Abstract

Metabolomics enables the simultaneous measurement of thousands of metabolites and has shown a broad range of applications within ecology, health, food and agricultural sciences. Measuring metabolic profiles allows for example, for the discovery of biomarkers or metabolic pathways related to disease, disentangling complex chemical underpinnings of microbial ecosystems, or assessing metabolic phenotypes of genetically modified plants intended for human consumption. Due to its high throughput and sensitivity, mass spectrometry is one of the most widely-used analytical platforms within metabolomics research. However, the unambiguous identification of chemical structures still represents a significant bottleneck. On average, only 2-5% of the data collected can be matched to known molecules. In this talk, several recently developed computational tools used to mine the metabolome will be introduced. For example, networks of mass spectral fragmentation patterns created through the GNPS platform can be deployed to propagate chemical structural information. These networks can be integrated with other metabolite mining tools, such as unsupervised substructure discovery through MS2LDA, a computational method inspired by text-mining that extracts common patterns of mass fragments and neutral losses. Furthermore, chemical structural relatedness can be explored to construct chemically-informed distance metrics. While conventional distance metrics used to measure pairwise dissimilarities across samples consider individual metabolites as independent entities, chemically-informed distance metrics account for the chemical structural relatedness and may thus improve subsequent multivariate statistics. The combined use of the computational tools introduced here allow for retrieving chemical structural information at a broad level of up to 50% of the data collected in a typical metabolomics experiment and thus significantly enhance and accelerate subsequent biological interpretation.