Vol. 11, No. 1, 2020
ISSN 2233-4203/ e-ISSN 2093-8950
ARTICLE www.msletters.org | Mass Spectrometry Letters

Evaluation of Recent Data Processing Strategies on Q-TOF LC/MS Based Untargeted Metabolomics

Ozan Kaplan and Mustafa Çelebier*

Department of Analytical Chemistry, Faculty of Pharmacy, Hacettepe University, Ankara, Turkey

Received November 26, 2019, Revised December 13, 2019, Accepted December 16, 2019
First published on the web March 31, 2020; DOI: 10.5478/MSL.2020.11.1.1

Abstract : In this study, some of the recently reported data processing strategies were evaluated and modified based on their capabilities and a brief workflow for data mining was redefined for Q-TOF LC-MS based untargeted metabolomics. Commercial pooled human plasma samples were used for this purpose. An ultrafiltration procedure was applied on sample preparation. Sample set was analyzed through Q-TOF LC/MS. A C18 column (Agilent Zorbax 1.8 μM, 50 × 2.1 mm) was used for chromatographic separation. Raw chromatograms were processed using XCMS - R programming language edition and Isotopologue Parameter Optimization (IPO) was used to optimize XCMS parameters. The raw XCMS table was processed using MS Excel to find reliable and reproducible peaks. Totally 1650 reliable and reproducible potential metabolite peaks were found based on the data processing procedures given in this paper. The redefined dataset was upload into MetaboAnalyst platform and the identified metabolites were matched with 86 metabolic pathways. Thus, two list were obtained and presented in this study as supplement files. The first list is to present the retention times and *m/z* values of detected metabolite peaks. The second list is the metabolic pathways related with the identified metabolites. The briefly described data processing strategies and dataset presented in this study could be beneficial for the researchers working on untargeted metabolomics for processing their data and validating their results.

Keywords : Metabolomics, metabolite profiling, Q-TOF LC/MS, XCMS, metabolic pathway analysis, data processing

Introduction

Metabolomics aims to identify and quantify the small molecules involved in metabolic reactions. In recent years, metabolomics has emerged as a key tool to understand the molecular basis of diseases, find diagnostic and prognostic biomarkers, and develop new treatment opportunities. The metabolomics studies can be arranged into two classes, which are targeted studies and untargeted studies known as metabolite profiling. Targeted studies are for absolute determination of already focused metabolites while metabolite profiling is to compare the relative amount of the metabolites changed according to environmental effects like age, diet, and diseases.[1]

Liquid chromatography-mass spectrometry (LC-MS) based untargeted metabolomics applications have almost become the leading method on biomarker discoveries for diseases.[2-4] Since the metabolites are not like DNA or proteins, it is hard to define a global procedure to analyze the whole metabolome in a system of a body through a single injection. The procedures on sample preparation, chromatographic separation, and ionization mode could affect the number of the potential metabolite peaks in a metabolite profiling study.[5-7] These could be expressed as analytical part on metabolomic studies. Data mining process is the another part of untargeted metabolomic studies to find the reliable metabolite peaks instead of false-positive results. Two major and well-known platforms, XCMS (http://xcmsonline.scripps.edu) and MZmine 2 (http://mzmine.github.io), offer easy to apply procedures on data mining processes on LC-MS based data. Both packages perform admirably in peak picking but they detect a problematic number of false-positive peaks.[8] To solve this problem, Isotopologue Parameter Optimization (IPO) software can be used to optimize XCMS parameters. The optimized parameters are subjected to XCMS.[9] Thus, the false-positive peaks are eliminated in the peak picking process on the data mining procedures. The other methodology to find the 'reliable' peaks is the preparation of consecutive dilution series of the samples intendedly.[10] This methodology uses the basic principle in analytical chemistry and the real peak areas reduce proportionally with the dilution factor. However, the noises or the areas of 'ghost'

*Reprint requests to Mustafa Çelebier
E-mail: celebier@hacettepe.edu.tr

peaks detected by software are not change or change but not related with the dilution factor. Thus, the correlation coefficient between the peak areas and the dilution factor could be used to rearrenge the data and discard false-positive peaks. In general, the peaks correlating ($r > 0.90$) with the dilution factor might be selected as 'reliable' metabolite peaks on such an appliacation.

