SPLASH, a hashed identifier for mass spectra

SPLASH, a hashed identifier for mass spectra

To the Editor:

Over the past few years, as the use of mass spectrometry (MS) has increased, multiple spectral libraries, databases and software frameworks have been created to enable sharing and searching of MS data. However, finding all the spectra that correspond to a specific compound across different databases continues to be a challenge. A spectral identifier that improves the exchange of mass spectra, as well as provenance and duplicate detection, would address these issues and enhance searchability.

MassBank[1] (http://www.massbank.jp and http://massbank.eu/MassBank/) has been the source of data for other open libraries, such as the Global Natural Products Social Molecular Networking[2] (GNPS) and Human Metabolome Database[3] (HMDB) libraries as well as the MetaboLights reference layer[4]. In turn, HMDB and community-contributed spectra from GNPS have also been imported into MassBank of North America (MoNA; http://mona.fiehnlab.ucdavis.edu/), while GNPS searches public MS data against the above-mentioned libraries as well as the National Institutes of Standards and Technology (NIST) spectral library[5]. The mzCloud (https://www.mzcloud.org/) library contains some spectra generated from the same raw data that were used to create MassBank records. As these examples show, the complexity and the cross-import of data are increasing, together with the number of mass spectra, such that these different resources can now contain identical, or near identical, spectra under different accession numbers. For example, the library entries PR100026 (MassBank, MoNA), 5464 (HMDB) and CCMSLIB00000222858 (GNPS) all refer to exactly the same mass spectrum of caffeine, originally sourced from MassBank. As the different libraries focus on different compound domains[7], users wishing to access mass spectra from all compounds must use several resources, some of which are not fully open access (e.g., NIST and mzCloud).

Mass spectra are highly variable, with one to potentially thousands of mass-to-charge (m/z) and intensity entries per spectrum, presenting a challenge in the design of an optimal identifier. However, other life science databases have faced a similar need. For databases with chemical structures, the InChI code and the hashed InChIKey[7,8] of fixed length, which have been broadly adopted as chemical identifiers, can be easily stored in databases, compared across resources and, for InChIKeys, searched on general-purpose search engines[9]. A hash is a one-directional mapping between a long, potentially complex object and a typically much shorter hash string with a fixed length of characters and numbers. For chemicals, the InChIKey is much easier to search than the (generally) much longer InChI, which contains special characters. Although it is not possible to

Table 1 SPLASH statistics for selected compounds[a]

	Alanine	Caffeine	Codeine	Clarithromycin
InChIKey first block	QNAYBMKLOCPYGJ	RYYVLZVUVIJVGH	OROGSEYTTFOCAN	AGOYDEPGAOXOCK
PubChem CID(s)	602, 5950, 71080	2519	2828, 5284371	894029
ChemSpider ID(s)	582, 5735, 64234	2424	2726, 4447447, 4642640	10342604
Monoisotopic mass (Da)	89.047676	194.080383	299.15213	747.476868
Number of spectra	58 (10 negative)	80	19	21
Coupling (GC/LC/neither)	6/37/15	14/52/14	0/19/0	0/21/0
Second/third/fourth blocks	10/7/43	16/13/67	6/9/19	6/13/21
List of second blocks (number)	0006 (32); 000i (10); 014i (6); 01b9, 00kf, 000f (2); 0f79, 0a4i, 00di, 0007 (1)	0002 (25); 000i (21); 0006 (9); 0536 (5); 052f (3); 0a4l, 05nf, 01x9, 00di, 001l, 000b (2); 01w0, 016u, 00dr, 000l, 000j (1)	0udi (8); 0uxr (4); 0lea, 015a, 0159 (2); 0uyi (1)	001i (6); 00di (4); 0a4j, 0a4i, 052e, 0006 (2); 0a59; 05o0, 053r (1)
List of third blocks (number)[b]	9000000000 (46); *0900000000 (5);* **9002000000 (2);** *6900000000 (2); 1940000000 (1); 1900000000 (1); 0910000000 (1)*	0900000000 (47); 1900000000 (8); 9100000000 (5); 3900000000 (5); 4900000000 (3); 2900000000 (3); 9800000000 (2); 6900000000 (2); 9500000000 (1); 9200000000 (1); 8900000000 (1); 7900000000 (1); 5900000000 (1)	0009000000 (6); 0973000000 (2); 0920000000 (2); 0910000000 (2); 0390000000 (2); 0139000000 (2); 1952000000 (1); 1940000000 (1); 1930000000 (1)	0000090000 (5); 9000000000 (4); 4900000000 (2); 9800000000 (1); 9300000000 (1); 9200000000 (1); 8900000000 (1); 3900020000 (1); 1900060800 (1); 1900030300 (1); 1900020500 (1); 0800070900 (1); 0000001900 (1)

