Reproducibility Analyses of Metabolomic Data

Xinjian Yan and Stephen Stein

Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, MD 20899-8380

1. Introduction

Although high-resolution ESI-LC-MS/MS is widely used for metabolomics studies, obtaining reliable and reproducible results remains challenging. This is due not so much to experimental variability but mostly to inherent difficulties in making reliable identifications of the components in the sample. Based on the results of replicate measurements, this work develops a software system that integrates the results of library search and LC-MS (MS1) data analysis to infer optimal metabolite assignments and provides descriptions of the various underlying uncertainties.

2. Methods

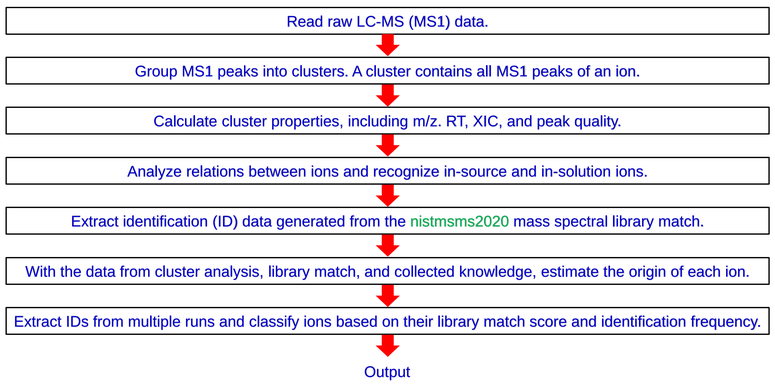
Spectral data of 108 raw files from six urine samples were acquired on a reverse phase C18 column coupled with a Thermo Lumos mass spectrometer in positive electrospray, data-dependent, and high-resolution modes. For each sample, 18 HCD spectrum replicates were acquired. We developed a software platform with these functions: (1) detection and clustering of MS1 peaks based on their measured m/z, retention time (RT), and abundance (XIC) metrics, (2) verification of metabolite identifications of spectral library search using MS1 data analyzed in Function 1, and (3) further investigation of consistency of the identifications across replicate runs. A final score is assigned to each identification according to the library match score, ion uniqueness, observed frequency, and quality of MS1 peaks as well as the knowledge collected previously.

3. Results

3.1 Integration of results from MS1 data analysis and spectral library search

A program for analyzing metabolite raw data of Thermo instruments is developed. This program reads and processes LC-MS (MS1) data and is combined with other NIST software to estimate sampled ions' identity and possible origin with uncertain information. The program's workflow is presented in Figure 1.

Figure 1. The workflow of the metabolite data analysis program.



3.2 Characterization of metabolite data

A. Highly consistent MS1 data

A systematic analysis of the 108 runs from six samples indicates there is a high degree of consistency in measured retention times and extracted ion chromatograms (XIC) across these replicate runs. For example, 90% of 532 ions (detected with > nine replicates of a single sample) in different runs showed standard deviations within 2 seconds. The XIC data showed a similar trend, where 85% of replicates have a relative standard deviation within 0.2 (= standard deviation of XIC / average XIC of replicates). These results laid a solid foundation for studying the reproducibility of identifications.

B. Inconsistent metabolite identifications

While most MS1 peaks can be correctly analyzed in this study, the identifications from replicate runs show various inconsistencies, which may be attributed to:

a. The existence of multiple isomers and lack of corresponding spectra in the NIST library. Hence, different isomers are assigned to the same identification (seeing Figure 2).

b. Isomers spectra are too similar to be distinguished, even if they are matched to library spectra with very high scores (seeing Table 1).

c. Highly coeluting isomers (seeing Figure 3).

Figure 2. This figure shows there are five acylcarnitine isomers at m/z 276.1442 in the urine sample 3671-3. At the given m/z, the nistmsms20 mass spectral library contains only one acylcarnitine, i.e., L-glutarylcarnitine. Because these isomer spectra have higher similarity, naturally, all the five sampled ions are identified as L-glutarylcarnitine, with higher library match scores (The green numbers in the figure. A complete match score is 999).

