

Structural annotation of small molecules by mass spectral libraries, substructure information and accurate mass GC-MS data

TOBIAS KIND; Gert Wohlgemuth; Yun Lu; Mine Palazoglu; Martin Scholz; Oliver Fiehn
UC Davis - Metabolomics, Davis, CA

Keywords: Automation; Chemometrics; Chromatography, Gas; Database; Exact Mass/Accurate Mass Measurement;

Novel Aspect: A workflow for automated structural annotation of unknown biomarkers from GC/MS is presented.

Introduction

Gas chromatography-mass spectrometry (GC-MS) is utilized for metabolic profiling of small molecules (less than 500 Dalton) in high-throughput manner since several years. Significant differences in such non-targeted screening approaches can reveal disease markers from tissue or plasma samples. This technique is also routinely used for screening of different plant genotypes and phenotypes. Such GC-MS setups use peak based alignment algorithms and require mass spectral libraries which also include retention time information. However the success rates for correct annotation of molecular structures to detected peaks is usually less than 50%. In order to perform a successful structure elucidation process of unknown peaks a multiple constraint filter cascade must be applied which uses multiple orthogonal input parameters.

Methods

Data were acquired by gas chromatography coupled to quadrupole and time-of-flight (TOF) mass spectrometers using electron impact (EI) and soft chemical ionization (CI) spectra (Agilent 5975, Leco Pegasus III and Waters GCT-Premier). Target peaks for structural elucidation were defined by listing unknown biomarkers from the in-house metabolomic database BinBase after identifying all known metabolites using mass spectral and retention index libraries. These unknown biomarkers were then structurally annotated by (1) selective derivatization using methoximation and ethoximation steps as well as different silylation reagents; (2) substructure information from expert algorithms; (3) retention index prediction algorithms; (4) accurate mass and isotope pattern measurements; (5) database search of possible isomer structures. Multiple constraints are combined for scoring the most likely structural annotations.

Preliminary results

Around 300 peaks per study are routinely exported by the in-house GC-TOF metabolomic database BinBase. One of the advantages of BinBase is that only metabolite spectra are exported that are positively detected in at least 80% of the biological replicate chromatograms of a given biomarker study, as defined by the in-house study design database SetupX. The BinBase algorithm effectively removes noisy and inconsistent spectra to limit the number of biomarkers that constitute differences between classes defined in SetupX. BinBase spectra are then matched against in-house quadrupole and TOF retention index/mass spectral libraries consisting of 1,200 spectra which directly identify approximately 120 metabolites per GC/MS metabolite study. A comparison of these libraries against the KEGG database has shown that the chemical space of known metabolites is very well covered and that many metabolite spectra are included that are not comprised in commercial or customized libraries. Despite these libraries, many putative biomarkers remain unidentified, hence, limiting the usefulness of such markers for biomedical research. To increase the metabolome coverage, efforts are presented for a chemically diverse set of metabolites from different compound classes. A workflow is given which combines algorithms for detection of substructures from mass spectra, data from retention index prediction algorithms (NIST), accurate mass measurements and accurate isotope abundance measurements for obtaining correct molecular formulas (Seven Golden Rules) and multiple selective derivatization strategies to detect functional groups. The algorithms are combined in a multi-constraint filter cascade which can also query isomer databases like Chemspider and PubChem. A hit score is finally assigned according to the correctness of the possible isomer structures.