OMICS A Journal of Integrative Biology
Volume 25, Number 6, 2021

DOI: 10.1089/omi.2021.0051

MS2Compound: A User-Friendly Compound Identification Tool for LC-MS/MS-Based Metabolomics Data

Santosh Kumar Behera,[i] Sandeep Kasaragod,[ii] Gayathree Karthikkeyan,[iii] Chinmaya Narayana Kotimoole,[iv] Rajesh Raju,[v] Thottethodi Subrahmanya Keshava Prasad,[vi] and Yashwanth Subbannayya[vii,*]

Abstract

Metabolomics is a leading frontier of systems science and biomedical innovation. However, metabolite identification in mass spectrometry (MS)-based global metabolomics investigations remains a formidable challenge. Moreover, lack of comprehensive spectral databases hinders accurate identification of compounds in global MS-based metabolomics. Creating experiment-derived metabolite spectral libraries tailored to each experiment is labor-intensive. Therefore, predicted spectral libraries could serve as a better alternative. User-friendly tools are much needed, as the currently available metabolomic analysis tools do not offer adequate provision for users to create or choose context-specific databases. Here, we introduce the MS2Compound, a metabolite identification tool, which can be used to generate a custom database of predicted spectra using the Competitive Fragmentation Modeling-ID (CFM-ID) algorithm, and identify metabolites or compounds from the generated database. The database generator can create databases of the model/context/species used in the metabolomics study. The MS2Compound is also powered with *mS-score*, a scoring function for matching raw fragment spectra to a predicted spectra database. We demonstrated that *mS-score* is robust in par with dot product and hypergeometric score in identifying metabolites using benchmarking datasets. We evaluated and highlight here the unique features of the MS2Compound by a re-analysis of a publicly available metabolomic dataset (MassIVE id: MSV000086784) for a complex traditional drug formulation called *Triphala*. In conclusion, we believe that the omics systems science and biomedical research and innovation community in the field of metabolomics will find the MS2Compound as a user-friendly analysis tool of choice to accelerate future metabolomic analyses.

Keywords: metabolomics, MS2Compound, bioinformatics, systems science, data analysis, computational biology, metabolite identification

Introduction

Metabolomics, the study of metabolites, is one of the leading crucial frontiers of systems science and omics technology innovation. Metabolomics provides novel postgenomic insights in cell biology and deepens our understanding of biological systems' metabolic phenotypes.

Advances in mass spectrometry (MS)-based metabolomics, combined with use of multiple fractionation techniques, have improved the identification of metabolic profile in a sample (Cheng et al., 2018). Global or untargeted metabolomics analysis results in vast amounts of data, which poses challenges during data analysis. The basic steps of metabolomics data analysis comprise identifying features by

Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore, India.
[i]ORCID ID (https://orcid.org/0000-0003-0807-6408).
[ii]ORCID ID (https://orcid.org/0000-0001-9499-3653).
[iii]ORCID ID (https://orcid.org/0000-0002-5965-2957).
[iv]ORCID ID (https://orcid.org/0000-0002-8454-2450).
[v]ORCID ID (https://orcid.org/0000-0003-2319-121X).
[vi]ORCID ID (https://orcid.org/0000-0002-6206-2384).
[vii]ORCID ID (https://orcid.org/0000-0002-3885-3514).
Current affiliation: Centre of Molecular Inflammation Research (CEMIR), Department of Clinical and Molecular Medicine (IKOM), Norwegian University of Science and Technology, Trondheim, Norway.

pre-processing MS-derived raw files, followed by a feature assignment to metabolites (Cambiaghi et al., 2017). The current strategies for metabolomics data analysis suffer from several limitations. For example, different algorithms result in different sets of pre-processed spectra, which increases the number of false-positive peaks (Myers et al., 2017). Moreover, a comparative study (Considine et al., 2017) highlighted the lack of reproducibility of metabolomics data due to limitations in reporting the analysis steps. The lack of standard pipelines and algorithms for pre-processing MS-derived spectra remains a bottleneck for the biomedical community.

The identification of compounds/metabolites from MS/MS spectra or features is a critical step in metabolomics data analysis. This step is achieved by comparing metabolite features with existing MS, and MS/MS spectra reference databases (Blazenovic et al., 2018) such as Human Metabolome Database (HMDB) (Wishart et al., 2018), METLIN (Guijas et al., 2018), and LIPIDMAPS (Sud et al., 2007). Although a large number of metabolites are curated in these repositories, the metabolite information is mostly restricted to those from well-studied species.

A recent metabolomic study identified the compounds of *Mycobacterium tuberculosis* (Mtb) by using an Mtb-specific compound library as the reference database (Collins et al., 2018). However, these Mtb compounds were not assigned in a previous study by the same group using LIPIDMAPS and METLIN (Frediani et al., 2014). In another study, Wang et al. (2019) used an *Astralagus*-specific database from SciFinder with NIST (https://www.nist.gov/srd/nist-standard-reference-database-1a-v17) and METLIN for the identification of metabolites present in *Astragalus mongholicus* and *Astragalus membranaceus*. Several compound databases have been developed for various species such as *E. coli Metabolome Database* (ECMDB) (Sajed et al., 2016), *Livestock Metabolome Database* (LMDB) (Goldansaz et al., 2017), and *Yeast Metabolome Database* (YMDB) (Ramirez-Gaona et al., 2017), which improved identification of the metabolites in the given species.

MZmine2 (Pluskal et al., 2010) and XCMS (Huan et al., 2017; Tautenhahn et al., 2012) are two widely used tools for analyzing metabolomics data. XCMS, which uses METLIN as a reference database for compound identification, provides in-built species-specific metabolites only for a limited number of organisms. In contrast, MZmine2 allows for use of any custom database; however, such a search is restricted to the precursor level. Besides, the HMDB web service (Wishart et al., 2007) can be used to identify the compounds from MS/MS spectra; however, it does not allow the use of a custom database of the user's choice.

Several studies have recently focused on metabolite identification approaches in the absence of reference databases (Allen et al., 2014; Djoumbou-Feunang et al., 2019; Gil-de-la-Fuente et al., 2019; Li et al., 2013). A hybrid search approach has been proposed to find the metabolites without any known spectral information, which combines direct peak matches along with neutral loss peak matches. This approach has shown an increase in the number of identifications (Cooper et al., 2019). MetDNA, a recursive algorithm based on metabolic reaction network (MRN), was developed to annotate metabolites without the back-end spectral library (Shen et al., 2019). This algorithm is based on finding the seed metabolites followed by annotation of metabolites with the reaction-paired neighbor metabolites in a recursive manner.