The one of the most important part of untargeted metabolomics is the identification procedure. At the end of data mining process, identification is the necessary step to correlate the results with the hypothesis of the research. To identify the metabolite peaks, high precision mass spectrometry data are uploaded into the databases. To simplify this procedure, some platforms like MetaboAnalyst (https://www.metaboanalyst.ca) could be used.[11] Since the experimental MS/MS data of the metabolites on databases are not sufficient and the accuracy of the in-silico MS/MS data for the metabolites are still discussed,[12] the only way to match the whole peaks with metabolites is to use MS data instead of MS/MS data. Human metabolome database (HMDB) has been grown up dramatically since it was announced in 2007.[13] This situation causes the matching of a single *m/z* value with more than one metabolites listed on HMDB. Although the precision of the recent LC-MS systems is better than previous models, the random errors on MS results still cover a wide range of metabolite lists and it is infeasible to match one peak with a single metabolite.

In this study, recent data processing strategies as follow were modified and applied on commercial pooled plasma samples: 1- Consecutive dilution method on sample preparation for validating peaks, 2- Using Isotopologue Parameter Optimization (IPO) on finding XCMS optimum parameters 3- Using XCMS on peak picking and grouping, 4- MS Excel data modification on raw XCMS output, 5- Using MetaboAnalyst to match the reliable peaks with metabolites, 6- Using MetaboAnalyst to match the pathways where these metabolites are involved in.

The application of these steps were briefly presented as a workflow in Figure 1 and the details were given in the experimental section. Two different lists were given in the supplement files (#1 and #2) based on the results obtained in the present study. The researchers working on Q-TOF LC/MS based untargeted metabolomics could use the output of the raw excel data on the supplement file #1 to process their results and they could use the supplement file #2 to validate their results. Besides describing a brief workflow, the results of the well-known basic procedures like using XCMS-Online version for data mining procedure and uploading the output data directly to HMDB were compared with the results obtained through the methodology in the present study.

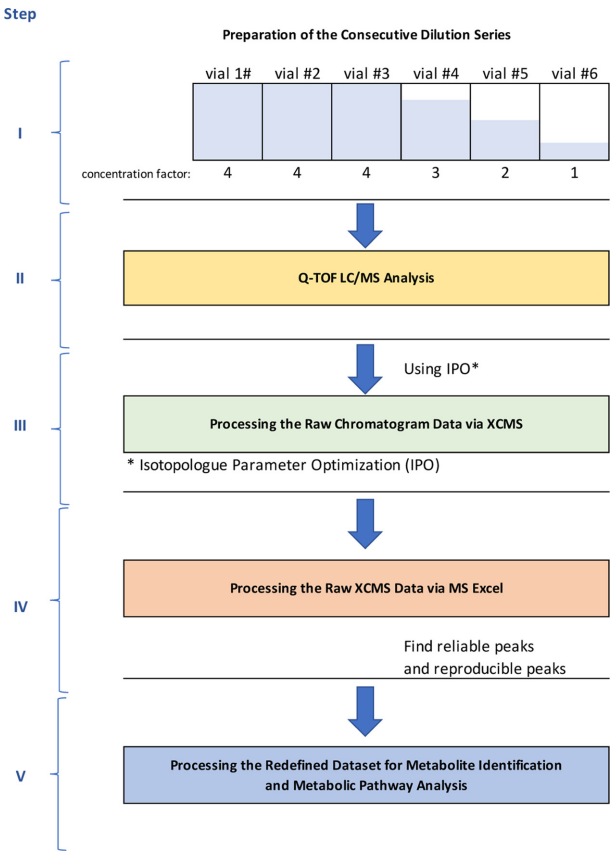

Figure 1. Schematic presentation of the workflow for untargeted metabolomics overviewed in this study.

