Vol. 9, No. 3, 2018
ISSN 2233-4203/ e-ISSN 2093-8950
LETTER www.msletters.org | Mass Spectrometry Letters

Identification Performance of Low-Molecular Compounds by Searching Tandem Mass Spectral Libraries with Simple Peak Matching

Boris L. Milman[1,*] and Inna K. Zhurkovich[2]

[1]*Institute of Experimental Medicine, ul. Akad. Pavlova 12, Saint Petersburg 197376, Russia*
[2]*Institute of Toxicology, ul. Bekhtereva 1, Saint Petersburg 192019, Russia*

Received July 4, 2018; Revised July 12, 2018; Accepted July 24, 2018
First published on the web September 30, 2018; DOI: 10.5478/MSL.2018.9.3.73

Abstract : The number of matched peaks (NMP) is estimated as the spectral similarity measure in tandem mass spectral library searches of small molecules. In the high resolution mode, NMP provides the same reliable identification as in the case of a common dot-product function. Corresponding true positive rates are (94±3) % and (96±3) %, respectively.

Keywords : Tandem mass spectrometry, mass spectral libraries, spectral similarity measure, chemical identification

Introduction

Non-target analysis is commonly performed by searching mass spectral libraries.[1,2] Within the last several years, these data resources have been increasingly generated for non-volatile compounds and high-resolution (HR) tandem mass spectrometry (MS^2).[1,3] In particular, these libraries have been steadily growing in the number and diversity of compounds. At the same time, computer programs for processing experimental mass spectra and their comparison to the reference data have been upgraded. Therefore, non-target determinations have become more widespread in environmental, pharmaceutical, and food analysis, as well as numerous –omics approaches, forensic investigations, and so on. The quality assurance of these analytical workflows closely connected with the emergence of new computer libraries/databases and softwares requires re-estimation of the performance rate of a library search and corresponding true/false identification rates.

A mass spectral library search is directed towards the retrieval of the best similarity between an experimental spectrum of unknown analytes and others from all the reference spectra. Usually, there is a two-dimensional similarity in both the *m/z* of ions and relative intensity of ion peaks. Corresponding measures/metrics/scores of the similarity for low-molecular compounds are commonly a dot-product function (DPF) and some other variables.[1,4]

We hardly need to mention that the simplest measure of similarity between mass spectra, i.e. ion mass/peak matching, is also widespread in mass spectrometry (MS). In high-resolution mass spectrometry (HRMS), a particular accurate mass search against various formula masses is widely used to generate molecular formulas of analytes. In peptide identification by (product) ion mass fingerprinting, a search for similarity in mass values within the predefined mass tolerance range is performed,[5] though DPF has been used with increasing frequency.[6] However, to the best of our knowledge, the number of matched peaks (NMP) similarity measure is usually not very common in mass spectral library searches of small molecules. Taking into account the overall progress in mass spectral libraries, it would be appropriate to study how that simplest similarity measure works for low-molecular compounds. This is particularly important for new libraries of *in silico* mass spectra because their peak intensities are predicted less well in comparison to corresponding product-ion masses.[7,8]

Here we present a brief report on the adequate case study of new versions of mass spectral collections and computer programs. In the work, we used the data from the MassBank of North America (MoNA) mass spectral repository.[9]

Methods

The file of various 44157 MS^2 spectra was downloaded and imported from MoNA into the NIST MS Search 2.3

*Reprint requests to Boris L. Milman
E-mail: bmilman@mail.rcom.ru

software.[10] According to our estimation, the file contained approximately $11 \cdot 10^3$ positive ion high-resolution mass spectra of a satisfactory quality. The "unknown" and reference spectra were originated from that conventional subfile and the entire initial file, respectively (Figure 1). A random sampling method[11] was used that enables to obtain rapid bias-free estimates.

Two raw subsets of 1% of initial spectra were randomly sampled to form the selection of spectra of "unknown" compounds. Both samples led to statistically indistinguishable results at 95% probability (see below) and the two search results were eventually combined. The samples were largely refined before the library search. One/two-peak spectra, very noisy ones, low-resolution (LR) and negative ion data, unique spectra of any compound, and some other spectra unsuitable for identification for a variety of reasons, were removed from the samples. The rest of the 217 items (the sum of 108 plus 109 extracts from two samples) was then considered as "unknown" spectra.

