Substance class annotation for long candidate lists in metabolite identification using structural ontologies

Sarah Scharfenberg[1], Janna Hastings[2,3], Pablo Moreno[2], Stephan Beisken[2,4],
Christoph Steinbeck[2,5] & Steffen Neumann[1]

[1]Dept. of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany
[2]Cheminformatics and Metabolism, EMBL-EBI, Hinxton, Cambridge, UK
[3]Current address: Epigenetics department, Babraham Institute, Cambridge, UK
[4]Current address: Discuva, Cambridge, UK
[5]Current address: Friedrich-Schiller-Universität Jena, Jena, Germany

As an integral part of Systems Biology, Metabolomics aims to detect and identify the chemical compounds that drive and participate in biological processes. In untargeted Metabolomics various sample types with a large number of different metabolites are characterized. The method of choice today is mass spectrometry because of its high sensitivity and the broad coverage of measurable metabolites. Tandem mass spectrometry, which reveals information about the compound structure, provides useful hints to identification. Despite the ongoing development of spectral- and compound databases, the main bottleneck in metabolomic research remains the annotation and identification of metabolites.

To identify a new compound, e.g., a biomarker it is necessary to collect a set of annotations that support a hypothesis [3], such as the mass, identified adducts or a substance class. Based on a tandem MS spectrum, MetFrag [1] reports a list of structurally related candidate compounds that are ranked according to their explanatory power in terms of the measured spectrum. Given such a list, BiNChE [5] performs an overrepresentational analysis on the structural ontology of ChEBI [4] and reports a list of possible substance classes that are overrepresented throughout the candidates with good scores.

This approach was successfully applied on two Benchmark datasets of plant metabolites and environmental standard compounds, respectively. The performance was compared for 'known unkown' and 'unkown unkown' compounds [6]. For a published dataset of *C. elegans* the suggested class annotations for possibly 'unkown unkowns' could be supported.

References:

[1] Ruttkies, Christoph, et al. "MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation." Journal of cheminformatics 8.1 (2016): 3.

[2] Gerlich, Michael, and Steffen Neumann. "MetFusion: integration of compound identification strategies." Journal of Mass Spectrometry 48.3 (2013): 291-298.

[3] Sumner, Lloyd W., et al. "Proposed minimum reporting standards for chemical analysis." Metabolomics 3.3 (2007): 211-221.

[4] Hastings, Janna, et al. "ChEBI in 2016: Improved services and an expanding collection of metabolites." Nucleic acids research 44.D1 (2016): D1214-D1219.

[5] Moreno, Pablo, et al. "BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology." BMC bioinformatics 16.1 (2015): 56.

[6] Wishart, David S. "Computational strategies for metabolite identification in metabolomics." (2009).

Topics: Chem- and Bioinformatics

Keywords: Metabolite identification, Overrepresentation Analysis, Metabolomics

Contact: Sarah.Scharfenberg@ipb-halle.de