Title

Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry

Authors

Kind, Tobias
Fiehn, Oliver

Publication Date

2017-09-01

DOI

10.1016/j.phytol.2016.11.006

Peer reviewed

HHS Public Access

Author manuscript

Phytochem Lett. Author manuscript; available in PMC 2018 September 01.

Published in final edited form as:

Phytochem Lett. 2017 September ; 21: 313–319. doi:10.1016/j.phytol.2016.11.006.

Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry

Tobias Kind[1] and **Oliver Fiehn**[1,2]

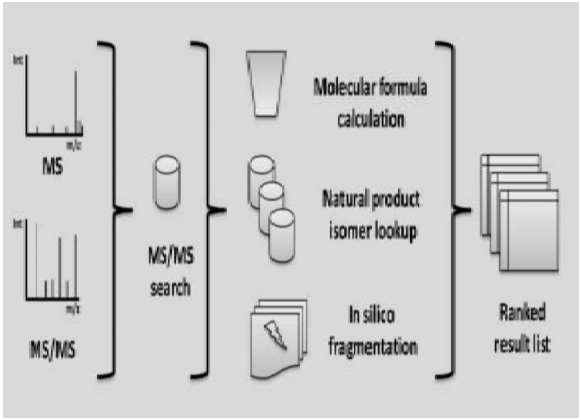
[1]West Coast Metabolomics Center, UC Davis, Davis 95616, California, U.S.A

[2]Biochemistry Department, King Abdulaziz University, Jeddah, Saudi-Arabia

Abstract

Complete structural elucidation of natural products is commonly performed by nuclear magnetic resonance spectroscopy (NMR), but annotating compounds to most likely structures using high-resolution tandem mass spectrometry is a faster and feasible first step. The CASMI contest 2016 (Critical Assessment of Small Molecule Identification) provided spectra of eighteen compounds for the best manual structure identification in the natural products category. High resolution precursor and tandem mass spectra (MS/MS) were available to characterize the compounds. We used the Seven Golden Rules, Sirius2 and MS-FINDER software for determination of molecular formulas, and then we queried the formulas in different natural product databases including DNP, UNPD, ChemSpider and REAXYS to obtain molecular structures. We used different in-silico fragmentation tools including CFM-ID, CSI:FingerID and MS-FINDER to rank these compounds. Additional neutral losses and product ion peaks were manually investigated. This manual and time consuming approach allowed for the correct dereplication of thirteen of the eighteen natural products.

Graphical abstract



Correspondence to: Tobias Kind (tkind@ucdavis.edu) or Oliver Fiehn (ofiehn@ucdavis.edu).

Keywords

Natural products; dereplication; tandem mass spectra; Seven Golden Rules; MS-FINDER

1. Introduction

Natural compound identification commonly employs multiple analytical techniques. Especially NMR is highly useful to determine the correct connection table and stereochemistry of compounds (Wolfender et al., 2001). It is however possible to perform a dereplication of known natural products looking up such compounds in natural product or public compound databases (Corley and Durley, 1994) and using only mass spectral information. We here discuss the manual methodology for annotation of known natural products by interpreting and utilizing high resolution tandem mass spectrometry information (Kind and Fiehn, 2010). Our approach included MS/MS search for fast compound annotation and for those compounds that are not covered in MS/MS databases a manual procedure. This involved the determination of molecular formulas, the subsequent query of these molecular formulas in natural product databases and retrieval of compound isomer structures. These were ranked by in-silico fragmentation algorithms and MS/MS spectra and rankings were refined by a manual and time consuming process.

2. Materials and methods

The challenges for the CASMI 2016 Category-1 contest are natural products from several organisms of different possible origin (plants, fungi, marine sponges, algae or micro-algae) acquired on QToF instruments from Waters and Agilent (see Figure 1). Based on the MS and MS/MS, the goal was to determine the correct molecular core structure (without any stereo information) at the given retention time using the spectral data and the additional information provided. The contest website lists detailed results and participant lists (http://www.casmi-contest.org/2016/results-cat1.shtml). The submitted structures are then ranked according to the absolute rank of the correct solution and for tied scores the average of tied ranks is used.