Further, machine learning-based approaches have been utilized to predict MS/MS fragment masses of metabolites. A random-forest-based model, SubFragment-Matching, can be used to predict MS/MS fragments based on known spectra of a compound with structural similarity (Li et al., 2020). An imputation-based mass-to-charge ratio (m/z) match within different datasets has been implemented to predict the biological role and pathway analysis of unknown features in the metabolomics experiment (Hsu et al., 2019). However, despite such efforts being undertaken toward identification of the metabolites without MS/MS spectral information, there are very few approaches toward the identification of compounds from the custom metabolite databases.

It is not economically viable to create an experimental spectral library for a particular experiment. In such cases, the prediction of MS/MS fragments can serve as a potential alternative. Advancement in the algorithms for the prediction of MS/MS spectra and their high accuracy rate has paved the way for creating custom compound libraries (Blazenovic et al., 2018). These custom databases can be used for the better identification of metabolites from corresponding data. SIRIUS, a tool for identifying metabolites from MS/MS information, has been developed by integrating the CSI: FingerID algorithm (Duhrkop et al., 2019).

In the current study, we used the Competitive Fragmentation Modeling-ID (CFM-ID) (Allen et al., 2014), an algorithm that predicts metabolite fragment masses, to create a custom database and developed a pipeline to analyze global or untargeted metabolomics data using custom databases. The CFM-ID algorithm uses probabilistic generative models for prediction of the MS/MS spectra for a given compound structure.

We also designed a new scoring function for MS/MS matches, which can improve the accuracy of compound identification. The developed scoring function showed a robust performance to the existing scoring strategies for MS/MS spectra match. The pipeline can be accessed with a user-friendly Graphical User Interface (GUI), named the MS2Compound (https://github.com/beherasan/MS2Compound, https://sourceforge.net/projects/ms2compound), without prior dependencies for easy and accurate compound identification. The MS2Compound will allow creating a custom database by using the CFM-ID algorithm and for the identification of compounds from metabolomics datasets.

Materials and Methods

Development of a spectral database for compounds found in different species

Compounds for the selected species (Supplementary Table S1) were downloaded from BioCyc (https://biocyc.org) (Caspi et al., 2016) in flat-file format. Alongside, a compound list containing 101 plant species was downloaded from the PlantCyc database (https://plantcyc.org). Unique IDs from BioCyc, along with their Simplified Molecular-Input Line-Entry System (SMILES), were used for the prediction of fragment ions using CFM-ID (version 2.4; https://sourceforge.net/projects/cfm-id) (Allen et al., 2014). SMILES ID stores the information of a chemical structure in a computer-readable format. This is one of the ways of

representing a chemical structure. The MS/MS fragmentation patterns were predicted in both positive and negative mode, in all three: low (10 V), medium (20 V), and high (40 V) collision energy levels.

Other public resources

The predicted and experimental MS/MS spectra from HMDB (released on 01-09-2019; https://hmdb.ca) (Wishart et al., 2018) were downloaded in Extensible Markup Language (XML) format, in addition to the databases mentioned earlier. Fragment ion information from both experimental and predicted data was extracted for the corresponding primary IDs of HMDB. A list of polyphenols contents in foods was downloaded from Phenol-Explorer Version 3.6 (http://phenol-explorer.eu) (Rothwell et al., 2013). The SMILES ID for all the compounds and extracted metabolites were used to predict the fragments. Similarly, the list of fragments predicted for phytochemical compounds present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) compound database (https://www.genome.jp/kegg-bin/get_htext?br08003.keg) (Kanehisa, 2019; Kanehisa et al., 2019) was added to the existing reference database.

Development of GUI

An interactive user interface was developed in C# with Visual Studio 2019 Community Edition. The query and the arguments were parsed to the command prompt, which calls a Perl or Python script for further analysis. The custom database module allows the user to predict the MS/MS fragments of compounds by using CFM-ID models.

In addition, it also allows the user to calculate the monoisotopic mass of a given SMILES identifier. The SMILES identifier of the compounds was parsed by using Pysmiles (https://github.com/pckroon/pysmiles). PyInstaller (https://pypi.org/project/PyInstaller) was used to convert the Python scripts into portable executables. The identification of compounds from MS and MS/MS data and the corresponding scoring was performed by Perl script. A portable version of Perl (v5.28.0) was provided to make the GUI dependency-free and facilitate easy installation for users with limited knowledge of computer applications. Also, DBD::SQLite (https://github.com/DBD-SQLite/DBD-SQLite) Perl module was included for executing SQLite functions.

MS search

The MS search panel allows the user to insert the query in batch mode. m/z Values were given as queries along with the other search parameters such as tolerance level and probable adduct. Users are allowed to upload tab-separated text files as query, and a custom database if required.

MS/MS search

The workflow for MS/MS search is described in Figure 1. A pre-processed raw file (in Mascot Generic Format [MGF] format) can be used as a query to match against a predicted spectra database for putative compound identifications. The precursor mass of the query was matched with the monoisotopic mass of the compounds at user-defined error tolerances (either in Da or parts per million [ppm]), thereby generating a list of candidate compounds. Fragment ions of the query were matched against the MS/MS fragments of the candidate compounds with user-provided tolerance for fragment match, and a score *mS-score* was calculated for each match.

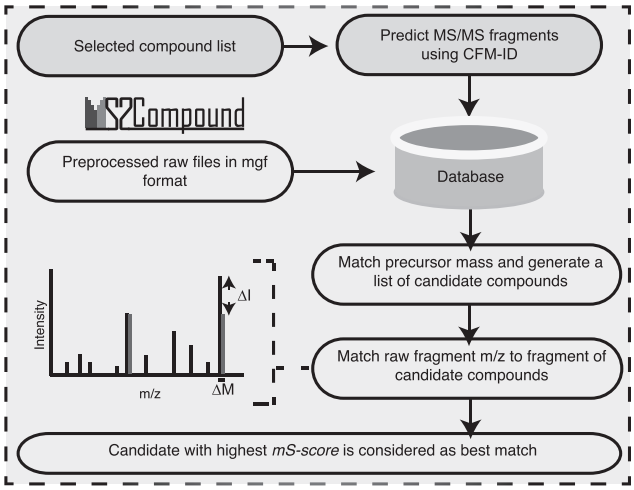

FIG. 1. MS2Compound workflow for the identification of compounds from MS/MS data using a custom metabolite database. The custom reference database was generated by using the CFM-ID MS/MS fragmentation algorithm. Raw spectra were matched at the precursor level to generate a preliminary list of candidate compounds. Fragment-level spectra were matched with generated predicted fragments for all the candidate compounds. A weighted *mS-score* was assigned to each positive match, and the assignments with the highest *mS-score* were considered high confidence matches. CFM-ID, Competitive Fragmentation Modeling-ID; MS, mass spectrometry.

