Identifying fentanyl with mass spectral libraries

Anthony J. Kearsley* and Arun S. Moorthy

National Institute of Standards and Technology

Gaithersburg, MD, USA

*anthony.kearsley@nist.gov

March 6, 2021

Abstract

Synthesis, distribution and abuse of fentanyl, a synthetic opioid, has led to a critical worldwide epidemic. Mass spectral library searching for opioids remains unresolved despite being central to law-enforcement involving identification, monitoring and prosecution of opioid related crimes. In this article, two model problems are presented to illustrate difficulties associated with fentanyl identification. A collection of both currently-employed similarity measures and intuitive measures of dissimilarity are employed to simulate identifying fentanyl analogs with mass spectral library searching.

Keywords: Fentanyl; Mass Spectral Library Searching; Pattern Similarity

1 Introduction

Fentanyl is a synthetic opioid originally developed as a prescription medication for treating severe pain [6, 5]. However, over the past several years, there has been explosive growth in the abuse of this extremely potent drug, leading to an unprecedented rise in the number of overdose deaths across the United States and around the world [13]. One key cause of this fentanyl epidemic is the illegal production of *designer fentanyl analogs*, drugs with distinct but similar chemical structure to that of fentanyl. These illegal compounds are often produced in clandestine laboratories and distributed through illegal channels at a rate that far exceeds law enforcement's ability to combat them. Particularly daunting tasks include the identification and scheduling of this complicated collection of compounds. Regulation of these types of compounds by federal U.S. drug policy as described by the Controlled Substances Act (CSA) seek to control the manufacture, importation, possession, use and distribution of substances like fentanyl. The U.S. Drug Enforcement Administration (DEA), for example has the task of scheduling, that is categorizing, drugs depending upon the drug's acceptable medical use and the drug's abuse

or dependency potential. Fentanyl is currently categorized as a Schedule II drug (together with more commonly known drugs like Adderall, Ritalin and oxycodone (OxyContin)) a substance or chemical agreed to have the high potential for abuse, with the possibility of leading to severe psychological or physical dependence. Furthermore, once distributed, fentanyl and its designer analogs are often used in very low quantities to increase potency and reduce the street cost of other drugs, like heroin. This significantly complicates the tasks of collection, identification and scheduling.

Electron Ionization Mass Spectrometry (EI-MS) is one of the primary tools used to identify drugs, in law-enforcement applications as well as emergency medical evaluation facilities where proper identification is often crucial. An analyte, a substance or molecule being identified or measured, is ionized through interactions with energetic electrons creating a *molecular ion*. Under normal conditions, this molecular ion will almost always fragment, resulting in *fragment ions*. These ions are separated according to mass-to-charge ratios, (m/z), and then observed electronically by a detector. The resulting detected signal, proportional to the number of ions, is plotted against corresponding (m/z) values in order to produce a *mass spectrum* (see [18]). Roughly speaking, the mass spectrum is the plot of intensity or abundance, proportional to the number of ions arriving in a given time, as a function of (m/z) values. In EI-MS, the charge of an ion is most often 1, and for this reason, (m/z) is often interpreted and interchangeably referred to as mass with units of *Daltons* (Da). Additionally, EI mass spectra are often *unit mass resolution*, hence the horizontal axis, mass values, are positive integers. A peak in the mass spectrum has height which corresponds to the positive abundance of ions detected at specific mass values. The height and location of these peaks that are, in turn, used to infer chemical composition and structural information. The mass spectrum does not directly yield information about the chemical structure of the analyte. While there are advanced alternative experiments (e.g. Tandem MS [7]) which can illuminate more information about the chemical structure associated with an analyte, these methods are less commonly employed in forensics applications and will not be discussed here.

Given an unidentified compound, its mass spectrum can be measured and compared to the mass spectra of reference compounds with the objective of chemical identification. Reference compounds are well characterized, often purchased, molecules with associated mass spectra obtained under controlled conditions. The mass spectra of these reference compounds can be assembled into searchable databases or libraries ([15]). In practice, the spectrum of the unidentified compound is compared with the mass spectra in a well-chosen library with the goal of identifying compounds with similar mass spectra. If successful this process results in the eventual identification of the compound from which the mass spectrum was originally measured. This process is referred to as *mass spectral library searching*. Typically, the mass spectral library search returns a well-ordered list of database entries determined to have mass spectra similar to that of the unidentified mass spectrum, commonly referred to as a *hit list*. The compounds associated with the hit list form a collection of candidate compound matches for the unidentified compound. Ideally, a hit list will provide an analyst adequate information to correctly infer

the identity of an unidentfied compound. It is important to note that the eventual classification of the analyte is still a human task - the burden of identification lies with the analyst.