Molecular formulas were determined with the Seven Golden Rules (Kind and Fiehn, 2007) and Sirius^2 (Böcker et al., 2009). In some cases the CASMI provided data was not sufficient and MS^1 and MS^2 data were extracted from the raw files using ProteoWizard (Kessner et al., 2008) and MZMine2 (Pluskal et al., 2010). Molecular formulas were queried in the Dictionary of Natural Products (http://dnp.chemnetbase.com/), the UNPD database (Gu et al., 2013) (http://pkuxxj.pku.edu.cn/UNPD/) as well as ChemSpider (Pence and Williams, 2010) (http://www.chemspider.com/) and REAXYS (http://www.reaxys.com) and Chemical Structure Lookup Service (CSLS) (Sitzmann et al., 2008) (https://cactus.nci.nih.gov/cgi-bin/lookup/search) to obtain molecular isomer structures. Obtained isomer candidates from the natural product databases were downloaded as SMILES or InChI (Heller et al., 2013) and InChIKey (Heller et al., 2015). The CSI:FingerID website (Dührkop et al., 2015) (http://www.csi-fingerid.org) and the freely available MS-FINDER software (Tsugawa et al., 2016) (http://prime.psc.riken.jp/Metabolomics_Software/) were utilized for in-silico fragmentation

and compound ranking. Additionally CFM-ID (Allen et al., 2014) (Allen et al., 2015) was used to generate in-silico MS/MS spectra. ChemAxon Molconvert and MSketch software (ChemAxon, 2016) and Open Babel 2.3.2 (O'Boyle et al., 2011) were used to view and convert compound structures. Finally all compound data was manually converted into MGF format and MS/MS spectra were submitted to NIST14 GUI MS/MS database search using a 5 ppm or 10 ppm precursor filter. For some cases additional neutral losses and characteristic product ion peaks were investigated with the MS-FINDER GUI.

3. Results

An overview of all the challenge compounds can be found in Figure 1 and compound names, molecular formula and database ranks are listed in Table 1.

Compound **1** was dibromophakellin (MKCFBJDWCJAOTN-GXSJLCMTSA-N) a bromine containing alkaloid that was first observed in the marine sponge *Phakellia flabellate* (Sharma and Burkholder, 1971). The compound was measured with LC-MS/MS and less than 5ppm mass accuracy on a QTOF instrument. MS-FINDER, CSI:FingerID and the Seven Golden Rules all confirmed the formula $C_{11}H_{11}Br_2N_5O$. Isomers were searched in four different databases (see Table 1). CSI:FingerID ranked the compound as #5 and MS-FINDER reported the compound as #3. A manual investigation after the automatic ranking showed a loss of 59 Da from *m/z* 387.9 to *m/z* 328.89 which could be attributed to the loss of a guanidine moiety H2N-CNH-NH2 from the dibromophakellin. However this useful information was not correctly included which lowered the assignment to rank #5 in the submission.

Compound **2** was oroidin (QKJAXHBFQSBDAR-OWOJBTEDSA-N) a bromo-pyrrole derivative first discovered in the sponge *Agelas oroides* (Forenza et al., 1971) with the structure later corrected (Aiello et al., 2007). MS-FINDER, CSI:FingerID and the Seven Golden Rules all confirmed the formula $C_{11}H_{11}Br_2N_5O$. Ten isomers were found in the Dictionary of Natural Products. The MS/MS spectrum was devoid of fragment ions with the exception of *m/z* 122.0708 in the product ion spectrum which was also correctly modelled by CFM-ID. CSI:FingerID and MS-FINDER correctly ranked the Oroidin isomer as first hit. No MS/MS database search could be utilized and use of retention time information was not possible. The correct structure was submitted as first hit.

Compound **3** was excluded from the challenge by the organizers.

Compound **4** was cytochalasin B (GBOGMAARMMDZGR-TYHYBEHEBX) which is a cytotoxic compound independently isolated by two different groups in 1966 from *Fungi imperfecti* and *Helminthosporium dematiaceous* (Peterson and Mitchison, 2002). The formula $C_{29}H_{37}NO_5$ was received as top hit with the Seven Golden Rules, CSI:FingerID and MS-FINDER. The core skeleton structure was correctly determined by CSI:FingerID and MS-FINDER, however with different stereochemistry. MS/MS spectra were simulated with CFM-ID with Cytochalasin B ranking highest in MS/MS dot-product search. The compound showed very rich fragmentations with over hundred product ion peaks and the correct solution (but slightly different stereochemistry) was submitted as top hit.

Compound **5** was cymoside (GMJDEOWOJAKXGU-JUSLMEEHSA-N) a recently discovered monoterpene indole alkaloid from the flowering plant *Chimarrhis cymosa* (Lémus et al., 2015). The formula was correctly determined by MS-FINDER as $C_{27}H_{34}N_2O_{10}$ and as fourth ranking formula in CSI:FingerID. The compound itself was not yet covered in the Dictionary of Natural Products (DNP), the CSLS DB or in Chemspider, only in PubChem. The MS/MS spectrum reveals a loss of 179 Da which can be related to a hexose moiety. Only DNP, Reaxys and CSLS DB were considered in this case for structure lookup. A SciFinder search and the inclusion of PubChem candidates could have led to the submission of a potentially correct candidate. Interestingly only one team out of seven participating Category-1 CASMI teams submitted a valid candidate. No correct candidate structure was submitted for this challenge.

