Article

Integration of Molecular Networking and In-Silico MS/ MS Fragmentation for Natural Products Dereplication

Pierre-Marie Allard, Tiphaine Péresse, Jonathan Bisson, Katia Gindro, Laurence Marcourt, Van Cuong Pham, Fanny Roussi, Marc Litaudon, and Jean-Luc Wolfender

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.5b04804 • Publication Date (Web): 16 Feb 2016

Integration of Molecular Networking and *In-Silico* MS/MS Fragmentation for Natural Products Dereplication

Pierre-Marie Allard[1], Tiphaine Péresse[2], Jonathan Bisson[3], Katia Gindro[4], Laurence Marcourt[1], Van Cuong Pham[5], Fanny Roussi[2], Marc Litaudon[2] and Jean-Luc Wolfender[1,*]

[1] *School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Quai Ernest-Ansermet 30, CH-1211 Geneva 4, Switzerland*

[2] *Institut de Chimie des Substances Naturelles CNRS UPR 2301, University Paris-Saclay, 1 Avenue de la Terrasse, 91198, Gif-sur-Yvette, France*

[3] *Center for Natural Product Technologies, Department of Medicinal Chemistry and Pharmacognosy College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612, United States*

[4] *Mycology and Biotechnology group, Institute for Plant Production Sciences IPS, Agroscope, Route de Duillier 50, P.O. Box 1012, 1260 Nyon, Switzerland*

[5] *Institute of Marine Biochemistry of the Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet road, Cau Giay, Hanoi, Vietnam*

ABSTRACT: Dereplication represents a key step for rapidly identifying known secondary metabolites in complex biological matrices. In this context, Liquid-Chromatography coupled to High Resolution Mass Spectrometry (LC-HRMS) is increasingly used and, via untargeted data-dependent MS/MS experiments, massive amount of detailed information on the chemical composition of crude extracts can be generated. An efficient exploitation of such datasets requires automated data treatment and access to dedicated fragmentation databases. Various novel bioinformatics approaches such as Molecular Networking (MN) and *in-silico* fragmentation tools have emerged recently and provide new perspective for early metabolite identification in Natural Product (NP) research. Here we propose an innovative dereplication strategy based on the combination of MN with an extensive *in-silico* MS/MS fragmentation database of NPs. Using two case studies we demonstrate that this combined approach offers a powerful tool to navigate through the chemistry of complex NPs extracts, dereplicate metabolites and annotate analogues of database entries.

In Natural Products (NPs) research, crude extracts of various origin (e.g. plants, marine organisms and microorganisms) containing thousands of metabolites have to be characterized, either as part of bioactivity guided isolation studies for drug discovery purposes or in the frame of metabolomics investigation for biomarker identification. Isolation and *de novo* structural elucidation of NPs is a tedious task and should ideally only be performed for new metabolites to avoid the costly re-isolation process of known molecules.[1] Unambiguous metabolite identification thus represents one of the major bottlenecks in metabolomics studies and in NPs chemistry.[2] The rapid identification of known metabolites by comparison of experimental spectral data to databases is referred to as dereplication. This dereplication process is now mandatory to efficiently guide the isolation of only valuable NPs or biomarkers within their complex biological matrices.[3] Notable improvements in metabolite profiling methods have been mainly related to the introduction of ultra-high performance liquid chromatography (UHPLC) with sub-2 µm particles columns and to the development of benchtop high-resolution mass spectrometry (HRMS) detectors. Detailed information of the chemical composition of crude natural extracts can now be efficiently obtained.[4] High-resolution MS data, when used in combination with orthogonal heuristic filters, such as isotopic pattern distribution, is able to lead to the correct molecular