Underpinning the success of a mass spectral library search is the selection of a similarity/dissimilarity or distance measure to compare mass spectra. A recent numerical study on the effectiveness of various measures of dissimilarity applied to replicate mass spectra, i.e. two different mass spectral measurements of the same compound, illustrated that all measures of dissimilarity between replicate mass spectra performed desirably (i.e. low dissimilarity, high similarity between replicate spectra) [10] . However, it was unclear how these measures would perform under the significantly more complex task of library searching, especially when the potential library contains very similar spectra coming from different compounds (or vice versa) as can be the case with fentanyl and designer fentanyl analogs.

This manuscript explores the effectiveness of various measures of dissimilarity in the context of compound identification using mass spectral library searching. In particular, we are interested in accurately identifying fentanyl, a pressing need. Like Gaul, the body of this manuscript is comprised of three parts. A brief description of the mechanics of library searching is presented in the next section. To compare measures of dissimilarity in the context of library searching, two model problems are constructed that illuminate some benefits and drawbacks associated with each of the dissimilarity measures. Finally, the results of our model problem are summarized in the last section together with some observations.

2   Library Searching

We consider a collection of mass spectra which, for simplicity, are all considered to be of equal size, say $n > 0$. Further, since we are only concerned with unit mass resolution spectra, the mass spectra can be denoted as vectors $x \in \Re^n$ with positive entries representing the intensity of measured signal with the indexed mass. All other entries are identically zero. These x are carefully measured in a controlled laboratory setting and bundled together into a library which is then distributed. In 1997 the (DEA) and the Office of National Drug Control Policy (ONDCP) co-sponsored the formation of an international group of forensic scientists and United Nations representatives known, since 1999, as the Scientific Working Group for the Analysis of seized Drugs (SWGDRUG). The (SWGDRUG) have collected and released mass spectra of interest, [3] to the forensics community, and a subset of spectra of fentanyl and related compounds was assembled into a mass spectral library which we refer to as L. This library is discussed in [11] and is publicly available facilitating comparison of different methods.

To imitate a situation encountered by law enforcement, we consider a mass spectrum measured from a presumed unidentified compound, often called a *query* spectrum, say, x_q. This spectrum is *not* contained in L, $x_q \notin L$ and is often attained under less than ideal circumstances. The enormous variation in instruments used to measure spectra, in sample
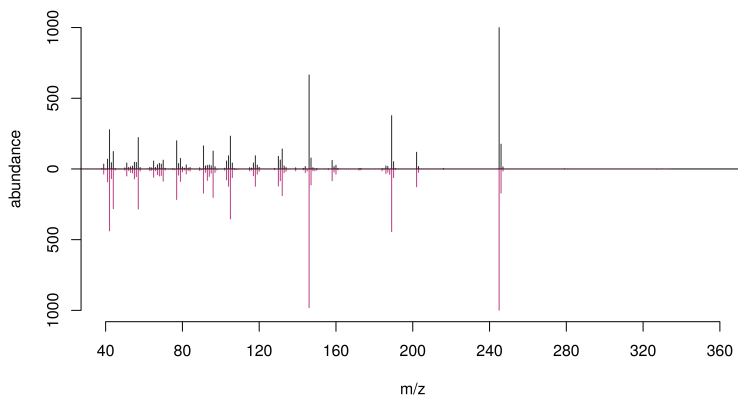
Figure 1: A *head-to-tail* plot comparing two mass spectra of fentanyl acquired from two different mass spectral libraries. In a head-to-tail plot, the first mass spectrum is plotted normally, and the reflection of a compared mass spectrum is plotted directly beneath the first such that the abundance of ions at equivalent mass-to-charge (m/z) values can be easily compared. This type of plot is commonly employed when comparing mass spectra.

preparation and in laboratory procedures, suggests that obtaining two identical spectra is a situation that, in practice, is virtually impossible for two measurements of the same compound, Fentanyl. (see Figure 1).

