Journal of Chromatography A, 1358 (2014) 155–164

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery

Jamie B. Coble [a], Carlos G. Fraga [b,*]

[a] Department of Nuclear Engineering, University of Tennessee, 210 Pasqua Engineering Bldg, Knoxville, TN 37996, USA
[b] Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA 99352, USA

ARTICLE INFO

Article history:
Received 6 March 2014
Received in revised form 9 May 2014
Accepted 18 June 2014
Available online 7 July 2014

Keywords:
Chemical forensics
Chemometrics
Metabolomics
Biomarkers
Impurity profiling
Metabolite profiling
GC/MS
LC/MS
MetAlign
MZmine
SpectConnect
XCMS

ABSTRACT

Preprocessing software, which converts large instrumental data sets into a manageable format for data analysis, is crucial for the discovery of chemical signatures in metabolomics, chemical forensics, and other signature-focused disciplines. Here, four freely available and published preprocessing tools known as MetAlign, MZmine, SpectConnect, and XCMS were evaluated for impurity profiling using nominal mass GC/MS data and accurate mass LC/MS data. Both data sets were previously collected from the analysis of replicate samples from multiple stocks of a nerve-agent precursor and method blanks. Parameters were optimized for each of the four tools for the untargeted detection, matching, and cataloging of chromatographic peaks from impurities present in the stock samples. The peak table generated by each preprocessing tool was analyzed to determine the number of impurity components detected in all replicate samples per stock and absent in the method blanks. A cumulative set of impurity components was then generated using all available peak tables and used as a reference to calculate the percent of component detections for each tool, in which 100% indicated the detection of every known component present in a stock. For the nominal mass GC/MS data, MetAlign had the most component detections followed by MZmine, SpectConnect, and XCMS with detection percentages of 83, 60, 47, and 41%, respectively. For the accurate mass LC/MS data, the order was MetAlign, XCMS, and MZmine with detection percentages of 80, 45, and 35%, respectively. SpectConnect did not function for the accurate mass LC/MS data. Larger detection percentages were obtained by combining the top performer with at least one of the other tools such as 96% by combining MetAlign with MZmine for the GC/MS data and 93% by combining MetAlign with XCMS for the LC/MS data. In terms of quantitative performance, the reported peak intensities from each tool had averaged absolute biases (relative to peak intensities obtained using instrument software) of 41, 4.4, 1.3 and 1.3% for SpectConnect, MetAlign, XCMS, and MZmine, respectively, for the GC/MS data. For the LC/MS data, the averaged absolute biases were 22, 4.5, and 3.1% for MetAlign, MZmine, and XCMS, respectively. In summary, MetAlign performed the best in terms of the number of component detections; however, more than one preprocessing tool should be considered to avoid missing impurities or other trace components as potential chemical signatures.

1. Introduction

Metabolomics [1] and chemical forensics [2] are two scientific disciplines that involve the discovery of chemical signatures, which can be molecular, ionic, elemental, or isotopic components or measurements that are characteristic to a specific phenomenon. In metabolomics, much of the applied research focuses on locating molecular biomarkers for the diagnoses of diseases. In chemical forensics much of the research focuses on determining chemical attribution signatures (CAS) such as trace impurities that match a chemical or mixture of interest to its source (e.g., synthesis route, starting material, or place of origin) for forensic applications. Each of these disciplines relies heavily on gas chromatography/mass spectrometry (GC/MS) and liquid chromatography/mass spectrometry (LC/MS) for the targeted and untargeted profiling of molecular compounds in biological or chemical samples that may be useful biomarkers or CAS, respectively. The samples analyzed are typically members of different groups. For example, in metabolomics, one sample group may be from healthy subjects and the other from a sickened group; in chemical forensics, a sample group may be from a different batch of a key toxicant precursor. Typically, the raw chromatography/mass spectrometry data are analyzed

* Corresponding author. Tel.: +1 509 371 7581; fax: +1 509 375 2227.
E-mail address: carlos.fraga@pnnl.gov (C.G. Fraga).

http://dx.doi.org/10.1016/j.chroma.2014.06.100
0021-9673/

by preprocessing tools to create a peak table, which ideally is a comprehensive and quantitative catalogue of all detected and non-detected chromatographic peaks across multiple samples from two or more sample groups. Ultimately, chemometric techniques are utilized on this preprocessed data (i.e., peak table) to determine those chromatographic peaks that best differentiate and characterize each of the sample groups.

