Phytochemistry Letters xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Phytochemistry Letters

journal homepage: www.elsevier.com/locate/phytol

# Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest

Arpana Vaniya[a], Stephanie N. Samra[a], Mine Palazoglu[a], Hiroshi Tsugawa[b], Oliver Fiehn[a,c,*]

[a] University of California Davis, West Coast Metabolomics Center, Genome Center, 451 Health Sciences Drive, Davis, CA 95616, USA
[b] RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan
[c] King Abdulaziz University, Biochemistry Department, Jeddah, Saudi Arabia

ARTICLE INFO

*Article history:*
Received 29 September 2016
Received in revised form 11 November 2016
Accepted 6 December 2016
Available online xxx

*Keywords:*
CASMI
Compound identification
MS-FINDER
Mass spectrometry
Tandem mass spectrometry
Natural products

ABSTRACT

In its fourth year, the CASMI 2016 contest was organized to evaluate current chemical structure identification strategies for 19 natural products using high-resolution LC–MS and LC–MS/MS challenge datasets using automated methods with or without the combination of other tools. These natural products originate from plants, fungi, marine sponges, algae, or micro-algae. Every compound annotation workflow must start with determination of elemental compositions. Of these 19 challenges, one was excluded by the organizers after submission. For the remaining 18 challenges, three software programs were used. MS-FINDER version 1.62 was able to correctly identify 89% of the molecular formulas using an internal database that comprised of 13 metabolomics repositories with 45,181 formulas. SIRIUS correctly identified 61% compositions using PubChem formulas and Seven Golden Rules correctly identified 83% by using the Dictionary of Natural Products as a targeted database. Next, we performed structural dereplication for which we used the consensus formula from the three software programs. We submitted two solution sets for these challenges. In the first solution set, avaniya001, we only used the internal MS-FINDER functions for predicting and ranking structures, correctly identifying 53% of the structures as top-hit, 72% within the top-3 structures, and 78% within the top-10 hits. For our second set, avaniya002, we used both MS-FINDER predictions as well as MS/MS queries against the commercial NIST 14, METLIN, and the public MassBank of North America libraries. Here we correctly identified 78% of the structures as top-hit and 83% within the top-3 hits. Three challenge spectra remained unidentified in either of our submissions within the top-10 hits.



**Contents**



## 1. Introduction

Identification and structural elucidation of unknown compounds, including small molecules and natural products, is a major bottleneck in untargeted metabolomics (Dunn et al., 2013). This is mostly due to the vast structural diversity of natural products. Without structural identifications, statistical findings in

* Corresponding author at: University of California Davis, West Coast Metabolomics Center, Genome Center, 451 Health Sciences Drive, Davis, CA 95616, USA.
*E-mail address:* ofiehn@ucdavis.edu (O. Fiehn).

http://dx.doi.org/10.1016/j.phytol.2016.12.008
1874-3900/

metabolomics studies cannot be interpreted in a biologically meaningful way. Mass spectrometry (MS) is the most widely used analytical technique for the analysis of small molecules, including natural products. Other analytical techniques, such as nuclear magnetic resonance (NMR), can also be used for more complete structural elucidation and connectivity data, but due to its lack of sensitivity MS is the dominant technique (Bjerrum, 2015). Precursor scans (MS1) provide data such as mass-to-charge ($m/z$) ratio of ions, while structural information must be obtained from fragmentation data using tandem mass spectrometry (MS/MS) or multi-stage mass analysis ($MS^n$). While older software like Seven Golden Rules (Kind and Fiehn, 2007) only relied on MS1 information to calculate molecular formulas, modern programs like MS-FINDER (Tsugawa et al., 2016) and SIRIUS (Böcker et al., 2009) implement both MS and MS/MS spectra for this task.

However, even with these advancements, it is impossible to yield correct identifications without using an underlying structure database. Enumeration of all chemically possible structures by brute force methods yield an enormous number of combinations even for a moderately sized elemental composition (Kind and Fiehn, 2006). After selecting a structure database, existing programs can predict spectra or generate fragments by *in silico* methods from structure collections and interpret, match, or rank, these structures against the experimental spectra. Programs use different approaches such as combinatorial or rule-based methods, but also include machine learning for the prediction of fragmentation spectra or fingerprints of the unknown compounds (Hufsky et al., 2014). Yet, the easiest way of matching spectra to structures is to simply match MS/MS spectra against the collection of publicly or licensed mass spectral reference libraries. Unfortunately, the chemical coverage of such libraries is very low in comparison to the number of known chemicals (Stein, 2012). Evaluations of these *in silico* tools and classic MS library searching require a benchmark dataset. Since 2012, the Critical Assessment of Small Molecule Identification (CASMI) contest has provided such dataset. In its fourth year, the CASMI 2016 contest included three categories. Category 1 included identifying the molecular structure of 19 natural products. Category 2 and 3 contained 208 challenges. Category 2 was restricted to using only *in silico* fragmentation software. Whereas Category 3, allowed the use of additional resources, such as databases or mass spectral libraries with any *in silico* software.

Here, we describe our method for determining the correct molecular structure, primarily carried out with MS-FINDER; a software for structure elucidation using MS and MS/MS spectra of unknown compounds. Other tools such as MetFrag were only used to yield candidate structures when neither MS-FINDER nor mass spectral matching yielded any hits. Searching experimental spectra against mass spectral reference libraries (MS library search) was also carried out for structural dereplication. We submitted two different solution sets in order to evaluate the accuracy of using MS-FINDER alone or in combination with MS library searching. Multiple candidates were submitted for 19 challenges and were ranked accordingly, while one candidate was later removed by the organizers of CASMI 2016.

