

Evaluation of NIST Library Search Software

AS Moorthy¹, AJ Kearsley², WG Mallard¹, WE Wallace¹, SE Stein¹

Abstract: This paper reports algorithm performance from the NIST mass spectral library search software (version 2.3 released June 2017). The “Normal-Identity” search was evaluated on its ability to correctly identify a molecule from its spectrum when searched against the NIST 17 Main Library. The search correctly identified 95% of query spectra within the top-5 hits of the hitlist, with the correct identification being the top-hit in 72% of the queries. The “Simple-Similarity” and “Hybrid-Similarity” searches were evaluated on their ability to generate hitlists that contain structurally similar molecules to the molecule producing the query spectrum. The Simple-Similarity search hitlists contained at least one structurally similar compound within the top-5 hits for 28% of the queries and the Hybrid-Similarity yielded at least one structurally similar compound within the top-5 hits for 43% of the queries. Performance of the similarity search algorithms will improve as reference libraries continue to become more comprehensive.

Keywords: Electron Ionization Mass Spectrometry (EI-MS), Hybrid Similarity Search, Mass Spectral Library Searching.

Introduction

Mass spectral libraries are an important resource to analytical chemists across a variety of applications. The National Institute of Standards and Technology (NIST) generates several highly curated libraries of mass spectral reference data [1–3]. Additionally, NIST produces search software for interacting with libraries [4]. Three commonly employed algorithms implemented in NIST MS Search v2.3 (2017), appropriate for searching electron-ionization mass spectra, are the “Normal-Identity”, “Simple-Similarity”, and “Hybrid-Similarity” searches.

All NIST mass spectral library search algorithms calculate a *match factor* between a *query* spectrum and a set of reference spectra. A match factor is typically an integer between 0 and 999 that approximates “similarity” between a pair of spectra – each search algorithm can be distinguished by how it computes match factors. The set of reference spectra can include entire libraries of spectra (no pre-search) but are often a well-selected subset of library spectra identified during preprocessing. The search algorithms return the reference spectra (and associated metadata) in order of descending match factor with the query in what is commonly referred to as a “hitlist”.

The objective of the Normal-Identity search algorithm is to return a hitlist that contains the correct identification of the query spectrum at, ideally, the top of the hitlist. Stein and Scott (1994) first described the (identity) match factor associated with “Normal-Identity” searches [5]. Stein (1999) further detailed empirical adjustments to the identity match factor that improved search

¹ Mass Spectrometry Data Center, National Institute of Standards and Technology, Gaithersburg, MD, USA

² Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

performance [6]. The objective of both the “Simple” and “Hybrid” similarity searches is to return a hitlist that can help an analyst propose a structure for their query. Stein and Scott first described the (simple) match factor associated with the Simple-Similarity search [5]. Moorthy et. al. described hybrid match factors associated with the Hybrid-Similarity search as an extension of simple match factors [7].

This paper reports the performance of the Normal-Identity, Simple-Similarity and Hybrid-Similarity search algorithms applied to a large set of query spectra from several external sources.

Method

The test set of electron ionization mass spectra were collected from commercial sources. These spectra were partitioned into two subsets. The “present” subset (P) consisted of spectra for compounds with representation in the NIST 17 EI-MS Main Library (here in referred to as the Reference Library). The “absent” subset (A) were the spectra of compounds not accounted for in the Reference Library.

The spectra from subset P were used to evaluate the performance of the Normal-Identity search algorithm. The performance of the Normal-Identity search algorithm can be directly quantified by observing the rank of the correct identification (determined by InChIKey) for every query spectrum.

The spectra from subset A were used to evaluate the performance of the similarity search algorithms. Quantifying performance of similarity algorithms is not as straight-forward as the identity algorithms. We adopt a measure of structural similarity using Tanimoto coefficients computed on 2D atom pair fingerprints [8–10] between the query and hitlist structures. The hits with Tanimoto coefficients greater than or equal to 0.5 were counted as “structurally similar”. These computations were completed using the ChemmineR package [11, 12] available for R [13].