Given a library L and a query mass spectrum x_q, completing a library search requires a selection of a dissimilarity measure. Consider ϕ that seeks to measure the dissimilarity between x_q and $x \in L$. Analytical chemist's Mass spectral library search programs employ these dissimilarity measures to construct a hit list. The mass spectrum associated with the first element in the hit list, say x^1, is the most similar (or least dissimilar) mass spectrum to x_q among mass spectra in L, the second most similar is x^2 and x^j is the jth. To assemble the hit list one solves,

$$x^1 = arg \min_{x \in L} \phi(x_q, x) \tag{1}$$

and the second

$$x^2 = arg \min_{x \in L \setminus x^1} \phi(x_q, x), \tag{2}$$

and so on, where $arg \min$ denots the argument-minimum.

Typically, an analytical chemist a trained user of EI-MS and library searching will interrogate compounds associated with the first several elements of a hit list, say n_H, $H = (x^1, x^2 \dots x^{n_H})$ where n_H varies for different applications. Taking $n_H = 10$ is a common practice. It is also worth noting that equation (1) need not have unique solutions. Mass spectra associated with distinct compounds can generate identical dissimilarity scores when compared to a query mass spectrum.

4

2.1 Dissimilarity Measures

A comparison of several dissimilarity measures applied to the comparison of replicate mass spectra were recently explored [10]. This included similarity measures traditionally considered in the field of mass spectral library searching, distances commonly considered in general pattern recognition tasks, and a distance measure often employed for the study of distances between histograms. For the completeness of this manuscript, we suggest and summarize a collection of dissimilarity measures.

2.1.1 Vector Norms

The ℓ_p vector norms are straight-forward measures of dissimilarity in Euclidean space. While these are not typically employed in practice, they are included here for illustrative purposes. For our model problem the ℓ_1 and ℓ_2 measures can be written

$$\phi_{\ell_1}(x_q, x) = \sum_i^n |(x_q[i] - x[i])| \tag{3}$$

and

$$\phi_{\ell_2}(x_q, x) = \left(\sum_i^n (x_q[i] - x[i])^2 \right)^{1/2} \tag{4}$$

where the the notation $[i]$ denotes the ith component of a vector.

2.1.2 Cosine similarity and derived vales

Mass spectra are quite sparse which suggests a reasonable measure of similarity is the *Cosine* or Orchini similarity function which measures the *cosine* of the angle between two non-zero vectors, in this case mass spectra.

$$\phi_C(x_q, x) = \frac{\sum_i^n (x_q[i] x[i])}{\sqrt{\sum_i^n x_q[i]^2} \sqrt{\sum_i^n x[i]^2}}. \tag{5}$$

This computationally inexpensive measure is often used in large-scale data-mining applications that involve sparse data. Sometimes referred to as a distance, it does not satisfy the triangle inequality or Cauchy-Schwarz inequality.

A variant of cosine similarity commonly employed in mass spectrometry is what has come to be called the *simple match factor*. This measure differs from cosine similarity by a series of three modifications: (i) elements of the mass spectrum vectors are replaced with the square roots of their values before computing the cosine similarity, (ii) the cosine function (5) is computed with the altered spectra inputs and the result is squared, and finally, (iii) the result is scaled by a constant. This sequence of modifications result in the often-used simple match factor [12], denoted here by ϕ_S,

$$\phi_S(x_q, x) = C_S \frac{\left(\sum_{i=1}^n x_q[i]^{1/2} \, x[i]^{1/2} \right)^2}{\sum_{i=1}^n x_q[i] \, \sum_{i=1}^n x[i]}, \tag{6}$$

where C_S is an arbitrary constant which, for historical reasons, is often taken to be 999. In commonly employed mass spectral library search algorithms and software that employ a simple match factor similarity, the value of ϕ_S is rounded to the closest integer. This practice has origins to when computer memory was more expensive but is still handled the same way today. A second and often used modification of the measure of similarity commonly employed in mass spectral library searching is the so called *identity match factor*. The idea was first introduced in [16] where it was referred to as the "composite score". This measure differs from the simple match factor by several modifications. A simplified variant of this commonly employed identity match factor is described and explored in this manuscript.

