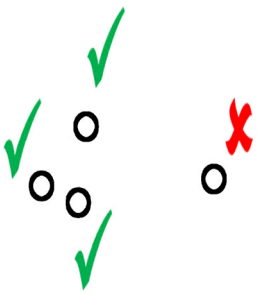
J. Am. Soc. Mass Spectrom. (2017) 28:733–738
DOI: 10.1007/s13361-016-1589-4

RESEARCH ARTICLE

Mass Spectral Library Quality Assurance by Inter-Library Comparison

William E. Wallace,[1] Weihua Ji,[1] Dmitrii V. Tchekhovskoi,[1] Karen W. Phinney,[2] Stephen E. Stein[1]

[1]Mass Spectrometry Data Center, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899-8362, USA
[2]Bioanalytical Science Group, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8314, USA



Abstract. A method to discover and correct errors in mass spectral libraries is described. Comparing across a set of highly curated reference libraries compounds that have the same chemical structure quickly identifies entries that are outliers. In cases where three or more entries for the same compound are compared, the outlier as determined by visual inspection was almost always found to contain the error. These errors were either in the spectrum itself or in the chemical descriptors that accompanied it. The method is demonstrated on finding errors in compounds of forensic interest in the NIST/EPA/NIH Mass Spectral Library. The target list of compounds checked was the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) mass spectral library. Some examples of errors found are described. A checklist of errors that curators should look for when performing inter-library comparisons is provided.

Keywords: Cross-check, Curation, Data validation, Electron ionization mass spectrometry, Quality control, Standard reference data

Received: 28 November 2016/Revised: 22 December 2016/Accepted: 23 December 2016/Published Online: 26 January 2017

Introduction

Standard reference libraries are widely used in mass spectrometry for the identification of unknown compounds [1]. It is a central task of the library curator to ensure that each entry is accurate [2]. Accuracy rests not only on having a high quality, representative mass spectra but also on having correct and self-consistent chemical identifiers for each library entry [3–6]. A critical means of ensuring entry reliability is to find agreement between two spectra of the same compound measured by two laboratories using different chemical sources [7, 8]. A useful way to expand this concept is through comparison of entries of the same compound among multiple libraries to identify anomalies and resolve any significant differences. An anomaly is defined as an entry whose mass spectrum departs from the consensus. It has long been known in analytical chemistry that with three or more entries, consensus building becomes a powerful tool to discover outliers. Through inter-library comparison, the quality assurance methods used in creating the individual reference libraries are implicitly applied in finding anomalies within other libraries. In addition to correcting spectrum anomalies, chemical identification errors must also be identified. This can be done by inter-library comparison as well as through the use of independent chemical identification resources. Both approaches are used in this work to ensure chemical identification accuracy. Ultimately, a well-crafted library relies both on good quality mass spectra and on complete and correct compound identification information.

The goal of this work was to use entries in other libraries to improve the quality of the entries in the NIST/EPA/NIH Mass Spectral Library (NIST Standard Reference Database 1A) that pertain to the identification of a specific class of compounds, seized drugs of forensics interest, but the approach could be applied to the library in its entirety. To accomplish this, a comprehensive target list of seized drugs was checked in an automated fashion against the NIST14 library and several other prominent, curated electron-ionization mass spectral libraries. The target list itself was the mass spectral library maintained by the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) [9]. The SWGDRUG library is compiled by the US Drug Enforcement Administration in cooperation with other widely recognized forensics laboratories. It

Correspondence to: William E. Wallace; *e-mail:* william.wallace@nist.gov

Table 1. Libraries Used in the Inter-Library Comparison

Library Name	Number of Compounds	Version Information
NIST/EPA/NIH Mass Spectral Library (NIST14)	242,466	Standard Reference Database 1A 2014 release
SWGDRUG Mass Spectral Library	2293	version 2.3 release date 04-01-15
Mass Spectra of Designer Drugs P. Roesner, Wiley, 2014 (DD2014)	16,343	2014 edition ISBN: 978-3-527-33795-8
Mass Spectral Library of Drugs, Poisons, Pesticides, Pollutants and Their Metabolites H.H. Maurer, K. Pfleger, A.A. Weber, Wiley, 2011 (DPPP2011)	8650	2011 edition ISBN: 978-3-527-32398-2
Cayman Spectral Library	748	Cayman Chemical Company release date 09-29-14
Identification of Essential Oil Components by Gas Chromatography/Mass Spectrometry R.P. Adams, Allured Publishing, 2007	2205	4th edition ISBN: 978-1932633214

was employed because it is an up-to-date, freely available seized drug library used by many forensic laboratories across the country.