Results and Discussion

The P subset of spectra appropriate for evaluating the Normal-Identity search consisted of 9663 spectra from two sources. In total, 95% of the queries had its correct identification within the top-5 hits of the hitlist, with 72% of the queries being correctly identified as the top hit (Table 1), the most desirable of outcomes. The source of the test spectra did affect the observed probabilities; however, the correct identification consistently appeared in the top-5 hits (95-96%) and often was the top hit (66-77%) regardless of source library. This performance is consistent with that reported in Stein and Scott (1994) who found the Normal-Identity search (there referred to as Composite Score search) correctly identified 95% of the queries in the top-5, with 76% identified as the top hit [5].

Table 1: Summary of NIST MS Search v2.3 "Normal-Identification" performance for query spectra from external sources searched against the NIST 17 Main Library.

source library	# of spectra in Ref. Library	probability of correct identification being within top n hits				
		$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
B	2281	0.66	0.79	0.88	0.93	0.96
C	7382	0.77	0.88	0.92	0.94	0.95
cumulative	9663	0.72	0.86	0.91	0.94	0.95

There are three reasons the normal identity search did not correctly identify all considered query molecule from their spectra within the top 5 hits of its hitlist.

1. The query molecule has several isomers or analogs that are not distinguishable by their mass spectra alone. In this scenario, the correct identification will still have a substantial match factor with the query spectrum, but it may not be elevated into the top 5. An example of this scenario is provided as Table 2.
2. The query spectrum and its representative in the Reference Library were measured inconsistently. If the considered spectra vary greatly, due to systemic differences in how the spectra were measured, the Normal-Identity search will not be successful/effective (Figure 1). Some systemic issues, such as contamination, can be mitigated using the Automated Mass spectral Deconvolution and Identification System (AMDIS) available with NIST MS Search to select a query spectrum from its chromatogram prior to library searching [14–17].
3. There was an error during curation. This issue was not common, but there were a few examples where the query spectrum and its purported structure were obviously incorrect.