Consider a vector, $r \in \Re^n$ computed by examining the product of heights at adjacent (m/z) values. For all $i = 2 \ldots n$, $r[i]$ is defined

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

where the vector $\gamma \in \Re^n$ is defined

$$\gamma[i] = \frac{x_q[i]}{x_q[i-1]} \frac{x[i-1]}{x[i]},$$

and we have $r[1] = \gamma[1] = 0$. If r has m_α non-zero entries, we denote the set of indices associated with those non-zero components of r as $\alpha \in \Re^{m_\alpha}$. With this available, a scalar modification term F is computed

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

The identity match factor is then computed

$$\phi_I(x_q, x) = C_I \frac{m_\alpha F + m_x \frac{\phi_S(x_q, x)}{C_S}}{m_\alpha + m_x}, \tag{7}$$

where m_x is the number of indices where elements of both x_q and x are positive at the same (m/z) value, and C_I is taken to be 999 for historical reasons as in (6). Despite the fact that both constants are taken here as 999, we distinguish them formally.

These three measures, ϕ_C, ϕ_S and ϕ_I, are commonly employed by the mass spectrometry community. They are inexpensive to compute and easily implemented but they are not true distances between specra so therefor are not metrics. It is worth observing that ϕ_I is not symetric, that is, it is not always the case that $\phi_I(x_q, x)$ produces the same value as $\phi_I(x, x_q)$. Both ϕ_C and ϕ_S are symetric but do not satisfy the triangle inequalty.

2.1.3 Wasserstein

The first Wasserstein distance [17] for discrete probability measures with finite support on $\Re^d$ are a natural candidate for this application. Some-

times called the earth movers distance or Mallows distance [8], the Wasserstein distance can be defined as the minimal cost of transforming x_q into x by shifting intensities. Let $\sigma(i,j)$ denote the amount of intensity located at $x_q[i]$ to be transported to location $x[j]$ in order to transform x into x_q. The most efficient means or optimal strategy for transforming x into x_q can be defined by,

$$\phi_W(x_q, x) = \min_{\sigma} \sum_i^n \sum_j^n \sigma(i,j) ||x_q - x|| \tag{8}$$

where $|| \cdot ||$ is a well-chosen norm often called the *ground distance*. For computational reasons, the ground distance is often selected to be ℓ_1. The value of $\phi_W(x_q, x)$ measures the Wassserstien distance between x_q and x and the optimal $\sigma \in \Re^{n \times n}$ gives an optimal flow of intensities from x_q and x.

We refrain from a thorough discussion of the Wasserstein distance but mention that the solution of (8) is an active research subject. For the purposes of our numerical comparison we normalized mass spectra before computing (8) despite the fact that there are mathematically elegant analogs to (8) which allow for vectors with different mean values to be compared. These changes may cause the resulting measure to fail to be a distance, in a mathematical sense. For the numerical comparisons presented here, it is worth noting that normalizing had little effect on the results. To solve (8) we implemented a transportation simplex solution approach based on [14] with an l_1 ground distance. The method ran quickly however, with an eye towards forensic applications where the number and size of comparisons could potentially grow, there is recent research on parallel algorithms for solving (8) which appear to be promising [9].

3 Model Problems

In the spirit of reproducing the situation faced by law-enforcement, one choice for a query spectrum is the mass spectrum of fentanyl taken from a different library and thus different from the fentanyl mass spectrum in L. Let $\bar{x}_F$ denote the mass spectrum of fentanyl taken from library [2] and x_F be the mass spectrum of fentanyl in library L. In this case $\bar{x}_F \neq x_F$.

3.1 Measurement Variation

A major obstacle in the identification of illegal substances by mass spectrometry is variations in measurement. Frequently in forensic applications, the abundance of seized unidentified compound is limited and it is virtually impossible that a library reference spectra and the query spectrum were acquired on the same instrument under identical circumstances. The sources of disparity in the assembly of mass spectra include things like variations in instrument resolution and sample preparations. Additional variations may be introduced by different instrument operators and environmental effects that arise from mass spectra being collected in different laboratories. A very coarse approximation to a sum of these uncertainties is to consider a synthetic query spectrum constructed from $\bar{x}_F$,

$$x_q[i] = \bar{x}_F[i] + \eta \text{ for } i = 1 \ldots n \tag{9}$$

where η is intended to simulate additive noise in the observed signal and is taken to be

$$\eta = \tau \bar{x}_F[i] \mathcal{N}(0,1) \text{ for } i = 1 \ldots n$$

where amount of noise in signal is controlled by $\tau > 0$, which is varied in our numerical experiments and $\mathcal{N}(0,1)$ is a standard normal distribution.

