Clinical Mass Spectrometry 3 (2017) 7–12

Contents lists available at ScienceDirect

Clinical Mass Spectrometry

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

Improvement of drug identification in urine by LC-QqTOF using a probability-based library search algorithm

Jennifer M. Colby [a,*], Jeffery Rivera [b], Lyle Burton [b], Dave Cox [b], Kara L. Lynch [c]

[a] Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA
[b] SCIEX, Concord, ON, Canada
[c] Department of Laboratory Medicine, University of California San Francisco and San Francisco General Hospital, San Francisco, CA, USA

ARTICLE INFO

Article history:
Received 16 December 2016
Received in revised form 13 April 2017
Accepted 13 April 2017
Available online 19 April 2017

Keywords:
Library matching
Algorithm
Mass spectrum
High resolution mass spectrometry
Toxicology

ABSTRACT

A common method for identifying an unknown compound involves acquiring its mass spectrum and then comparing that spectrum against a spectral database, or library. Accurate comparison and identification is dependent on the quality of both the library and the test spectrum, but also the search algorithm used. Here, we describe a redesigned probability-based library search algorithm (ProLS) and compare its performance against two predicate algorithms, AMDIS from NIST (NIST) and LibraryView/MasterView (LV/MV), on human urine samples containing drugs of interest that were analyzed by quadrupole-time of flight (QqTOF) mass spectrometry. Each algorithm was used to compare the spectral data collected against an in-house spectral library. ProLS outperformed both NIST and LV/MV in efficiency of drug detection. Additionally, it demonstrated a scoring profile that resulted in an increased likelihood of low match scores for compounds that were absent from a sample. Increased scoring accuracy has the potential to reduce the time that analysts spend manually reviewing match data. Although search algorithms tend to be underappreciated, since they are not typically part of the end-user interface, this work illustrates how a redesigned algorithm can impact the accuracy of identification of small molecules in a biological matrix, and influence the overall utility of a bioanalytical method.

1. Introduction

Mass spectrometry-based techniques have an extensive history of use in the identification of small molecules in biological samples. Gas chromatography–mass spectrometry (GC–MS) is often referred to as the gold standard technique for broad spectrum drug screening in clinical and forensic applications because GC–MS fragmentation patterns contain structural information about the analyte. Molecular identification is achieved by comparison of fragmentation patterns to a database of reference spectra, also known as a spectral library. Regardless of the pre-introduction separation methodology used, mass spectrometers such as triple quadrupoles (QqQ), quadrupole linear ion traps (QqLIT) and quadrupole time-of-flight (QqTOF) instruments are all capable of producing robust fragmentation spectra though collision induced dissociation. Recently, bioanalaytical laboratories have begun transitioning to liquid chromatography–mass spectrometry (LC–MS). Although LC–MS product ions are produced through a slightly different process than the ions in a GC–MS spectrum, LC–MS product ion fragmentation spectra still contain structural information about the original molecule and can be used for identification.

Modern computer-based algorithms have drastically decreased the time required to match fragmentation patterns against spectral libraries [1]. While a variety of these algorithms have been developed, one of the most commonly employed is the dot-product algorithm [2–7]. However, the use of dot-product algorithms on spectra collected from LC–MS systems, especially systems that measure mass with high accuracy, has highlighted several shortcomings. The primary issue stems from its initial design purpose, which was for use with nominal-mass electron-ionization spectra. High-resolution accurate-mass spectra, on the other hand, require adjustments of the mass tolerance function to account for error inherent in mass measurements, and in the intensity function to account for variations in intensity that occur in collision-induced dissociation spectra. One proposed solution to address the discrepancies of these settings between nominal and high resolution mass analysis is to utilize a continuous probability function that transi-

* Corresponding author.
E-mail address: jennifer.colby@vanderbilt.edu (J.M. Colby).

http://dx.doi.org/10.1016/j.clinms.2017.04.001

tions smoothly from 1.0 (good match) to 0.0 (no match), rather than using a discrete tolerance.

To examine the potential benefits of a continuous function on spectral matching, we compared the performance of three library search algorithms, all based on the dot-product algorithm. Two algorithms that use discrete tolerances – NIST, a composite algorithm that makes use of peak averaging, and LibraryView/MasterView (LV/MV), the predicate SCIEX algorithm – were compared against Probability-based Library Search (ProLS), a newly designed algorithm that implements a continuous tolerance. We used an LC-QqTOF to collect full-scan data on 41 human urine samples, with information-dependent acquisition of product ion fragmentation spectra. These urine samples were extensively characterized and served as known samples for this study. We then tested the ability of each algorithm to score matches between collected spectra and an in-house spectral library. The performance of each algorithm was evaluated based on scores assigned to compounds known to be present in, or absent from, the sample, as well as overall efficiency of compound detection. We found that the ProLS algorithm outperformed both the NIST and LV/MV algorithms, displaying an improved scoring profile and the highest overall efficiency. We attribute the increase in performance to the enhanced selectivity of the probabilistic approach when aligning mass and intensity features.