Compound **6** was again dibromophakellin (MKCFBJDWCJAOTN-GXSJLCMTSA-N) ($C_{11}H_{11}Br_2N_5O$), but this time measured with 10ppm mass accuracy on a different vendor instrument. The experimental isotopic abundance error for *m/z* 387.9 was 19.48%. However an isotopic abundance of 10% was falsely assumed and the formula was consequently wrongly determined as $C_{12}H_{11}Br_2N_3O_2$. The remaining peaks in the LC-MS/MS chromatogram were not used to determine the isotopic abundance error. The MS/MS spectrum was not comparable to that from compound **1** but only confirmed two bromine compounds. Due to the false formula assignment no correct isomer structure was submitted for this challenge.

Compound **7** was again oroidin (QKJAXHBFQSBDAR-OWOJBTEDSA-N), however measured on a 10ppm instrument. The molecular formula was correctly determined as $C_{11}H_{11}Br_2N_5O$. The MS/MS spectrum was not comparable to that from compound **2**, only *m/z* 122.07 was matching. The only other two product ion peaks at *m/z* 249.8 and 251.8 were both around 5% abundance. The peak at 122 Da is not a diagnostic peak and can refer to multiple elemental compositions and multiple isomer structures. Over 500 product ion peaks from different compounds were found when performing a sequential MS/MS database search. It was falsely assumed that the contest designers will not give out duplicate challenge candidates. Also the manual investigation yielded no useful insight into the correct structure candidate. Therefore five final compounds were intentionally submitted with the same score and the correct compound was ranked 3rd.

Compound **8** was again cytochalasin B (GBOGMAARMMDZGR-TYHYBEHESA-N) this time measured on a different instrument with 10 ppm mass accuracy. The formula was correctly determined as $C_{29}H_{37}NO_5$. Eleven structures were submitted with the same score, nine of these compounds were annotated as different stereoisomers of Cytochalasin B, only differentiated by a different second block of the InChIKey stereochemistry layer. The correct structure was submitted as top hit and the average rank was calculated as #2.

Compound **9** was brucine (RRKTZKIUPZVBMF-IBTVXLQLSA-N) is a compound related to strychnine and was discovered by Joseph Pelletier and Joseph Caventou in 1817 (Delepine, 1951) and named after the traveler James Bruce who collected samples of tree *Brucea antidysenterica* when travelling to Egypt (Hepper, 1980). The compound with the

formula $C_{23}H_{26}N_2O_4$ was directly discovered from the MS/MS spectra in MassBank and NIST14. The correct structure was submitted ranking first.

Compound **10** was creatinine (DDRJAANPRJIHGJ-UHFFFAOYSA-N) and is a compound found in high concentrations in urine and blood (Narayanan and Appleton, 1980). The formula of the compound is $C_4H_7N_3O$ and the compound was found by MS/MS search in MassBank and NIST14 and subsequently confirmed by MS-FINDER. The correct structure was submitted as first hit.

Compound **11** was anthrone (RJGDLRCDCYRQOQ-UHFFFAOYSA-N) a compound commonly found in plants (Perkin and Hummel, 1894). Interestingly no reference MS/MS spectrum was found in common mass spectral databases, despite over 2,400 published research papers. The formula $C_{14}H_{10}O$ and structure was determined correctly by MS-FINDER and submitted as first hit.

Compound **12** was flavone (VHBFFQKBGNRLFZ-UHFFFAOYSA-N) an anthracene derivate and plant dye, which was synthesized over 100 years ago (Feuerstein and v. Kostanecki, 1898) by the group of von Kostanecki in Bern (Tambor, 1912). The compound with the formula $C_{15}H_{10}O_2$ was found with MS/MS search in the NIST14 database and correctly submitted after MS-FINDER confirmation as first hit.

Compound **13** was medroxyprogesterone (FRQMUZJSZHZSGN-UHFFFAOYSA-N) a steroid derivative and progestin drug first synthesized by Syntex S.A. in Mexico in 1956 (GB 868303, 6a-Alkyl-4-pregnene and later Upjohn with both companies known for strong interest in steroid synthesis (Djerassi, 1992; Hogg, 1992). The compound with the formula $C_{22}H_{32}O_3$ was found by MS/MS search in the NIST database and the correct structure was submitted.