3.2 Contamination and masking

To disguise or avoid detection, fentanyl is often mixed with other compounds. To simulate this evasion technique, consider a query spectrum cobbled together as a *convex combination* of fentanyl and a second, perhaps legal, compound,

$$x_q(\lambda) = (1 - \lambda)\bar{x}_F + \lambda x_M, \tag{10}$$

where $0 \leq \lambda \leq 1$ and x_M is the representative mass spectrum for a legal masking compound not contained in our library of spectra, $x_M \notin L$. In our numerical experiments, masking agents are selected intentionally to be compounds not represented in the library.

We selected quinine $(C_{20}H_{24}N_2O_2)$ as a first simulated masking compound. It is a chemical used to treat malaria and a compound that is often employed in mixtures with fentanyl to avoid detection. Quinine is bitter, as is heroin, and is thus used to imitate a higher quality product to drug abusers. The second simulated masking compound we selected was serotonin $(C_{10}H_{12}N_2O)$, a biochemical produced in the human body and is found in foods like chocolate. Serotonin is often attributed to feelings of euphoria when consumed though it is not a compound associated with fentanyl.

4 Numerical Results

When the query spectrum $x_q = \bar{x}_F$ is searched against library L, the rank of the reference fentanyl in hit lists generated with each dissimilarity measures is as follows:

Dissimilarity Measure	Library Fentanyl Rank
ϕ_C	3
ϕ_S	1
ϕ_I	1
ϕ_{ℓ_1}	5
ϕ_{ℓ_2}	7
ϕ_W	36

It is unsurprising that the dissimilarity measures traditionally employed in mass spectral library searching, ϕ_S and ϕ_I, produce hit lists with fentanyl ideally ranked as the first element in the hit list. These less traditional similarity measures were designed specifically for the purpose of

mass spectral library searching and include mathematically unintuitive but chemically advantageous modifications.

The performance of library searches conducted using ϕ_C, ϕ_{ℓ_1} and ϕ_{ℓ_2}, while not ideal, still placed the reference fentanyl within the top-10 entries of the hit list, meaning that a practitioner is still likely to review its spectrum before making an identification of the query or perhaps pursing additional tests. The library search conducted using ϕ_W to compare spectra resulted in the reference fentanyl being listed as the 36th entry of the hit list and thus would be unlikely reviewed by a practitioner. Fentanyl, had it been present, would likely have gone undetected.

Simulated measurement variance In Table 1, the numerical results of 10 library searches where the query spectra are constructed using $\bar{x}_F$ and (9) with several *reasonable* values τ are presented. The rank of the reference fentanyl in hit lists generated using ϕ_S and ϕ_I is almost uniformly 1 regardless of the value of τ. Similarly, the change in hit list positions of fentanyl using ϕ_C, ϕ_{ℓ_1}, and ϕ_{ℓ_2} is minimal, with standard deviations of the rank evaluating to less than 3 in every case. Searches using ϕ_W are, not surprisingly, very sensitive to τ.

Figures 2-3 demonstrate an extended assessment of simulated noise, where $0 \leq \tau \leq 1$ and 100 computations of similarity or distance are computed between the constructed query mass spectra and the reference fentanyl, i.e. $\phi_i(x_q, x_F)$ where $i = \{C, S, I, \ell_1, \ell_2, W\}$. The figures illustrate the distribution of dissimilarity measure for a given value of τ. 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. Outliers scores greater than 1.5 IQR of the 1st or 3rd quartile are shown as open circles. Note that $\tau > 0.3$ is bordering on impractical variation within a single peak on replicate measurements, however, it is worth considering for a model problem.