Experimental

Preparation of pooled plasma sample consecutive dilution series

Ultrafiltration cartridges with 3kDa pores (Amicon® Ultra 0.5 mL Centrifugal Filters, Merck, Darmstadt, Germany) were used to prepare the samples. 300 μL of pooled plasma samples and 150 μL of acetonitrile were placed into the ultrafiltration cartridge and centrifuged at 15000 rpm for 45 minutes. After centrifugation, the liquid phase (metabolite phase) passed through the filter was taken and dried under vacuum centrifuge at 4°C. Then, three vials were prepared in identical concentrations (40 μL sample (re-dissolved with mobile phase) + 40 μL mobile phase) to check the reproducibility and eliminate the random errors from injections. Other three vial prepared on dilution series type to undergo regression analysis to the peaks (4th vial 30 μL sample + 50 μL mobile phase, 5th vial 20 μL sample + 60 μL mobile phase, 6th vial 10 μL sample + 70 μL mobile phase). If a peak is for real, its intensity should be identical in the first three vials and the intensity should be decreased with following dilutions (from 4th to 6th vials).

Q-TOF LC/MS analysis of the samples

Mass spectrometry analyses were carried on Agilent

6530 LC/MS Q-TOF instrument (Agilent Technologies, 184 Santa Clara, CA). C18 column (Agilent Zorbax 1.8 μM, 50.0 × 2.1 mm) was used as the chromatography column. Mobile phases were water and acetonitrile and both of them consisting of 0.1% formic acid. Flow (0.20 mL min^{-1}) started with 90% H_2O until 1st minute, the ACN ratio was increased linearly to 90% ACN until the 15th minute. The chromatographic conditions were later turned back to starting conditions linearly till 20th minute and 5-minute post-run was applied for further injections. The scan range for MS device was 100-1700 *m/z*. All samples were injected into the system as two replicates in a random order.

Processing the raw chromatogram data via XCMS

Chromatograms taken from LC-MS instruments are raw data and the bioinformatics data inside of it should be extracted. For this purpose, raw data files were converted to .mzml format via ProteoWizard software (http://proteowizard.sourceforge.net).[14] Peak picking, grouping, and comparison parts were performed (metabolite profiling) via XCMS (https://xcmsonline.scripps.edu/) software.[15] XCMS is an "R software" based, freeware program used for peak picking, grouping and comparing the findings. XCMS has many parameters for optimization. Isotopologue Parameter Optimization (IPO) is a software that automatically optimizes XCMS parameters.[9] Thus, the optimized parameters obtained through IPO were subjected to XCMS to find the potential metabolite peaks.

Processing the Raw XCMS Data via MS Excel

XCMS results were modified in MS Excel. The first step was to find the 'reliable' metabolite peaks using consecutive dilution series. As it is given in supplement file 1, the peak areas (from 3rd to 6th vials) properly corelated with dilution factor ($r > 0.90$) were indicated as 'reliable' metabolite peaks. To investigate the reproducibility of the peaks, the obtained peak areas for concentrated vials (from 1st to 3rd vials) were multiplied with the multiplication factors (1/3, 1/2 and 1/1) and the redefined peak areas were correlated with the multiplication factors. The recalculated peak areas in a correlation ($r > 0.90$) with multiplication factors were indicated as reproducible peaks. The intersection of reliable and reproducible peaks was found and these peaks were used for metabolite identification.

Processing the redefined dataset for metabolite identification and metabolic pathway analysis

As it is already known, the *m/z* value for a specific peak could refer more than one metabolite on database search due to the fact that the organic compounds might have different chemical formulas but identical masses. Recent Q-TOF LC/MS instruments are capable of working less than 10 ppm errors on mass accuracy. Even such low and acceptable errors could cause multiple matches on metabolite identifications. In the present study, the intersection of reliable and reproducible peaks as mentioned above were uploaded into MetaboAnalyst to match the *m/z* values of the peaks with metabolites. 'MS Peaks to Pathways' utility was used for this purpose. The KEGG codes [KEGG: Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/)] of the final metabolite list were uploaded into MetaboAnalyst again and 'Pathway Analysis' utility of the MetaboAnalyst was used to find the metabolic pathways related with the uploaded list of the metabolites.