Methods

The initial step was to find for each compound in the target list all spectra in all libraries. This was done by generating and then matching InChIKeys [10] and associated spectra in all libraries. In concept, this is similar to what Oberacher and coworkers have done in comparing two tandem mass spectral libraries [11]. Calculation of the InChIKey was based on the compound structure provided by each library. Compounds having the same InChIKey were grouped and inter-compared for all entries across all libraries. Of course, an incorrect structure would lead to an InChIKey that does not represent the actual compound measured. Such errors were typically discovered in the mass spectrum comparison step where they tended to yield low match factors, say below 650. Additionally, it should be noted that the InChIKeys used here did not contain stereoisomer information because most libraries used did not provide this information. For example, the InChIKeys of stereoisomers cocaine, allococaine, pseudoallococaine, and allopseudococaine are identical. Mitigating this effect is the fact that stereoisomers typically have indistinguishable electron ionization mass spectra, so they are effectively treated as measurement replicates for the purposes of this study.

The NIST Mass Spectral Search Program (ver. 2.2) [12–15] was used in an automated batch mode for making spectrum comparisons. The program uses a modified vector dot product to calculate a match factor that scales from zero (no match) to 999 (identical spectra). All libraries used were available in the NIST data format, simplifying the task of comparison. An inventory of all libraries involved is shown in Table 1. In cases where three or more entries were available, a consensus determination could be made and the NIST spectrum was classified as consistent or inconsistent depending on its match scores. In cases where only one instance of a compound was available (i.e., the entry only existed in the SWGDRUG library), or when only two entries were available, no consensus determination was possible. Inconsistent spectra were flagged for further scrutiny using NIST's MS Interpreter program (ver. 2.0). MS Interpreter uses estimates of bond dissociation energies and well-understood reaction paths to assign the major peaks in a mass spectrum. Note that although spectra can be found to be consistent with the various heuristic fragmentation rules, spectra cannot, at present, be predicted to a useful level of accuracy.

Chemical information curation relied on mutual intercomparison of the identification data for each target compound

Table 2. Resources Used in Verifying Chemical Identifiers

Resource	Curator	Web Address
NIST Chemistry WebBook (Standard Reference Database 69)	National Institute of Standards and Technology	webbook.nist.gov/chemistry
SciFinder	Chemical Abstracts Service of the American Chemical Society	scifinder.cas.org
ChemSpider	Royal Society of Chemistry	www.chemspider.com
PubChem	National Center for Biotechnology Information of the National Institutes of Health	pubchem.ncbi.nlm.nih.gov
Forendex	Southern Association of Forensic Scientists	forendex.southernforensic.org