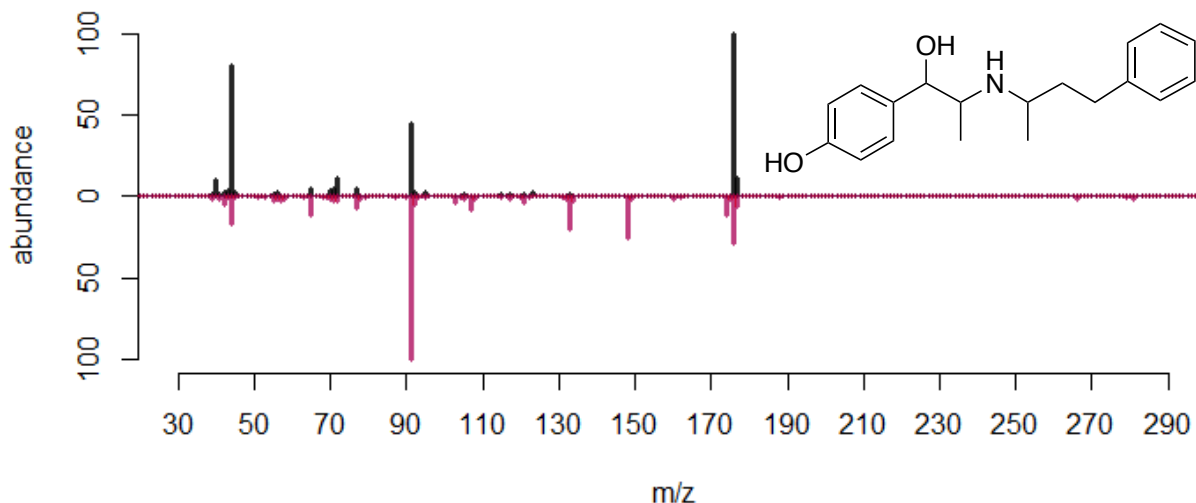
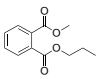
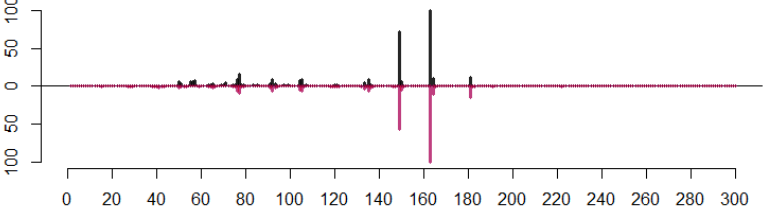
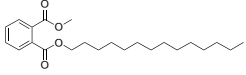
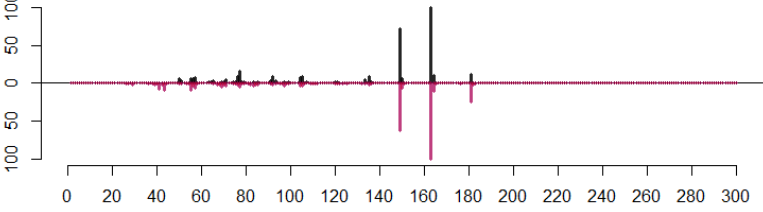
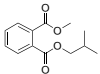
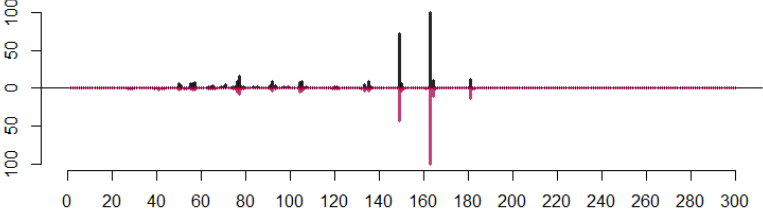
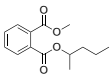
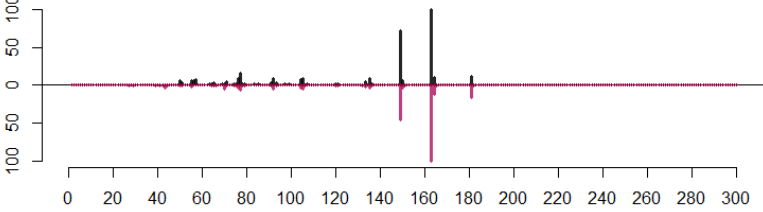
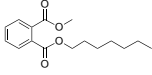
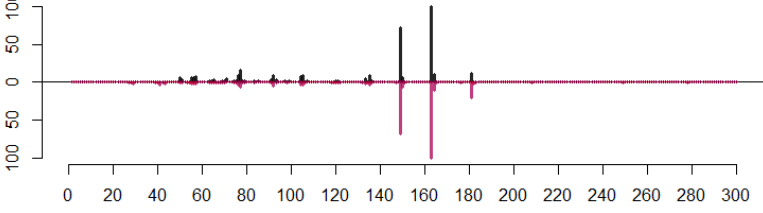
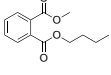
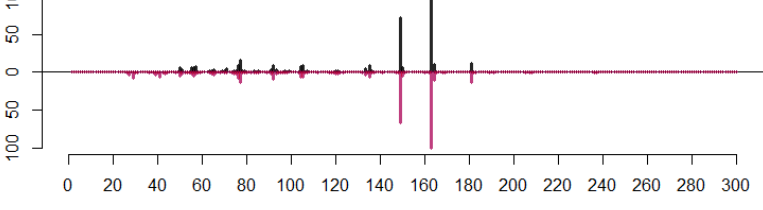


Figure 1: An example of two spectra of the same molecule measured under different conditions. The identity match factor between these two spectra is 601.

Table 2: Example hitlist containing molecules with indistinguishable mass spectra. The correct identification of the query spectrum is entry 6 in the hitlist.

