Pattern similarity measures applied to mass spectra

Arun S. Moorthy and Anthony J. Kearsley

Abstract Mass spectrometry is a core analytical chemistry technique for elucidating the structure and identity of compounds. Broadly, the technique involves the ionization of an analyte and analysis of the resulting *mass spectrum*, a representation of ion intensity as a function of mass to charge ratios. In this article, the notion of *similarity* as it applies to mass spectra is explored. In particular, several modes of approximating distances and similarities in patterns are touched upon: ℓ_1 and ℓ_2 distances, the Wasserstein metric (earth mover's distance) and cosine similarity derived measures. Concluding the manuscript is a report on the performance of the similarity measures on a small test set of data, followed by a discussion of mass spectral library searching and prospects for quantifying uncertainty in compound identifications leveraging mass spectral similarity.

1 Background and Motivation

When discussing industrial mathematics, it is natural for one to think of the modeling, simulation, and optimization of industrial processes (operations research), a field in which some of the great mathematicians of the 20^{th} century made their mark. More recently, industrial mathematicians have been tasked with making sense of the abundance of data that exists on public and private servers (data science) which has spawned beautiful algorithms in statistical and machine learning. From the operations research specialist to the data scientist, all industrial mathematicians leverage mathematical thinking with application-specific knowledge to solve problems of industrial importance.

Arun S. Moorthy
National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD, 20899-8362 USA, e-mail: arun.moorthy@nist.gov

Anthony J. Kearsley
National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD, 20899-8910 USA, e-mail: anthony.kearsley@nist.gov

One task of significant industrial importance is the characterization of complex material composition. These materials may include natural products, foods and drugs, fuels, biological fluids, plastics, etc. As examples of the importance of material composition characterization, consider the implications of a fuel containing impurities causing inefficient performance, or of a food containing contaminants negatively affecting the health of consumers. It is of utmost importance that the composition of materials is accurately characterized and, to this end, one of the most commonly employed tools to approach this task is mass spectrometry.

The objective of this manuscript is to introduce the notion of estimating similarity between measurements obtained through mass spectrometry. In particular, the similarity of mass spectra, the resulting measurements from analysis using mass spectrometry, for pure compounds is explored. There are two primary reasons why estimating the similarity of mass spectra is of great importance. (1) Measures of spectral similarity are leveraged in mass spectral library searching, the process of sorting through curated libraries of mass spectra of known compounds to aid in the identification of an analyte from its mass spectrum. (2) Accurate measures of mass spectral similarity are necessary for quantifying the uncertainty of compound identification using mass spectrometry.

The manuscript is organized as follows. In Section 2, a brief overview of mass spectrometry and mass spectral library searching is provided, followed by details of several pattern similarity measures in Section 3. Concluding the manuscript is a report on the efficacy of each similarity measure for an illustrative test set of mass spectra, and a larger discussion about the implications of similarity measures to mass spectral library searching and uncertainty quantification in Section 4.

2 Mass Spectrometry and Mass Spectral Library Searching

Mass spectrometry has been a prominent tool in the analysis of matter for over one hundred years. Broadly, the technique involves the ionization of an analyte, through one of a variety of methods, followed by detecting the intensity of ions across a mass-to-charge (m/z) range. After processing, the output of a mass speectromety analysis is a mass spectrum. A comprehensive discussion of the technology is outside the experience of the authors and thus the scope of this manuscript, however, good introductory texts [22] and historical review articles [10] can be readily found in the literature. The discussion in this manuscript will focus exclusively on *unit-mass resolution* mass spectra of pure compounds (molecules) obtained through electron ionization (EI) mass spectrometry.[1]

A mass spectrum of the molecule caffeine is shown in Figure 1. The measure along the x-axis is mass-to-charge (m/z) and the y-axis indicates the relative abundance at each m/z. For reference, a standard 2-dimensional representation of the structure

[1] The specification of "unit-mass resolution" indicates that the mass-to-charge ratio of ions will always be positive integer values. This resolution of electron ionization mass spectra are commonly used in many industrial applications.