2. Materials and methods

2.1. Library spectra

Library spectra were collected on a SCIEX 5600 TripleTOF QqTOF mass spectrometer operated in product ion scanning mode. Each of the 169 reference standards was acquired from one of the following vendors: Cerilliant, Round Rock, TX; Alltech/Grace, Columbia, MD; Cayman Chemicals, Ann Arbor, MI; Toronto Research Chemicals, Toronto, ON, CAN. Standards were diluted to 500 ng/mL with 50:25:25 water:methanol:acetonitrile and injected onto the LC-QqTOF. Ionization was performed using the following settings: ion source gas 1, 30 psi; ion source gas 2, 30 psi; curtain gas, 25 psi; temperature 500 °C; ion spray voltage floating, 5500 V; declustering potential, 100 V. Data was collected from 50 to 700 *m*/*z*. The mass spectrometer was calibrated using an automatic calibrant delivery system that injected calibration solution (SCIEX, Framingham, MA) every five samples. Spectra were collected with a rolling collision energy [8]. The collected spectra were averaged across the chromatographic peak and the averaged spectrum was added to the library. This process was performed for each of the 169 drugs included in this analysis.

2.2. Test spectra

Test spectra were acquired in human urine samples that were subjected to extensive toxicological analysis using a variety of methods and found to contain drugs of abuse. We considered the presence of a drug to be confirmed if it was validated by prescription records or by another analytical technique (e.g., LC–MS/MS, GC–MS). The test spectra were collected using the method described by Thoren et al. [8]. Briefly, 41 remnant urine samples sent for routine clinical toxicology analysis were subjected to chromatographic separation prior to introduction into a SCIEX QqTOF mass spectrometer. Full-scan data was collected from 50 to 700 *m*/*z*, with information-dependent acquisition of product ion spectra. Data files were analyzed using the MasterView function within PeakView (version 2.0, SCIEX, Framingham, MA). Full-scan QqTOF data was compared against an extracted ion chromatogram (XIC) database containing expected mass, chromatographic retention time, and isotope pattern for each of the 169 drugs. Product ion spectra were compared against the in-house spectral library using each of the three test algorithms: NIST, LV/MV, and ProLS. All three algorithms were implemented in MasterView 1.1; the versions for ProLS and NIST were custom-built for the purpose of this study. A precursor window of 0.4 Da and a MS/MS window of 0.4 Da were used for library searches with all three algorithms.

2.3. Search algorithms

Three different algorithms based on the dot-product backbone were compared; the NIST algorithm (AMDIS build 121.86), the SCIEX LV/MV algorithm, and the SCIEX ProLS algorithm. Although alternative algorithms are available, most require the user to purchase software. We selected NIST because it is commonly used in toxicology, and LV/MV and ProLS are the algorithms used by the data analysis software provided by our mass spectrometer vendor (SCIEX).

Briefly, dot-product algorithms treat each spectrum as a vector in multi-dimensional space; the dimensions represent the *m*/*z* values present in the unknown spectrum and the reference spectrum. Each *m*/*z* value is mass-aligned and compared with the same *m*/*z* value from the reference spectrum. The cosine of the angle between the two vectors is calculated, and possibly mathematically transformed, to generate a score reflecting the likelihood that the unknown spectrum and the library spectrum reflect the same compound. Accurately setting the mass accuracy and intensity tolerance factors is crucial to algorithmic performance. Considering mass alignment, if the tolerance is set for a higher degree of mass accuracy than the instrument is currently achieving, theoretically identical *m*/*z*, which are measured as just outside the tolerance, will not be aligned; if the tolerance is set below the mass accuracy the instrument is achieving, then different *m*/*z* may be considered to be the same. A similar issue occurs with intensity matching. The overall probability that two peaks match is the product of the separate mass and intensity probabilities:

$$p = P_{mz}(\Delta m) \cdot P_{intensity}(\text{intensity ratio})$$

where P_{mz} is the mass matching probability, which is a function of Δm (the measured mass difference), and $P_{intensity}$ is the intensity-ratio probability, which is a function of the intensity ratio (the ratio of the relative observed and library intensities).

The overall library score, which reflects the likelihood of a match between the unknown and reference spectrum, is the square of the dot-product (with each vector normalized to unit length) and is calculated as:

$$\text{score} = 100.0\% * \left(\sum U_i L_i\right)^2 / \left(\left(\sum U_i U_i\right)\left(\sum L_i L_i\right)\right)$$

where L_i is the intensity of the i'th library peak and U_i is the intensity of the i'th unknown peak. L_i and U_i are always non-negative numbers; if background subtraction is performed, any 'negative' intensities are replaced by zero. The U_iL_i sum is calculated using peaks common between the two spectra; the other sums (i.e., U_iU_i and L_iL_i) are calculated using all unknown and library peaks.

All three algorithms tested are based on the dot-product algorithm; a critical difference between the three is the type of probability function each uses. The NIST and LV/MV algorithms use dot-product algorithms with discrete tolerances, meaning the probability of matching drops immediately to zero when the tolerance is exceeded. The ProLS algorithm uses a dot-product algorithm with a continuous probability function, meaning the probability of matching decreases smoothly to zero when the tolerance is exceeded (Fig. 1). For example, P_{mz} is equal to 1.0 if Δm is within a relatively stringent mass tolerance; is linearly interpolated if