#	structure	identity match factor	head-to-tail plot of query and library spectra
1		875	
2		865	
3		862	
4		860	
5		857	
6		854	

The subset *A* of spectra appropriate for evaluating the similarity searches (Simple and Hybrid) consisted of 12833 spectra from two sources. In total, 27% of the hitlists returned by Simple-Similarity searches had at least one structurally similar molecule in the top-5 of the hitlist, with this molecule being the first hit in 18% of the hitlists (Table 3). In comparison, the Hybrid-Similarity searches returned at least one structurally similar entry in the hitlist top-5 for 41% of the queries (Table 3). When considering the probability of hitlists having multiple (2 or 3) structurally similar molecules to the query in the top-5 of the hitlist, the Hybrid-Similarity search consistently performed twice as well as the Simple-Similarity search (Table 3).

Table 3: Summary of NIST MS Search v2.3 "Simple-Similarity" and "Hybrid-Similarity" performance for query spectra from external sources searched against the NIST 17 Main Library.

source library	# of spectra not in Ref. Library	search	probability of at least 1 similar structure being within top <i>n</i> hits					
			<i>n</i> = 1	<i>n</i> = 2	<i>n</i> = 3	<i>n</i> = 4	<i>n</i> = 5	
A	727	Simple	0.16	0.22	0.27	0.29	0.31	
		Hybrid	0.20	0.27	0.31	0.35	0.36	
B	12106	Simple	0.18	0.22	0.25	0.26	0.28	
		Hybrid	0.28	0.35	0.39	0.42	0.43	
cumulative	12833	Simple	0.18	0.22	0.25	0.26	0.28	
		Hybrid	0.28	0.35	0.39	0.42	0.43	
			probability of at least 2 similar structure being within top <i>n</i> hits					
			search	<i>n</i> = 1	<i>n</i> = 2	<i>n</i> = 3	<i>n</i> = 4	<i>n</i> = 5
A	727	Simple	-	0.06	0.08	0.09	0.11	
		Hybrid	-	0.09	0.12	0.15	0.17	
B	12106	Simple	-	0.06	0.08	0.10	0.11	
		Hybrid	-	0.12	0.17	0.20	0.22	
cumulative	12833	Simple	-	0.06	0.08	0.10	0.11	
		Hybrid	-	0.12	0.17	0.20	0.22	
			probability of at least 3 similar structure being within top <i>n</i> hits					
			search	<i>n</i> = 1	<i>n</i> = 2	<i>n</i> = 3	<i>n</i> = 4	<i>n</i> = 5
A	727	Simple	-	-	0.03	0.04	0.05	
		Hybrid	-	-	0.05	0.06	0.07	
B	12106	Simple	-	-	0.02	0.04	0.05	
		Hybrid	-	-	0.06	0.09	0.11	
cumulative	12833	Simple	-	-	0.02	0.04	0.05	
		Hybrid	-	-	0.06	0.09	0.11	

Structural similarity is an indirect measure of spectral similarity search algorithm performance. It is possible that a hitlist with structurally similar molecules is not actually helpful to an analyst. The best way to assess similarity search algorithms is through manual inspection of hitlists for real applications. An example of hitlist inspection is detailed in Moorthy et. al. [7] for fentanyl analogs. Several successful examples of using the Hybrid Search with electrospray ionization tandem mass spectra can be found in the literature [18–22]. That said, similarity search performances will improve as the reference library becomes more comprehensive.

Conclusions

Mass spectral libraries are an invaluable resource to analytical chemists. They can be used for identifying common molecules using an Identity Search algorithm, and for proposing likely structures for less common molecules from using similarity algorithms. This paper provides an

updated summary of algorithm performance using the latest NIST Mass Spectral Library (NIST 17) and search software (NIST MS Search v2.3).