Compound **14** was abietic acid (RSWGJHLUYNHPMX-ONCXSQPRSA-N) a diterpene first discovered by Kelbe in 1880 (Kelbe, 1880) and later also found in *Callitris quadrivalvis* conifers (Henry, 1901). The formula from the MS1 spectrum was correctly annotated by MS-FINDER as $C_{20}H_{30}O_2$. A database search in DNP revealed 669 potential isomer candidates for this formula. An MS/MS search revealed isopimaric acid had a very similar matching spectrum and was falsely submitted as top ranking solution. MS-FINDER ranked the structure on position 128 in the selected dataset. Interestingly none of the competitors was able to rank the substance correctly or positions for this compound were extremely low. In the final submission Abietic acid only ranked 292nd (average rank based on tied scores) and therefore was ranked the worst for all category one submissions.

Compound **15** was estrone-3-(beta-D-glucuronide) (FJAZVHYPASAQKM-JBAURARKSA-N) a mammalian steroid metabolite which is now screened frequently because of high concentrations in wastewater and concerns about hydrolysis back to the active steroidal form (Shrestha et al., 2012). The compounds formula $C_{24}H_{30}O_8$ was correctly identified and Estrone 3-glucuronide was ranked highest in MS-FINDER. The compound identity was also confirmed by MS/MS search in the NIST14 database. The correct solution was submitted as top hit.

Compound **16** was alizarin (RGCKGOZRHPZPFP-UHFFFAOYSA-N) and is a red colored anthraquinone derivate found in roots of *Rubia tinctorum* plants (Angelini et al., 1997). The correct formula $C_{14}H_8O_4$ was annotated by MS-FINDER and the compound ranked 3rd in the in-silico fragmentation when compared to the experimental MS/MS spectrum. There are a number of very similar dihydroxyanthraquinones including Dantron, Quinizarin and Xanthopurpurin that only differ in the position of the hydroxy substitution. The compound was confirmed by MS/MS search and the correct candidate was submitted.

Compound **17** was thyroxine (XUIIKFGFIJCVMT-UHFFFAOYSA-N) an iodine containing thyroid compound characterized by Kendall in 1918 after processing 2 tons of thyroid glands (Kendall, 1918). The formula was correctly found by MS-FINDER as $C_{15}H_{11}I_4NO_4$. The MS/MS spectrum was information poor as it only contained a very abundant *m/z* 126.9 which is related to iodine and two minor peaks with 2% abundance. No matching MS/MS was found. Of the six isomers in ChemSpider with this formula only three distinct structures are covered. MS-FINDER scored the correct compound first and this structure was submitted as top hit.

Compound **18** was purpurin (BBNQQADTFFCFGB-UHFFFAOYSA-N) another anthraquinone dye which was first isolated by Robiquet and Colin in 1827 (Wisniak, 2013). The discovery of synthetic pathways for dye stuffs lead to a breakdown of the natural dye industry (Travis, 1994) and to a huge uprising of chemical enterprises producing synthetic dyes (Decelles, 1949). The formula $C_{14}H_8O_5$ for purpurin was correctly identified by MS-FINDER and the compound ranked on 10th position.. The compound was confirmed by MS/MS search and the correct structure was submitted as first hit.

Compound **19** was monensin (GAOZTHIDHYLHMS-GDMSFIFLSA-N) an anticoccidial antibiotic discovered in 1967 (Agtarap et al., 1967) with formula $C_{36}H_{62}O_{11}$. The MS/MS spectrum consisted of only one abundant precursor ion and a number of very low abundant fragment ions. It must be argued that without MS/MS reference spectra it is impossible to elucidate such low information tandem mass spectra. Only eleven isomers are covered in natural product databases, with monensin the most prominent compound, thus leading to the correct annotation. MS-FINDER ranked the compound as second hit and an additional MS/MS reference database search confirmed the correct hit which was then submitted as top hit.

4. Discussion

The first and easiest step for compound annotation is to check if the compound can be directly matched in a MS/MS database. If there is no good spectral match, the compound needs to be annotated by first determining the molecular formula. This step is crucial for the correct selection of potential isomer candidates. Isotopic abundances and elemental restriction information as well as MS/MS information should be used in this step. For large molecular weight components multiple molecular formula candidates have to be considered. For higher probability in correct formula identification we recommend to use multiple tools such as the Seven Golden Rules, Sirius, CSI:FingerID and MS-FINDER. A critical step is to correctly determine the experimental mass accuracy and isotopic abundance error. For this

contest the mass accuracies were given by the CASMI organizers but isotopic abundance errors were not. False isotopic abundance errors lead to false assumptions on our side and subsequently to false annotation for compound **6**. We extracted some of the peaks from the raw data to confirm mass accuracies and adducts, but such an approach is quite time consuming.