Results and Discussion

Data processing on untargeted metabolomics is to handle with metabolites having no specific knowledge of them. The physical properties and chemical structures of the metabolites are totally different within each other. The only common properties of the metabolites are their molecular masses (less than 1.5 kDa). Therefore, it is hard to offer a global and standardized methodology to identify the peaks in an untargeted metabolomics study and match these peaks with metabolites. In this study, a Q-TOF LC/MS based untargeted metabolomics approach was performed on commercial pooled human plasma samples. An ultrafiltration procedure was applied on sample preparation. C18 column (Agilent Zorbax 1.8 μM, 50 × 2.1 mm) was used as the chromatography column. Mobile phases were water and acetonitrile and both of them consisting of 0.1% formic acid. A gradient elution program was used for separation and positive MS mode was applied. Recent data processing strategies including using ProteoWizard software to convert the raw data to a numerical format, using consecutive dilution series to discard false-positive peaks, using XCMS R version combined with IPO to optimize the parameters on finding as much as peaks possible, modifiying data on MS Excel based on the results of regression equations correlated with concentration factor and multiplication factor were used to process the raw data. The final results were presented as a processed dataset (supplement file 1) for the researchers working on untargeted metabolomics. The ready to use formulas on MS Excel file given as supplementary files could be helpful for researchers to process their data and to compare their results with the results of this study.

As it is seen on the supplement file 1, the total number of the potential metabolite peaks found in the pooled human plasma samples was 2287. When the peak picking and grouping was applied through XCMS on-line platform having an user friendly interface in contrast to R programming language edition, it was found 7047 peaks for concentrated vials. This situation shows that using IPO to optimize the parameters of XCMS allows researchers to find 27% of the number of peaks found by the default parameter (parameters for ultra performance liquid chromatography columns) of XCMS online. This situation

might be explained by the fact that XCMS - R programming language edition combined with IPO allows users to find 'real' peaks through optimum parameters and the consecutive dilution series processed with IPO migh cause to be found less number of peaks. In an ordinary metabolomics study, the XCMS output could be directly uploaded to HMDB in order to match the *m/z* values of the peaks with the metabolites. However, this strategy matches high number of peaks (including false-positives) with high number of metabolites. As it is given in literature,[10] using consecutive dilution technique discards non-reliable peaks. In this study, the results of the data processing based on the concentration factors via MS Excel allowed us to discard 563 non-reliable peaks. Finally, 1724 peaks remained and it showed that almost 25% of the peaks could cause unreal results on metabolic pathway analyses unless they are discarded. In the present study, it was also described a novel methodology to find reproducible peaks. This strategy is very close to consecutive dilution technique but it works using multiplication factors on table instead of dilution factors. In this strategy, the peak areas of the identical vials (concentrated three vials) are multiplied with multiplication factors (1/3, 1/2 and 1/1) and the regression equation for multiplication factor and redefined peak areas are used to find the reproducible peaks as mentioned in experimental section. According to the results of this strategy, it was found that 74 peaks were not reproducible and they were discarded. Thus, the final number for the peaks on the list (given in supplement file 1) decreased 1650 and these reliable and reproducibe peaks were used on metabolic pathway analysis. The overlapped base peak chromatograms of consecutive dilution series are given in Figure 2. The metabolic pathway analysis results are given in supplement file 2.

Finally, a brief workflow for untargeted metabolomics and a processed data of Q-TOF LC/MS based metabolite profiling application on pooled human plasma samples were presented. The researchers working on LC-MS based metabolite profiling could use the described procedures and obtained dataset in this study for the following purposes:

Using the described workflow to profile metabolites in human plasma samples

Using the described procedures for process the raw data

Using the described methodology to redefine their dataset to find reliable and reproducible peaks

Using the obtained datasets on supplementary files for validating their results

To compare the plasma metabolome for two different groups like healthy subjects and patients, authors strongly recommend applying peak normalization[16] before processing the raw XCMS data.