## 2. Materials and methods

For CASMI 2016, data for the 19 challenges were acquired on three different instruments. Challenges 001 – 004 were collected on an Agilent 6540 quadrupole time-of-flight (Q-TOF) with less than 5 parts per million (ppm) mass accuracy, Challenges 005 – 009 were collected on the Waters Synapt G2i Q-Tof with less than 10 ppm mass accuracy, and Challenges 010 – 019 were acquired on Thermo Scientific Q Exactive Plus Orbitrap with less than 5 ppm mass accuracy. These challenges were natural products that originated from plants, fungi, marine sponges, algae, or micro-algae. The data for Challenges 010 – 019 were from the dataset for Category 2 and 3. This set not only included natural products, but included data for endogenous metabolites and one synthetic hormone. For each challenge, raw data files and peak lists with $m/z$ values and relative abundances were provided for both MS and MS/MS spectra. Metadata for each challenge included the retention time, the type of molecular species (i.e. $[M+H]^+$ or $[M-H]^-$), and the $m/z$ value of the precursor ion for the MS/MS spectrum.

Molecular formulas were determined with MS-FINDER version 1.62, SIRIUS version 3.1, and Seven Golden Rules. For MS-FINDER, text formats for both MS and MS/MS spectra were used and the following data; precursor $m/z$, ion mode, mass accuracy of instrument, and precursor type, were provided in the metadata of each challenge. Default parameter settings were used for both formula and structure finder functions (Lewis/Senior rules checked, element probability checked, element ratio check at common range (99.7%), besides C and H which are always considered in the formula generation process other elements that were selected were O, N, S, P, F, Cl, Br, I, and the maximum report number was set to 100 for both functions, tree depth set to 2, relative abundance cut off set to 1, selected *Never use it* for PubChem Online setting, and all 13 local databases were selected). While the current MS-FINDER software contains a total of 14 databases, the version used in for CASMI 2016 contained 13 databases not including the STOFF repository (http://risk-ident.hswt.de/pages/de/links.php). The only parameters different to the default values for the formula finder function were that of isotopic ratio tolerance and mass tolerance which were adjusted to a combination of either 3% and 5 ppm or 5% and 10 ppm. To avoid long computational run times a batch job was submitted to process both formula and structure finder calculations on the top 500 candidates. Results from MS-FINDER were exported as separate text files and formula candidates with hits in any of the 13 local structure databases were further investigated. First, formula candidates were ranked from highest to lowest then subsequent structure candidates belonging to each formula were then ranked from highest to lowest.

For formula determination by the Seven Golden Rules algorithm, only the MS1 spectrum was used. The $m/z$ value for the $[M+H]^+$ or $[M-H]^-$ ion was used to calculate the neutralized accurate mass. The isotopic abundance was also extracted from the MS1 spectrum. For some challenges, additional $m/z$ values and isotope abundances were extracted from the raw data files using SeeMS graphical user interface found in ProteoWizard version 3.0 (Kessner et al., 2008). Like MS-FINDER, isotopic ratio tolerance and mass tolerance were used in a combination of either 3% and 5 ppm or 5% and 10 ppm. For element searches C, H, N, O, P, S, F, Cl, and Br were selected and the Dictionary Natural Products (DNP) (http://dnp.chemnetbase.com/) was used as a targeted database. For SIRIUS version 3.1, the text files of both MS and MS/MS spectra were used. To compute the molecular formulas, the parent mass, ionization, instrument type, and mass accuracy were retrieved from the metadata. C, H, N, O, P, S, Br, Cl, and F were selected for element searches, and the number of candidates was set to 10. Top ranking formulas from both tools were compared to MS-FINDER.

In MS-FINDER, formulas and structures are queried together, using parameters as given above. When MS-FINDER did not yield viable structures or when results were ambiguous, then MetFrag web interface was used for alternative solutions. For MetFrag, the MS/MS spectrum was used and the parent ion, mode, and charge were selected based on the metadata. The default parameters were used for the following; “search PPM” was set to 10, “Limit # of structures” was set to 100, “Mzabs” was set to 0.01, and “Mzppm” was set to 10. The neutral exact mass was automatically calculated by MetFrag. The only parameters different from the MetFrag

Please cite this article in press as: A. Vaniya, et al., Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest, Phytochem. Lett. (2016), http://dx.doi.org/10.1016/j.phytol.2016.12.008

default settings was the selection of PubChem as the target database and the deselection of "Only biological compounds". Results from MetFrag were downloaded as an Excel file.

All challenges were also searched against multiple mass spectral libraries to find the best spectral match. MS/MS spectra was converted to NIST Mass Search format (MSP) file and searched against three public libraries that are now combined in MassBank of North America (http://mona.fiehnlab.ucdavis.edu/), specifically MassBank (Horai et al., 2010), ReSpect (Sawada et al., 2012), LipidBlast (Kind et al., 2013), and two licensed libraries, METLIN (Smith et al., 2005) and NIST 14 (http://www.sisweb.com/software/ms/nist.htm). We used the NIST MS Search 2.0 tool (http://chemdata.nist.gov/) to perform the matches between library and experimental spectra. Candidate hits with a reverse dot product score (Rev-Dot) of 500 and above were confirmed by manually examining the match of query MS/MS spectrum to library reference MS/MS spectrum. Candidate hits were ranked from highest to lowest Rev-Dot; the score in NIST ranges from 0 to 1000 where a score of 1000 is a perfect match. Two submissions, *avaniya001* and *avaniya002* were submitted to assess the influence of different methods. In *avaniya001* only candidates from MS-FINDER and MetFrag were included and in *avaniya002* candidates from MS library searching were combined to the list from *avaniya001*. For each challenge in the each solution set, duplicates were removed and SMILES for multiple candidates with scores were ranked from highest to lowest and reported as a text file for submission. Scores from the different software were not normalized.