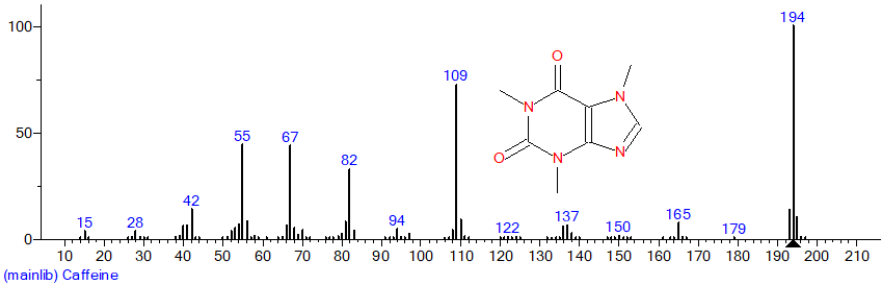

Fig. 1 A representative electron ionization mass spectrum of caffeine from the NIST17 Demo Library [2] with the structure of caffeine overlaid on the spectrum.

of caffeine is overlaid on the mass spectrum. Note that in EI mass spectrometry, the charge of an ion is most often 1, and so m/z is often interpreted and referred to as mass. Ideally, every peak in a mass spectrum can be explained exclusively as a *molecular ion* or a *fragment ion*.

A molecular ion, as its name implies, is an ion of the intact molecule being analyzed. In the example of caffeine with nominal molecular mass 194 Da, the molecular ion peak appears at m/z 194 (see Figure 1). The peaks occurring with m/z values greater than 194 are molecular ions where the molecule is constructed with heavier isotopes (e.g. Carbon-13 instead of Carbon-12). For some molecules, molecular ions will contain weak bonds that cause it to fragment prior to reaching the detector, resulting in the molecular ion being unobserved in the mass spectrum.

Under normal conditions, an ionized molecule will almost always fragment. The portion of the molecule that remains charged after fragmentation is referred to as a fragment ion and the portion(s) that are neutral charged are referred to as neutral losses. Fragment ions are recorded in mass spectra, neutral losses are not. Most ionized molecules can fragment in several ways. Accordingly, mass spectra will typically contain a number of fragment ion peaks; however, there are some cases where a very stable ion - either the molecular ion or a fragment ion - will lead to mass spectra of very few peaks (see Figure 2).

Since a mass spectrum summarizes the mass of a molecule and its fragments, it is possible for an analytical chemist to identify a molecule directly through interpretation of its mass spectrum. This is particularly true for simple molecules with limited mechanisms for fragmentation. For more complex molecules, identification through interpretation is impractical if not impossible. An important resource used by many analysts to aid in identifying molecules are *mass spectral libraries*. These carefully curated databases of mass spectra of known molecules can be sorted through by comparing to the mass spectrum of the analyte, a process referred to as mass spectral library searching, potentially finding a match to the analyte spectrum or providing information that supports further investigation. For interested readers, a seminal report on the topic of mass spectral library searching was provided by Stein and Scott [21].

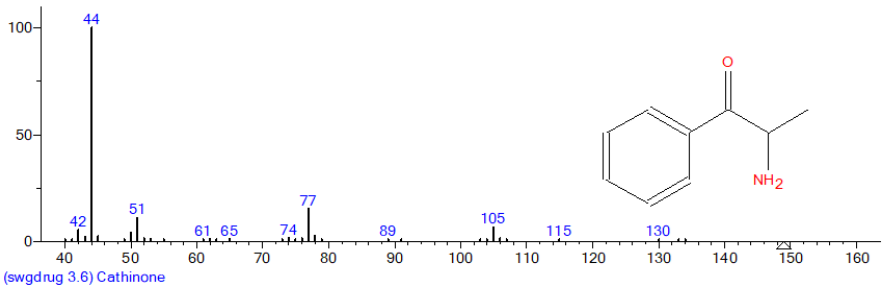

Fig. 2 A representative electron ionization mass spectrum of cathinone from the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) mass spectral library [3] with the structure of cathinone overlaid on the spectrum.

3 Pattern Similarity Measures

All unit-mass resolution EI mass spectra can be easily represented as vectors of equivalent length, where the index of each element in the vector is representative of the mass-to-charge of an ion and the value of each element being the corresponding relative abundance from the spectrum. It is between these vectors that a similarity estimate is computed. This section first describes similarity measures that are traditionally used in mass spectral library searching, followed by similarity measures used in general pattern recognition tasks and a final measure motivated by the study of histograms.