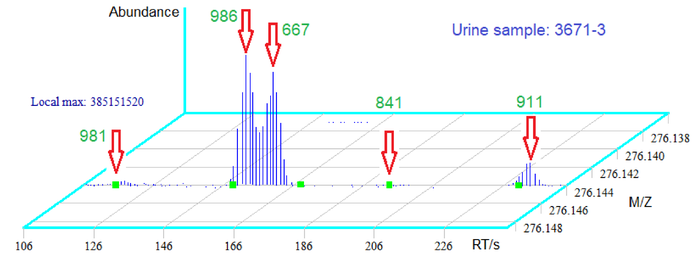
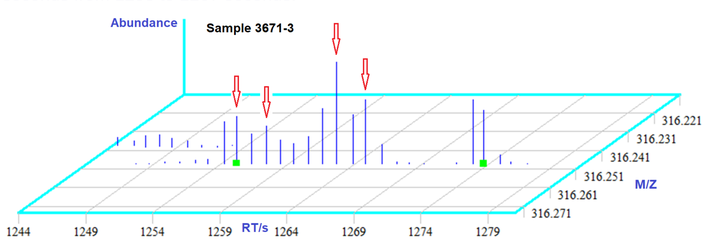


Table 1. This table presents the identification of two isomers: isobutyryl-L-carnitine (IsoBC) and (R)-butyrylcarnitine (BC). By considering their chemical structures and observed retention times, it was determined that IsoBC eluted at 281s, while BC exhibited a later elution at 300s, in line with existing literature. Upon inspecting the identifications from the nistmsms20 library search, it was found that IsoBC was misidentified as BC in 12 out of 18 urine runs with a top score of 987, and correctly identified as IsoBC 10 times with a score of 989. This was attributable to their near-identical spectra. The findings thus indicate that reliance on the library match score alone may not yield reliable identifications.

Isomeric metabolite	Retention time (s)	Average XIC	Identified by library search	Replicates	Max match score	Correctness
BC	281	7.69E+08	BC	12	987	×
BC	281	7.71E+08	IsoBC	10	989	√
IsoBC	300	9.77E+06	BC	14	772	√

Note: The data are processed from 18 replicate runs of NIST SRM 3667.

Figure 3. This figure indicates the possible existence of four acylcarnitine isomers in a range of 14 seconds from 1253 to 1267 seconds.



3.3 Approaches to improve the quality of identifications

A. Description of possible inconsistencies of identifications

The above examples indicate that the mass spectral data obtained from metabolomic studies are complex and many identifications are of large uncertainties. Thus, functions are developed to provide relevant information. For example, isomer information is provided for the cases described in Figures 2 and 3 and Table 1. Cases like Figure 3 are recorded for reporting the existence of coeluting isomer group (CIG) at a given m/z and RT. A knowledge base has been developed to provide chemical information accumulated from previous data analyses and literature.

B. Use retention time to improve the annotation of metabolites

Our analysis reveals that RT data from different samples may be reliably correlated if the correct references are used. In this study, the reference ions were selected based on that they were detected more than nine times in 18 replicate runs, library match (MF) score > 850, XIC > 10000000, without isomers or with only fully distinguishable isomers. Highly correlated relations are obtained for our urine samples (Figure 4) and urine with other samples (for example plasma, Figure 5). The closed RT relation of the data from different samples is invaluable for confirming identifications, distinguishing different isomers, and recognizing different species.

Figure 4. It shows the RT correlation of 84 reference ions identified from samples 3667 and 3672. The RT value of each reference ion is the mean of replicated runs' RT data.

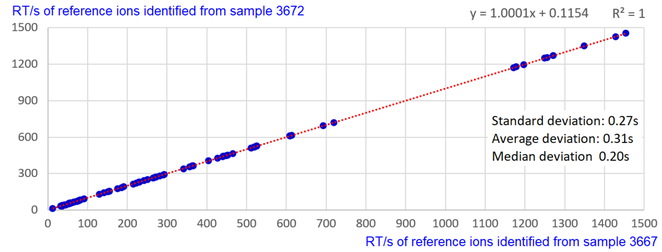
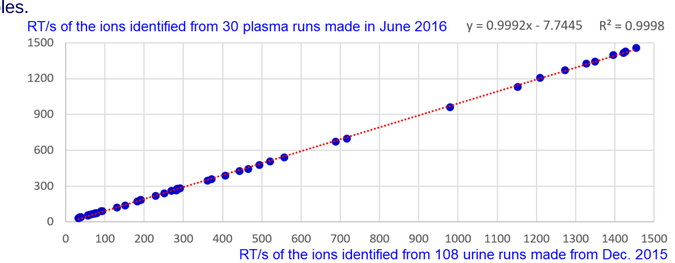


Figure 5. This graph presents the corresponding RT data of 49 metabolites identified from plasma runs made in June 2016, and urine runs made in Dec. 2015, indicating that a well correlated RT relation can be obtained for data acquired at different times and for different samples.



4. Summary

The present research highlights the difficulties associated with identifying isomeric metabolites solely using spectral library search of LC-MS/MS data. We developed a computer platform to provide users with information on the uncertainties caused by the isomers. Additionally, we demonstrate that retention times of different runs or even samples can be highly correlated.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure identified adequately. Such identification is not intended to imply recommendation or endorsement by [the National Institute of Standards and Technology], nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.