BIORANKINGS®
TECHNICAL REPORT SERIES

Finding untargeted metabolomics intensity differences

Metabolomics researchers want to know how the differences in metabolite compositions impact disease and health in humans and animals (Barnes et al., 2016, Courant et al., 2014). However, current methods of analyzing untargeted metabolomics LC/MS (liquid chromatography and mass spectroscopy) data are time-intensive and subject to bias due to subjectivity in the process. This technical report introduces a new approach for analyzing LC/MS untargeted metabolomics data that is automatic and unbiased. By automating the process to find m/z x RT regions where intensities significantly differ between sample groups, our algorithms will save researchers time and money, while providing results that are valid and unbiased. We start by briefly reviewing untargeted LC/MS metabolomics data, show the challenge of current LC/MS data analysis pipelines, introduce our method, and illustrate its application to publicly available data.

Challenge of untargeted LC/MS metabolomics data analysis

Untargeted LC/MS metabolomics quantifies the amount of known and unknown metabolites in samples with the purpose of finding the metabolites that are associated with health status. **Figure 1** is a schematic of this process taken from Vinayavekhin et al. starting on the left with sample preparation, followed by high-throughput data collection by LC/MS, and finally data analysis comparing ion intensity levels across samples (Vinayavekhin et al., 2010).

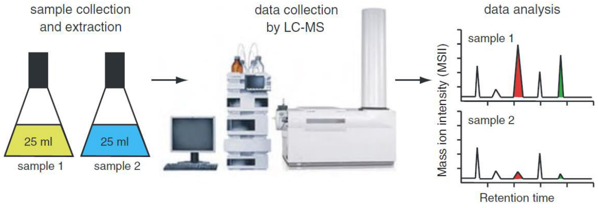

Figure 1

The LC/MS metabolomics data consists of retention time (RT, the time the metabolite took passing through the LC column), mass-to-charge ratio (m/z, a physical measurement obtain by MS), and ion intensity level. RT ×m/z combinations represent a unique metabolite, or in some cases a mixture of metabolites, with the amount of that metabolite in the sample indicated by the intensity level. In our analyses the researchers and bioinformaticians have already processed the raw data for each sample and generated the m/z, RT, and intensity data for each sample.

A single sample of LC/MS can generate a gigabyte of data with 10's to 100's of millions of RT × m/z combinations, making this a *big data* problem. A standard approach currently used in the data analysis is to first identify RT x m/z combinations where intensities look different, and then test those differences using univariate statistics. This is an inefficient approach and introduces statistical problems. First, identifying the m/z x RT regions where intensity peaks appear to be different takes significant researcher time to manually process the data. This can be days or weeks of added work to the experiment. Second, identifying the peaks *manually* is highly subjective making the results potentially biased, unreliable, and non-reproduceable. Third, testing the regions that appear different is statistically invalid and introduces bias. This is known as HARKing, or Hypothesizing After Results Known (Kerr et al., 1998), which increases both the false positive and negative rates.

The first challenge of analysis on LC/MS data is *data pre-processing*. **Figure 2** from Smith et al. shows a typical pipeline of LC/MS data analysis, with some of the steps outlined below.

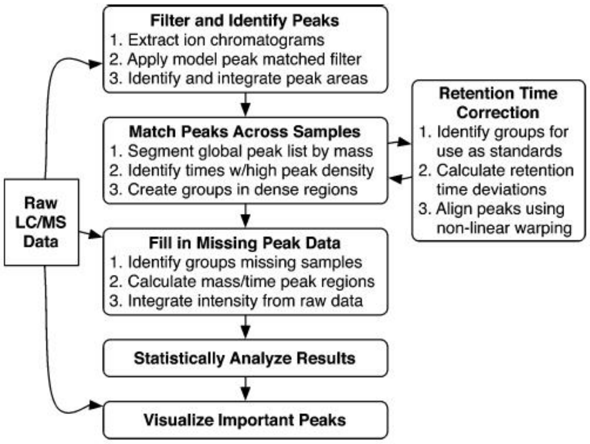
Filter and Identify Peaks
1. Extract ion chromatograms
2. Apply model peak matched filter
3. Identify and integrate peak areas
Match Peaks Across Samples
1. Segment global peak list by mass
2. Identify times w/high peak density
3. Create groups in dense regions
Retention Time Correction
1. Identify groups for use as standards
2. Calculate retention time deviations
3. Align peaks using non-linear warping
Raw LC/MS Data
Fill in Missing Peak Data
1. Identify groups missing samples
2. Calculate mass/time peak regions
3. Integrate intensity from raw data
Statistically Analyze Results
Visualize Important Peaks