Using the standard dot-product formula we can define the cosine similarity between two non-zero vectors, $\mathbf{x}$ and $\mathbf{y}$ as ξ_1,

$$\xi_1 = \frac{\sum_{i=1}^{n} \mathbf{x}[i]\,\mathbf{y}[i]}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}[i])^2}\,\sqrt{\sum_{i=1}^{n} (\mathbf{y}[i])^2}}\,, \tag{1}$$

where n is the length of the vectors, and the $\mathbf{x}[i]$ notation indicates the ith element of the vector $\mathbf{x}$. A commonly employed variant of cosine similarity is the *simple match factor*. It differs from standard cosine similarity by three modifications: (i) the values of the elements of the input vectors are replaced with their square roots, (ii) the resulting cosine similarity measure is squared, and (iii) the resulting value is scaled by a constant. This sequence of modifications result in a simple match factor, ξ_2,

$$\xi_2 = C\frac{\left(\sum_{i=1}^{n} (\mathbf{x}[i])^{1/2}\,(\mathbf{y}[i])^{1/2}\right)^2}{\sum_{i=1}^{n} \mathbf{x}[i]\;\sum_{i=1}^{n} \mathbf{y}[i]}\,, \tag{2}$$

where C is, for historical reasons, 999. In mass spectral library search programs that use simple similarity, the value is further rounded to the nearest integer, also for

historical reasons. A further modified measure of similarity commonly employed in mass spectral library searching is the *identity match factor*. It was first introduced in [21] and is referred to as the "composite score". This measure differs from the simple similarity match factor in that it is modified by a ratio based on relative abundances at adjacent m/z values. An alternate identity match factor is computed in this manuscript. A vector, $\mathbf{r}$ is computed with each element defined

$$\mathbf{r}[i] = \begin{cases} \gamma[i] & \text{if } \mathbf{x}[i]\mathbf{x}[i-1]\mathbf{y}[i]\mathbf{y}[i-1] > 0, \\ 0 & \text{if } \mathbf{x}[i]\mathbf{x}[i-1]\mathbf{y}[i]\mathbf{y}[i-1] = 0, \end{cases}$$

where

$$\gamma[i] = \frac{\mathbf{x}[i]}{\mathbf{x}[i-1]} \frac{\mathbf{y}[i-1]}{\mathbf{y}[i]}.$$

The set of indices marking non-zero values of $\mathbf{r}$ is denoted α. A modification term, F, is computed

$$F = \frac{\sum_i^{m_1} \alpha[i] \cdot \min\left(\mathbf{r}[\alpha[i]], 1/\mathbf{r}[\alpha[i]]\right)}{\sum_i^{m_1} \alpha[i]}$$

where m_1 is the number of elements in the set α. The modified identity match factor, ξ_3, is then computed as

$$\xi_3 = C \frac{m_1 F + m_2 \frac{\xi_2}{C}}{m_1 + m_2}, \tag{3}$$

where m_2 is the number of indicies where elements of both $\mathbf{x}$ and $\mathbf{y}$ have non-zero values, and C is 999 as in (2).

For general pattern recognition tasks, the set of ℓ_p distances are often employed. For consistency and completeness, we present the ℓ_1 distance, ξ_4, and ℓ_2 distance, ξ_5, between non-zero vectors,

$$\xi_4 = \sum_{i=1}^{n} | \mathbf{x}[i] - \mathbf{y}[i] | , \tag{4}$$

$$\xi_5 = \left(\sum_{i=1}^{n} (\mathbf{x}[i] - \mathbf{y}[i])^2 \right)^{1/2} . \tag{5}$$

An intriguing measure of similarity comes from viewing the mass spectra as discrete probability distributions with finite support on a metric space. The Wasserstien metric, a commonly employed method in computer vision and often called the earth mover's distance [19], can be viewed as a distance that represents the minimum cost

associated with transforming a reference mass spectrum into a second to which it is being compared. We denote this distance as ξ_6,

$$\xi_6 = \mathrm{EMD}\,(\mathbf{x}, \mathbf{y}) \ , \tag{6}$$

and note that it is a far more complicated computation than ξ_i, where $1 \leq i \leq 5$. The numerical results presented in this paper result from (6) being evaluated using a transportation simplex algorithm with a ground distance computed using an ℓ_1 metric. It is worth noting that the larger cost associated with evaluating (6) has resulted in significant research on reducing the computational costs including approximations to the metric [20] and the development of parallel programming algorithms [15].