Simulated contamination The results of library searches of simulated masked query spectra were constructed as convex combinations of $\bar{x}_F$ and the spectrum of Quinine (a masking compound) using (10) are presented in Table 2. The results of a library search of a spectrum of just fentanyl ($\lambda = 0$), and the results of a library search using just Quinine ($\lambda = 1$) are included for reference. Walking away from fentanyl, from $\lambda = 0$ to $\lambda = 0.2$, we observe that the rank of fentanyl improves in the hit lists generated using ϕ_C, ϕ_{ℓ_1} and ϕ_W to measure spectral dissimilarity. We see this again when walking from $\lambda = 0.2$ to 0.4, as the rank of fentanyl improves in the hit list using both ϕ_{ℓ_2} and ϕ_W as measures of spectral dissimilarity. This non monotonic behavior was not expected prior to reviewing the results. Similar unexpected behavior can be seen

9

Table 1: Results of a library search when the queried spectrum is constructed using (9) to simulate signal variance in a query mass spectrum of Fentanyl.

	Gaussian Noise, τ				
	0.10	0.15	0.20	0.25	0.30
Mean rank[a] of fentanyl					
ϕ_C	2.80	3.30	2.70	2.70	4.60
ϕ_S	1.00	1.00	1.00	1.00	1.10
ϕ_I	1.00	1.00	1.00	1.00	1.00
ϕ_{ℓ_1}	4.10	3.80	3.60	2.70	4.70
ϕ_{ℓ_2}	5.60	5.70	5.00	3.90	5.80
ϕ_W	34.40	34.10	30.90	21.00	34.30
Standard Deviation rank[a] of fentanyl					
ϕ_C	1.55	2.06	1.49	2.06	2.55
ϕ_S	0.00	0.00	0.00	0.00	0.32
ϕ_I	0.00	0.00	0.00	0.00	0.00
ϕ_{ℓ_1}	2.60	2.57	2.41	2.36	2.79
ϕ_{ℓ_2}	2.27	2.31	2.83	2.92	2.53
ϕ_W	4.35	6.67	13.02	16.46	11.48

(a) Mean and standard deviation of rank computed over 10 library searches.

in the results of library searches where query spectra were constructed as convex combinations of $\bar{x}_F$ and the spectrum of Serotonin (Table 3).

5   Discussion

We begin this discussion by highlighting some of the limitations of the outlined model problems. Measurement variance was simulated by constructing a synthetic query mass spectrum using (9). In doing so, we implicitly assumed that the abundance of each peak in the mass spectrum is independent of the others. Such a model was selected for its ease of implementation and, subsequently, interpretation. However, there is likely to be some correlation between observed measurements within a mass spectrum. Recall that the peaks in an EI mass spectrum can be interpreted as the mass of a fragment ion from the molecule. Molecules and, subsequently, their fragments are composed of atoms which naturally occur at several isotopic masses. For example, one can find stable atoms of carbon in nature with molecular mass 12 Da or 13 Da. Though "carbon-12" is far more common, a molecule of fragment constructed of several carbons is likely to have at least a few "carbon-13" components. Accordingly, the heights of fragment ion peaks that only differ by isotopic mass are likely to be correlated. That is the abundances, in this case, are likely to be correlated. Additionally, several fragment ions recorded in a mass spectrum may originate from the same *fragmentation pathway.* Roughly speaking, a molecular ion may fragment into several fragment
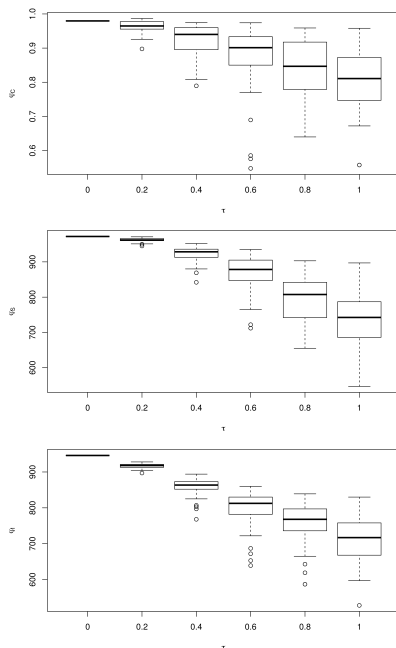
Figure 2: Box and whisker plots summarizing the distribution of similarity measures ϕ_i where $i = \{C, S, I\}$ computed between a query mass spectrum, x_q constructed using a mass spectrum of fentanyl and (9) across $0 \leq \tau \leq 1$, and the mass spectrum of fentanyl contained in the reference library.

ions that may further fragment into several more fragment ions. The pathway connecting the measured fragment ion to the molecular ion is a fragmentation pathway. The abundance of fragment ions that originate from the same pathway may be correlated as well.