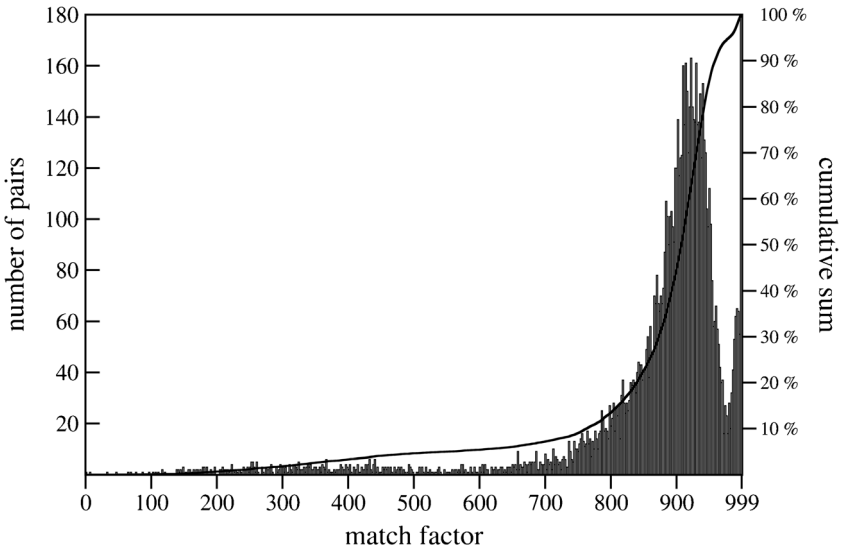


Figure 1. A histogram of the match factors when each element in the target list is compared with the same chemical compound in all the libraries (not including the SWGDRUG library, which provided the target list). Left axis: number of pairs with a given match factor. Right axis: solid line shows the cumulative sum of pairs moving from left to right across the match factor axis

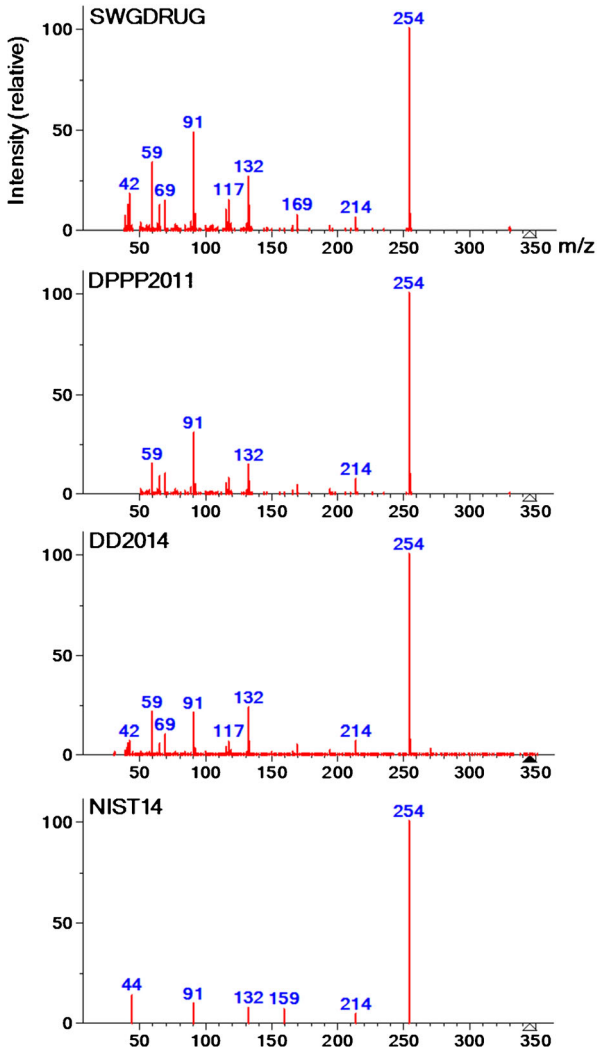
entry found in the NIST14 library. If an anomaly in chemical identification could not be resolved by inspection, outside sources of chemical information as found in Table 2 were used. Chemical structure drawing software was also used, in some cases, to check for correct chemical formulas and molecular masses.

Results

Figure 1 shows a histogram of the resulting scores when each element in the target list is compared with all the other libraries (with the exception of the SWGDRUG library, which is itself the target list). Several salient observations can be made. The