After solutions were released by the CASMI 2016 organizers further analysis was done for Challenges 009, 016, 018, and 019 to understand why lower rankings occurred for correct candidates. In order to determine the similarity for different candidates the Tanimoto similarity scores were calculated with ChemMine Tools (Backman et al., 2011) and PubChem BioAssay Tools (Wang et al., 2009). PubChem Tanimoto similarity scores were calculated to compare to the scores calculated by ChemMine Tools. The web interface was used for both ChemMine Tools and PubChem BioAssay Tools. For ChemMine Tools compounds were added by uploading SMILES of each candidate from different challenges. ChemMine Tools was used to calculate the following; atom pair (AP) Tanimoto similarity scores, maximum common substructure (MCS) Tanimoto similarity scores, binning clusters and multidimensional scaling (MDS) clusters. For Challenge 009, the similarities of top-10 candidates in *avaniya001* were determined using binning and MDS clustering methods because the number of compounds being compared was greater than 3. For the binning cluster and MDS calculations, the similarity cutoff of 0.6 was used and 2D was selected for the MDS dimensions. Results were downloaded as either comma separated values (CSV) file or text file. For Challenge 016 and 018, AP and MCS Tanimoto similarity scores were calculated for two candidates with the same score in *avaniya001*. Using the Similarity Workbench, each compound was selected and scores were automatically calculated. For Challenge 019, AP and MCS Tanimoto similarity scores were calculated for three candidates with the same score in *avaniya001*. Again using the Similarity Workbench the scores were calculated for each compound pair, since only two compounds can be submitted at a time for calculations. For Challenge 016, 018, and 019; PubChem Tanimoto similarity scores were also calculated using the PubChem BioAssay Tools. PubChem Compound Identifier (CID) was used calculate a 2D Tanimoto similarity tree. The data matrix containing the Tanimoto similarity scores was downloaded as a CSV file.

## 3. Results

Correct solutions and structures for CASMI 2016 challenges are shown in Supplementary data, page 1 (Fig. S1). In Category 1, six out of nineteen challenges were actually the same compound measured on different instruments. Challenges 001, 002, and 004 were all measured on an Agilent 6540 Q-TOF with less than 5 ppm mass accuracy. Challenges 006, 007, and 008 were all measured on a Waters Synapt G2i Q-Tof with less than 10 ppm mass accuracy. Challenges 010 through 019 were all measured on a Q Exactive Plus Orbitrap with less than 5 ppm mass accuracy. Challenge 003 was excluded from the contest due to the selection of the incorrect precursor ion for MS/MS data acquisition. MS-FINDER correctly identified 89% of the molecular formulas, SIRIUS correctly identified 61%, and Seven Golden Rules identified 83% by using a targeted database, DNP. Without using MS/MS spectral searches, MS-FINDER correctly identified 53% of the structures as top-hit, 72% within the top-3 structures, and 78% within the top-10 hits.

**Table 1**
Molecular formula results for Seven Golden Rules, SIRIUS 3.1, and MS-FINDER. Scores and ranks are given for each challenge. Additional boosted ranks are given for results from Seven Golden Rules using a targeted database, DNP. The formula for Challenge 006 was not determined by any software. Seven Golden Rules was also unable to calculate the correct formula for Challenge 017. SIRIUS 3.1 was unable to calculate the correct formula for Challenges 005, 007, 008, and 017. MS-FINDER was able to correctly determine the molecular formula for 16 out of 18 challenges.

| | | Seven Golden Rules | | | SIRIUS 3.1 | | MS-FINDER | |
|---|---|---|---|---|---|---|---|---|
| Challenge | Molecular Formula | Score | Overall Rank | DNP Rank | Score | Rank | Score | Rank |
| 001 | C11H11Br2N5O | 97.17 | 2 | **1** | −0.96 | 7 | 1.41 | 60 |
| 002 | C11H11Br2N5O | 96.57 | 50 | **1** | 0.12 | **1** | 1.71 | **1** |
| 003 | *Excluded* | – | – | – | – | – | – | – |
| 004 | C29H37NO5 | 97.66 | **1** | **1** | 118.12 | **1** | 3.06 | **1** |
| 005 | C27H34N2O10 | 94.92 | 115 | **1** | n.a. | n.a. | 3.41 | **1** |
| 006 | C11H11Br2N5O | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. |
| 007 | C11H11Br2N5O | 97.93 | 33 | **1** | n.a. | n.a. | 2.02 | **1** |
| 008 | C29H37NO5 | 93.04 | 56 | **1** | n.a. | n.a. | 3.25 | **1** |
| 009 | C23H26N2O4 | 96.07 | 6 | **1** | 14.42 | 2 | 3.56 | **1** |
| 010 | C4H7N3O | −28.17 | **1** | **1** | 12.73 | **1** | 3.02 | **1** |
| 011 | C14H10O | 88.21 | **1** | **1** | 46.93 | **1** | 3.13 | **1** |
| 012 | C15H10O2 | 98.15 | **1** | **1** | 43.22 | **1** | 2.60 | **1** |
| 013 | C22H32O3 | 98.69 | **1** | **1** | 214.78 | **1** | 3.60 | **1** |
| 014 | C20H30O2 | 82.54 | **1** | **1** | 121.66 | **1** | 3.55 | **1** |
| 015 | C24H30O8 | 98.69 | **1** | **1** | 14.39 | **1** | 3.94 | **1** |
| 016 | C14H8O4 | 82.78 | 12 | 2 | 26.44 | **1** | 3.29 | **1** |
| 017 | C15H11I4NO4 | n.a. | n.a. | n.a. | n.a. | n.a. | 3.32 | **1** |
| 018 | C14H8O5 | 82.30 | 16 | **1** | 15.29 | **1** | 3.20 | **1** |
| 019 | C36H62O11 | 99.10 | **1** | **1** | 15.42 | **1** | 3.64 | **1** |

Please cite this article in press as: A. Vaniya, et al., Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest, Phytochem. Lett. (2016), http://dx.doi.org/10.1016/j.phytol.2016.12.008

These results were submitted as set *avaniya001*. Including MS/MS searches, we correctly identified 78% of the structures as top-hit and 83% within the top-3 hits in the solution set *avaniya002*.