4 Results and Discussion

To demonstrate the performance of the similarity measures outlined in Section 3, pairs of replicate spectra[2], and non-replicate spectra[3] were chosen at random from two highly regarded commercial libraries. The distribution of similarity measures generated with each method are shown in Figure 3 as box and whisker plots, generated using the default *boxplot* function as implemented in base-R [17]. Each box and whisker object describes the distribution of similarity measures computed on a set of mass spectra. The set of non-replicate spectra is labeled on the x-axis of each plot as "other", and the set of replicate spectra is labeled "replicate". For each distribution, the outlined box is the computed Inter Quartile Range (IQR) with the 2nd quartile (median) marked as a darkened line within the box, and the bottom and top edges of the box indicating the 1st and 3rd quartile measurements, respectively. The upper whisker indicates either the maximum measured value in the distribution or the maximum similarity measure within 1.5 IQR of the 3rd quartile value. Similarly, the lower whisker is minimum similarity measure or the minimum measure within 1.5 IQR of the first quartile value. Outlier scores greater than 1.5 IQR of the 1st or 3rd quartile are shown as open circles.

The three cosine similarity derived measures traditionally employed in mass spectral library searching are summarized in panels a-c of Figure 3. It is clear that the distribution of these measures differ notably between the sets of replicate and non-replicate spectra. That is to say, the cosine similarity derived measures are performing as desired. There is, however, still overlap between computed measures between the two sets. The maximum similarity measure computed between a pair of non-replicate spectra is greater than the minimum similarity measure computed between a pair of replicate spectra. We do see that the modifications from cosine

[2] The term "replicate spectra" is used here to indicate spectra of one compound sourced from two different commercial libraries, differing from the usual convention of a repeated measurement by a single individual/source.

[3] The term "non-replicate spectra" is used here to indicate a pair of spectra from two different compounds, each spectrum from a different library

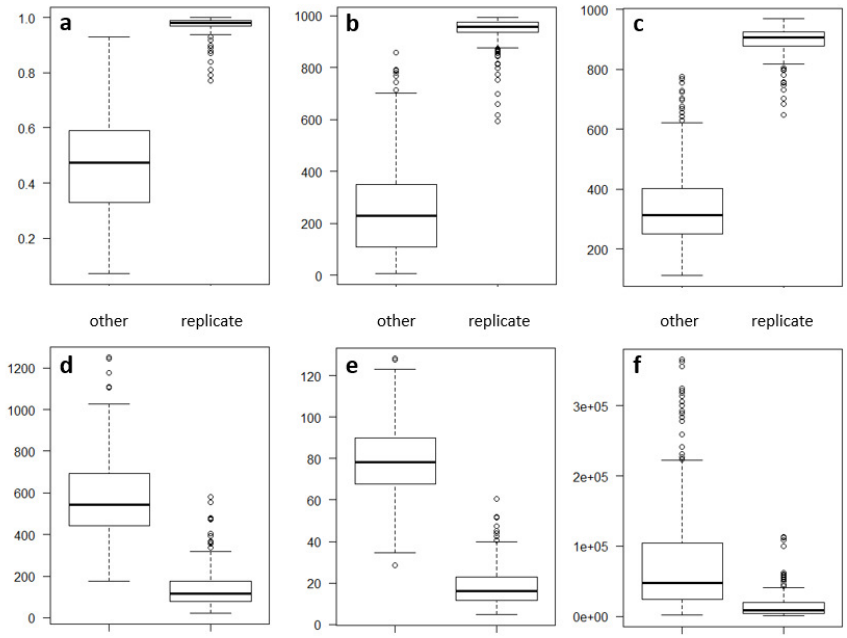

Fig. 3 Distribution of similarity and distance measures when considering spectra of the same molecules (replicate) and of different molecules (other). Note that the scale of the ordinate axis varies for each similarity measure. (a) Similarity as approximated by a cosine similarity. (b) Similarity as approximated by a simple match factor. (c) Similarity as approximated by an identity match factor. (d) Distance as approximated by an ℓ_1 norm. (e) Distance as approximated by the ℓ_2 norm. (f) Distance as approximated by the earth mover's distance.

similarity (Figure 3a) to simple match factor (Figure 3b), and then identity match factor (Figure 3c) do improve the separation between similarity scores computed on replicate and non-replicate spectra, with only outlier measures overlapping using identity match factors.

