SimMet®: Informatics Tool for LC-MS and MS/MS based Large Metabolomic Data Processing and Analysis

Ningombam Sanjib Meitei[1], Arun Apte[2], Aiko Barsch[3]
[1]PREMIER Biosoft, Indore, India, [2]PREMIER Biosoft, Palo Alto, U.S.A, [3]Bruker Daltonik GmbH, Bremen, Germany

PREMIER Biosoft

Abstract

Liquid chromatography- mass spectrometry (LC-MS), with high sensitivity and requirement for only low sample amounts, is one of the leading analytical platforms applied for metabolite profiling [1-3]. However, this method generates large data which are not feasible for manual interpretation. Multiple software tools are often required to analyze such data. For example, data analysis through MetaboAnalyst [4] needs XCMS [5]. Besides, as these tools are cloud based, they are associated with downsides of cloud computing [6] such as (i) Reduced control over the distribution of the computation, (ii) Data transfer problem since network bandwidth issues make the transfer of large data sets into and out of the cloud or between clouds impractical, and (iii) Privacy concerns relating to the hosting of data sets on publicly accessible servers. In order to address the challenges, we have developed SimMet® (PREMIER Biosoft, http://www.premierbiosoft.com/), a standalone software tool supporting complete metabolomic workflow including feature detection, retention time alignment, metabolite identification, annotation, statistical analysis, and data visualization. All results and images can be exported into MS Excel, html or CSV files.

Methods

MS: compact (Q-TOF MS, Bruker Daltonik GmbH). ESI(+) with MS and autoMS/MS modes. Scan range: m/z 75-1000. Acquisition rate: 3 Hz.
HPLC: U3000 RSLC(Thermo Scientific). Column: 50 x 2.1 mm BEH C18, 1.7 um column (Waters) Column temp. 30 °C. Flow rate: 0.45 mL/min. Injection volume: 5 µL. Mobile phase: A = H2O, B =MeOH (each containing 0.1% HCOOH). Gradient: linear gradient 2 – 98% B in 5 min, hold 1 min.
Sample: Capsules of 13 different types of coffee (espresso and lungo varieties from different blends and geographical regions) were extracted using 35 ml of water on a standard coffee capsule machine (Krups XN 301T Nespresso Pixie). Two replicates of each type were prepared. Extracts were diluted 1:50 in water prior to analyzing 3 replicates for each extract by UHPLC-MS.
Data Processing: SimMet® software tool (www.premierbiosoft.com).

SimMet® Software Data Analysis Workflow

Raw Data File Formats: Bruker's native files viz., .fid, .baf and .yep. Profile data or Line data types are supported.
Import Data From Hundreds of Raw Data Files in Batch (Figure 1(a)) using different filters (Figure 1(b)).
Precursor m/z Selection for MS/MS Scans: A proprietary algorithm that involves identification of isotope cluster from the MS scans containing the observed precursor m/z value.

Model Experimental Design Through Intuitive GUI Based Windowed Interactive Setup Assistance Dialogs
A software-wizard that guides users to model experimental design by assigning raw data files to their respective biological/technical replicates, assign color code, shape and custom description for each of the biological/technical replicates (Figure 2(a)).
Define data analysis pipeline: Peak detection and picking, feature detection, retention time alignment, metabolite identification using MS and MS/MS database search, and statistical analysis such as Principal Component Analysis (PCA), Partial Least Square Discriminant Analysis (PLS-DA), Hierarchical Cluster Analysis etc.
Data Normalization: Select proper data pretreatment method (Figure 2(b)) [7].

Generate Peaklists in Batch: Peaks detected in LC timescales for hundreds of raw data files.
Separate Isobaric and Isomeric Compounds: Use second derivative Savitzky-Golay smoothing Technique.
Data Reduction: Combines all ions belonging to the same compound (peaks corresponding to isotopes, charge states, adducts and common neutral losses such as, NH4, Na, Li, K etc.).
Compound ID: A unique ID for each detected compound. All MS/MS scans corresponding to ions of this ID are also clustered. The program automatically checks if precursor m/z values of the clustered MS/MS scans correspond to mono-isotopic (M+0) peak. Otherwise, observed precursor m/z values are replaced by corresponding mono-isotopic peak m/z values.
Retention Time Alignment: Either RANSAC or Gale-Shapely techniques [8-11].
Review Peaks and Fill Up Missing Values: Remove unwanted peaks, fetch intensity from raw data files for missing peaks using the start and end LC timescale of detected compounds in other peaklists.

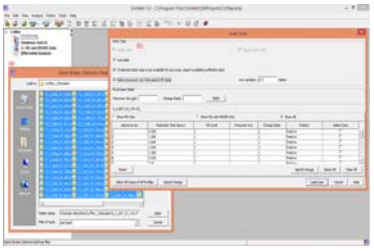

Figure 1: Typical SimMet® software GUI windows showing (a) raw data file browser for importing hundreds of files in batch mode, and (b) choice of data types such as line data or profile data in addition to option to perform precursor m/z selection from parent MS scans.