A modified distance function has been proposed as *mS-score*, which considers the difference in matched m/z values and corresponding intensities. As the back-end database used consists of predicted spectra, our scoring function provides more importance to the intensity of raw/experimental m/z values as a normalization factor β. For N number of fragments matched at a defined tolerance level, M^t and M^e are fragment m/z values of theoretical and experimental spectra, respectively. I^t and I^e are fragment intensity values of theoretical and experimental spectra, respectively. The distance function for ith candidate can be calculated as follows:

$$S_i = \frac{\sqrt{\sum_{j=1}^{N} \Delta M_j \times \beta_j \times \Delta I_j}}{N^2}$$

where, $\Delta M_j = \left|M_j^t - M_j^e\right|$ and $\Delta I_j = \left|I_j^t - I_j^e\right|$, and $j \in \{Set\ of\ matched\ fragments\}$. The normalization factor β was calculated by $100/\left(I_j^e\right)$. The *mS-score* from the distance function was calculated as follows:

$$mS_i = \begin{cases} -\ln(S_i + 1), & N = 1 \\ -\ln(S_i), & N > 1 \end{cases}$$

The assigned compounds are ranked based on the decreasing order of *mS-score* for given spectrum. The candidate with the highest *mS-score* is considered as the probable compound for the corresponding spectrum.

Performance of mS-score

The performance of the proposed scoring function (*mS-score*) for the compound identification was compared with existing scoring functions such as dot product, hypergeometric score, and fit score. This comparison was performed by using data from the contest of Critical Assessment of Small Molecule Identification (CASMI; http://casmi-contest.org/2017/index.shtml).

Raw data of challenges (in MGF format) and the corresponding solutions (compounds) were downloaded from CASMI organized during 2017. Molecular weight and SMILES information of the solutions with PubChem CID were extracted from the PubChem database (https://pubchem.ncbi.nlm.nih.gov). Monoisotopic mass of the compounds without PubChem CIDs was calculated manually in ACD/ChemSketch (Freeware, version 2019.2.1; https://www.acdlabs.com/resources/freeware/chemsketch) by using SMILES identifier.

A custom reference database was created by the prediction of MS/MS fragments of ''solution compounds.'' The ''solution compounds'' are the true-positive hits for the given challenge spectra. The fragment prediction was performed by CFM-ID using SMILES information. The fragment prediction was carried out for positive and negative modes at three different energy levels (https://github.com/beherasan/MS2Compound/tree/master/benchmark).

The challenges provided by CASMI were divided into two groups based on the data acquisition mode. Compound identification was performed separately for positive and negative ion modes. The list of adducts used for the identification of candidate compounds is provided in Supplementary Table S2. Raw spectra of challenges were matched to fragments of the corresponding candidate compounds. Both precursors and fragments were matched with a tolerance level of 0.05 Da. Each match is scored with four scoring functions; dot product, hypergeometric score, fit score, and *mS-score*. Dot product and hypergeometric function were implemented, as previously described (Yen et al., 2011); the fit score was taken from the scoring function of MyCompoundID, as previously described (Huan et al., 2015). The details of all the scoring functions are provided in Supplementary Methods in Supplementary Data.

Validation of the tool by re-analysis of public data

To validate the utility of the developed tool, we reanalyzed a public dataset (MassIVE id: MSV000086784) and compared the identification with previously reported results. A global profiling of MS/MS metabolomics data of *Triphala* (Subbannayya et al., 2018) was considered for re-analysis. The raw data (.wiff) were converted to mzML format by using MSConvert (Chambers et al., 2012). These files were further pre-processed in MZmine2 with the parameters provided in Supplementary Table S3. Data acquired in the positive and negative modes were pre-processed separately, and this resulted in two pre-processed raw files.

The MS2Compound was used to identify the compounds for the pre-processed features. Phenol-Explorer and Phytochemical compounds from KEGG were selected as a reference database for the MS/MS search. The features with only MS information were used for the identification of compounds at MS-level with the same resources as the reference database. The precursor match was performed with 0.05 Da of m/z tolerance, and 0.5 Da tolerance was used for fragment match. Data acquired in the positive ion mode were searched with M+H, M+Na, M + 2H, M + 2Na, M + 3H, and M + 3 Na adducts. Similarly, the data acquired in the negative ion mode were searched with M–H, M – H_2O–H, M – 2H, and M – 3H adducts. The final list of compounds was compared with the compounds identified in the previously published results.

Comparison of compounds identified from MS2Compound and MZmine2

Triphala metabolomic data acquired in the positive mode (as described in the previous section) were used to compare the compounds identified from the MS2Compound and MZmine2. The raw files were pre-processed, as described in the previous section. The features were mapped to the KEGG database with 0.05 Da tolerance and [M+H]+ adduct through the ''Online database search'' module in MZmine2. The same MS/MS features were searched against ''KEGG phytochemical compounds'' in the MS2Compound with 0.05 and 0.5 Da tolerances for precursor and fragment level, respectively. The identified compounds were compared by mapping the compound names that resulted from two searches.

Results

The workflow for compound identification from a given metabolomic data is shown in Figure 1. A total of 5183 and 3366 non-redundant compounds were collected from BioCyc (for the 10 selected species and their strains) and PlantCyc database, respectively. Among these, 1081 compounds are found to be shared between these 2 datasets. Among the 10 species that contributed to the dataset, *M. tuberculosis* and *Brucella melitensis* have 353 and 180 unique compounds, respectively, compared with all the other species in the list.

For this back-end dataset, CFM-ID in the MS2Compound predicted fragment ions for 1785 and 2482 compounds in both positive and negative ion modes, respectively (Supplementary Table S1). For another dataset from KEGG phytochemicals, the MS2Compound predicted 2489 and 2756 theoretical fragments in positive mode and negative mode, respectively. From Phenol-Explorer, 738 and 756 compounds were fragmented in positive and negative ion mode, respectively. In total, the current version of the MS2Compound back-end data has 5372 and 5849 predicted fragment spectra in positive and negative ion mode, respectively. A dataset from CASMI 2017 was used to check the performance of *mS-score* in compound identification. In a positively acquired dataset from CASMI, a total of 121, 120, 104, and 81 compounds were assigned correctly with rank one by *mS-score*, fit-score, hypergeometric score, and dot product, respectively.