Table 3. A Checklist of Mass Spectral Library Error Categories

Error	Causes and comments
Spectrum errors (assuming the compound is correctly described)	
Missing peaks	• Reaction, chemical ionization, or unexpected decomposition of the sample in the inlet system or ion source
	• Inadequate instrument resolution or poor peak shape (missing peaks are often small peaks adjacent to much larger peaks)
	• Low signal-to-noise ratio and/or instrument noise threshold set too high (the latter sometimes detectable as distorted isotope ratios)
Spurious peaks and/or incorrect peak intensities	• Impurity in the measured compound, especially when a separation method is not used before ionization or when co-elution of compounds occurs
	• Reaction, chemical ionization, or unexpected decomposition of the sample in the inlet system or ion source
	• Peak splitting due to incorrect determination of peak position (often the result of poor peak shape)
	• Noise (often the result of electronics or instrument environment problems)
	• Lack of, or errors in, correction for background (often air, water, or chromatographic column bleed which can be reduced by proper instrument care and the use of peak deconvolution software)
	• Detector saturation
Incorrect mass assignments	• Incorrect mass calibration of instrument (often the result of infrequent calibration or large room temperature variations)
	• Arbitrary assignment of multiply charged ions not occurring at an integer mass to an adjacent integer mass
	• Incorrect rounding or truncation of mass values from decimal values to integer values during data processing
Compound description errors (assuming the spectrum is accurate)	
Incorrect chemical structure	• Mislabeled sample
	• Incorrect structure assigned to a correct chemical identifier
	• Structure does not specify isomer
Incorrect formula and/or molecular mass	• Chemical structure, chemical formula, and molecular mass must be self-consistent
Incorrect compound name	• Multiple inherently complex chemical nomenclature systems and many trivial and product names with variations according to language in common usage
Other incorrect identifier	• Wrong Chemical Abstracts Service registry number
	• Incorrect InChI code or InChIKey
Curation errors	
Identical spectrum repeated in library under different chemical identifier	• Failure to identify and remove redundant entries
Corruption of the data file	• Inadvertent loss or alteration of information that occurs when copying, transmitting, or otherwise processing data files



Match Factor Matrix	SWGDRUG	DPPP2011	DD2014	NIST14
SWGDRUG	999	925	859	574
DPPP2011	925	999	882	664
DD2014	859	882	999	674
NIST14	574	664	674	999

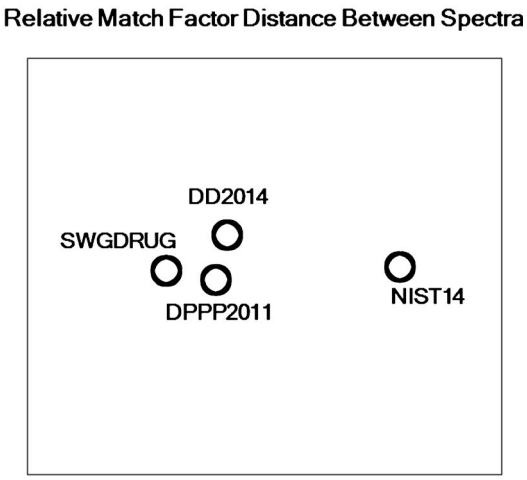


Figure 2. An example of the discovery of an anomalous spectrum in the NIST14 library (see text for details)

histogram has a peak at a match factor of approximately 925. This arises from the fact that good quality spectra from curated libraries show excellent reproducibility leading to high inter-library match factors. The histogram also has a peak at very high match factors. This arises from SWGDRUG entries also appearing in other libraries. Ideally these entries should all have match factors of 999 but subtle differences in how mass and intensity data are treated by various curation schemes slightly lower the match factors, giving values from 975 to 999. Redundant entries were never used to determine outlier status of any library entry since redundant entries add no new information. As points of reference, it can be observed from the cumulative (running) sum shown in Figure 1 that about 92% of the spectrum pairs have match factors above 750, and about 87% of the spectrum pairs have match factors above 800, two touch points in determining spectral match quality. Finally, the low match factor tail contains the outliers whose presence was determined by visual inspection of spectra. These may be due to either poor quality mass spectra or due to improper chemical structures recorded in the library. Investigation of each individual case is necessary to determine which error has occurred, although in a few cases a determination could not be made.

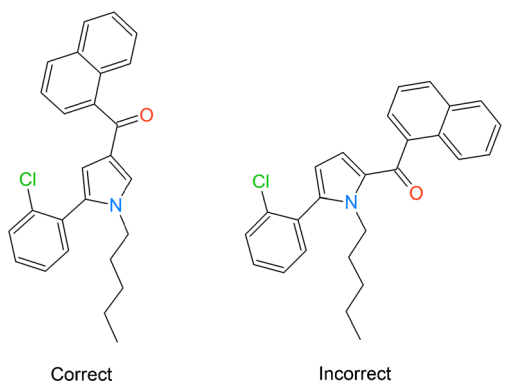


Figure 3. The correct chemical structure of JWH-369 (left), and an erroneous structure found in the NIST14 library (to be corrected in the NIST17 library). Their respective InChIKeys are SUELCWQJMQQCTF-UHFFFAOYSA-N, and HGUBBSHGGLHABY-UHFFFAOYSA-N