The second step is to query the molecular formula in natural product databases. Here we can use the large compound databases including CAS SciFinder and PubChem. However in the case of dereplication of natural product databases a more targeted approach towards natural products is recommended hence, including only DNP, UNPD and REAXYS, because they will not return large numbers of isomers. In some cases such as compound **5** that can lead to misses because at the time cymoside was not included in DNP or UNDP or Chemspider. The REAXYS database and the Chemical Structure Lookup Service are convenient to use because they provides spreadsheet downloads of all compounds including their structures. Merging results from multiple database lookups and organizing them according to InChiKeys and their natural product sources is however a very time consuming but required step.

The third step is to utilize the natural product structures from their InChI and SMILES codes and perform in-silico fragmentation or manual fragmentation analysis. We observed that CFM-ID performed well for assignment of low molecular weight compounds; however the calculation of large molecular weight compounds took a long time or did not lead to correct assignments in all cases. CSI:FingerID was not able to process negative mode spectra at the time and compounds with complex rearrangements were ranked high but not always correctly. The online availability and automated step-by-step guide by CSI:FingerID was the fastest and most convenient annotation used for some of the natural products. The desktop software MS-FINDER required manual prepossessing of input data, which can be time consuming and was not fully parallelized which increases computational times. MS-FINDER also scores compounds based on the occurrence in chemical databases. If a compound is found more often it is subsequently ranked higher. In many cases very similar fragmentation scores were calculated, leading to no direct distinction of compound rankings. At the time no objective scoring advantage was observed between MS-Finder, CSI:FingerID and MS-FINDER. Here a manual ensemble approach of combining individual results and manual interpretation was used to assign rank scores. In the future it will be possible to automate many of these workflows.

The fastest and most accurate approach for compound annotation we took was MS/MS database search (see Table 2). For the first ten CASMI candidates such an approach was not possible because no MS/MS spectra were acquired or shared by the community. Whenever an MS/MS hit was retrieved it was scored higher than the in-silico fragmentation results. Freely available tools such as MS-Dial, NIST MS Search and databases like NIST, MassBank or Mona can be utilized if MS/MS spectra are available. We saw two extreme situations where MS/MS annotations and reference spectrum searches failed. Product ion spectra that only contain two or three peaks do not contain sufficient information to potentially differentiate between the many isomers found in structure databases. Here the fact that only few isomer structures are actually known and some are more prominent or

better described in the literature often led to the correct identification. Conversely, when the product ion spectra were very fragmentation rich, annotations also did not automatically lead to higher rankings in distinguishing isomers, because rearrangement reactions are hard to rationalize if a pool of possibly thousands of isomers need to be considered or cross-checked manually.

Most mass spectra from natural product publications are still shared in the very inefficient way of listing very few *m/z* values and abundances in text format or publishing scanned pictures of spectra in PDF format (Kind et al., 2009). In many cases peaks are even removed or not listed, leading to errors and misassignments. But complex tandem mass spectra can have up to 100 product ion peaks. Many publications still only provide the molecular ion mass in electron ionization (EI) or fast-atom-bombardment (FAB) format. However the shift clearly goes to a higher utilization of electrospray (ESI) ionization and such spectra and even chromatograms of complex mixtures could be shared publicly to allow for collaborative investigations using platforms such as GNPS (https://gnps.ucsd.edu) (Wang et al., 2016), MetaboLights (http://www.ebi.ac.uk/metabolights/) (Haug et al., 2012), or the Metabolomics Workbench (www.metabolomicsworkbench.org) (Sud et al., 2015). Electronic data sharing solutions outside traditional paper publications are already in place to keep up with the tremendous flow of information and allow fast access to novel compound information. Here the MassBank database (http://www.massbank.jp/) laid the foundation for mass spectral data sharing (Horai et al., 2010). The public MassBank of North America (MoNA) mass spectral database (http://mona.fiehnlab.ucdavis.edu/) now provides upload functions for all types of mass spectra as long as the compound structure is added.

The winner of the CASMI 2016 Category-1 contest Dejan Nikolic (University of Illinois at Chicago) stated “the ability to rationalize as many fragment ions as possible as well as the overall experience in working with a particular class of compounds” helps to score high in compound identification processes. The winning team identified 15 of the 18 compounds, the second placed team (Team Vaniya at UC Davis) identified 14 of the 18 compounds. The methodology described here identified 13 of the 18 compounds correctly, leading to the third place. Our approach was very time consuming (see Table 2) and required the use of multiple tools and databases. The extraction and validation of raw spectra from the chromatographic runs was another negative time factor. However the existence of raw files was quite important to investigate the correctness of certain data. A modified and more automated approach with the MS-FINDER software, MS/MS database search and isomer database lookup lead to victory in the Category-3 challenge with 159 correctly identified compounds out of 208 unknowns. As a final conclusion one can state that it is now possible even for non-natural product researchers to dereplicate known natural compounds (omitting the stereo information) with the help of high-resolution tandem mass spectral data if the compound is known and contained in a database.