**Challenge 001 and Challenge 006** were both dibromophakellin, an alkaloid found in marine sponge *Phakellia flabellata* (Sharma and Burkholder, 1971). For Challenge 001, only Seven Golden Rules reported the correct molecular formula as a top-hit, whereas SIRIUS and MS-FINDER ranked the formulas at #7 and #60, respectively (Table 1). Nevertheless, the structure itself was ranked #1 by MS-FINDER in both solutions sets (Tables 2 and 3). MetFrag was used for Challenge 001 to find additional candidates due to ambiguous results in comparison to Challenge 006. For Challenge 006, the correct formula was not found by any of the three tools. MS-FINDER was unable to find any candidate structures for this challenge. In this unique case, MetFrag was used to find additional candidates. Solutions submitted by MetFrag for both challenges were incorrect. Challenge 006 could not be solved due to the poor mass accuracy of the instrument used to acquire the MS/MS spectrum. The mass error for the neutral mass calculated from the $m/z$ value of the $[M+H]^+$ ion was 14.75 ppm compared to 3.38 ppm for Challenge 001.

**Challenge 002 and 007** were both oroidin, an alkaloid found in marine sponge *Agelas* sp. (Zidar et al., 2014). For this challenge, the correct molecular formula was found as #1 using the Seven Golden Rules algorithm (Table 1) by querying DNP. SIRIUS and MS-FINDER ranked the correct formula at #1 (Table 1). In both sets, MS-FINDER ranked the correct structure at #1 (Tables 2 and 3). For Challenge 007, the correct molecular formula was again ranked #1 using Seven Golden Rules by querying DNP (Table 1). SIRIUS was unable to find the correct molecular formula (Table 1). MS-FINDER was also able to rank the correct formula and structure at #1 (Tables 1–3). MetFrag was used for both challenges to find additional candidates due to ambiguous results in comparison to each other.

**Challenge 004 and 008** were both cytochalasin B, a fungal metabolite produced by *Helminthosporium dematioideum* (Prescott et al., 1972). For Challenge 004, the correct molecular formula (Table 1) was confirmed as the top ranking candidate with all three software programs. The correct structure was ranked at #1 using MS-FINDER in both sets (Tables 2 and 3). For Challenge 008, Seven Golden Rules ranked the correct formula as #1 using DNP query. SIRIUS was unable to determine the correct formula. The correct formula and structure were both ranked #1 by MS-FINDER (Table 1–3). MetFrag was used for both challenges to find additional candidates due to ambiguous results in comparison to each other.

**Table 2**
Results from *avaniya001* which used the *in silico* tools MS-FINDER and MetFrag. The range of isomers or number of candidates ranged from 5 to 387. Results from MetFrag were all incorrect when used to find alternative candidates. MS-FINDER did not find the correct solutions for Challenge 005 and 006. Challenge 016 and 018 had ranks of #1.5 and #2.5 because two candidates were scored the same in MS-FINDER.

| Challenge | Rank | Number of Candidates | Score | Source |
|---|---|---|---|---|
| 001 | **1** | 43 | 2.14 | MS-FINDER |
| 002 | **1** | 44 | 2.76 | MS-FINDER |
| 004 | **1** | 299 | 4.32 | MS-FINDER |
| 005 | – | 5 | – | MS-FINDER |
| 006 | – | 35 | – | MS-FINDER/MetFrag |
| 007 | **1** | 46 | 4.30 | MS-FINDER |
| 008 | **1** | 285 | 5.45 | MS-FINDER |
| 009 | 9 | 189 | 5.51 | MS-FINDER |
| 010 | **1** | 2 | 4.42 | MS-FINDER |
| 011 | **1** | 13 | 4.99 | MS-FINDER |
| 012 | 2 | 24 | 4.43 | MS-FINDER |
| 013 | 40 | 108 | 5.36 | MS-FINDER |
| 014 | 67 | 387 | 5.51 | MS-FINDER |
| 015 | **1** | 83 | 7.70 | MS-FINDER |
| 016 | 2.5 | 144 | 7.98 | MS-FINDER |
| 017 | **1** | 17 | 5.17 | MS-FINDER |
| 018 | 1.5 | 56 | 7.81 | MS-FINDER |
| 019 | 3 | 20 | 8.21 | MS-FINDER |

**Table 3**
Results from *avaniya002* which used the combination of *in silico* tools and MS library searching. The range of isomers or number of candidates ranged from 5 to 389. This strategy improved and boosted the ranks of many challenges compared to the ranks in *avaniya001*. No correct candidates were found for Challenges 005 and 006.

| Challenge | Rank | Number of Candidates | Score | Source |
|---|---|---|---|---|
| 001 | **1** | 43 | 2.14 | MS-FINDER |
| 002 | **1** | 44 | 2.76 | MS-FINDER |
| 004 | **1** | 299 | 4.32 | MS-FINDER |
| 005 | – | 5 | – | MS-FINDER |
| 006 | – | 35 | – | MS-FINDER/MetFrag |
| 007 | **1** | 46 | 4.30 | MS-FINDER |
| 008 | **1** | 285 | 5.45 | MS-FINDER |
| 009 | **1** | 189 | 848 | MS Library Search |
| 010 | **1** | 2 | 999 | MS Library Search |
| 011 | **1** | 13 | 4.99 | MS-FINDER |
| 012 | **1** | 24 | 999 | MS Library Search |
| 013 | **1** | 108 | 934 | MS Library Search |
| 014 | 68 | 389 | 5.51 | MS-FINDER |
| 015 | **1** | 87 | 999 | MS Library Search |
| 016 | 2 | 144 | 966 | MS Library Search |
| 017 | **1** | 17 | 5.17 | MS-FINDER |
| 018 | **1** | 56 | 936 | MS Library Search |
| 019 | **1** | 17 | 683 | MS Library Search |

**Challenge 005** was cymoside, a hexacyclic monoterpene indole alkaloid which is found in the leaves of *Chimarrhis cymosa* (Lémus et al., 2015). Challenge 005 was measured on a Waters Synapt G2i Q-Tof with less than 10 ppm mass accuracy. The correct formula, $C_{27}H_{34}N_2O_{10}$ was ranked at #1 by MS-FINDER and the Seven Golden Rules algorithm using DNP query. However, SIRIUS was unable to calculate the correct formula (Table 1). Five candidate structures were reported by MS-FINDER for the correct formula. The top candidate was found to be 3β-isodihydrocadambine (FCECVXQMCZMWDG-QLIJHQAKSA-N) with a score of 5.28. However, these solutions including the other candidates submitted from MS-FINDER were incorrect due to poor mass accuracy of the data (Tables 2 and 3).