Several preprocessing tools that incorporate different approaches and options are available for chromatography/mass spectrometry data, as described in recent review articles [3–5]. The core functions performed by most tools are filtering, peak detection, and peak matching. Filtering and peak detection focus on finding real chromatographic peaks in each data file (i.e., sample) while peak matching focuses on locating and cataloguing detected peaks (i.e., chemical components) that are present in more than one data file. Peak matching typically requires retention-time alignment of each data file to correct for run-to-run retention time shifts. The main output of a preprocessing tool is a peak table where each detected chromatographic peak is characterized by a specific retention time and m/z value, and the intensity of that peak in each sample (i.e., data file) is reported. If a preprocessing tool consistently misses a peak (having a specific m/z and retention time) or fails to properly match it in several data files, then the component associated with that peak (assuming it has no other reported characteristic peaks) will have no chance of becoming a useable chemical signature. In signature discovery, raw chromatography/mass spectrometry data must first be comprehensively profiled for component signals in order to increase the chances of finding those chemical components that are useful signatures. Herein, four automated preprocessing tools for chromatography/mass spectrometry data were primarily evaluated to determine what tool or combination of tools facilitates the discovery of the greatest number of chemical components consistently present in a sample group and absent in method blanks. The selected preprocessing tools were MetAlign [6], MZmine [7], XCMS [8], and SpectConnect [9]. All four tools are freely available, their algorithms published, and their source code accessible; they are also highly cited, except for SpectConnect which was included because it uses an approach unique to the other three.

MetAlign, MZmine, SpectConnect, and XCMS were evaluated based on qualitative and quantitative performance, that is, their thoroughness for detecting chemical components through the detection of chromatographic peaks, and the accuracy and precision of the detected chromatographic peak intensities. The data used for this evaluation were nominal mass GC/MS data and accurate mass LC/MS data previously collected from the analysis of replicate samples from multiple stocks of a nerve-agent precursor and method blanks. In this study, the stock samples were one group and the method blanks were another. After parameter optimization, each tool generated a single peak table. Each peak table was then processed to determine the number of impurity components reported in every replicate sample per stock (5 stocks for GC/MS and 10 stocks for LC/MS) and reported absent in the method blanks; this peak-table processing permitted tool comparisons and provided a more relevant assessment of tool performance. For instance, not every tool reports the same m/z chromatographic peak(s) corresponding to a specific component therefore component detection is a better metric than peak detection. In addition, a component reported in all replicate samples (from a given stock) is likely a more persistent signature than one reported in a fraction of the samples. After processing each peak table, a comprehensive component table was generated by combining the results from each tool. This comprehensive table was used to determine the percentage of component detections for each tool or combination of tools relative to the total number of component detections. The accuracy and precision for each tool was measured by the averaged bias and its standard deviation for the intensities of peaks that were detected by all the preprocessing tools. The true peak intensities for the bias measurements were determined using the instrument data software and by visual verification of the peak integration boundaries and heights.

All the preprocessing tools were new to the user (i.e., the first author) and therefore no preconceived notions or experience with any of the tested tools existed prior to the study. Analysis by a user initially unfamiliar with all of the tools helped prevent an unintentional bias toward any of the tools while also making it more relevant to novice or future users. Although the user gained considerable experience with each tool prior to tool comparisons, an individual (such as a tool developer) who is highly experienced with one of the tools would likely produce better results with that tool. While this approach was not infallible, it was practical and provided an additional level of objectivity to the results. While previous reviews of data preprocessing tools compared usability [10], performance on a single task (e.g., retention time alignment [11,12] or spectral deconvolution [13]), or simple enumeration of the available capabilities [14], this assessment extends these prior efforts by measuring and comparing the overall qualitative and quantitative performances of each tool on both nominal and accurate mass chromatography/mass spectrometry data.

1.1. MetAlign

MetAlign is the third most cited preprocessing tool of 21 preprocessing tools in metabolomics papers published between 1995 and 2013 as searched using SciFinder (www.cas.org/products/scifinder) and Web of Science (www.thomsonreuters.com/web-of-science). MetAlign was developed at RIKILT Institute of Food Safety to identify statistically significant differences between classes of full scan LC/MS or GC/MS data sets [6,15]. MetAlign inherently allows for comparison between two conditions, such as the stock samples and the method blanks. When both groups are defined, MetAlign attempts to select peaks present in one group but not the other. MetAlign 3.0 (www.metalign.nl) was used in this study.

1.2. MZmine

MZmine is the second most cited preprocessing tools (XCMS is first) in metabolomics papers according to our literature search. MZmine supports several stages of data preprocessing, including spectral filtering, peak detection, alignment, and normalization [7]. MZmine supports multiple data file types, including open source formats (mzML, mzXML, mzData, NetCDF) and select proprietary formats (e.g., Thermo RAW). MZmine includes visualization tools, both plots and tables, to allow for easy comparison of data across multiple samples. MZmine 2.5 (originally downloaded from http://mzmine.sourceforge.net) was used for the analyses presented here.

1.3. SpectConnect

SpectConnect was built on the Gemoda discovery algorithm in Python by researchers at the Massachusetts Institute of Technology (MIT) to systematically detect conserved metabolites across samples without a reference library [9]. SpectConnect detects conserved components by comparing each spectrum in a sample with every spectrum in every other sample. This exhaustive search allows SpectConnect to find the components that are conserved across replicate samples and the components that differentiate sample conditions or stocks. SpectConnect requires replicate samples for each condition to account for systematic error in peak detection and deconvolution software.