Figure 2

1) *Chromatographic peak detection* finds the peaks that are generated by real metabolites but not those generated by random ions or noise. This algorithm depends on the relevant parameters such as the width of the peak and signal to noise ratio threshold (Tautenhahn et al., 2008). These parameters need to be set up for each experiment and are based on

subjective assessments, reducing the reproducibility of the results. Changing these parameters also results in different peaks being identified. Plotting the raw data and visually examining the detected peaks is also subjective.

2) *Peak matching* aligns the peaks identified in individual samples. The current method first sets an arbitrary m/z bin border for grouping the peaks. Then, the peaks from different samples in each bin are matched. Last, the groups which contain peaks from fewer than half the samples are eliminated. Here again, this method presents some problems: the first step of peak matching requires prior knowledge of deviations in retention time, and the last step may eliminate peaks which are significantly different between the groups (Smith et al. 2006). For example, a metabolite present in 50% of cases would be eliminated based on not being in more than half of the samples. Statistically this can lead to incorrect conclusions in cases where more than one metabolite might contribute to disease. If half the samples with disease have one metabolite and the other half the other metabolite, both disease causing metabolites would be eliminated from further analyses.

3) *Retention time alignment* aims to adjust for the variation in time of eluting analytes in the chromatography between samples by shifting intensities along the retention time axis (Tomasi et al., 2004, Prince et al., 2006). There are many alignment algorithms, however, all of them depend on a warping function (e.g. (COW) (Tomasi et al., 2004)), OBIWarp (Prince et al., 2006) which requires specifying a bin size of m/z ratio. The bin size choice is subjective, requires prior knowledges of the experiment, and can often lead to poor alignment.

4) A significant number of peaks can be missed in steps 1-3 (Smith et al. 2006). which is a problem for robust statistical analysis (Kwaket al., 2017). One approach is to fill in missing peak data by rescanning the raw spectra and reintegrating (Smith et al. 2006). This approach requires a user defined range for searching the local highest maximum intensity around the missing location (Katajamaa et al., 2005) which is highly subjective and has risk of filling in inexistent peaks.

In summary, the preprocessing procedure is very complicated, requires a lot of prior subjective thinking, and is subject to individual interpretation. The process can involve significant computing time and many hours of technician time. In addition, after the preprocessing procedure, a classical standard statistical test (e.g. t-test) is used to test whether the areas under selected peaks are significantly different between groups (Smith et al. 2006). This process of first searching for visually different peaks and then applying statistical tests to the peaks after they have been found is HARKing (Kerr et al., 1998). Making more 'false positive' calls wastes time and money in downstream experiments investigating metabolites with no clinical relevance. In the next section we introduce our method that avoids these problems.

Automating the LC/MS metabolomics data analysis

BioRankings has developed an automatic algorithm described below that eliminates these problems in pre-processing and offers a robust statistical analysis of the data.

An Amazon Web Services (AWS) cloud-based analysis software platform has been created to apply a patented method that automates the combined process of peak detection, alignment, and hypothesis testing. This method analyzes metabolomic intensity data in its natural form as functions (curves), rather than discrete segments at m/z x RT locations. By modeling the intensity *surface* researches gain more insight into the level of differences in metabolites across groups instead of a simple present/absent call. In addition, researchers can be agnostic about their data and identify all peak difference, not just differences for a pre-selected subset of metabolites. A typical intensity surface is shown in **Figure 3**.