Of the 2283 target list spectra examined, 605 were not in the NIST14 library and another 89 had only one additional

measurement making inter-comparison impossible. Many of those not in the NIST14 library have already been measured and will appear in the upcoming release of the library, NIST17. About 1% of the compounds in the target list required further examination. Each proved to have unique anomalous behavior, often with multiple errors; nevertheless, the types of errors that were found can be generalized according to categories, which could serve as a useful checklist for library quality control and curation as shown in Table 3. The first level divides errors into three broad categories: those with the library spectrum, those arising from identity information, and those having to do with curation or record keeping issues. For each broad category of error, typical characteristics, symptoms, or examples are given.

What follows is an example of an anomalous entry where the NIST spectrum was outside the consensus. Figure 2 shows four different library entries for phentermine, a prescription-only appetite suppressant, which has been derivatized with heptafluorobutyric anhydride to aid in gas chromatography separation. The NIST14 entry in the lower left of the figure stands out as being different from the other three. The upper right corner of Figure 2 shows the match factor similarity matrix for the four spectra. This is a diagonally symmetric matrix of the mutual match factors for the four spectra. Note that a match factor of 999 lies along the matrix diagonal since each spectrum matches itself perfectly. In the lower right corner of Figure 2, the similarity matrix has been converted by multi-dimensional scaling (MDS) [16, 17] into a two-dimensional spatial representation of the similarities between spectra. By visual inspection, the NIST14 spectrum was determined to be an outlier. Closer examination of the NIST library spectrum provides two obvious indications leading to its anomaly status: it contains peaks at *m/z* 159 and *m/z* 44 not found in the others, and it is missing a peak at *m/z* 59 common to the others. These anomalies can be traced to the original published spectrum [18]. The spectrum of heptafluorobutyric anhydride-derivatized phentermine will need to be remeasured to be included in future releases of the NIST library.

Regarding errors concerning chemical identification, it was found that the NIST library had 14 entries with incorrect structures and two entries with incorrect compound names. The misidentification of positional isomers was often the source of chemical structure errors. For example, in Figure 3 the structure of the cannabinoid JWH-369 is shown. The left structure in the figure is correct, the right structure is as it appeared in the NIST14 library (to be corrected in the NIST17 library) showing a positional isomer of JWH-369. The correct structure is a common cannabinoid type with the pendant groups at the 2 and 4 positions of the pyrrole ring. The erroneous structure has the groups pendant at the 2 and 5 positions. References to such 2,5-structures have been patented [19] but were apparently never commercialized. They were not found in any seized-drug database consulted.

A common observation in this work was that chemists and non-chemists alike struggle with proper unambiguous naming of seized drug compounds. A correct IUPAC name for JWH-369 is [5-(2-chlorophenyl)-1-pentyl-1H-pyrrol-3-yl](naphthalene-1-yl)methanone, but it may be referred to as 1-pentyl-2-(2-chlorophenyl)-4-(1-naphthoyl) pyrrole [20]. A well-curated library would have each of these names linked to a structure and associated mass spectrum as synonyms. A synonym field allows for multiple names for the same compounds but just as importantly allows for the same common name to point toward multiple structures. This helps the analyst tease out naming issues but is very labor-intensive for the curator. Of course, identifying all entries primarily by the InChIKey virtually removes any name ambiguity issues. In this case, the name JW-369 was selected as the principal name for this compound since it is familiar to analysts most involved in its analysis. For other users, the chemical structure and alternate names will provide needed structural information.

Conclusion

A multi-library comparison method was used to find spectrum anomalies and chemical identity errors in a target list of chemical compounds of forensic interest. Multi-library inter-comparison reaches beyond the traditional pairwise comparison often used in library quality assurance. Anomalous spectra could be traced typically to missing or spurious peaks. Compound identity information errors were most often traced either to incorrect structures or ambiguous naming conventions. A checklist of common errors has been provided for MS library curators.