References

1. Stein, S.: Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Anal. Chem.* 84, 7274–7282 (2012). doi:10.1021/ac301205z
2. Wallace, W.E., Ji, W., Tchekhovskoi, D. V, Phinney, K.W., Stein, S.E.: Mass Spectral Library Quality Assurance by Inter-Library Comparison. 733–738 (2017). doi:10.1007/s13361-016-1589-4
3. Yang, X., Neta, P., Stein, S.E.: Quality control for building libraries from electrospray ionization tandem mass spectra. *Anal Chem.* 86, 6393–6400 (2014). doi:10.1021/ac500711m
4. Stein, S.E.: NIST/EPA/NIH Mass Spectral Library (NIST 17) and NIST Mass Spectral Search Program (Version 2.3) User Manual. (2017)
5. 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)
6. Stein, S.E.: An Integrated Method for Spectrum Extraction. 0305, (1999)
7. 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. *Anal. Chem.* 89, (2017). doi:10.1021/acs.analchem.7b03320
8. Chen, X., Reynolds, C.H.: Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.* 42, 1407–1414 (2002). doi:10.1021/ci025531g
9. Bajusz, D., Rácz, A., Héberger, K.: Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7, 1–13 (2015). doi:10.1186/s13321-015-0069-3
10. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996 (1998). doi:10.1021/ci9800211
11. Cao, Y., Charisi, A., Cheng, L.-C., Jiang, T., Girke, T.: {ChemmineR}: a compound mining framework for R. *Bioinformatics.* 24, 1733–1734 (2008). doi:10.1093/bioinformatics/btn307
12. Cao, Y., Backman, T., Horan, K., Girke, T.: ChemmineR : Cheminformatics Toolkit for R. 1–46 (2014)
13. R Core Team: R: A Language and Environment for Statistical Computing, <https://www.r-project.org/>, (2016)
14. Mallard, W.G., Reed, J.: Automated Mass Spectral Deconvolution & Identification System AMDIS — USER GUIDE. Program. (1997)
15. Halket, J.M., Przyborowska, a, Stein, S.E., Mallard, W.G., Down, S., Chalmers, R. a: Deconvolution gas chromatography/mass spectrometry of urinary organic acids--potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.* 13, 279–284 (1999). doi:10.1002/(SICI)1097-0231(19990228)13:4<279::AID-RCM478>3.0.CO;2-I
16. Meyer, M.R., Peters, F.T., Maurer, H.H.: Automated mass spectral deconvolution and identification system for GC-MS screening for drugs, poisons, and metabolites in urine.

- Clin. Chem. 56, 575–584 (2010). doi:10.1373/clinchem.2009.135517
17. Mallard, W.G., Andriamaharavo, N.R., Mirokhin, Y.A., Halket, J.M., Stein, S.E.: Creation of libraries of recurring mass spectra from large data sets assisted by a dual-column workflow. *Anal. Chem.* 86, 10231–10238 (2014). doi:10.1021/ac502379x
 18. Burke, M.C., Mirokhin, Y.A., Tchekhovskoi, D. V, Markey, S.P., Heidbrink Thompson, J.L., Larkin, C., Stein, S.E.: The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. *J. Proteome Res.* acs.jproteome.6b00988 (2017). doi:10.1021/acs.jproteome.6b00988
 19. Remoroza, C.A., Mak, T.D., Leoz, M.L.A. De, Mirokhin, Y.A., Stein, S.E.: Creating a Mass Spectral Reference Library for Oligosaccharides in Human Milk. *Anal. Chem.* 90, 8977–8988 (2018). doi:10.1021/acs.analchem.8b01176
 20. 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. *Mol. Nutr. Food Res.* 1801184, 1801184 (2018). doi:10.1002/mnfr.201801184
 21. Barupal, D.K., Fan, S., Fiehn, O.: Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. *Curr. Opin. Biotechnol.* 54, 1–9 (2018). doi:10.1016/j.copbio.2018.01.010
 22. Blaženović, I., Kind, T., Ji, J., Fiehn, O.: Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites.* 8, (2018). doi:10.3390/metabo8020031