Interestingly, none of the true-positives shared the same score with false-positive results in the case of *mS-score* and dot product. However, with the fit-score and hypergeometric score, 2 and 13 true assignments shared the same score with false-positive hits, respectively (Fig. 2A). We have plotted the Receiver Operating Characteristic (ROC) curve, which is a graphical representation with two parameters: true-positive and false-positive hits for comparison of all the four scores considered in our study. The ROC curve shows that the *mS-score* can distinguish the true-positives from the false-

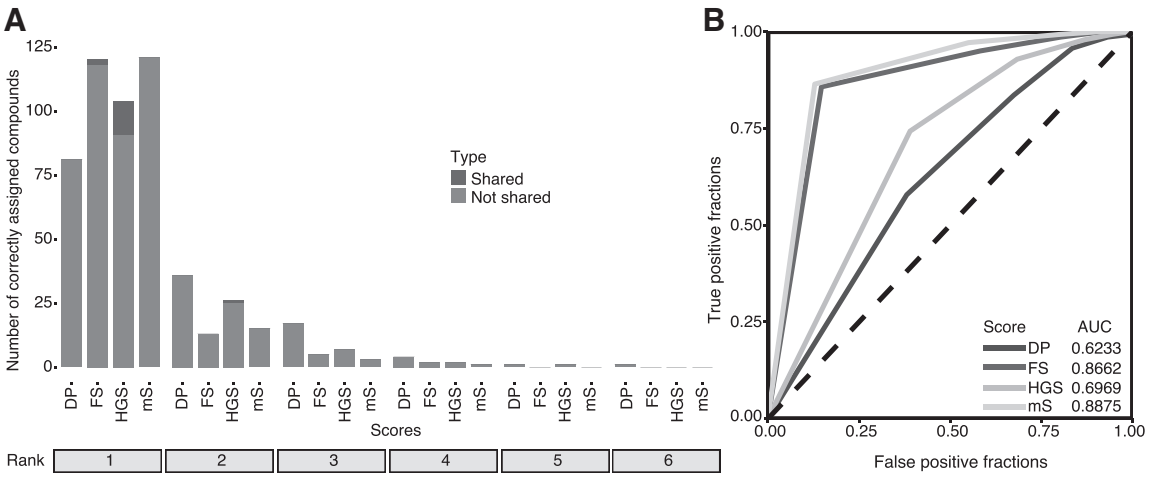

FIG. 2. Comparison of *mS-score* with other scoring functions for compound identification using CASMI 2017 challenge datasets (http://casmi-contest.org/2017/index.shtml). Scores including DP, FS, HGS, and *mS-score* (mS) for compounds identified from positive data acquisition mode have been depicted. **(A)** The number of true-positive hits from all the scoring functions. The *mS-score* was found to assign higher scores to true-positive hits with a higher rank (Rank 1) and provided more unique identifications. **(B)** ROC curve demonstrating the distinction of true-positive hits from false-positive hits in all the scoring functions. CASMI, Critical Assessment of Small Molecule Identification; DP, dot product; FS, fit score; HGS, hypergeometric score; ROC, Receiver Operating Characteristic.

positives with the area under the curve (AUC) of 0.8875, which was higher than dot product and hypergeometric score, and comparable with fit-score (Fig. 2B). However, our score performs similar to that of fit-score as suggested by AUC values; *mS-score* is able to distinguish the false-positive hits from the true-positive hits better than the fit-score.

Contrary to the positive ion mode acquired data, the fit-score assigned a greater number of true-positives in rank one than the *mS-score* in the negative ion mode acquired data. A total of 85 true-positives were assigned correctly by fit-score with the highest score; however, 83 such cases resulted from *mS-score*. Out of 85, in 3 cases, fit-score shared the same score with false-positive hits (Supplementary Fig. S1A). The fit-scores are similar in multiple matches, as this scoring is based on the set of matched raw m/z and intensity to the reference spectra. Sharing the same score with false-positive hits hinders the distinguishing of true-positives. However, the *mS-score* was found to distinguish the false-positives from true-positive hits. This was supported by the AUC of 0.8307 for *mS-score*, which was comparatively better than the other scoring functions (Supplementary Fig. S1B).

We further manually inspected the fragments matched and corresponding *mS-score* and fit-score for such matches. The raw spectrum of challenge005 was matched to the predicted spectrum of solutions of challenge005 and challenge006 (Fig. 3A, B). Challenge005 matched six fragments when mapped to true-positive hits (challenge005), and it matched with five fragments compared with a false-positive hit (challenge006). The match score should be high for a match with a true-positive hit; however, the fit-score gives a better score to the false-positive hit. One of the fragment m/z values is high in the matched spectrum list for Challenge006, which resulted in a high fit-score value. More such examples are shown in Figure 3C–H and Supplementary Figure S2.

In contrast, the *mS-score* is based on the difference between the matched m/z values and the number of matches. In another instance, *mS-score* cannot distinguish the true-positive hits from the false-positive hits, as the number of fragments match is more in false-positives compared with the true hits (Supplementary Fig. S3). This instance shows that when provided with the correct predicted or experimental spectra, the *mS-score* can clearly distinguish the true-positive from similar false-positive spectra.

The validation of the developed tool was performed by re-analysis of a publicly available dataset on *Triphala*, a well-known Ayurvedic formulation. This traditional medicine formulation is a complex of three plant species and is rich in secondary metabolites or phytochemicals (Parveen et al., 2018; Russell et al., 2011). The complexity of this dataset was ideal for testing our tool. A previous study by Subbannayya et al. (2018) identified metabolites of *Triphala* only at MS level. However, predicted spectra of plant phytochemicals could have been used as a reference to identify the metabolites at MS/MS level. The MS2Compound will substantially facilitate compound identification at MS/MS level.

In addition, one can also limit the size and nature of the database according to the context of the study (e.g., secondary metabolites; lactic acid cycle; fatty acids, and so on). Re-analysis of *Triphala* metabolomics dataset against a specific database of ''Phenol-Explorer'' and ''Phytochemical compounds from KEGG'' resulted in the assignment of 596 features at MS level to 358 metabolites and 333 features at MS/MS level to 255 metabolites. The MS2Compound enabled the identification of 255 metabolites in *Triphala* at MS/MS level, which substantially improves the quality of data.

In addition, 242 and 220 metabolites were newly identified at MS and MS/MS level, when compared with previous findings by Subbannayya et al. (2018), by searching against a small and context-specific database (Supplementary Fig. S4A). This demonstrates the use of the MS2Compound as a complementary approach to identify additional metabolites with better identification, thereby improving the