Removing Noise using Blank Samples

1. LC-MS run of blank extracts subjected to peak detection and picking and then aligned based on retention time with other sample peaklists.
2. All the peaks that are aligned with peaks detected in the blank extracts are removed from further analysis.

Hence, unwanted peaks are removed without increasing the risk of removing compounds that have low abundances with poor signal to noise ratios.

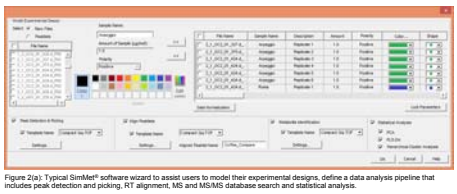

Figure 2(a): Typical SimMet® software wizard to assist users to model their experimental designs, define a data analysis pipeline that includes peak detection and picking, RT alignment, MS and MS/MS database search and statistical analysis.

Figure 2(b): Typical SimMet® software windows allowing users to specify data normalization, reference sample, data centering and scaling and data transformation.

Handling Dynamic Range of Analyte Concentrations in Complex Biological Samples

Detection of Alanine and Trigonelline from a selected Coffee Sample:
Figure 3 shows the XICs of the low concentrated alanine that was detected with an intensity of 88 cts versus the trigonelline peak that has an intensity of 2012591 cts. The ratio 2012591/ 88 = 2.2 X 10^4 > 4 orders of magnitude.

This observation demonstrates the unique capability of the compact QTOF to detect target compounds on an LC timescale across a dynamic range.

Figure 3: XICs of the alanine with an intensity of 88 cts versus the trigonelline peak that has an intensity of 2012591 cts detected in a selected coffee sample.

Metabolite Differential Analysis

Principal Component Analysis: Aligned peaklists from13 different coffee extracts

PCA Score Plot: The 2 "biological" and 3 technical replicates for each sample type (highlighted by using the same colour and symbol) formed clusters in the PCA scores plot as seen in Figure 4(a).

PCA Loadings Plot: Showing analytes with m/z values 195.088 and 138.0552 mainly contributing to the separation of samples in the PCA scores plot (Figure 4(a)). As m/z 195.088 corresponds to caffeine, we removed it from the model and re-run the PCA data analysis. Two compounds labeled X with m/z value 124.0394 and Y with m/z value 138.0561 are detected to have a high content in strong and weak coffee samples, respectively (Figure 4(b)).

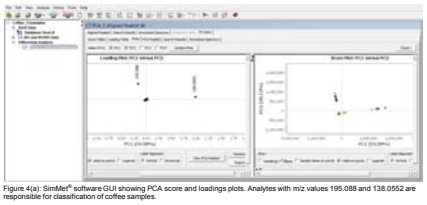

Figure 4(a): SimMet® software GUI showing PCA score and loadings plots. Analytes with m/z values 195.088 and 138.0552 are responsible for classification of coffee samples.

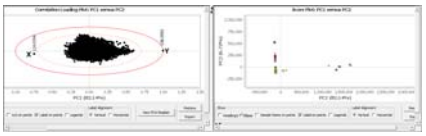

Figure 4(b): SimMet® software GUI showing PCA score and loadings plots after removing the compound with m/z value 195.088 (caffeine). Compounds labeled X with m/z values 124.0394 and Y with 138.0552 are picked for structural elucidation.

Metabolite Identification

Basis of Compound Identification: Exact mass database search, *in silico* fragmentation and matching of expected and observed fragments in batch mode.

Identify Candidate Metabolites: Search metabolites with precursor m/z as the search predicate and use 5 ppm as the tolerance.

Identify the Most Likely Structure of Any Candidates: All candidates are then fragmented *in silico* and the resulting fragment masses are matched to the measured fragment ions with user specified tolerance. A proprietary ranking algorithm differentiate isobaric compounds based on the number of matched observed ions and intensities of those matched ions.

Identification of Caffeine using MS and MS/MS Data

Goal: Test SimMet® software's ability to accurately identify metabolite using MS and MS/MS data
Caffeine MS/MS Data: The QC sample data subjected into SimMet's MS and MS/MS database search workflow.
Compounds Identified: Caffeine structure was correctly identified and ranked 1st. Enprofylline was the 2nd ranked structure.
MS/MS Annotation Showing Fragmentation Patterns of IDed Compounds: Caffeine (Figure 5(a)) enprofylline in Figure 5(b). It is evident that many important diagnostic ions such as m/z 138, 110, 42 etc. are left unmatched in 5(b).
Scoring Mechanism: A propriety algorithm that assigns penalty for fragment ions that can not be matched to in-silico fragments wherein the amount of penalty is decided based on the relative intensity of the non interpreted ions. The higher a penalty a structure receives, the lower the likelihood that the structure corresponds to the MS/MS spectrum.
Portable Reports: MS excel, CSV and HTML files.