Acknowledgements

Discussions on the SWGDRUG library with Jason A. Bordelon, Drug Enforcement Administration (Southwest Laboratory) were highly informative. The authors thank Anzor Mikaia and Ed White V, both of the NIST Mass Spectrometry Data Center, as well as Peter Linstrom of the NIST Office of Data and Informatics, for critical evaluation of the manuscript.

References

1. Stein, S.: Mass Spectral Reference Libraries: an ever-expanding resource for chemical identification. Anal. Chem. **84**, 7274–7282 (2012)
2. Dasu, T., Johnson, T.: Exploratory data mining and data cleaning. Wiley-Interscience, Hoboken (2003)
3. Milne, G.W.A., Budde, W.L., Heller, S.R., Martinsen, D.P., Oldham, R.G.: Quality-control and evaluation of mass spectra. Org. Mass Spectrom. **17**, 547–552 (1982)
4. Terwilliger, D.T., Behbehani, A.L., Ireland, J.C., Budde, W.L.: The status and evaluation of a mass-spectral database. Biomed. Environ. Mass Spectrom. **14**, 263–270 (1987)
5. Lias, S.G.: Numerical databases for chemical analysis. J. Res. Natl. Inst. Stand. Technol. **94**, 25–35 (1989)
6. Yang, X.Y., Neta, P., Stein, S.E.: Quality control for building libraries from electrospray ionization tandem mass spectra. Anal. Chem. **86**, 6393–6400 (2014)
7. Stein, S.E., Ausloos, P., Lias, S.G.: Comparative evaluations of mass-spectral databases. J. Am. Soc. Mass Spectrom. **2**, 441–443 (1991)
8. Ausloos, P., Clifton, C.L., Lias, S.G., Mikaya, A.I., Stein, S.E., Tchekhovskoi, D.V., Sparkman, O.D., Zaikin, V., Zhu, D.: The critical evaluation of a comprehensive mass spectral library. J. Am. Soc. Mass Spectrom. **10**, 287–299 (1999)
9. Available at: http://www.swgdrug.org/ms.htm. Accessed 1 Apr 2015
10. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I.: InChI – the worldwide chemical structure identifier standard. J. Cheminformatics **5**, 9 (2013)

11. Oberacher, H., Whitley, G., Berger, B.: Evaluation of the sensitivity of the 'Wiley registry of tandem mass spectral data. MSforID' with MS/MS data of the 'NIST/NIH/EPA mass spectral library'. J. Mass Spectrom. **48**, 487–496 (2013)
12. Stein, S.E.: Estimating probabilities of correct identification from results of mass-spectral library searches. J. Am. Soc. Mass Spectrom. **5**, 316–323 (1994)
13. Stein, S.E., Scott, D.R.: Optimization and testing of mass-spectral library search algorithms for compound identification. J. Am. Soc. Mass Spectrom. **5**, 859–866 (1994)
14. Stein, S.E.: Chemical substructure identification by mass-spectral library searching. J. Am. Soc. Mass Spectrom. **6**, 644–655 (1995)
15. Stein, S.E., Heller, D.N.: On the risk of false positive identification using multiple ion monitoring in qualitative mass spectrometry: large-scale inter-comparisons with a comprehensive mass spectral library. J. Am. Soc. Mass Spectrom. **17**, 823–835 (2006)
16. Seber, G.: Multivariate observations. John Wiley and Sons, Inc., Hoboken (1984)
17. France, S.L., Carroll, J.D.: Two-way multidimensional scaling: a review. IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev. **41**, 644–661 (2011)
18. Thurman, E.M., Pedersen, M.J., Stout, R.L., Martin, T.: Distinguishing sympathomimetic amines from amphetamine and methamphetamine in urine by gas-chromatography mass-spectrometry. J. Anal. Toxicol. **16**, 19–27 (1992)
19. Laforest, J., Bonnet, J., Bessin, P.: New pyrrole derivatives, process for their preparation and therapeutic applications thereof. U.S. Patent 4,194,003, 18 Mar 1980
20. Huffman, J.W.: In: Reggio, P.H. (ed.) The Cannabinoid Receptors, pp. 49–94. Humana Press, New York (2009)