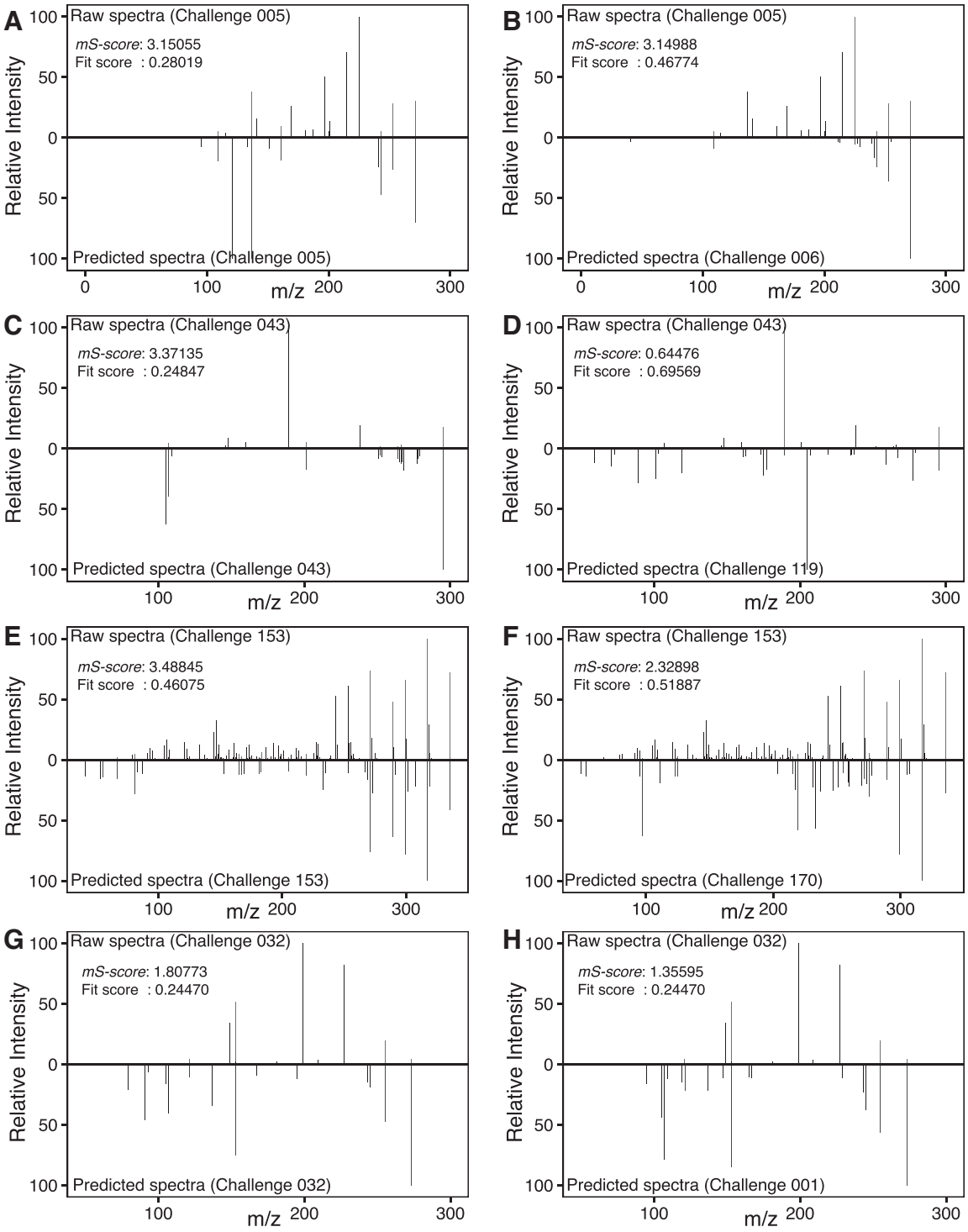

FIG. 3. Examples of a spectral match showing the ability of *mS-score* to identify true-positive hits compared with other scoring functions. **(A, C, E, G)** Spectral match for a true-positive hit, and **(B, D, F, H)** spectral match for corresponding false-positive hit. *mS-score* assigned a high score to true-positive match compared with the false-positive hit; however, fit-score assigns the score inversely. In **(B)** some of the high m/z values are mapped to the reference database, which increases the fit-score. m/z, mass-to-charge ratio.

outcome of any metabolomic investigation. The details of the compounds identified from *Triphala* metabolomics data are summarized in Table 1. Compounds such as gallic acid derivatives and quercetin derivatives were assigned at MS/MS level. The MS/MS match for gallic acid and 4-coumaroylshikimate is shown in Figure 4. Compounds such as nortrachelogenin, fargesin, feruloyl glucose, and hesperetin 7-O-glucoside were not identified in the previous study. The complete list of compounds identified is provided as Supplementary Tables S4 and S5.

Table 1. Details of Mass-Spectrometry and Mass-Spectrometry/Mass-Spectrometry Features and Corresponding Compound Assignment for the *Triphala* Metabolomics Dataset

MS level	*Data acquisition mode*	*No. of features*	*No. of features assigned*	*No. of metabolites identified*	*Selected compounds*
MS	Positive	737	319	219	Syringic acid; gallic acid 3-O-gallate; apigenin; quercitrin
	Negative	966	277	173	Quercetin 3-sulfate; cembrene; myricetin 3-O-arabinoside; quercetin3-O-(6″-acetyl-galactoside) 7-O-rhamnoside
Total (MS)		1703	596	358 (Non-redundant)	
MS/MS	Positive	1060	252	202	Gallic acid; gallic acid ethyl ester; 3,4-O-dimethylgallic acid; 4-coumaroylshikimate; epicatechin 7-O-glucuronide
	Negative	631	81	60	Ellagic acid; quercetin 3-O-(6″-malonyl-glucoside) 7-O-glucoside; quercetin 3,3'-bissulfate; [6]-gingerol
Total (MS/MS)		1691	333	255 (Non-redundant)	

MS, mass spectrometry.

Pre-processed raw features from positively acquired data of *Triphala* sample were mapped to KEGG database by using MZmine2 software. This resulted in the assignment of 346 and 525 MS and MS/MS features to a putative compound (Supplementary Tables S6–S8). A total of 263 and 316 non-redundant compounds were identified from the MS and MS/MS features, respectively. We compared the compounds identified from MS/MS features to the compounds identified from the MS2Compound. We have re-searched the MS/MS features with ''KEGG phytochemical compounds'' as a reference database with [M+H]+ as the probable adduct. The MS2Compound assigned 160 MS/MS features to 135 putative compounds. A total of 191 features at MS level were assigned to probable compounds with an error tolerance of 0.05 Da. Comparison of all the compounds identified in MZmine2 and the MS2Compound resulted in 434 (one feature identified to multiple compounds) common identifications (Supplementary Fig. S4B).

Finally, the MS2Compound tool also provides the unassigned spectra in MGF format after completion of the search. These spectra can be further matched to public databases for additional identification of compounds.