**Challenge 009** was brucine, an alkaloid found in the seed and bark of *Strychnos nux-vomica* (Frédérich et al., 2003). Brucine was measured on a Waters Synapt G2i Q-Tof with less than 10 ppm mass accuracy. Despite this poor mass accuracy, the molecular formula for this challenge had ranked #1 using the Seven Golden Rules algorithm with DNP query, a rank of #2 using SIRIUS, and rank of #1 using MS-FINDER (Table 1). The correct molecular structure when only using MS-FINDER was ranked at #9 in *avaniya001* (Table 2), but the rank was boosted to #1 in *avaniya002* when combining MS-FINDER with MS library searching with a hit found in NIST 14 library. The Rev-Dot score for the hit was 848 (Table 3).

**Challenge 010** was creatinine, an imidazolinone found in the muscle as a byproduct of creatine (Allen, 2012). The correct molecular formula for this challenge was ranked at #1 by all three software solutions (Table 1). The correct structure was ranked at #1 in both submissions (Tables 2 and 3). Creatinine had a score of 4.42 in MS-FINDER and a hit found in MassBank had a Rev-Dot score of 999.

**Challenge 011** was anthrone, an anthraquinoid reduced from the natural product anthraquinone (Yen et al., 2000). The correct molecular formula for this challenge was ranked at #1 by all three software programs and in both solutions sets the correct structure was ranked at #1 using MS-FINDER (Tables 1–3).

**Challenge 012** was flavone, a natural product belonging to a class of compounds known as flavonoids which is found in most plants. It has been known to be isolated from the entire plant of

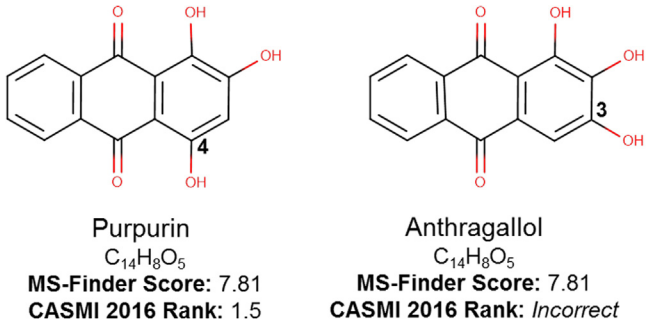


**Fig. 1.** Top ranking candidates in Challenge 018 from MS-FINDER with the same score. In *avaniya001*, purpurin had a lower rank of #1.5 due to the inability to distinguish between positional isomers.

*Analphalis lacteal* (Wang et al., 2004), the leaves of *Ginkgo biloba* L. (Joyeux et al., 1995), and leaves of *Feijoa sellowiana Berg.* (Ayoub et al., 2009). The correct molecular formula for this challenge was ranked at #1 by all three software (Table 1). The correct molecular structure in *avaniya001* was ranked at #2 (Table 2), but the rank was boosted to #1 in *avaniya002* (Table 3). A hit was found in the licensed METLIN library with a Rev-Dot score of 999.

**Challenge 013** was medroxyprogesterone, a synthetic drug belonging to the class of steroid hormones known as progesterone (Block et al., 1981; Gilloteaux et al., 1997). The correct molecular formula for this challenge was ranked at #1 by all three software (Table 1). The correct molecular structure in *avaniya001* was ranked at #40 (Table 2), but the rank was significantly boosted to #1 in *avaniya002* due to a hit found in NIST 14 with a Rev-Dot score of 934 (Table 3).

**Challenge 014** was abietic acid, a diterpene isolated from the leaves of *Pimenta racemosa var. grissea* and *Pygeum africanum* (Fernandez et al., 2001). The correct molecular formula for this challenge was ranked at #1 by all three software programs, the Seven Golden Rules algorithm, SIRIUS, and MS-FINDER (Table 1). The correct structure found by MS-FINDER in both sets with a score of 5.51 and was ranked #67 in *avaniya001* and #68 in *avaniya002* (Tables 2 and 3). The difference to account for the lowered rank in *avaniya002* was due to the higher number of candidates.

**Challenge 015** was estrone-3-(beta-D-glucronide), is a steroid glucuronide, which is a metabolite of estradiol (Barnard et al., 1989). The correct molecular formula for this challenge was ranked at #1 by all three software programs (Table 1). Both submission sets *avaniya001* and *avaniya002*, ranked the correct structure at #1 (Tables 2 and 3). Estrone-3-(beta-D-glucronide) had a score of 7.70 in MS-FINDER and a Rev-Dot score of 999 with a hit found in NIST 14 library.

**Challenge 016** was alizarin, an anthraquinoid produced in the roots, stem, and leaves from *Rubia cordifolia* (Vankar et al., 2008). The correct molecular formula for this challenge was ranked at #1 by SIRIUS and MS-FINDER only. The Seven Golden Rules algorithm had an overall rank of #12 without database query and a boosted DNP rank of #2 (Table 1). This result showed the importance of using (small) molecular formula target databases even for molecular formula searches. In *avaniya001* the correct molecular structure was ranked at #2.5 (Table 2), with two different candidate structures ranked at #2 with the same score. The rank was slightly improved to #2 in *avaniya002* because there was a hit found in METLIN with a Rev-Dot score of 966 (Table 3).

