway as the top hit with match coefficient > 0.7. These results suggest that Jaccard Coefficient > 0.7 is a good indicator for the confidence in metaPathwayMap pathway neighborhood prediction. Nonetheless, users can set this threshold in the script parameters.

**Usage**

metaPathwayMap can be run on Unix as standalone Python scripts downloadable from our GitHub page (XX). This tool is also embedded with the SolCyc project (XX). Pathway similarity networks for SolCyc, BrachypodiumCyc and PlantCyc are default, however, the GitHub page lists steps for making networks of new MetaCyc pathways.

**References**

Bennett,A.A. *et al.* (2021) Untargeted metabolomics of purple and orange-fleshed sweet potatoes reveals a large structural diversity of anthocyanins and flavonoids. *Sci Rep*, **11**, 1–13.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, **36**, D344–D350.

Djoumbou Feunang,Y. *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, **8**, 61.

Dührkop,K. *et al.* (2020) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology*, 1–10.

Feldman,H.J. *et al.* (2005) CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Letters*, **579**, 4685–4691.

Hawkins,C. *et al.* (2021) Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *Journal of Integrative Plant Biology*, **63**, 1888–1905.

Horai,H. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.

Kruse,L.H. *et al.* (2021) Computational metabolomics illuminates the lineage-specific diversification of resin glycoside acylsugars in the morning glory (Convolvulaceae) family.

Landis,J.B. *et al.* (2021) Migration through a major Andean ecogeographic disruption as a driver of genetic and phenotypic diversity in a wild tomato species. *Mol Biol Evol*.