Discussion

User-friendly tools are much needed, as the currently available metabolomic analysis tools do not offer adequate provision for users to create or choose context-specific databases. Here, we introduced and evaluated the MS2Compound, a metabolite identification tool that can be used to generate a custom database of predicted spectra by using the

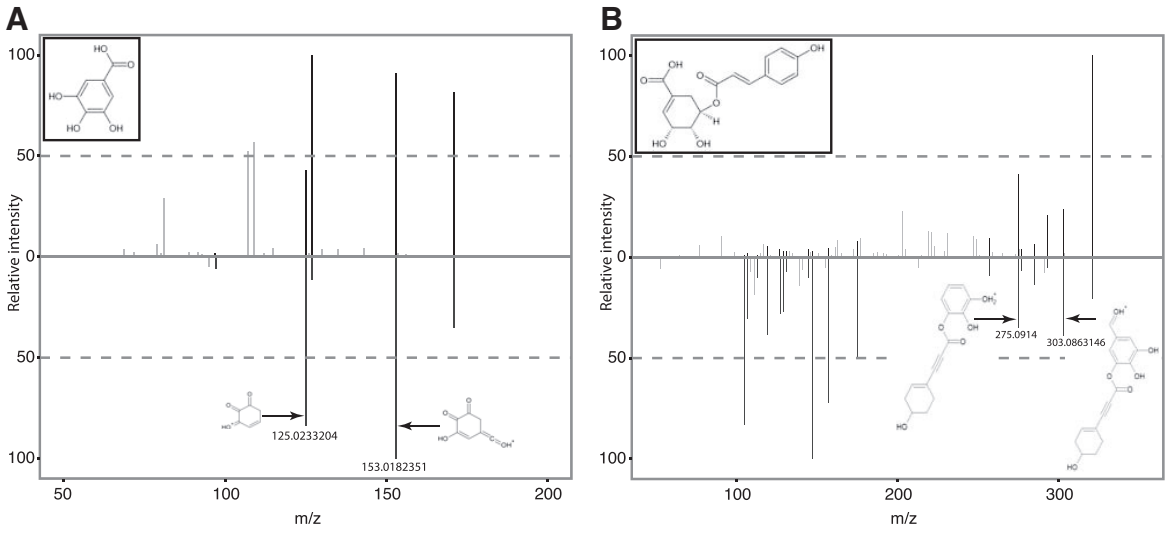

FIG. 4. Identification of gallic acid **(A)** and 4-Coumaroylshikimate **(B)** at the MS/MS level. The experimental fragments are shown in the upper half, whereas the predicted fragments are shown in the lower half. The matched spectra are shown in the *dark gray* and the intact structures of both molecules are shown inside a *box* in the *top left*. The structures of selected fragment ions are depicted in the image.

CFM-ID algorithm, and identify metabolites or compounds from the generated database.

The MS2Compound is a fully open-source metabolomics data search tool and it can be used by biologists even without any prior scripting expertise. The predicted spectra can be used as custom database for compound identification from a given metabolomic data. A modified scoring scheme has been proposed for MS/MS matches. The current version of the MS2Compound comprises a back-end database with compounds gathered from BioCyc, PlantCyc, KEGG phytochemicals, and Phenol-Explorer.

Our study is based on the assumption that using a context-relevant custom-made reference database will improve the confidence of compound identification from MS/MS metabolomics data by reducing the number of false-positives. Recent advances in spectra prediction algorithms and their accuracy to predict the spectra allow us to create a custom database for a list of compounds. The raw spectra of a given sample can be compared with these custom databases to improve the confidence of compound identification. These reasons motivated us to develop an easy-to-use and fully open-source tool MS2Compound, where a user can choose databases of interest.

Notably, the current version of the MS2Compound contains the compound database for ten species and their different strains. To our knowledge, this is the first of its kind search tool for metabolomic data analysis, which allows users to identify compounds from different reference databases of interest.

For the benefit of the biomedical scientists broadly and for those who may not have experience in handling scripts, a user-friendly GUI was also provided for the selection of a user-defined reference database. The MS2Compound contains a simple user interface for MS and MS/MS search, which has been developed in C#. The back-end execution of the program is achieved by using Perl scripts. Python module was used in only one instance to generate the chemical formula of given SMILES identifiers. The required modules and other dependencies were compiled to make it highly portable and easy to install.

The MS2Compound uses *mS-score* to score the spectra match at MS/MS level. *mS-score* is a modified distance function for matching raw spectra to a predicted spectra database. Unlike the other existing scoring functions, the *mS-score* assumes that compound identification is based on how the m/z and intensities are matched at fragment levels. A normalization factor β has been introduced for every pair of matched m/z and intensity, which is inversely proportional to the raw intensity of the matched fragment. A fragment with high raw intensity will be given less normalization value compared with the fragment with low intensity.

As the spectra in reference database consists of predicted spectra, the normalization factor β is dependent only on the intensity of the raw/experimental spectra. This makes the score an asymmetric function; however, it gives better results in such matches during compound identification. The score also assumes that the provided intensities are relative intensities (ranging from >0 to 100), therefore the normalization factor β is defined with a constant term 100 in the numerator. *mS-score* also considers the number of matched fragments at a given tolerance level. An increased number of matched fragments increases the confidence in compound identification. The current version of MS2Compound is restricted to the compound identification step, so it lacks any pre-processing steps for the raw data. Therefore, it uses a pre-processed (pre-processing such as removing noises from the spectra and other necessary steps) raw file as an input for MS/MS search.

In addition, other software such as MZmine2 can be used for pre-processing of the raw spectra and then export the MS/MS features as MGF format. The performance of the *mS-score* was compared with the three other scoring functions considering two parameters; the number of correctly assigned compounds with the best score (i.e., with rank 1) and the number of false-positive hits sharing the same score with the true-positive hits. *mS-score* was found to assign a greater number of true-positive hits compared with other scoring functions. Our score considers the difference in m/z and corresponding intensity for each match; however, other scores were based on the matched list of m/z and intensity, allowing the *mS-score* to distinguish the best match from other matches.

The functioning of MS2Compound was demonstrated with a re-analysis of a complex metabolomics dataset from a traditional medicine formulation-*Triphala*. It is one of the well-known ayurvedic formulations in traditional practices with anticancer, antimicrobial, and immunomodulatory activities (Belapurkar et al., 2014; Biradar et al., 2008; Vadde et al., 2015). This formulation comprised dried extracts from three plants; *Phyllanthus emblica*, *Terminalia bellirica*, and *Terminalia chebula*, and they are rich in phytochemicals such as phenols, flavonoids, tannins, saponins, and others. The study of metabolites at the MS/MS level pose challenges due to the lack of MS/MS information for plant secondary metabolites. Prediction of the MS/MS spectra for these metabolites serves the purpose of the reference database in compound identification.

Metabolite assignment of *Triphala* metabolomics data, searched against phytochemicals and phenols database, resulting in 558 metabolite matches. Previous data were generated by searching against a comprehensive KEGG database, because there was no opportunity to select context-specific datasets in MZmine2. However, using the MS2Compound, we could select a context-specific dataset of MS/MS spectra from KEGG phytochemicals. Approximately 37% and 20% of features were assigned to putative compounds in both MS and MS/MS-level, respectively. We found MS/MS evidence for the signature metabolites present in *Triphala*. Derivatives of gallic acid, quercetin, myricetin, and epicatechin were also assigned at MS/MS level this time. We also found other pharmacologically active compounds, including gibberellin A15, feruloyl glucose, and hesperetin 7-O-glucoside, with a good score at MS/MS level.