The results of measuring similarity by the ℓ_1 and ℓ_2 distance are shown in Figures 3d-e. In general, both distance measures do perform as desired, with replicate spectra having smaller computed distances than non-replicate spectra. The separation, however, is not as pronounced as was the case in the cosine similarity derived measures. Using earth mover's distance to measure similarity (Figure 3f), replicate spectra have smaller computed distances than non-replicate spectra across the total distribution of similarity measures. However, a substantial number of non-replicate spectra are similar according to this metric.

In general, the superior efficacy of the identity match factor as compared to the distance measures is due to the nature of variability in mass spectra and the modifications built-in to the identity match factor. Consider the replicate spectra of *cocaine*, taken from two publicly available mass spectral libraries, presented in

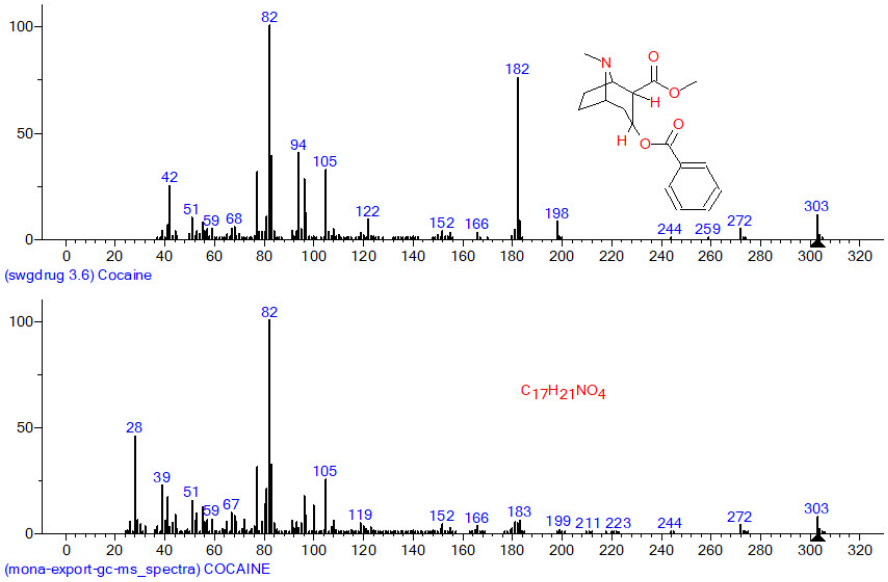

Fig. 4 Representative electron ionization mass spectra of cocaine from (top) the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) mass spectral library [3] with structure overlaid and (bottom) the Mass Bank of North America [1] with chemical formula overlaid.

Figure 4. Both spectra are valid measurements of cocaine, yet vary substantially in the relative abundance of certain fragment ions (see m/z 182). The processes of taking the square root of relative abundances as in simple match factors (2) and accounting for the ratio of relative abundances for adjacent peaks are able to mitigate some of the natural variability observed in mass spectra. The use of similar modifications to improve the separation of similarity measures between replicate and non-replicate spectra using the ℓ_p and earth mover's distances is on-going work.

It is also worth discussing the limitations of the numerical assessments described in this manuscript. The random selection of non-replicate mass spectra for assessment limits the interpretation of the results. Some of the non-replicate mass spectra may have come from vastly different molecules, and others may have come from molecules of very similar nature, such as positional isomers or chemical analogs. Creating subsets of mass spectra for testing, based on either properties of chemicals (e.g. molecule size, molecule type, etc.) or the mass spectra themselves (e.g. small numbers of peaks, large numbers of peaks, etc.), may illuminate insights about what similarity means in different situations. For example, we may find that similarity of mass spectra with few peaks is better captured by one similarity measure than the others.

