

Mapping of Chemical and Biochemical Relationships of Mass Spectrometry-based Metabolomics Data

Dinesh Kumar; Tobias Kind; Oliver Fiehn

UC Davis, Davis, CA

Introduction

Mass spectrometry-based metabolomic studies generate quantitative data of the differential regulation of metabolites in response to genetic, environmental or physiological perturbations. Improved algorithms, larger mass spectral libraries and better instrumentations enable the identification of a larger number of metabolites in unbiased GC-MS or LC-MS runs. However, the evaluation of such data sets must not stop at the level of applying statistical tests to the data sets and delivering tables of regulated metabolites. Instead, metabolomics results must additionally convey a biological view of the data sets, enabling biochemical interpretations and structure the metabolites with respect to their chemical, biochemical and physiological roles in an analogous manner to gene ontologies in microarray studies.

Methods

Publicly available data sets of metabolite profiles were used that were downloaded from the SetupX/BinBase database (http://fiehnlab.ucdavis.edu:8080/m1/main_public.jsp). This database annotates metabolites from mass spectral deconvolution results using a Leco Pegasus IV GC-TOF mass spectrometer. BinBase currently hosts over 2,200 unique metabolic spectra which are matched against in-house libraries of over 1,100 spectra of known compounds.

A variety of open access tools from KEGG, Reactome, MetaCyc and ARM were used to map differential regulation of metabolites to biochemical network graphs. Pathway networks were downloaded as SBML and Biopax formats or created in Cytoscape SIF network format. Database mapping was achieved by cross-referencing between different database identifiers. Similarity indices were computed from chemical structures and visualized by their resulting similarity distances. All calculations were performed using the R-project and MS-office tools. The attribute files were generated for studies on the effect of environmental tobacco smoke on lung metabolism in rat fetuses, highlighting the statistical differences between treated and control rats.

Results

Depending on the samples, 250-450 metabolic signals can be reliably and consistently detected in GC-TOF chromatograms. 132 unique compounds were identified in a GC-TOF/BinBase data set of lung metabolites of rat fetuses that had been exposed to 1mg/m³ environmental tobacco smoke (ETS) for 20 days during gestation for 8h/day. When mapping these identified metabolites to biochemical reaction pairs in pathway databases (MetaCyc, KEGG, Reactome, Edinburgh Human Metabolic Network), only 50-80% of the identified compounds were retrieved. For some databases such as MetaCyc, almost all compounds were included but often lacked annotated enzymatic reactions. For most other databases, certain metabolites or metabolites classes were absent. Most databases had poor representation of lipid structures.

Three different approaches were compared for visualizing results from mass spectrometry-based metabolomic data sets: mapping to standard biochemical databases, visualization by chemical structure similarities and a hybrid approach employing atomic reconstruction of metabolism (ARM). The visualization results were then evaluated by the number of metabolic nodes that were represented in network graphs and the clarity of network clusters that should aid the interpretation of differential expression of metabolites.

It was found that for any biochemical mapping approach, highly connected metabolites (e.g. ATP, CoA) have first to be removed to yield meaningful clusters. However, since biochemical pathways comprise thousands of compounds, direct visualizations to overall pathway maps did not result in meaningful overviews even when ATP, NAD(H), Pi and CoA were omitted from network graphs. We can further conclude that graphs based on shortest-path mapping between two metabolite nodes do not yield biologically relevant clusters.

The most suitable biochemical representation of GC-TOF MS identified metabolites was found using the atomic reconstruction of metabolism (ARM) approach that imposes chemical topology on biochemical pathway information. ARM maps resulted in metabolite clusters according to known biochemical modules (lipids, carbohydrates, TCA cycle intermediates, aromatics, amino acids and urea cycle compounds) (fig.

1, lower panel). Nevertheless, ARM presentations comprised only 90 of 132 identified metabolites, leading to a loss of information. The only approach to fully utilize relationships between all identified metabolites is to use the chemical structures themselves, foregoing biochemical databases. Metabolite structures can be decomposed from their machine-readable InChI codes into chemical substructures. Chemical similarities between all identified metabolites were then directly computed using matrices of present and absent substructures. The Tanimoto similarity index was used as distance measure. Interestingly, similar clusters were obtained for networks based on chemical structure Tanimoto distances as found for ARM graphs (figure 1). Tanimoto graphs enable a quick overview over dysregulation in complex metabolic networks without losing any identified metabolite which might not yet

be assigned to enzymatic conversions in pathway databases such as MetaCyc, KEGG or ARM.