We found metabolites such as nortrachelogenin and fargesin with known medicinal properties, which have not been captured in the previous article. Nortrachelogenin was known as one of the pharmacologically active compounds present in medicinal plants (Kato et al., 1979; Kaunda and Zhang, 2017). Fargesin was known to have anti-inflammatory properties (Pham et al., 2017). This analysis of *Triphala* dataset using the MS2Compound suggests that this tool can be effectively used to mine several such

publicly available datasets (such as data acquired by Banerjee et al., 2020 and Karthikkeyan et al., 2020) to successfully identify significant and context-relevant metabolites pertaining to the model being studied.

MS/MS matches from MS2Compound and MZmine2 show 62.8% similar identification using the positively acquired *Triphala* raw data. The MZmine2 identifications were performed only at MS level; however, the MS2Compound provided the evidence at MS/MS level. This improves the confidence of compound identification. However, we missed some of the identification as we have used only phytochemicals of KEGG database and at least three fragment matches for MS/MS searches.

Alongside other tools, Global Natural Products Social Molecular Networking (GNPS) (Wang et al., 2016) and XCMS are among the widely used software for compound identification. GNPS and XCMS provide multiple features with respect to the compound identification from metabolomics data. The GNPS allows library search against known spectra for standard compounds compiled from various sources (https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp).

Similarly XCMS uses METLIN as the back-end library for compound identification. However, these databases are comprehensive; they do not allow context-specific compound identification for a given experiment, which might lead to a high number of false-positive identifications. Instead, the current tool provide the luxury of selecting own databases or creating a custom database for a particular study, thereby finding the compounds that are most likely to present in the given sample.

The MS2Compund offers multiple advantages over other existing tools by:

(i) allowing the use of databases of choice;
(ii) generation of MS/MS spectra of compounds powered with CFM-ID algorithm;
(iii) search at MS/MS level (for experiments with MS/MS information); and
(iv) powered with a better scoring function *mS-score.*

The MS2Compound also provides the unassigned spectra as MGF file output, which will make it easy to integrate it in further analysis by incorporating searches against additional databases and algorithms of choice for deeper mining of metabolomic datasets. We believe that the MS2Compound will become an effective and complementary tool and will transform the metabolomic data analysis pipelines in the times to come. The current version of the MS2Compound does not allow for data pre-processing, which will be considered for next updates.

Conclusions

The MS2Compound tool developed in the study is poised to improve the quality of biomedical research by improving the metabolomic data analysis pipeline. It will easily complement the current search algorithms, which often search metabolites only at MS level. Implementation of a novel *mS-score* minimizes false-positives in metabolite-spectral matching. By allowing a user-specified custom database, the MS2Compound can transform the way the metabolomics data are handled currently. Powered with a robust prediction algorithm of potential MS/MS spectra for a database of metabolites, the MS2Compound will accelerate the generation of MS/MS spectral library for metabolites in the near future. It will facilitate the identification of new metabolites, which have not been identified at MS/MS level so far. By providing unassigned spectra as MGF files, which then can be easily plugged to subsequent data analysis pipelines, the MS2Compound will serve as a complementary tool to the existing tools of metabolomics.

Data Availability

Raw files for *Triphala* metabolomics study are available at MassIVE (MassIVE id: MSV000086784).

Acknowledgments

S.K.B. is a recipient of a Junior Research Fellowship from the Department of Biotechnology-Bioinformatics National Certification (DBT-BINC), Government of India. S.K. is a recipient of a Senior Research Fellowship from the Indian Council of Medical Research (ICMR), Government of India. G.K. was a recipient of Senior Research Fellowship from the Council of Scientific and Industrial Research (CSIR), Government of India (2014–2019), and is currently a recipient of KSTePs DST-PhD. Fellowship from the Department of Science and Technology-Karnataka Science and Technology Promotion Society, Government of Karnataka (2020–2021). R.R. is a recipient of the Young Scientist Award (YSS/2014/000607) from the Science and Engineering Research Board, Department of Science and Technology (DST), Government of India. The authors thank Ms. Nupur Agarwal for designing the MS2Compound logo for the GUI.

Author Disclosure Statement

The authors declare they have no conflicting financial interests.

Funding Information

The authors thank the Department of Biotechnology, Government of India for research support to Yenepoya (Deemed to be University) under ''Bioinformatics Center and National Network Program for Skill Development in Mass Spectrometry-based Metabolomics Technologies'' (40202). They also thank Karnataka Biotechnology and Information Technology Services (KBITS), the Government of Karnataka, for the support to the Center for Systems Biology and Molecular Medicine at Yenepoya (Deemed to be University) under the Biotechnology Skill Enhancement Programme in Multiomics Technology (BiSEP GO ITD 02 MDA 2017).

Supplementary Material

Supplementary Data
Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3

Supplementary Table S4
Supplementary Table S5
Supplementary Table S6
Supplementary Table S7
Supplementary Table S8

References

Allen F, Pon A, Wilson M, et al. (2014). CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 42, W94–W99.

Banerjee S, Kar A, Mukherjee PK, et al. (2020). Immunoprotective potential of Ayurvedic herb Kalmegh (*Andrographis paniculata*) against respiratory viral infections—LC-MS/MS and network pharmacology analysis. Phytochem Anal. DOI: 10.1002/pca.3011.

Belapurkar P, Goyal P, and Tiwari-Barua P. (2014). Immunomodulatory effects of triphala and its individual constituents: a review. Indian J Pharm Sci 76, 467–475.

Biradar YS, Jagatap S, Khandelwal KR, et al. (2008). Exploring of antimicrobial activity of Triphala Mashi-an ayurvedic formulation. Evid Based Complement Alternat Med 5, 107–113.

Blazenovic I, Kind T, Ji J, et al. (2018). Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. Metabolites 8, 31.

Cambiaghi A, Ferrario M, and Masseroli, M. (2017). Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief Bioinform 18, 498–510.

Caspi R, Billington R, Ferrer L, et al. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 44, D471–D480.

Chambers MC, Maclean B, Burke R, et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30, 918–920.

Cheng J, Lan W, Zheng G, et al. (2018). Metabolomics: a high-throughput platform for metabolite profile exploration. Methods Mol Biol 1754, 265–292.

Collins JM, Walker DI, Jones DP, et al. (2018). High-resolution plasma metabolomics analysis to detect *Mycobacterium tuberculosis*-associated metabolites that distinguish active pulmonary tuberculosis in humans. PLoS One 13, e0205398.

Considine EC, Thomas G, Boulesteix AL, et al. (2017). Critical review of reporting of the data analysis step in metabolomics. Metabolomics 14, 7.

Cooper BT, Yan X, Simon-Manso Y, et al. (2019). Hybrid search: a method for identifying metabolites absent from tandem mass spectrometry libraries. Anal Chem 91, 13924–13932.

Djoumbou-Feunang Y, Fiamoncini J, Gil-De-La-Fuente A, et al. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminform 11, 2.