**Challenge 017** was thyroxine, a thyroid hormone produced by the thyroid gland (Braverman et al., 1970). The correct formula for this challenge was ranked at #1 by MS-FINDER only and was found by neither the Seven Golden Rules algorithm nor SIRIUS (Table 1). Seven Golden Rules was unable to determine the correct molecular formula because iodine is not option as an element that can be used for molecular formula calculation. In both submissions MS-FINDER was able to only find one candidate structure which ultimately was the correct structure ranked at #1 (Tables 2 and 3). MetFrag was used to find additional candidates due to the single result found in MS-FINDER.

**Challenge 018** was purpurin, another anthraquinoid also produced in *Rubia cordifolia* (Vankar et al., 2008). MS-FINDER and SIRIUS ranked the correct formula at #1 and the Seven Golden Rules also ranked the formula #1 with DNP database query (Table 1). In *avaniya001*, the correct molecular structure was ranked at #1.5 (Table 2). This rank was a result of two candidates ranked at #1 with the same score. The rank was boosted to #1 in *avaniya002* with a hit found in METLIN with a Rev-Dot score of 936 (Table 3).

**Challenge 019** was monensin, an ionophore produced by a strain of *Streptomyces cinnamonensis* (Duffield and Bagg, 2000). The correct molecular formula for this challenge was ranked at #1 by all three software programs (Table 1). In *avaniya001*, the correct molecular structure was ranked at #3 (Table 2). This rank was a result of three candidates ranked at #2 with the same score. A hit in NIST 14 library boosted the rank to #1 in *avaniya002* (Table 3).

**Table 4**
Tanimoto similarity scores for purpurin and anthragallol in Challenge 018. ChemMine Tools was used to calculate the atom pair (AP) score of 0.81 and the maximum common substructure (MCS) score of 0.90. PubChem BioAssay Tools was also used to calculate the Tanimoto similarity score of 0.98.

| Approach | Tanimoto Similarity Score |
|---|---|
| Atom Pair | 0.81 |
| Maximum Common Substructure | 0.90 |
| PubChem | 0.98 |

## 4. Discussion

For structural dereplication the first step is determining the molecular formula. Current software programs yield the correct formulas when including isotope ratios (Seven Golden Rules) or MS/MS fragment information (MS-FINDER or SIRIUS). Importantly, using high mass resolving power alone does not always guarantee accurate mass (Henke and Kelleher, 2016) and even high mass accuracy of less than 1 ppm does not always give exactly one candidate molecular formula (Kind and Fiehn, 2006). In general, the use of multiple sources of information is important for determinations of both formulas and structures. We compared several programs for determining molecular formulas against MS-FINDER that was able to correctly identify 89% of all the elemental compositions. Results for CASMI 2016 challenges differed due to the use of diverse algorithms, scoring functions, and the use of target molecular databases. The Seven Golden Rules algorithm was the only software that implements DNP as its target database, thus significantly increasing the number of top ranking formulas from 44% to 83%. Second, the quality of input data can lead to an incorrect identification. Challenges 001 and 006 were the same compound collected on two different instruments (i.e. 5 ppm vs. 10 ppm mass accuracy) (Fig. S1). This inaccuracy in data acquisition impacted the isotopic pattern accuracy. Ultimately, the number and type of halogens was not discernible even by manual inspection of Challenge 006. No tool was able to correctly identify the molecular formula for Challenge 006 (Table 1).

The second step included searching the MS/MS data provided against *in silico* fragmentation software and multiple reference mass spectral libraries. *In silico* fragmentation software programs all have advantages and disadvantages. For instance, machine learning is promising because it can learn patterns to eliminate manual time-consuming analysis for compound identification however it is confined by the molecules used for training and is

Please cite this article in press as: A. Vaniya, et al., Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest, Phytochem. Lett. (2016), http://dx.doi.org/10.1016/j.phytol.2016.12.008

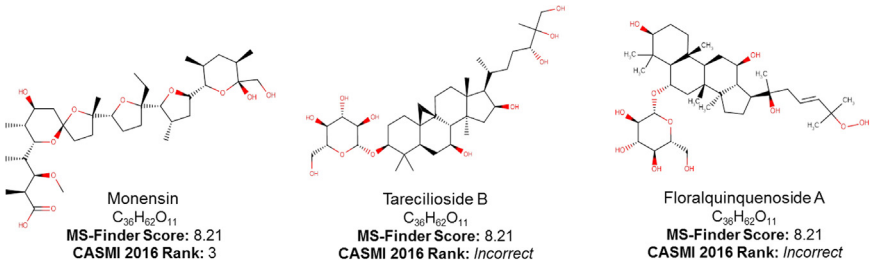


**Fig. 2.** Candidates that had the same score had monensin in Challenge 019. The rank for monensin was lowered to #3 in *avaniya001*. Tarecilioside B and floralquinquenoside A show a higher similarity than each of these candidates compared monensin.

**Table 5**
Tanimoto similarity scores for the three candidates in Challenge 019 that shared the same score of 8.21 in MS-FINDER. Each compound was compared to another to find the AP and PubChem Tanimoto similarity scores. Floralquinquenoside A and tarecilioside B show high similarity because both the AP and PubChem scores are the highest when compared to the other two groups of comparisons.

| | Tanimoto Similarity Score | |
|---|---|---|
| Approach | Atom Pair | PubChem |
| **Compounds** | | |
| Monensin to Tareciloside B | 0.45 | 0.56 |
| Monensin to Floralquinquenoside A | 0.41 | 0.52 |
| Floralquinquenoside A to Tareciloside B | 0.53 | 0.79 |

often inaccurate to predict accurate spectra outside that training realm. Reviewing the results between our two solution sets, *avaniya01* and *avaniya02*, it becomes clear that combining *in silico* software with mass spectral library searching outperforms the use of *in silico* software alone. This was seen through the increase of correctly identified challenges from 53% to 78% when we used a combination approach. This approach of combining *in silico* software with mass spectral libraries approach agrees with Level 2 of the Metabolomics Standards Initiative; where annotation of compounds are confirmed by two different sources of evidence (Fiehn et al., 2007).