In comparison, clustering of network graphs is more eminent using chemical structure similarities than by biochemical networks. In addition, chemical structure networks do not require manual interaction, i.e. removal of hub nodes, which are required for any biochemical pathway network graph. On the other hand, for some compounds such as fumaric acid, the biochemical graph presentation is better because it is otherwise structurally more similar to fatty acids than to TCA hydroxyl acids. Nevertheless, it can be concluded that graph visualizations based on cytoscape are a suitable way to represent metabolic changes in a comprehensive overview.

Acknowledgments

This research was funded by grant 5R01ES13932 of the U.S. National Institute of Environmental Health Sciences, granted to Oliver Fiehn.

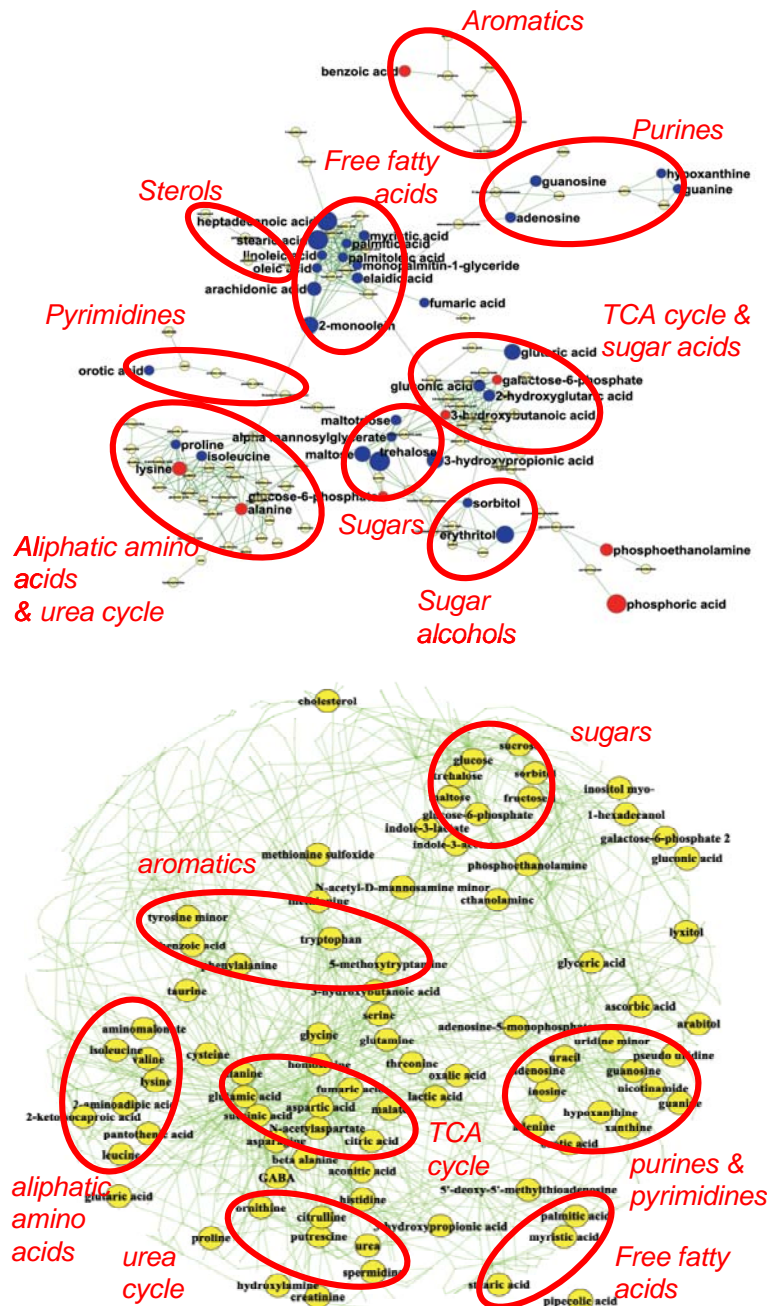


Fig. 1. Mapping identified metabolites detected in lungs of rat fetuses after 20 d exposure to environmental tobacco smoke during gestation
Upper panel: Tanimoto chemical similarity map.
Blue/red nodes: lower /higher metabolites levels under smoke exposure.
Size of nodes reflect the significance values.
Lower panel: Atomic reconstruction biochemical network graph (ARM).