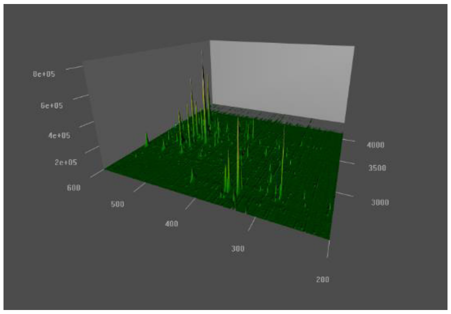

Figure 3

MetabSmooth automatically finds all m/z x RT regions where intensities are significantly different between groups. Briefly, for an m/z slice, say m/z = 300.1 – 300.2, all RT and intensity values from each sample is collected. For each sample the Intensity (Y) x RT (X) data are modeled and represented by a functional form. The functional representations for each sample are then combined and statistical analyses used for signal-to-noise calculations. This eliminates the need for prior peak detection and RT alignment. Comparing peak differences across groups is then done by permutation testing that identifies all peak differences between two groups and their m/z x RT coordinates that the chemist can curate to identify the metabolite.

The method doesn't require data pre-processing – it takes for the m/z, RT, and ion intensity values for each sample -- saving labor and computer time, and avoiding error inflation and bias in data analysis caused by subjective assessments. In the next section we illustrate the method using a publicly available dataset and compare the results to those obtained using the XCMS software package for LC/MS pre-processing and data analysis.

Description of faahKO data

The faahKO data was originally reported in "Assignment of Endogenous Substrates to Enzymes by Global Metabolite Profiling" Biochemistry (Saghatelian et al., 2004). The faahKO data is stored in package *faahKO* (https://bioconductor.org/packages/release/bioc/manuals/xcms/man/xcms.pdf) in open source R software. We selected this dataset to illustrate **MetabSmooth** since anyone can access the data to run their preferred analysis and compare those results with ours. The m/z ratio values, retention times, and intensity values data analyzed were extracted ion chromatograms from 12 sample CDF formatted files. The XCMS help file describes this as "quantitated LC/MS peaks from the spinal cords of 6 wild-type and 6 FAAH knockout mice. The data is a subset of the original data from 200-600 m/z and 2500-4500 seconds RT. It was collected in positive ionization mode."

Results

The XCMS method performed a series of data pre-processing methods, including peak detection, retention time alignment, and peak grouping (see **Figure 2**). After these procedures, XCMS found 4,721 peaks and 403 peak groups. A t-test was applied to the integrated intensity of each peak, finding 33 peaks with significantly different intensity levels between KO and WT groups.

In comparison, our method found 161 intensity regions with significant differences between the two groups. Of these, 31 were the same as found by XCMS and 130 were detected by our method only. We examined the two peaks that were found by the XCMS method only and believe them to be false positives. As a reminder, our analysis was done automatically and run in a few hours with m/z slices run in parallel on AWS.

The results are illustrated as the **Figure 4**. All the plots of the four scenarios described in the last paragraph are shown in the separate attached files.

The figure includes three panels: the first is the raw data, the second is the average functional curves with 95% confident intervals, and the third is the functional permutation test plot. In the top two panels, red represents KO and black represents WT mice. The bottom panel shows the results of the functional permutation test, and the two groups are significantly different in the region where the blue line is above the dotted black line.

Figure 4. A peak found by both methods: The plot shows the intensity difference between KO and WT mice at M/Z value of 300.2. The area at the center of the middle panel where black and red regions are non-overlapping indicates significant difference in intensity between the two groups. The plot indicates significant difference between the two groups for retention times 3378-3401 at m/z of 300.2. This region was found by both methods.