The inability to distinguish stereochemical information with current *in silico* tools is a limitation that influences the candidate ranks when using only *in silico* software. Though mass spectrometry alone cannot distinguish isomers, it is sometimes possible to use ion abundances from the MS/MS spectrum to deduce positional isomer information. For *avaniya001*, there were four challenges where candidates had lower ranking due to the impact of isomers. In Challenge 018, there were two candidates ranked at #1 with the same score, lowering the rank of the correct structure to #1.5. The only difference between the two candidates was the position of the hydroxyl group (Fig. 1). Challenge 016, was a similar case where the correct candidate had a lower rank due to positional isomers. The Tanimoto similarity scores determine the similarity of different candidates and the score ranges from 0 to 1, where a score of 1 implies high similarity. Between purpurin and anthragallol in Challenge 018 the Tanimoto similarity score ranged from 0.81 to 0.98 indicating high similarity (Table 4). In Challenge 019, there were three candidates from MS-FINDER ranked at #2 with the same score which lowered the rank of monensin to #3 (Fig. 2). The Tanimoto similarity scores indicate that there is a higher similarity between floralquinquenoside A to tarecilioside B, the two structures that were incorrect (Table 5). For Challenge 009, the correct structure was ranked #9. Figure 3, shows the similarity of the top-10 candidates in *avaniya001* is a 2-dimensional scatter plot where high similarity is represented by a small distance between two compounds. The clusters grouped the correct candidate structure ranked at #9 (which we reported) together with the incorrect structures ranked at #1 and #8. Candidate structures with MS-FINDER ranks #4 and #7 were also grouped together, but candidates with MS-FINDER ranks #2, #3, #5, #6, and #10 did not group with any other structures (Fig. 3). It remains unclear why MS-FINDER ranked structure #9 (the correct candidate) so far lower than structure ranked at #1, despite the high chemical similarity.

The method and approach described here identified 14 out of 18 challenges correctly in this year's CASMI 2016 Category 1 contest. The winner, Dejan Nikolic from the University of Illinois correctly identified 15 out of 18 challenges, using a manual approach. With

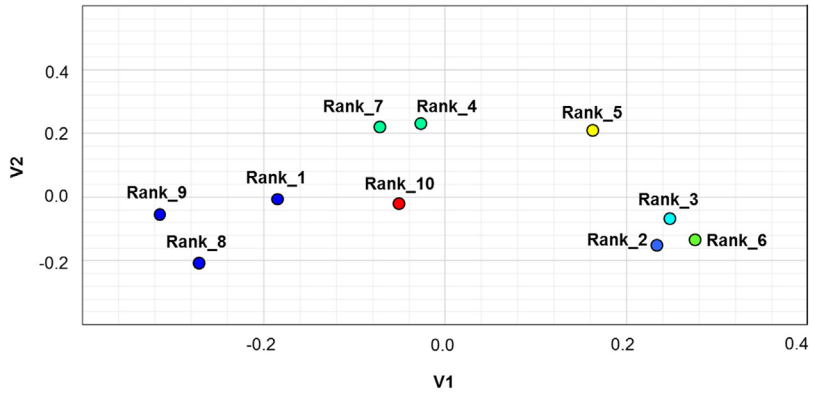


**Fig. 3.** Multidimensional scaling cluster of the top-10 candidates in Challenge 009 from solution set *avaniya001*. High similarity is determined by distance, the smaller the distance the higher the similarity between two compounds or a group of compounds. V1 and V2 are unit-less coordinates that represents the similarity of each compound in the first (V1) and second (V2) dimensions. The correct structure for this challenge was ranked #9. Candidates ranked at #1, #8, and #9 where in one group and candidates ranked at #4 and #7 were in another. All other candidates ranked #2, #3, #5, #6, and #10 did not group with any other candidate.

Please cite this article in press as: A. Vaniya, et al., Using MS-FINDER for identifying 19 natural products in the CASMI 2016 contest, Phytochem. Lett. (2016), http://dx.doi.org/10.1016/j.phytol.2016.12.008

two solution sets submitted, the combination of using *in silico* fragmentation software and MS library searching outperformed using only *in silico* software with more top ranking candidates. As a result, we came in second place with *avaniya002*. It is important to note that to this day, no team or software was able to correctly identify *all* challenges with candidate structures as top-rank, partly due to inaccurate data acquisitions (e.g., Challenge 006). Efforts made for CASMI 2016 has shown that *in silico* software such as MS-FINDER have advanced such that interpretation of MS and MS/MS spectra from unknown compounds can be performed in a manageable time. However, the challenges of unknown compound identification are not yet solved by using only *in silico* programs. The combination of *in silico* fragmentation software and MS library searching for compound identification is important for the accuracy of identifying unknown compounds. The tasks of identifying of "unknown unknowns" still remains an obstacle which has yet to be tested in future CASMI contests.

## Acknowledgements

We appreciate funding this work through National Science Foundation grants MCB 1139644 and MCB 1611846, and National Institutes of Health grant DK097154.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.phytol.2016.12.008.