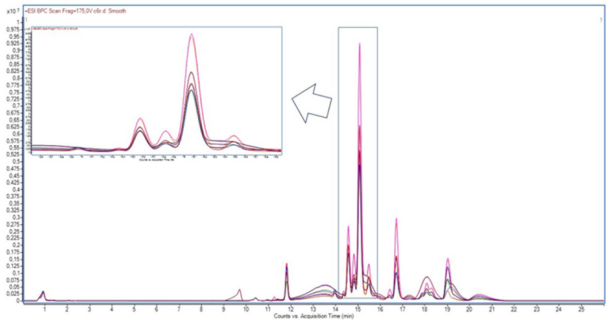

Figure 2. Overlapped base peak chromatograms.

Conclusions

In this study, recent data processing strategies in the literature were evaluated and a brief procedure was described based on the combination of different methodologies given in various articles. A novel approach using multiplication factors was also offered for the researchers. While the researchers applying untargeted metabolomics with pooled plasma samples, the non-reliable peaks and random errors occurring non-reproducible peaks cause them find false-positive results. The results in this study shows that using the user-friendly interface of the XCMS online could save time on data processing part on metabolomic studies but recent data processing strategies presented in this study through the rearrangement and modification of the data via MS Excel can lead researchers to perform metabolic pathway analysis with a cleaned-up data. As it is seen, at least more than 25% of the peaks were eliminated before metabolic pathway analysis.

The results in this study shows the importance of data processing strategies and it also shows that metabolic pathway analysis is hard to achieve without recent tools developed for this purpose. MetaboAnalyst platform having an online interface provides a useful tool to match the *m/z* values of the peaks with KEGG codes. These KEGG codes could be used on metabolic pathway analysis through 'Pathway analysis' utility of the same platform. Although, it looks like the recent development allows untargeted metabolomics and metabolic pathway analysis reachable for researchers, the human factor with a logical approach is still needed to make the final results valuable.

Supporting Information

Supplementary Information is available at https://drive.google.com/file/d/1_ab5ZxzM7vVAjbUyBD6q4s8HDvw2h8o2/view

References

1. Dunn, W. B. *Phys. biol.* **2008**, 5, 011001.
2. Zhou, B.; Xiao, J. F.; Tuli, L.; Ressom, H. W. *Mol. BioSyst.* **2012**, 8, 470.
3. Fang, Z.-Z.; Gonzalez, F. J. *Arch. Toxicol.* **2014**, 88, 1491.

4. Bajad, S.; Shulaev, V. *LC-MS-based metabolomics. Metabolic Profiling*: Springer: New York, **2011**.
5. Álvarez-Sánchez, B.; Priego-Capote, F.; de Castro, M. L. *TrAC- Trends Anal. Chem*. **2010**, 29, 111.
6. Álvarez-Sánchez, B.; Priego-Capote, F.; de Castro, M. L. *TrAC- Trends Anal. Chem*. **2010**, 29, 120.
7. Patti, G. J. *J. Sep. Sci*. **2011**, 34, 3460.
8. Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. *Anal. Chem*. **2017**, 89, 8689.
9. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Peiber, T.; Magnes, C. *BMC Bioinform*. **2015**, 16, 118.
10. Eliasson, M.; Rännar, S.; Madsen, R.; Donten, M. A.; Marsden-Edwards, E.; Moritz, T.; Shockcor, J. P.; Johansson, E.; Trygg, J. *Anal. Chem*. **2012**, 84, 6869.
11. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. *Nucleic Acids Res*. **2018**, 46, W486.
12. Blaženović I.; Kind T.; Torbašinović, H.; Obrenović, S.; Mehta, S. S.; Tsugawa, H.; Wermuth, T.; Schauer, N.; Jahn, M.; Biedendieck, R.; Jahn, D.; Fiehn, O. *J. Cheminform.* **2017**, 9, 32.
13. Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C,; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A; Liu, Y,; Mandal, R.; Neveu, V.; Pon, A,; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. *Nucleic Acids Res*. **2017**, 46, D608.
14. Holman, J. D.; Tabb, D. L.; Mallick, P. *Curr. Protoc. Bioinformatics* **2014**, 46, 13.24.1
15. Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem*. **2012**, 84, 5035.
16. Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Orešič, M. *BMC Bioinform*. **2007**, 8, 93.