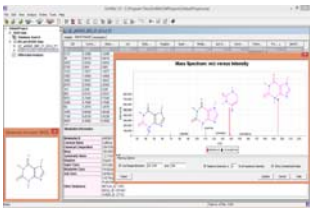

Figure 5(a): Typical SimMet® software result page displaying accurate identification of caffeine as the compound eluted at 1.9257 min of LC timescale. The result page displays identified compounds, structures, metabolite information such as ID number of public databases, common name, other database links besides annotated MS/MS spectrum with matched ions.

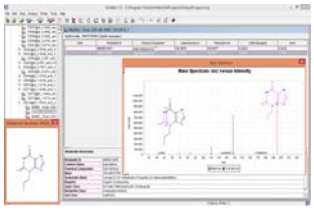

Figure 5(b): SimMet® identifies enprofylline as the second ranked structure for the caffeine MS/MS data. Note the unmatched ion with m/z values 138, 110 and 42.

Identify Compounds X and Y Observed in PCA Score Plot

Candidate Structure: Exact mass search with 5 ppm tolerance in metabolite databases such as HMDB, YMDB, Metlin
Compounds ID using Minimum Delta Mass only: Compounds X tentatively identified as Isonicotinic acid and compound Y as 4-Fluoro-L-threonine
MS/MS Data: We use MS/MS data from QC sample.
Averaging MS/MS scans for the Compound X: The MS/MS scans observed between 0.4-0.6 minute, the LC-timerange in which compound X is eluting (Figure 6).
MS/MS based Identification of compound X: 1st Ranked Structures: Nicotinic acid, 2nd Ranked: Picolinic acid and 3rd Ranked: Nitrobenzene
MS/MS spectra annotation of compound X: see Figure 7.

Using the workflow described above, the molecular formula for compound Y was identified as C7H8NO2 ([M+H]+). In-silico fragmentation identified the analyte to be trigonelline. The ID of compound X as nicotinic acid and compound Y as trigonelline was confirmed by comparison to the authentic standards.

SimMet® Software Highlights

1. Import data from hundreds of native file formats such as .baf, .yep and .fid in batch.
2. Peak detection, peak picking, feature detection for hundreds of raw files in batch.
3. Retention time alignment, gap filling, remove unwanted peaks etc.
4. Statistical analyses such as PCA along with confidence ellipses.
5. Identify thousands of metabolites using MS and MS/MS data.
6. Export results into portable reports such as MS excel, HTML and CSV file formats.

Figure 6: SimMet® software GUI for averaging MS or MS/MS scans. Filters based on scan numbers, retention time, precursor m/z are provided in order to facilitate easy selection of the target scans.

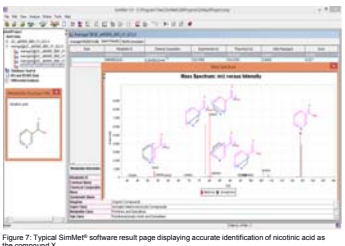

Figure 7: Typical SimMet® software result page displaying accurate identification of nicotinic acid as the compound X.

Conclusion

The compact QTOF provides unrivaled dynamic range (> 5 orders of magnitude as previously reported [12]) in combination with mass accuracy, sensitivity, MS/MS performance and robustness enabling this instrument to be the tool of choice for analyzing batches of highly complex metabolomics samples. Together with SimMet®, a high throughput sophisticated software for comprehensive LC-MS and MS/MS metabolomics data analysis pipeline, it enabled accurate detection of peaks, quick pinpointing of relevant compounds contributing to coffee intensity, identification of two selected target compounds which are characteristic for weak and strong coffee samples. The complete data analysis of the complete data set could be achieved within 35 hours based on a single software solution.
This reliable proposal of compound identities helped to save analysis time and money spent for purchasing multiple references in order to confirm the identity of the target compounds.

References

1. Castrillo et al. Phytochemistry. 2003; 62: 929–37.
2. Theodoridis et al. Mass Spectrom Rev. 2011; 30: 884–906.
3. Bajad et al. Methods Mol Biol. 2011; 708: 213–28.
4. Xia et al. Nucl. Acids Res. 2012; 37: W652-660.
5. Gowda et al. 2014; 86 (14): 6931-6939.
6. Eric et al. Nat Rev Genet. 2010; 11(9): 647–657.
7. Berg et al. BMC Genomics. 2006; 7:142.
8. Fischler and Bolles. Comm Of the ACM. 1981; 24:381-395.
9. Cleveland and Devlin. J Am Stat Assoc. 1998; 83(403):596-610.
10. Pluskal et al. BMC Bioinformatics. 2010; 11:395.
11. Voss et al. Bioinformatics. 2011; 27(7):987-93.
12. Bruker Application Note # LCMS-79