[a]The lower two rows show the variety of different spectra per compound. The combination of second and third blocks is selective (e.g., 0a41-1940000000 and 01ea-1940000000 for alanine and codeine). [b]Data for alanine show how derivative spectra (italics) and suspicious (bold) database entries can be detected with the third block.

Table 2 Example of most common second block, third block and combinations of blocks from SPLASH, together with the approximate distribution of compounds

Place[a]	2nd block	Number of spectra[b]	% Spectra	Number of structures[c]	3rd block	Number of spectra[b]	% Spectra	Number of structures[c]	Second+ third block	Number of spectra[b]	% Spectra	Number of structures[c]
1	0006	36,323	6.82	21,990	9000000000	49,553	9.30	17,920	0006-9000000000	6,569	1.23	2,930
2	0a4i	33,191	6.23	17,529	0900000000	36,724	6.89	7,375	0a4i-9000000000	4,771	0.90	2,023
3	00di	28,888	5.42	14,008	9100000000	19502	3.66	13,435	001i-0900000000	3,438	0.65	1,288
4	014i	28,213	5.30	15,278	9200000000	14988	2.81	11,693	000i-0900000000	3,287	0.62	1,111
5	000i	26792	5.03	12,965	0090000000	14,724	2.76	3,507	00di-0900000000	3,251	0.61	1,173
6	001i	25438	4.78	11,697	1900000000	13,351	2.51	8,679	0002-9000000000	3,020	0.57	1,161
7	004i	24893	4.67	11,728	2900000000	13,201	2.48	10,196	014i-0900000000	2,791	0.52	1,062
8	0002	24,247	4.55	12,543	3900000000	13,046	2.45	10,737	0002-0900000000	2,744	0.52	1,096
9	0udi	21,556	4.05	10,389	9300000000	12,504	2.35	10,380	001i-9000000000	2,683	0.50	1173
10	03di	19,913	3.74	9,748	4900000000	11,438	2.15	9,701	004i-9000000000	2,605	0.49	929
20	004l	2,444	0.46	1,966	9800000000	6,461	1.21	5,810	014i-9000000000	1,855	0.35	904
30	0fb9	1,843	0.35	1,385	0390000000	2,289	0.43	1,701	0002-0090000000	1,238	0.23	537
40	00fr	1,700	0.32	1,362	9510000000	1,512	0.28	1482	0006-0090000000	1,024	0.19	426
50	00xr	1,600	0.30	1256	8910000000	1336	0.25	1,306	000i-0009000000	909	0.17	380
100	0abc	949	0.18	806	9630000000	585	0.11	580	000i-9200000000	541	0.10	376
200	0fmi	218	0.04	195	9350000000	250	0.05	249	014i-9400000000	255	0.05	239
500	0ac3	76	0.01	76	9102000000	60	0.01	59	0udi-0590000000	94	0.02	87

[a]Place indicates how common the combination is (1 = most common, 200 = 200th most common). [b]The number of spectra and substances (estimated by first block of the InChIKey) with the 'most common' second, third and second+third SPLASH blocks, calculated on a subset of the validation data set containing 532,675 spectra with compound information. [c]The number of structures is an estimate; missing structure information was filled in automatically using the Chemical Translation Service (http://cts.fiehnlab.ucdavis.edu/).

obtain the original object back purely from the hash value, hash keys provide easy access to the original data within a data collection.