Implications for Mass Spectral Library Searching

A mass spectral library search algorithm sifts through a reference library of mass spectra and produces a list, commonly referred to as a "hits list" or "hitlist", of mass spectra that are presumably similar to a query spectrum. From this hitlist, a chemist will either propose an identity for the analyte or will conclude that further investigation is needed. A thorough numerical evaluation of how each described dissimilarity measure would affect mass spectral library searching requires defining several specific and involved identification tasks. This and the subsequent selection of search parameters is outside the scope of this manuscript. General evaluations of similarity measures in library searches where the objective is to return a hitlist where the top entry is the correct identification of the analyte can be found in the literature [21, 12, 13, 14].

In general, for the purpose of sifting through the library, any of the similarity measures outlined in Section 3 can be used and one of the top few library mass spectra appraised to be most similar to the queried spectrum will be measurements of the same compound. One concern with using any similarity measure, especially the Earth Mover's distance, is that some of the reference spectra that appear in the hitlist may not be measurements of the same compound producing the query spectrum, thus challenging the identification of the analyte. However, in this case it is possible that a hitlist containing non-replicate spectra can also aid the identification process. In 2017, a measure of spectral similarity referred to as "Hybrid Similarity" was introduced [7, 16, 9], as a means of searching through libraries when it was suspected that a reference spectrum of the analyte was not contained in the library (e.g. a novel designer drug with no representative mass spectrum contained in the library). This measure is able to capture the similarity between mass spectra of molecules that differ by a single modification, that does not significantly alter the fragmentation mechanism of common fragment ions, resulting in spectra differing by predictable shifts in ion m/z values - these types of molecules are now referred to as *cognates*. A hitlist containing several cognates of the query can provide a chemist adequate information to propose additional investigations or, potentially, even propose a possible identity of the compound generating the query spectrum. There have been several recent publications describing applications of hybrid similarity [18, 6, 5, 4, 11, 8].

The popularity of hybrid similarity in mass spectral library searching is evidence supporting the continued exploration of novel similarity measures for mass spectra. Though it is not immediately clear whether the similarity measures outlined in this manuscript will be beneficial, it is possible that one or more of these measures will be useful in identifying particular types of molecules or in very specific situations (e.g. identifying a molecule based on a spectrum with poor signal-to-noise as is often the case when samples are collected in less-than-ideal circumstances). Further investigation of these similarity measures with specific test cases is an on-going course of work.

Towards Uncertainty Quantification

The last section of this manuscript is a discussion of leveraging similarity measures for quantifying uncertainty with compound identifications using mass spectrometry. As was noted in Section 1, the accurate identification of pure compounds, that may be contained within complex materials, is of incredible industrial importance. At present, using only the numerical value of any mass spectral similarity measure to identify a molecule from its spectrum should be avoided - there is adequate overlap in similarity measures that incorrect identifications are possible. It is worth noting that using mass spectrometry alone for identifying compounds is generally not recommended. Rather, using a series of complimentary measurements is prudent, such as gas chromatography retention times [23] to simultaneously identify compounds.

As noted previously, conducting strict systematic investigations with subsets of mass spectra may provide useful insights about the types of mass spectra that generate high similarity measures. For example, if we find that good similarity scores using the ℓ_2 distance only occur between molecules that are isomers, this suggests the uncertainty of an identification using only an ℓ_2 measure of similarity will be a function of the number of possible isomers of the proposed molecular identity. Another approach that may support uncertainty quantification is the simultaneous analysis of multiple similarity measures. If two spectra are completely identical, then all computed similarity measures should return their optimal value. If two mass spectra are deemed similar using the ℓ_2 distance, but the two spectra appear significantly dissimilar when measured by cosine similarity, this casts significant doubt upon the identification of the compound that generated the mass spectra. Developing composite measures or schemes for evaluating mass spectra may be fruitful endeavors that greatly reduce uncertainty and enhance the efficacy of mass spectrometry based compound identification procedures.

Acknowledgements The first author would like to thank Prof. Peregrina Quintella Estevez (Universidade de Santiago de Compostella) for coordinating Industry Day at the International Congress of Industrial and Applied Mathematics 2019 meeting. The meeting has generated meaningful relationships and discussion that will greatly benefit future work in this field. The authors also acknowledge Christopher Schanzle (National Institute of Standards and Technology) for an implementation of the earth mover's distance, and Dr. Gary Mallard (National Institute of Standards and Technology) for his guidance in preparing this manuscript.