Contamination was simulated by constructing a synthetic query spectrum using (10). In doing so, we implicitly assumed that the two molecules in a mixture would ionize with equal probability, thus the fragmentation pathways observed from the ionization of pure molecules will be conserved in the mixture. Although this is highly likely to be true, there is no guarantee that ionization will occur equally among the molecules. It is conceivable that a mass spectrum taken of a mixture will not be so directly modeled as a convex combination of the mass spectra of its components. It is also worth noting that in most applications, a *separation* step such as gas chromatography will be employed, and so the values of contamination we considered are far outside of those that are likely be observed in practice.

In this demonstrative article, we focused on the task of identifying fentanyl from its mass spectrum using a very small and specifically curated library of mass spectra, with the knowledge that a reference spectrum for
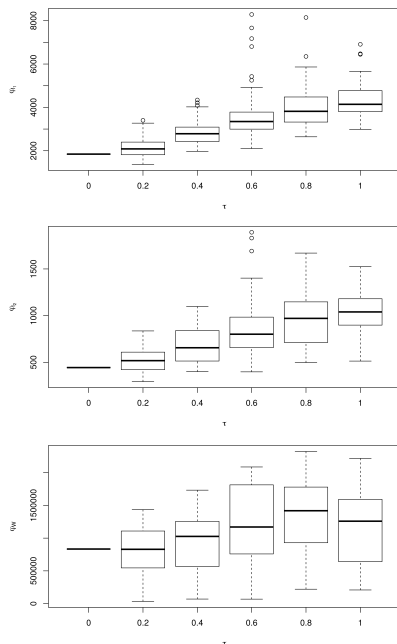
11

Figure 3: Box and whisker plots summarizing the distribution of distance measures ϕ_i where $i = \{\ell_1, \ell_w, W\}$ computed between a query mass spectrum, x_q constructed using a mass spectrum of fentanyl and (9) across $0 \leq \tau \leq 1$, and the mass spectrum of fentanyl contained in the reference library

.

fentanyl was contained in the library. This scale of analysis is appropriate for *targeted* searches as are common in forensics as well as quality control applications (e.g. flavor and fragrance molecules). However, *untargeted* searches, where the analyte is not only unidentified but potentially completely novel, are not an uncommon task in modern scientific applications (e.g. metabolomics, lipidomics, etc.). When conducting an untargeted analysis, it is far more common to use a general-purpose mass spectral library (or several libraries). These libraries contain upwards of a million mass spectra covering a wide variety of molecules [2, 1] and thus allow for a far wider net to be cast while trying to identify an analyte. Naturally, building, maintaining and searching libraries of such magnitude introduces several logistical and computational challenges.

It is also worth discussing the library search objective considered in this article. If used for *automated* identification of an analyte, it is desirable for the library spectrum of the correct identification to be the first entry of a returned hit list. However, there are scenarios where it may be desirable to have several similar compounds make up the top several entries of the hit even at the expense of the correct identification being

12

Table 2: Results of a library search when the queried spectrum is constructed using (10) to model Fentanyl contaminated with Quinine.

	Quinine Contamination, λ					
	0	0.2	0.4	0.6	0.8	1.0
	Hit list position of fentanyl					
ϕ_C	3	2	2	3	2	9
ϕ_S	1	1	1	1	3	16
ϕ_I	1	1	1	1	1	23.5^a
ϕ_{ℓ_1}	5	2	1	7	32	40
ϕ_{ℓ_2}	7	7	6	7	23	42
ϕ_W	36	26	1	20	40	38

(a) Non integer hit list rank suggests that another library candidate produced the same dissimilarity score.

Table 3: Results of a library search when the queried spectrum is constructed using (10) to model Fentanyl contaminated with Serotonin.