Acknowledgments

We thank ChemAxon for a free research license of the Marvin cheminformatics tools. Funding for T.K. and O.F. was supported by NSF MCB 1139644, NSF MCB 1611846, NIH P20 HL113452 and U24 DK097154.

References

Agtarap A, Chamberlin JW, Pinkerton M, Steinrauf LK. Structure of monensic acid, a new biologically active compound. Journal of the American Chemical Society. 1967; 89:5737–5739. [PubMed: 5622366]

Aiello, A., Fattorusso, E., Menna, M., Taglialatela-Scafati, O. Modern Alkaloids. Wiley-VCH Verlag GmbH & Co. KGaA; 2007. A Typical Class of Marine Alkaloids: Bromopyrroles; p. 271-304.

Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics. 2015; 11:98–110.

Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic acids research. 2014; 42:W94–W99. [PubMed: 24895432]

Angelini LG, Pistelli L, Belloni P, Bertoli A, Panconesi S. Rubia tinctorum a source of natural dyes: agronomic evaluation, quantitative analysis of alizarin and industrial assays. Industrial crops and products. 1997; 6:303–311.

Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics. 2009; 25:218–224. [PubMed: 19015140]

ChemAxon. J Chem Base sofware; v16.4.1; ChemAxon. 2016

Corley DG, Durley RC. Strategies for database dereplication of natural products. Journal of natural products. 1994; 57:1484–1490.

Decelles C. The story of dyes and dyeing. J. Chem. Educ. 1949; 26:583.

Delepine M. Joseph Pelletier and Joseph Caventou. J. Chem. Educ. 1951; 28:454.

Djerassi C. Steroid research at Syntex: “the pill” and cortisone. Steroids. 1992; 57:631–641. [PubMed: 1481227]

Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. Proceedings of the National Academy of Sciences. 2015; 112:12580–12585.

Feuerstein W, v Kostanecki S. Synthese des Flavons. Berichte der deutschen chemischen Gesellschaft. 1898; 31:1757–1762.

Forenza S, Minale L, Riccio R, Fattorusso E. New bromo-pyrrole derivatives from the sponge Agelas oroides. Journal of the Chemical Society D: Chemical Communications. 1971:1129–1130.

Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. PLoS one. 2013; 8:e62839. [PubMed: 23638153]

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic acids research, gks1004. 2012

Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI-the worldwide chemical structure identifier standard. Journal of cheminformatics. 2013; 5:1. [PubMed: 23289532]

Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. Journal of Cheminformatics. 2015; 7

Henry TA. CXXIII.—The constituents of the sandarac resins. Journal of the Chemical Society, Transactions. 1901; 79:1144–1164.

Hepper FN. On the botany of James Bruce's expedition to the source of the Blue Nile 1768–1773. Journal of the Society for the Bibliography of Natural History. 1980; 9:527–537.

Hogg JA. Steroids, the steroid community, and Upjohn in perspective: a profile of innovation. Steroids. 1992; 57:593–616. [PubMed: 1481225]

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K. MassBank: a public repository for sharing mass spectral data for life sciences. Journal of mass spectrometry. 2010; 45:703–714. [PubMed: 20623627]

Kelbe W. Zur Kenntniss der Abiëtinsäure. Berichte der deutschen chemischen Gesellschaft. 1880; 13:888–891.

Kendall EC. The active constituent of the thyroid: Chemical groups that are responsible for its physiologic activity. Journal of the American Medical Association. 1918; 71:871–873.

Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008; 24:2534–2536. [PubMed: 18606607]

Kind T, Fiehn O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC bioinformatics. 2007; 8:1. [PubMed: 17199892]

Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanalytical reviews. 2010; 2:23–60. [PubMed: 21289855]

Kind T, Scholz M, Fiehn O. How large is the metabolome? A critical analysis of data exchange practices in chemistry. PloS one. 2009; 4:e5440. [PubMed: 19415114]

Lémus C, Kritsanida M, Canet A, Genta-Jouve G, Michel S, Deguin B, Grougnet R. Cymoside, a monoterpene indole alkaloid with a hexacyclic fused skeleton from Chimarrhis cymosa. Tetrahedron Letters. 2015; 56:5377–5380.