Duhrkop K, Fleischauer M, Ludwig M, et al. (2019). SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16, 299–302.

Frediani JK, Jones DP, Tukvadze N, et al. (2014). Plasma metabolomics in human pulmonary tuberculosis disease: a pilot study. PLoS One 9, e108854.

Gil-De-La-Fuente A, Godzien J, Saugar S, et al. (2019). CEU Mass Mediator 3.0: a metabolite annotation tool. J Proteome Res 18, 797–802.

Goldansaz SA, Guo AC, Sajed T, et al. (2017). Livestock metabolomics and the livestock metabolome: a systematic review. PLoS One 12, e0177675.

Guijas C, Montenegro-Burke JR, Domingo-Almenara X, et al. (2018). METLIN: a technology platform for identifying knowns and unknowns. Anal Chem 90, 3156–3164.

Hsu YH, Churchhouse C, Pers TH, et al. (2019). PAIRUP-MS: pathway analysis and imputation to relate unknowns in profiles from mass spectrometry-based metabolite data. PLoS Comput Biol 15, e1006734.

Huan T, Forsberg EM, Rinehart D, et al. (2017). Systems biology guided by XCMS Online metabolomics. Nat Methods 14, 461–462.

Huan T, Tang C, Li R, et al. (2015). MyCompoundID MS/MS search: metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. Anal Chem 87, 10619–10626.

Kanehisa M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci 28, 1947–1951.

Kanehisa M, Sato Y, Furumichi M, et al. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Res 47, D590–D595.

Karthikkeyan G, Pervaje R, Subbannayya Y, et al. (2020). Plant omics: metabolomics and network pharmacology of liquorice, Indian Ayurvedic Medicine Yashtimadhu. OMICS 24, 743–755.

Kato A, Hashimoto Y, and Kidokoro M. (1979). (+)-Nortrachelogenin, a new pharmacologically active lignan from Wikstroemia indica. J Nat Prod 42, 159–162.

Kaunda JS, and Zhang YJ. (2017). The genus Carissa: an ethnopharmacological, phytochemical and pharmacological review. Nat Prod Bioprospect 7, 181–199.

Li L, Li R, Zhou J, et al. (2013). MyCompoundID: using an evidence-based metabolome library for metabolite identification. Anal Chem 85, 3401–3408.

Li Y, Kuhn M, Gavin AC, et al. (2020). Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. Bioinformatics 36, 1213–1218.

Myers OD, Sumner SJ, Li S, et al. (2017). Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data. Anal Chem 89, 8689–8695.

Parveen R, Shamsi TN, Singh G, et al. (2018). Phytochemical analysis and in-vitro biochemical characterization of aqueous and methanolic extract of Triphala, a conventional herbal remedy. Biotechnol Rep (Amst) 17, 126–136.

Pham TH, Kim MS, Le MQ, et al. (2017). Fargesin exerts anti-inflammatory effects in THP-1 monocytes by suppressing PKC-dependent AP-1 and NF-kB signaling. Phytomedicine 24, 96–103.

Pluskal T, Castillo S, Villar-Briones A, et al. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395.

Ramirez-Gaona M, Marcu A, Pon A, et al. (2017). YMDB 2.0: a significantly expanded version of the yeast metabolome database. Nucleic Acids Res 45, D440–D445.

Rothwell JA, Perez-Jimenez J, Neveu V, et al. (2013). Phenol-Explorer 3.0: a major update of the Phenol-Explorer

database to incorporate data on the effects of food processing on polyphenol content. Database (Oxford) 2013, bat070.

Russell LH, Jr., Mazzio E, Badisa RB, et al. (2011). Differential cytotoxicity of triphala and its phenolic constituent gallic acid on human prostate cancer LNCap and normal cells. Anticancer Res 31, 3739–3745.

Sajed T, Marcu A, Ramirez M, et al. (2016). ECMDB 2.0: a richer resource for understanding the biochemistry of *E. coli*. Nucleic Acids Res 44, D495–D501.

Shen X, Wang R, Xiong X, et al. (2019). Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun 10, 1516.

Subbannayya Y, Karthikkeyan G, Pinto SM, et al. (2018). Global metabolite profiling and network pharmacology of Triphala identifies neuromodulatory receptor proteins as potential targets. J Proteins Proteomics 9, 101–114.

Sud M, Fahy E, Cotter D, et al. (2007). LMSD: LIPID MAPS structure database. Nucleic Acids Res 35, D527–D532.

Tautenhahn R, Patti GJ, Rinehart D, et al. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 84, 5035–5039.

Vadde R, Radhakrishnan S, Reddivari L, et al. (2015). Triphala extract suppresses proliferation and induces apoptosis in human colon cancer stem cells via suppressing c-Myc/cyclin D1 and elevation of Bax/Bcl-2 ratio. Biomed Res Int 2015, 649263.

Wang M, Carver JJ, Phelan VV, et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34, 828–837.

Wang Y, Liu L, Ma Y, et al. (2019). Chemical discrimination of *Astragalus mongholicus* and *Astragalus membranaceus* based on metabolomics using UHPLC-ESI-Q-TOF-MS/MS approach. Molecules 24, 4064.

Wishart DS, Feunang YD, Marcu A, et al. (2018). HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46, D608–D617.

Wishart DS, Tzur D, Knox C, et al. (2007). HMDB: the human metabolome database. Nucleic Acids Res 35, D521–D526.

Yen CY, Houel S, Ahn NG, et al. (2011). Spectrum-to-spectrum searching using a proteome-wide spectral library. Mol Cell Proteomics 10(7), M111.007666.

Address correspondence to:
Yashwanth Subbannayya, PhD
Center for Systems Biology and Molecular Medicine
Yenepoya Research Centre
Yenepoya (Deemed to be University)
University Road, Deralakatte
Mangalore 575018, India

E-mail: yashwanth.subbannayya@gmail.com

Thottethodi Subrahmanya Keshava Prasad, PhD
Center for Systems Biology and Molecular Medicine
Yenepoya Research Centre
Yenepoya (Deemed to be University)
University Road, Deralakatte
Mangalore 575018, India

E-mail: keshav@yenepoya.edu.in

Abbreviations Used

AUC = area under the curve
CASMI = Critical Assessment of Small Molecule Identification
CFM-ID = Competitive Fragmentation Modeling-ID
DP = dot product
FS = fit score
GNPS = Global Natural Products Social Molecular Networking
GUI = Graphical User Interface
HGS = hypergeometric score
HMDB = Human Metabolome Database
KEGG = Kyoto Encyclopedia of Genes and Genomes
LC-MS/MS = liquid chromatography-tandem mass spectrometry
MGF = Mascot Generic Format
MS = mass-spectrometry
Mtb = *Mycobacterium tuberculosis*
m/z = mass-to-charge ratio
ROC = Receiver Operating Characteristic
SMILES = Simplified Molecular-Input Line-Entry System