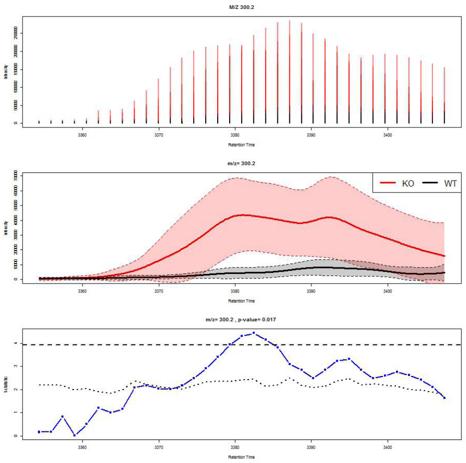

Figure 4

Conclusions

MetabSmooth avoids data preprocessing and directly analyzes the raw data. The results confirm that our method is more accurate than the currently available methodology. MetabSmooth software is available on AWS and a researcher only needs to upload data to get results. By automating the process to find m/z × RT regions where intensities significantly differ between sample groups, our algorithms save researchers time and money while providing results that are valid and unbiased.

BioRankings' Technical Report Series

BioRankings' mission is to help biomedical researchers move their technology from the lab to clinical applications using statistically valid analytical tools for efficient study designs, correct data analyses and conclusions, and rigorous and objective decision making for designing follow-up studies and eventual FDA approval.

To help achieve its mission, BioRankings publishes a Technical Report series focused on applying various statistical methods to real data analyses. Written for understanding by scientists and administrators, these reports will provide an intuitive understanding of the analyses leaving the statistical details to other publications.

For more information, contact BioRankings
at 314-704-8725 or bill@biorankings.com.

References

Barnes S, Benton HP, Casazza K, Cooper SJ, Cui X, Du X, Engler J, Kabarowski JH, Li S, Pathmasiri W, Prasain JK, Renfrow MB, Tiwari HK. Training in metabolomics research. I. Designing the experiment, collecting and extracting samples and generating metabolomics data. Journal of mass spectrometry : JMS. 2016;51(7):461-75. Epub 2016/07/21. doi: 10.1002/jms.3782. PubMed PMID: 27434804; PMCID: PMC4964969.

Courant F, Antignac JP, Dervilly-Pinel G, Le Bizec B. Basics of mass spectrometry based metabolomics. Proteomics. 2014;14(21-22):2369-88. Epub 2014/08/30. doi: 10.1002/pmic.201400255. PubMed PMID: 25168716.

Vinayavekhin N, Saghatelian A. Untargeted metabolomics. Current protocols in molecular biology. 2010;Chapter 30:Unit 30.1.1-24. Epub 2010/04/08. doi: 10.1002/0471142727.mb3001s90. PubMed PMID: 20373502.

Kerr NL. HARKing: hypothesizing after the results are known. Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc. 1998;2(3):196-217. Epub 2005/01/14. doi: 10.1207/s15327957pspr0203_4. PubMed PMID: 15647155.

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical chemistry. 2006;78(3):779-87. Epub 2006/02/02. doi: 10.1021/ac051437y. PubMed PMID: 16448051.

Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008;9(1):504. doi: 10.1186/1471-2105-9-504.

Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics. 2004;18(5):231-41. doi: 10.1002/cem.859.

Prince JT, Marcotte EM. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. Analytical chemistry. 2006;78(17):6140-52. doi: 10.1021/ac0605344.

Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean Journal of Anesthesiology. 2017;70(4):407-11. doi: 10.4097/kjae.2017.70.4.407. PubMed PMID: PMC5548942.

Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2005;6. doi: 10.1186/1471-2105-6-179.

Saghatelian A, Trauger SA, Want EJ, Hawkins EG, Siuzdak G, Cravatt BF. Assignment of Endogenous Substrates to Enzymes by Global Metabolite Profiling. Biochemistry. 2004;43(45):14332-9. doi: 10.1021/bi0480335.