	Serotonin Contamination, λ					
	0	0.2	0.4	0.6	0.8	1.0
	Hit list position of fentanyl					
ϕ_C	3	1	1	1	1	2
ϕ_S	1	1	1	1	1	4
ϕ_I	1	1	1	1	1	2
ϕ_{ℓ_1}	5	1	6	6	14	28
ϕ_{ℓ_2}	7	1	3	7	7	14
ϕ_W	36	27	36	40	39	38

first. As noted earlier in this manuscript, the ultimate identification of the analyte is made by an human analyst and so having search results that are easily interpreted and justifiable to a chemist carries significant value. A recent search procedure that attempts to provide analysts with higher-value hit lists is the *Hybrid Similarity Search*[4, 12]. If information about the molecular mass of the analyte and the reference compounds is known *a priori*, the Hybrid Similarity Search procedure promotes compound that have mass spectra that differ in a manner justifiable by the difference in molecular mass. In many cases, however, little information is known about the analyte prior to a library search limiting the applicability of this type of method.

Using mass spectrometry for the chemical identification of fentanyl in forensics or healthcare is accomplished through the combination of experimental and mathematical procedures that are then coupled to intellectual interrogation. Taken together this information aids an analysts decision-making. The reliability of existing approaches is coupled to the quality of measurements taken of an analyte and reference compounds. The construction of high-quality transferable reference libraries and establishment of more finely developed mathematical techniques are both ongoing and

active research areas.

Disclaimer

Certain commercial products are identified to improve clarity of the manuscript; this does not imply endorsement or recommendation by NIST.

References

[1] Wiley registry: Mass spectral library 11th edition.

[2] NIST Mass Spectral Libraries, Online: Accessed December, 2019.

[3] *Scientific Working Group for the Analysis of Seized Drugs (SWG-DRUG) Mass Spectral Libraries*, Online: Accessed December, 2019).

[4] Meghan C Burke, Yuri A Mirokhin, Dmitrii V Tchekhovskoi, Sanford P Markey, Jenny Heidbrink Thompson, Christopher Larkin, and Stephen E Stein. The hybrid search: A mass spectral library search method for discovery of modifications in proteomics. *Journal of proteome research*, 16(5):1924–1935, 2017.

[5] David Cooper, M Jacob, and A Allen. Identification of fentanyl derivatives. *Journal of Forensic Science*, 31(2):511–528, 1986.

[6] Paul AJ Janssen. A review of the chemical features associated with strong morphine-like activity. *BJA: British Journal of Anaesthesia*, 34(4):260–268, 1962.

[7] Tobias Kind and Oliver Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews*, 2(1-4):23–60, 2010.

[8] Elizaveta Levina and Peter Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE, 2001.

[9] Wuchen Li, Ernest K Ryu, Stanley Osher, Wotao Yin, and Wilfrid Gangbo. A parallel method for earth mover's distance. *Journal of Scientific Computing*, 75(1):182–197, 2018.

[10] Arun S. Moorthy and Anthony J Kearsley. Pattern similarity measures applied to mass spectra (in press). In Manuel Cruz, Carlos Pares, and Peregrina Quintela, editors, *Progress in Industrial Mathematics: Success Stories*, chapter 4. Springer International Publishing, Switzerland, 2021.

[11] Arun S Moorthy, Anthony J Kearsley, W Gary Mallard, and William E Wallace. Mass spectral similarity mapping applied to fentanyl analogs. *Forensic Chemistry*, page 100237, 2020.

[12] Arun S Moorthy, William E Wallace, Anthony J Kearsley, Dmitrii V Tchekhovskoi, and Stephen E Stein. 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.

[13] Jayne B Morrow, Jeri D Ropero-Miller, Megan L Catlin, Agnes D Winokur, Amy B Cadwallader, Jessica L Staymates, Shannan R Williams, Jonathan G McGrath, Barry K Logan, Michael M McCormick, et al. The opioid epidemic: moving toward an integrated, holistic analytical response. *Journal of analytical toxicology*, 43(1):1–9, 2019.

[14] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[15] Stephen E Stein. Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical chemistry*, 84(17):7274–7282, 2012.

[16] Stephen E Stein and Donald R Scott. 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.

[17] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[18] J Throck Watson and O David Sparkman. *Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation*. John Wiley & Sons, 2007.