Narayanan S, Appleton H. Creatinine: a review. Clinical chemistry. 1980; 26:1119–1126. [PubMed: 6156031]

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of cheminformatics. 2011; 3:1. [PubMed: 21214931]

Pence HE, Williams A. ChemSpider: an online chemical information resource. Journal of Chemical Education. 2010; 87:1123–1124.

Perkin AG, Hummel JJ. LXXV.-The colouring principles of Ventilago madraspatana. Journal of the Chemical Society, Transactions. 1894; 65:923–944.

Peterson JR, Mitchison TJ. Small molecules, big impact: a history of chemical inhibitors and the cytoskeleton. Chemistry & biology. 2002; 9:1275–1285. [PubMed: 12498880]

Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC bioinformatics. 2010; 11:1. [PubMed: 20043860]

Sharma G, Burkholder P. Structure of dibromophakellin, a new bromine-containing alkaloid from the marine sponge Phakellia flabellata. Journal of the Chemical Society D: Chemical Communications. 1971:151–152.

Shrestha SL, Casey FX, Hakk H, Smith DJ, Padmanabhan G. Fate and transformation of an estrogen conjugate and its metabolites in agricultural soils. Environmental science & technology. 2012; 46:11047–11053. [PubMed: 22967238]

Sitzmann M, Filippov IV, Nicklaus MC. Internet resources integrating many small-molecule databases. SAR and QSAR in Environmental Research. 2008; 19:1–9. [PubMed: 18311630]

Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic acids research, gkv1042. 2015

Tambor J. Stanislaus von Kostanecki. Berichte der deutschen chemischen Gesellschaft. 1912; 45:1683–1707. [16. April 1860 – 15. November 1910]

Travis AS. Between broken root and artificial alizarin: Textile arts and manufactures of madder. History and Technology, an International Journal. 1994; 12:1–22.

Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Analytical Chemistry. 2016; 88:7946–7958. [PubMed: 27419259]

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crusemann M, Boudreau PD, Esquenazi E, Sandoval-Calderon M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya PCA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V,

Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodriguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Muller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutierrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotech. 2016; 34:828–837.

Wisniak J. Pierre-Jean Robiquet. educación química. 2013; 24:139–149.

Wolfender JL, Ndjoko K, Hostettmann K. The potential of LC-NMR in phytochemical analysis. Phytochemical Analysis. 2001; 12:2–22. [PubMed: 11704957]

Highlights

- Dereplication of natural products utilizing tandem mass spectra
- Combination of database search, formula predication and in-silico fragmentation software
- Manual ranking of potential candidates utilizing software scores