## References

Allen, P.J., 2012. Creatine metabolism and psychiatric disorders: does creatine supplementation have therapeutic value? Neurosci. Biobehav. Rev. 36, 1442–1462.
Ayoub, N., Hussein, S., Hashim, A., Hegazi, N., Linscheid, M., Harms, M., Wende, K., Lindequist, U., Nawwar, M., 2009. Bone mineralization enhancing activity of a methoxyellagic acid glucoside from a Feijoa sellowiana leaf extract. Die Pharm.-Int. J. Pharm. Sci. 64, 137–141.
Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A., 2009. SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics 25, 218–224.
Backman, T.W., Cao, Y., Girke, T., 2011. ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res. 39, W486–W491.
Barnard, G., Kohen, F., Mikola, H., Lövgren, T., 1989. Measurement of estrone-3-glucuronide in urine by rapid, homogeneous time-resolved fluoroimmunoassay. Clin. Chem. 35, 555–559.
Bjerrum, J.T., 2015. Metabonomics: Analytical Techniques and Associated Chemometrics at a Glance. Springer.
Block, A.J., Wynne, J.W., Boysen, P.G., Lindsey, S., Martin, C., Cantor, B., 1981. Menopause, medroxyprogesterone and breathing during sleep. Am. J. Med. 70, 506–510.
Braverman, L.E., Ingbar, S.H., Sterling, K., 1970. Conversion of thyroxine (T4) to triiodothyronine (T3) in athyreotic human subjects. J. Clin. Invest. 49, 855.
Dictionary of Natural Products, 2016. Dictionary of Natural Products. http://dnp.chemnetbase.com/ (accessed 04/01/2014).
Duffield, T.F., Bagg, R.N., 2000. Use of ionophores in lactating dairy cattle: a review. Can. Vet. J. 41, 388.
Dunn, W.B., Erban, A., Weber, R.J., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9, 44–66.
Fernandez, M., Tornos, M., Garcia, M., Heras, B.d.l., Villar, A., Saenz, M., 2001. Anti-inflammatory activity of abietic acid, a diterpene isolated from Pimenta racemosa var. grissea. J. Pharm. Pharmacol. 53, 867–872.
Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., 2007. The metabolomics standards initiative (MSI). Metabolomics 3, 175–178.
Frédérich, M., Choi, Y.H., Verpoorte, R., 2003. Quantitative analysis of strychnine and brucine in Strychnos nux-vomica using 1H-NMR. Planta Med. 69, 1169–1171.
Gilloteaux, J., Karkare, S., Don, A.Q., Sexton, R.C., 1997. Cholelithiasis induced in the Syrian hamster: evidence for an intramucinous nucleating process and down regulation of cholesterol 7α-hydroxylase (CYP7) gene by medroxyprogesterone. Microsc. Res. Tech. 39, 56–70.
Henke, M.T., Kelleher, N.L., 2016. Modern mass spectrometry for synthetic biology and structure-based discovery of natural products. Nat. Prod. Rep. 33, 942–950.
Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714.
Hufsky, F., Scheubert, K., Böcker, S., 2014. New kids on the block: novel informatics methods for natural product discovery. Nat. Prod. Rep. 31, 807–817.
Joyeux, M., Lobstein, A., Anton, R., Mortier, F., 1995. Comparative antilipoperoxidant, antinecrotic and scavanging properties of terpenes and biflavones from Ginkgo and some flavonoids. Planta Med. 61, 126–129.
Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24, 2534–2536.
Kind, T., Fiehn, O., 2006. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinf. 7, 1.
Kind, T., Fiehn, O., 2007. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinf. 8, 1.
Kind, T., Liu, K.-H., Lee, D.Y., DeFelice, B., Meissen, J.K., Fiehn, O., 2013. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods 10, 755–758.
Lémus, C., Kritsanida, M., Canet, A., Genta-Jouve, G., Michel, S., Deguin, B., Grougnet, R., 2015. Cymoside, a monoterpene indole alkaloid with a hexacyclic fused skeleton from Chimarrhis cymosa. Tetrahedron Lett. 56, 5377–5380.
MoNA – Mass Bank of NorthAmerica, 2016. MoNA MassBank of North America. http://mona.fiehnlab.ucdavis.edu/ (accessed 04 January 2015).
NIST MS Search Program, 2016. NIST MS Search Program. http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:ms-search (accessed 4 January 2015).
Prescott, D., Myerson, D., Wallace, J., 1972. Enucleation of mammalian cells with cytochalasin B. Exp. Cell Res. 71, 480–485.
RISK IDENT, 2016. Risk Ident. http://risk-ident.hswt.de/pages/de/links.php (accessed 09/27 September 2016)..
Sawada, Y., Nakabayashi, R., Yamada, Y., Suzuki, M., Sato, M., Sakata, A., Akiyama, K., Sakurai, T., Matsuda, F., Aoki, T., 2012. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. Phytochemistry 82, 38–45.
Sharma, G., Burkholder, P., 1971. Structure of dibromophakellin, a new bromine-containing alkaloid from the marine sponge Phakellia flabellata. J. Chem. Soc. D 151–152.
Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751.
Stein, S., 2012. Mass spectral reference libraries: an ever-expanding resource for chemical identification. Anal. Chem. 84, 7274–7282.
The NIST 14 Mass Spectral Library & Search Software NIST 2014/EPA/NIH), 2014. The NIST 14 Mass Spectral Library & Search Software (NIST /EPA/NIH). http://www.sisweb.com/software/ms/nist.htm (accessed 04 January 2014).
Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., Saito, K., Fiehn, O., Arita, M., 2016. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Anal. Chem. 88, 7946–7958.
Vankar, P.S., Shanker, R., Mahanta, D., Tiwari, S., 2008. Ecofriendly sonicator dyeing of cotton with Rubia cordifolia Linn using biomordant. Dyes and Pigm. 76, 207–212.
Wang, A.-X., Zhang, Q., Jia, Z.-J., 2004. A new furobenzopyranone and other constituents from Anaphalis lactea. Die Pharm.-Int. J. Pharm. Sci. 59, 807–811.
Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H., 2009. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 37, W623–W633.
Yen, G.-C., Duh, P.-D., Chuang, D.-Y., 2000. Antioxidant activity of anthraquinones and anthrone. Food Chem. 70, 437–441.
Zidar, N., Montalvão, S., Hodnik, Ž., Nawrot, D.A., Žula, A., Ilaš, J., Kikelj, D., Tammela, P., Mašič, L.P., 2014. Antimicrobial activity of the marine alkaloids, clathrodin and oroidin, and their synthetic analogues. Mar. Drugs 12, 940–963.