All of the remaining spectra, approximately $44 \cdot 10^3$ ones, including 11 thousand of the most relevant spectra, constituted the reference library (Figure 1). In library searches using the above-mentioned MS Search program, we selected the *Identity, MS/MS* search option. Two *m/z* tolerance ranges, ± 0.5 Da (by default) and ± 0.01 Da (the practical range previously tested[12]), were set for matching both precursor and product ions. Those ranges simulated the search in the LR and HR mode, respectively. The dot product of the "unknown" and library spectrum (*DotProd* in the MS Search program) and the number of "unknown" spectrum peaks matched to those in the reference spectrum

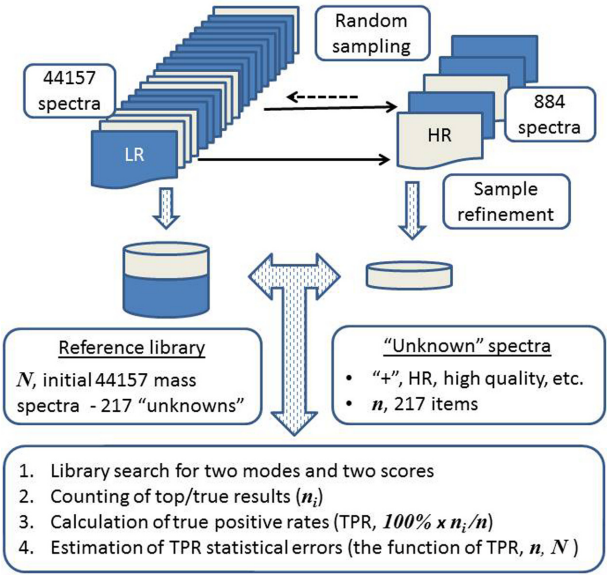

Figure 1. The schematic showing the formation of "unknown" spectra selection (sampled by using random numbers) and reference library, library searches, and the statistical analysis.

Table 1. True positive rates

Spectral similarity measure	Mass tolerance range*	
	±0.5 Da	±0.01 Da
DPF	(93±4) %	(96±3) %
NMP	(59±7) %	(94±3) %

*Overall statistical estimates and their confidence intervals are given in the parentheses

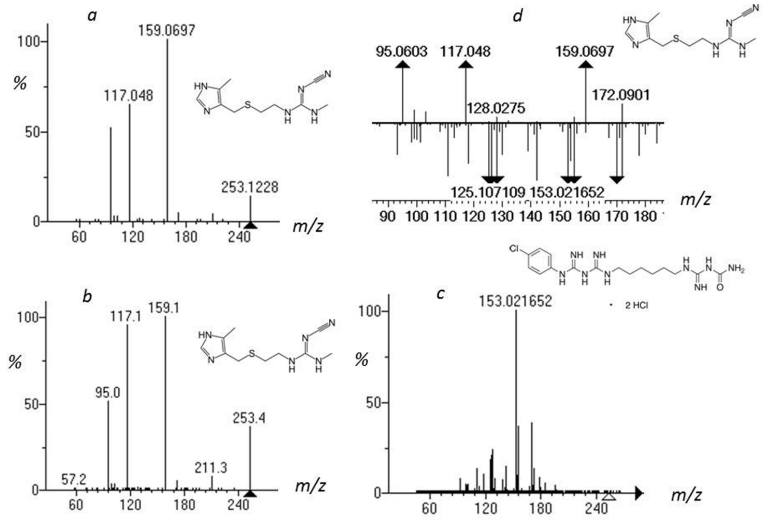

Figure 2. The LR mode. The "unknown" tandem mass spectrum of cimetidine (*a*) and the reference spectra with the most DPF (989 of 999, *b*, cimetidine) and NMP (19 of 19 or 100%, *c*, chlorhexidine dihydrochloride) values. The comparison (*d*) of spectra (*a*) and (*c*) demonstrates that the false identification is due to the occasional matching of "unknown" peaks to low background signals of the reference spectrum.

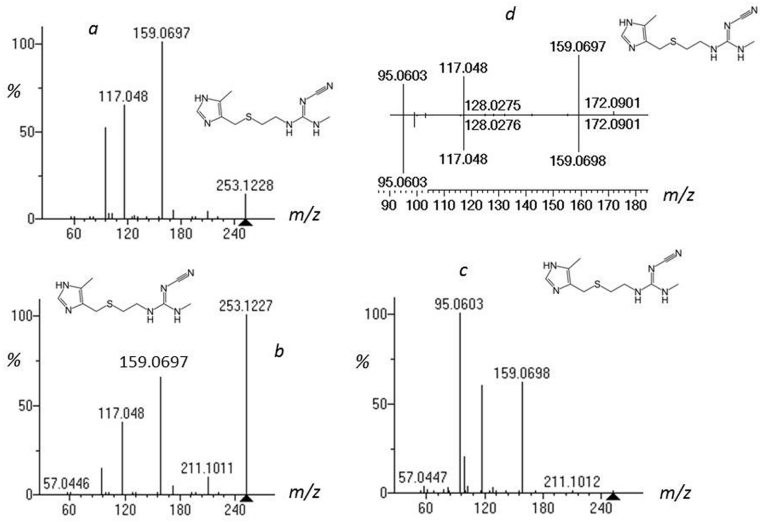

Figure 3. The HR mode. The "unknown" tandem mass spectrum of cimetidine (*a*) and the reference spectra with the most DPF (974 of 999, *b*, cimetidine) and NMP (16 of 19 or 84%, *c*, cimetidine) values. The comparison (*d*) of spectra (*a*) and (*c*) demonstrates that the true identification is due to the correct matching of "unknown" peaks to corresponding reference signals.