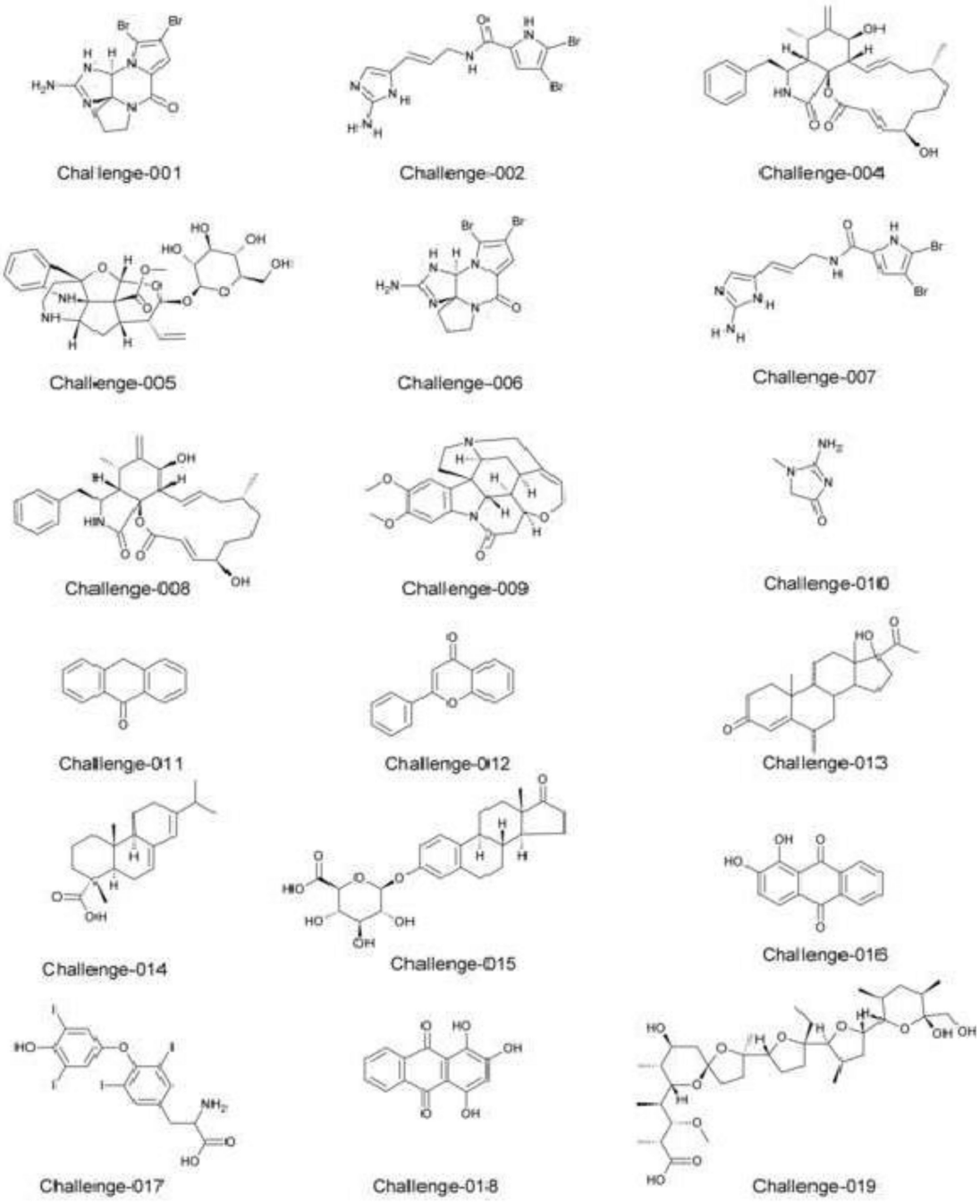


Figure 1.

The CASMI (Critical Assessment of Small Molecule Identification) 2016 contest in category one (natural products) provided the accurate MS and MS/MS spectra of eighteen compounds. Based on the mass spectral information and other metadata correct structural information needed to be assigned.

Table 1

Average rank of the CASMI 2016 Challenge-1 compounds submitted and the related number of isomer structures found for specific molecular formulas in Dictionary of Natural Products (DNP), CSLS, Reaxys database with natural product filter and in ChemSpider.

#	Name	Rank	Formula	DNP	CSLS	Reaxys	ChemSpider
1	Dibromophakellin	5	$C_{11}H_{11}Br_2N_5O$	10	18	16	59
2	Oroidin	1	$C_{11}H_{11}Br_2N_5O$	10	18	16	59
3	Excluded	-	-	-	-	-	-
4	Cytochalasin B	1	$C_{29}H_{37}NO_5$	9	206	39	293
5	Cymoside	-	$C_{27}H_{34}N_2O_{10}$	5	92	16	22
6	Dibromophakellin	-	$C_{11}H_{11}Br_2N_5O$	10	18	16	59
7	Oroidin	3	$C_{11}H_{11}Br_2N_5O$	10	18	16	59
8	Cytochalasin B	2	$C_{29}H_{37}NO_5$	9	206	39	293
9	Brucine	1	$C_{23}H_{26}N_2O_4$	22	250	22	6207
10	Creatinine	1	$C_4H_7N_3O$	1	250	2	178
11	Anthrone	1	$C_{14}H_{10}O$	3	250	2	88
12	Flavone	1	$C_{15}H_{10}O_2$	8	250	11	114
13	Medroxyprogesterone	1	$C_{22}H_{32}O_3$	112	250	115	626
14	Abietic acid	292	$C_{20}H_{30}O_2$	669	250	667	1103
15	Estrone-3-(beta-D-glucuronide)	1	$C_{24}H_{30}O_8$	82	250	112	204
16	Alizarin	1	$C_{14}H_8O_4$	14	250	9	73
17	Thyroxine	1	$C_{15}H_{11}I_4NO_4$	4	123	2	6
18	Purpurin	1	$C_{14}H_8O_5$	29	249	14	52
19	Monensin	1	$C_{36}H_{62}O_{11}$	9	58	19	24

Table 2

Time effort for the dereplication of a single natural product using high resolution MS and MS/MS data averaged 5 hours total. Detailed tasks are listed below with MS/MS search being the fastest and most accurate option.

Task	Tool	Time for one compound
Data downloading, formatting conversion, literature lookup	ProteoWizard, MZMine, MS-FINDER	1h
Molecular formula determination and validation	Seven Golden Rules and Sirius2	30 min
Quick Investigation of MS/MS	NIST MS Search	20 min
Manual MS/MS investigation including neutral losses, fragment analysis	NIST MS Search and MS-FINDER	2 hours
In-silico MS/MS generation	CFM-ID	30 min
Molecular formula calculation and isomer ranking	MS-FINDER and CSI:FingerID	2 min
MS/MS database search	NIST MS Search	1 min