We designed the SPLASH (SPectraL hASH) as an unambiguous, database-independent spectrum identifier that fulfills the criteria outlined above and offers some additional functionality. Inspired by the broad applicability of the InChIKey across cheminformatics and like the InChIKey (which encodes skeleton, stereochemistry and charge), SPLASH contains separate blocks that define different layers of information, separated by dashes. As an example, the full SPLASH of the caffeine spectrum above is splash10-0002-0900000000-b112e4e059e1ecf98c5f. The first block is the SPLASH identifier, the second and third are summary blocks, and the fourth is the hash block.

To calculate a SPLASH, spectra are converted into a canonical text representation: the intensities are normalized to an integer value between 0 and 100, with m/z values given in exactly six decimal places. To ensure consistent handling between different software and implementations, entries with zero intensities are included, but empty ("N/A") values are eliminated before creating the SPLASH. The first block (splash10) encodes the SPLASH identifier, starting with letters for semantic web compatibility, followed by a number representing the measurement type (1 for MS, 2 and above for other data types to be included in the future) and the SPLASH version number, starting at 0, to allow future specification updates. Thus, splash10 is a SPLASH identifier for MS, version 0.

Both the second and third blocks are spectral summaries, which serve to prefilter and restrict searches. In the second and third blocks, intensities are summed over fixed (but different) bin sizes and wrapped over ten bins. The wrapped bin (zero-based) index for a given ion is computed as floor (m/z ÷ BinSize) modulo 10. This wrapping strategy accommodates all possible spectral mass ranges while maintaining fixed-length summary blocks. The second block (0002) is formed using a reduced spectrum (the top ten or fewer ions greater than 10% of the base peak). This reduced spectrum is summed over bins of 5 Da. Each bin is then scaled to a single-digit integral value in base 3 (0–2), and the resulting 10 digit histogram is converted to a base 36 number, resulting in a 4-digit block. In the third block (0900000000) the intensities are summed over 100-Da bin sizes, each bin is then scaled to a single-digit, integral base-10 digit (0–9).

The fourth block (b112e4e059e1ecf98c5f) is a hash of the full spectrum in Secure Hash Algorithm[10] SHA256 (numbers and lowercase letters only), calculated in hexadecimal notation and truncated to 20 characters. The full spectrum string of m/z and relative intensity pairs are sorted by ascending m/z and then by descending intensity. The m/z value is multiplied by 10^6, cast to a long (64-bit) integer and joined with the normalized intensity as strings separated by a colon. The resulting ion pairs are then joined, delimited by a single space. Specification document and reference implementations have been created for several programming environments (Python, Scala, C++, C#, R, Ruby and Java) under a BSD-3 license as well as a REST interface (**Supplementary Code**); additional information is available at http://splash.fiehnlab.ucdavis.edu/ and all code is also available on GitHub at https://github.com/berlinguyinca/spectra-hash.

The SPLASH concept was developed and refined on a data set of 563,902 mass spectra from MassBank[1], GNPS[2], HMDB[3], ReSpect[11], FiehnLib[12] and NIST 14 (ref. 5); all but the NIST spectra (which cannot be released publically) are available on MoNA (http://mona.fiehnlab.ucdavis.edu/). This data set is a mix of many types of mass spectra and the SPLASH was designed to account for this, plus be easily searchable in general-purpose search engines, offer a unique identifier (through the hash) and basic pre-filter and similarity functionality (through the second and third blocks).