(*NumMP* in MS Search 2.3) within the mass tolerance, were of a two-dimensional and one-dimensional measure, respectively, for spectral similarity. The latter, NMP score, was expressed as the percentage in relation to its maximum value, i.e. the number of peaks in the "unknown" spectrum matched to itself. In the library search, the formal 1st rank of the reference spectrum of the same compound as "unknown" meant the true positive result (TP) of the search and, consequently, true identification. In some cases, the software calculated the same scores for reference spectra of several compounds and the foreign compound ranked first. This is one of reasons for a formally false result. Following all of the searches, the percentage rate of TP (TPR, sensitivity) was calculated (Figure 1). The statistical errors of result rates[13] as well as the significance of their differences[14] were estimated at 95% probability.

Results and Discussion

The searches for the LR mode predictably demonstrated that DPF resulted in far more true outcomes (TPR 93%) than NMP (59%) (see Table 1). The main origin of erroneous identification with the use of the NMP measure was the occasional matching of "unknown" peaks to low background signals of reference spectra when those were insufficiently cleared (the example in Figure 2).

The HR mode provided far more reliable identification in the case of NMP (94%) (see Table 1). It is evident that here occasional matching becomes less probable (the example in Figure 3). Another score, DPF, resulted in a bit truer outcome (TPR 96%) and the two rates seem to be statistically indistinguishable providing the statistical error (Table 1). Both rates are at a level of results of the interlaboratory comparison performed for HRMS^2 with the special measure of mass spectral similarity.[12]

The above NMP rate may be increased with an improvement in the quality of reference spectra and by increasing the number of replicate ones. On the other hand, the rate may diminish for the reference library with higher coverage of isomers and isobar compounds often having similar spectra. The library built here is not full of such candidates for identification: three or more compounds having the same or similar (within ± 0.01 Da range) molecular mass, were recorded in only about 40% of searches.

Conclusions

Therefore, with HRMS^2, the library search exploiting the NMP one-dimensional score seems to be acceptable for identification workflows at a screening level. We anticipate at least three variants of the use of this measure. First, the NMP measure of spectral similarity would be suitable for searching *in silico* mass spectral libraries[8] without regard to peak intensities. Second, (old) reference mass spectral data available in the literature as mass lists without peak intensities[15] can be adequately involved in a library search. Last, but not least, identification of low-molecular compounds and peptides could be further combined in corresponding software.

References

1. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. *Mass Spec. Rev.* **2018**, 37, 513.

2. Milman, B. L.; Zhurkovich, I. K. *TrAC Trends Anal. Chem.* **2017**, 97, 179.
3. Milman, B. L.; Zhurkovich I. K. *TrAC Trends Anal. Chem.* **2016,** 80, 636.
4. Milman, B. L. *Chemical identification and its quality assurance,* Springer: Heidelberg, **2011**.
5. Matrix Science. Mascot database; See http://www.matrixscience.com.
6. Shao, W.; Lam, H. *Mass Spec. Rev.* **2017**, 36, 634.
7. Kind, T.; Fiehn, O. *Bioanal. Rev.* **2010**, 2, 23.
8. Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; Wolan, D. W.; Spilker, M. E.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2018**, 90, 3156.
9. West Coast Metabolomics Center and Genome Center University of California Davis. Mass Bank of North America; See http://mona.fiehnlab.ucdavis.edu/downloads.
10. The National Institute of Standards and Technology. NIST Mass Spectral Search Program, version 2.3; See https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17.
11. Thompson, S. K. *Sampling*, Wiley: New York, **1992**.
12. Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszarc, E.; Kofeler, H. C. *J. Mass Spectrom.* **2009**, 44, 494.
13. Creative Research Systems. Sample Size Calculator; See https://www.surveysystem.com/SSCALC.HTM.
14. EasyCalculation.com. Statistical Significance Calculator; See https://www.easycalculation.com/statistics/statistical-significance.php.
15. Milman, B. L.; Zhurkovich, I. K. *Anal. Chem. Res.* **2014**, 1, 8.