References

1. Mass Bank of North America (MoNA). https://mona.fiehnlab.ucdavis.edu/. Accessed: 2019-11-26
2. NIST 2017 Mass Spectral Library (demo). https://chemdata.nist.gov. Accessed: 2019-11-26
3. Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) Mass Spectral Library v.3.6. https://swgdrug.org. Accessed: 2019-11-26
4. Barupal, D.K., Fan, S., Fiehn, O.: Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. Current opinion in biotechnology **54**, 1–9 (2018)

5. Blaženović, I., Kind, T., Ji, J., Fiehn, O.: Software tools and approaches for compound identification of lc-ms/ms data in metabolomics. Metabolites **8**(2), 31 (2018)
6. Blaženović, I., Oh, Y.T., Li, F., Ji, J., Nguyen, A.K., Wancewicz, B., Bender, J.M., Fiehn, O., Youn, J.H.: Effects of gut bacteria depletion and high-na+ and low-k+ intake on circulating levels of biogenic amines. Molecular nutrition & food research **63**(4), 1801184 (2019)
7. Burke, M.C., Mirokhin, Y.A., Tchekhovskoi, D.V., Markey, S.P., Heidbrink Thompson, J., Larkin, C., Stein, S.E.: The hybrid search: A mass spectral library search method for discovery of modifications in proteomics. Journal of proteome research **16**(5), 1924–1935 (2017)
8. Burke, M.C., Zhang, Z., Mirokhin, Y.A., Tchekovskoi, D.V., Liang, Y., Stein, S.E.: False discovery rate estimation for hybrid mass spectral library search identifications in bottom-up proteomics. Journal of proteome research **18**(9), 3223–3234 (2019)
9. Cooper, B., Yan, X., Simón-Manso, Y., Tchekhovskoi, D., Mirokhin, Y., Stein, S.: Hybrid search: A method for identifying metabolites absent from tandem mass spectrometry libraries. Analytical chemistry **91**(21), 13924–13932 (2019)
10. Griffiths, J.: A brief history of mass spectrometry. Anal Chem **80**(15), 5678–5683 (2008)
11. Jang, I., Lee, J.u., Lee, J.m., Kim, B.H., Moon, B., Hong, J., Oh, H.B.: Lc–ms/ms software for screening unknown erectile dysfunction drugs and analogs: artificial neural network classification, peak-count scoring, simple similarity search, and hybrid similarity search algorithms. Analytical Chemistry (2019)
12. Kim, S., Koo, I., Wei, X., Zhang, X.: A method of finding optimal weight factors for compound identification in gas chromatography–mass spectrometry. Bioinformatics **28**(8), 1158–1163 (2012)
13. Kim, S., Zhang, X.: Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. Computational and mathematical methods in medicine **2013** (2013)
14. Koo, I., Kim, S., Zhang, X.: Comparative analysis of mass spectral matching-based compound identification in gas chromatography–mass spectrometry. Journal of chromatography A **1298**, 132–138 (2013)
15. Li, W., Ryu, E.K., Osher, S., Yin, W., Gangbo, W.: A parallel method for earth mover's distance. Journal of Scientific Computing **75**(1), 182–197 (2018)
16. Moorthy, A.S., Wallace, W.E., Kearsley, A.J., Tchekhovskoi, D.V., Stein, S.E.: Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification. Analytical chemistry **89**(24), 13261–13268 (2017)
17. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). URL https://www.R-project.org/
18. Remoroza, C.A., Mak, T.D., De Leoz, M.L.A., Mirokhin, Y.A., Stein, S.E.: Creating a mass spectral reference library for oligosaccharides in human milk. Analytical chemistry **90**(15), 8977–8988 (2018)
19. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
20. Shirdhonkar, S., Jacobs, D.W.: Approximate earth mover's distance in linear time. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
21. Stein, S.E., Scott, D.R.: Optimization and testing of mass spectral library search algorithms for compound identification. Journal of the American Society for Mass Spectrometry **5**(9), 859–866 (1994)
22. Watson, J.T., Sparkman, O.D.: Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation. John Wiley & Sons (2007)
23. Wei, X., Koo, I., Kim, S., Zhang, X.: Compound identification in gc-ms by simultaneously evaluating the mass spectrum and retention index. Analyst **139**(10), 2507–2514 (2014)