Ensuring that all these features are present in one short text string requires compromise; the SPLASH is not intended to replace more sophisticated database-specific functions, but does offer simple cross-database functionality. The second block was chosen from 136 different potential block formats as the most appropriate short, web-search-compatible way to reduce the mass spectral search space.

To determine the best-performing second block, we queried a subset of 19,435 spectra against the full 563,902 data set. The second block that we selected for use reduced the search space by 94% or above (36,107 spectra or less) in all cases, while returning 87% of all spectra within a similarity score of 700 (using the NIST cosine similarity score[5,13]) of the queried spectra. In contrast, other tested formats for this block returned more spectra (maximum 93.4%), but too many spectra (up to 100,000 or 1 in 5 spectra) remained in the search space so that the search space reduction was insufficient. The third block provides a visual summary (shown in **Table 1** for selected compounds) and a simple text-based summary and basic similarity search that can be used in search engines or spreadsheets. More information on the most common second and third blocks, as well as the most common combinations and the approximate distribution of compounds (not all spectra are annotated with structures in the validation set) is given in **Table 2**.

Although the mapping from object to hash should ideally be unique, hash collisions (where two totally different objects have the same hash, or fourth block of the SPLASH) may occur, depending on the hash algorithm and length of the hash string. Testing the fourth block for hash collisions on the full data set of 53,250,921 spectra (563,902 from the validation set and 52,687,019 from BinBase[14]) revealed that identical SPLASHes arose only from mass spectra containing a single ion of the same mass, where the SPLASH is identical by definition due to intensity normalization. The theoretical probability for a collision[15] with any given hash is approximately 10^{-31} for a database containing 10^9 spectra and is further reduced by the presence of two preceding spectral summary blocks. Thus, the SPLASH fulfills its role as a unique identifier while offering simple summary and searching functionality.

The SPLASH has already been implemented in MassBank[1], MoNA (http://mona.fiehnlab.ucdavis.edu/) GNPS[2], HMDB[3], MetaboLights[4] and mzCloud (https://www.mzcloud.org/), as well as software tools including MZmine[16], MS-DIAL[17], RMassBank[18], BinBase[14], Bioclipse[19] and the Mass Spectrometry Development Kit (MSDK; https://msdk.github.io/).

The format of the SPLASH allows direct access to spectra on database websites and searching using general purpose search engines. Spectral libraries with more restrictive licenses (e.g., mzCloud and possibly NIST) could also use the SPLASH to provide summarized information about their spectra. SPLASH enables an easier calculation of spectral overlap between libraries, to detect and remove exact duplicate spectra and perform provenance operations. Through the second and third blocks, SPLASH empowers quick searches for similar spectra within or between libraries, using a variety of search methods. The SPLASH algorithm has been kept independent of metadata, similar to the InChIKey, because an extension to include and distinguish metadata (e.g., analytical conditions or chemical information) would rapidly become complex and reduce the applicability of the identifier. Instead, the SPLASH is designed to facilitate quick queries and subsequent metadata retrieval. The widespread adoption of the SPLASH as a standard spectral identifier allows automated, cross-resource spectral exchange and enables enhanced searchability and data processing across MS platforms.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3689).

Editor's note: This article has been peer-reviewed.

ACKNOWLEDGMENTS

G.W., S.S.M., D.P. and O.F. were supported by National Institute of Health U24 DK097154 and National Science Foundation MCB 1139644; M.W., P.C.D. and N.B. by the National Institutes of Health 5P41GM103484 for the Center for Computational Mass Spectrometry; E.L.S. by SOLUTIONS (European Union's Seventh Framework Programme Grant Agreement No. 603437); P.C.D. by the European Union's Horizon2020 program under the Grant Agreement No. 634402 (METASPACE); R.F.M. was funded by the Database Integration Coordination Program of the National Bioscience Database Center, Japan. European MassBank is supported by the NORMAN Association (France) and hosted by the Helmholtz Centre for Environmental Research. Discussions with anonymous parties, T. Hofstetter and the reviewer feedback are gratefully acknowledged.

Gert Wohlgemuth[1], Sajjan S Mehta[1], Ramon F Mejia[2], Steffen Neumann[3], Diego Pedrosa[1], Tomáš Pluskal[4], Emma L Schymanski[5], Egon L Willighagen[6], Michael Wilson[7], David S Wishart[7], Masanori Arita[2,8], Pieter C Dorrestein[9,10], Nuno Bandeira[9,11,12], Mingxun Wang[11,12], Tobias Schulze[13], Reza M Salek[14], Christoph Steinbeck[14], Venkata Chandrasekhar Nainala[14], Robert Mistrik[15], Takaaki Nishioka[16] & Oliver Fiehn[1,17]

[1]West Coast Metabolomics Center and Genome Center University of California Davis, Davis, California, USA. [2]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan. [3]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany. [4]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts, USA. [5]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland. [6]Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, the Netherlands. [7]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada. [8]National Institute of Genetics, Mishima, Shizuoka, Japan. [9]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, California, USA. [10]Departments of Pharmacology and Pediatrics, School of Medicine, UC San Diego, La Jolla, California, USA. [11]Computer Science and Engineering, UC San Diego, La Jolla, California, USA. [12]Center for Computational Mass Spectrometry, UC San Diego, La Jolla, California, USA. [13]Department of Effect-Directed Analysis, UFZ Helmholtz Centre for Environmental Research GmbH, Leipzig, Germany. [14]European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [15]HighChem Ltd., Bratislava, Slovakia. [16]Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Kyoto, Japan. [17]Biochemistry Department, King Abdulaziz University, Jeddah, Saudi Arabia.

e-mail: G.W. (wohlgemuth@ucdavis.edu), E.L.S. (emma.schymanski@eawag.ch) or O.F. (ofiehn@ucdavis.edu).

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.3689).

1. Horai, H. *et al. J. Mass Spectrom.* **45**, 703–714 (2010).
2. Wang, M. *et al. Nat. Biotechnol.* **34**, 828–837 (2016).
3. Wishart, D.S. *et al. Nucleic Acids Res.* **41**, D801–D807 (2013).
4. Haug, K. *et al. Nucleic Acids Res.* **41**, D781–D786 (2013).
5. Stein, S.E. *et al.* NIST Mass Spectral Search Program version 2.2 and NIST/EPA/NIH Mass Spectral Library, June 2014. (National Institute of Standards and Technology, US Secretary of Commerce, USA, 2014).
6. Vinaixa, M. *et al. Trends Analyt. Chem.* **78**, 23–35 (2016).
7. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. *J. Cheminform.* **5**, 7 (2013).
8. Heller, S.R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. *J. Cheminform.* **7**, 23 (2015).
9. Southan, C. *J. Cheminform.* **5**, 10 (2013).
10. National Institute of Standards and Technology. *Secure Hash Standard*, FIPS PUB 180–4, http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf (2015) (accessed 8 June 2016).
11. Sawada, Y. *et al. Phytochemistry* **82**, 38–45 (2012).
12. Kind, T. *et al. Anal. Chem.* **81**, 10038–10048 (2009).
13. Stein, S.E. & Scott, D.R. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
14. Skogerson, K., Wohlgemuth, G., Barupal, D.K. & Fiehn, O. *BMC Bioinformatics* **12**, 321 (2011).
15. Preshing, J. http://preshing.com/20110504/hash-collision-probabilities/ (2011) (accessed 8 June 2016).
16. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. *BMC Bioinformatics* **11**, 395 (2010).
17. Tsugawa, H. *et al. Nat. Methods* **12**, 523–526 (2015).
18. Stravs, M.A., Schymanski, E.L., Singer, H.P. & Hollender, J. *J. Mass Spectrom.* **48**, 89–99 (2013).
19. Spjuth, O. *et al. BMC Bioinformatics* **8**, 59 (2007).