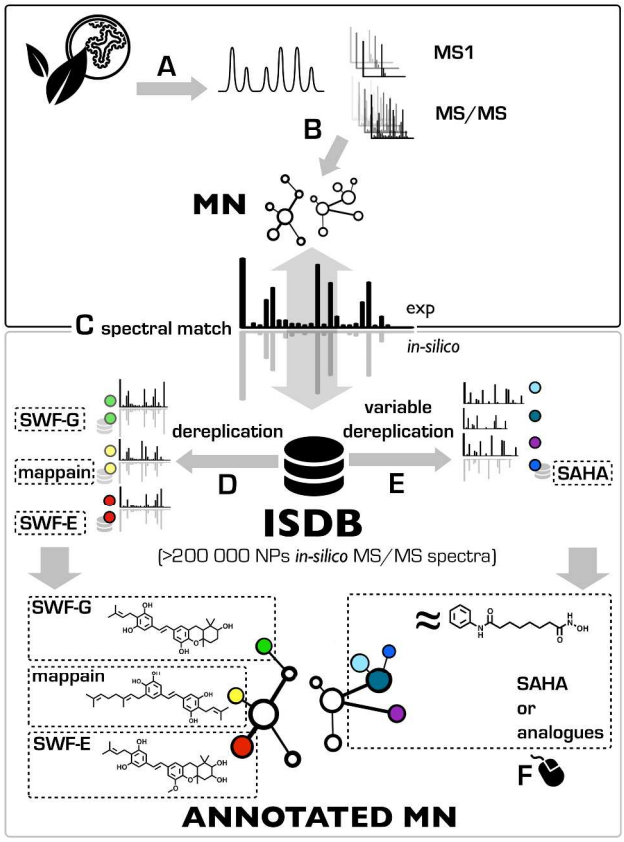


Figure 1: Conceptual scheme of Molecular Network (MN)-based dereplication using the *In-Silico* MS/MS DataBase (ISDB). Crude extracts are profiled and untargeted MS/MS data are acquired (**A**). Parent ions are organized as MN according to their MS/MS data (**B**). The spectral library search (**C**) can be made in two different modes. Individual nodes are dereplicated against the ISDB using the parent ion mass as pre-filter and the MS/MS data (**D**, also illustrated in **Fig. 2**). Alternatively, individual nodes can be dereplicated against the ISDB in a modification tolerant spectral library search called variable dereplication (**E**, also illustrated in **Fig. 3**). By a simple right click, dereplication results are directly visualized as chemical structures on top of dereplicated nodes in the MN (**F**).

formula of the analytes in many cases.[5,6] Nevertheless, even with the correct molecular formula, isomers can not be resolved and additional spectral information are then needed in order to discriminate between the potential candidates. Tandem MS/MS offers structural insights by breaking the analyzed ion into fragment ions and measuring their *m/z* ratio. Tandem MS/MS data is thus more discriminant in a dereplication process than the parent mass alone.[7] However the manual inspection of individual MS/MS spectra is a tedious task and the complexity and amount of data generated by LC-MS/MS analysis of complex extracts makes automated methods preferable.

Recently, various bioinformatics approaches have been developed to organize or interpret large sets of MS/MS fragmentation data. For example solutions such as MAGMa (MS Annotation based on *in silico* Generated Metabolites) allow matching of multistage fragmentation data (MSn) against candidate molecules substructures and were successfully applied on complex NPs extracts.[8,9] Among these new approaches, Molecular Networking (MN) is a particularly effective one to organize MS/MS fragmentation spectra. MN compares all MS/MS spectra in a given extract or a set of extracts and groups them according to their similarity via the establishment of a modified cosine score.[10] Since the MS/MS spectra is linked to the chemical structure of the fragmented metabolites, MN is able to group molecules according to their structural similarities. This approach thus provides new ways to navigate in the metabolome of biological samples by providing key information on analogies among the detected metabolites.[11] Besides it's ability to classify MS/MS data, MN has been shown to be a useful tool for dereplication[12] and various applications of MN in NPs chemistry have been reported in the last years.[13–16]

Two strong advantages of the MN-based dereplication strategy can be highlighted when comparing it to dereplication based on MS1 data only. The first advantage consists in the pooling of ions according to their structural similarity. This structural grouping thus helps with assessing the relevance of the dereplication hit within a cluster. Indeed, if a dereplicated node belongs to a particular structural type, it may help identifying its neighbors in the network by giving to similar scaffolds a better score than to unrelated ones. Thus, annotations can be propagated trough connected nodes.[17] The second advantage offered by MN-based dereplication is the possibility to search for structural analogues. Since the spectral matching stage allows shifts in the parent mass as well as searching for common neutral losses, molecules with different parent masses but with similar fragmentation patterns can be connected. Analogues can be linked during the MN generation but also during the dereplication process. This variable dereplication is a radically new way to characterize the chemistry of complex extracts and may be regarded as a paradigm-changing tool for the NPs chemist. When using MS1-based approaches, only isomers of entries of the database can be dereplicated. With a variable dereplication approach, it is the whole chemical space surrounding these entries that can be mapped, and unknown entities can be putatively identified or at least highlighted.

Despite these advantages, MN-based dereplication, which is dependent upon comparison with tandem MS/MS data, is currently limited by the size of available fragmentation data libraries. Existing MS/MS libraries include MassBank[18], Metlin[19], ReSpect[20], NIST[21] and the GNPS-Community libraries.[22] All together, these libraries represent less than 20 000 individual compounds for which experimental MS/MS data are searchable.[23]

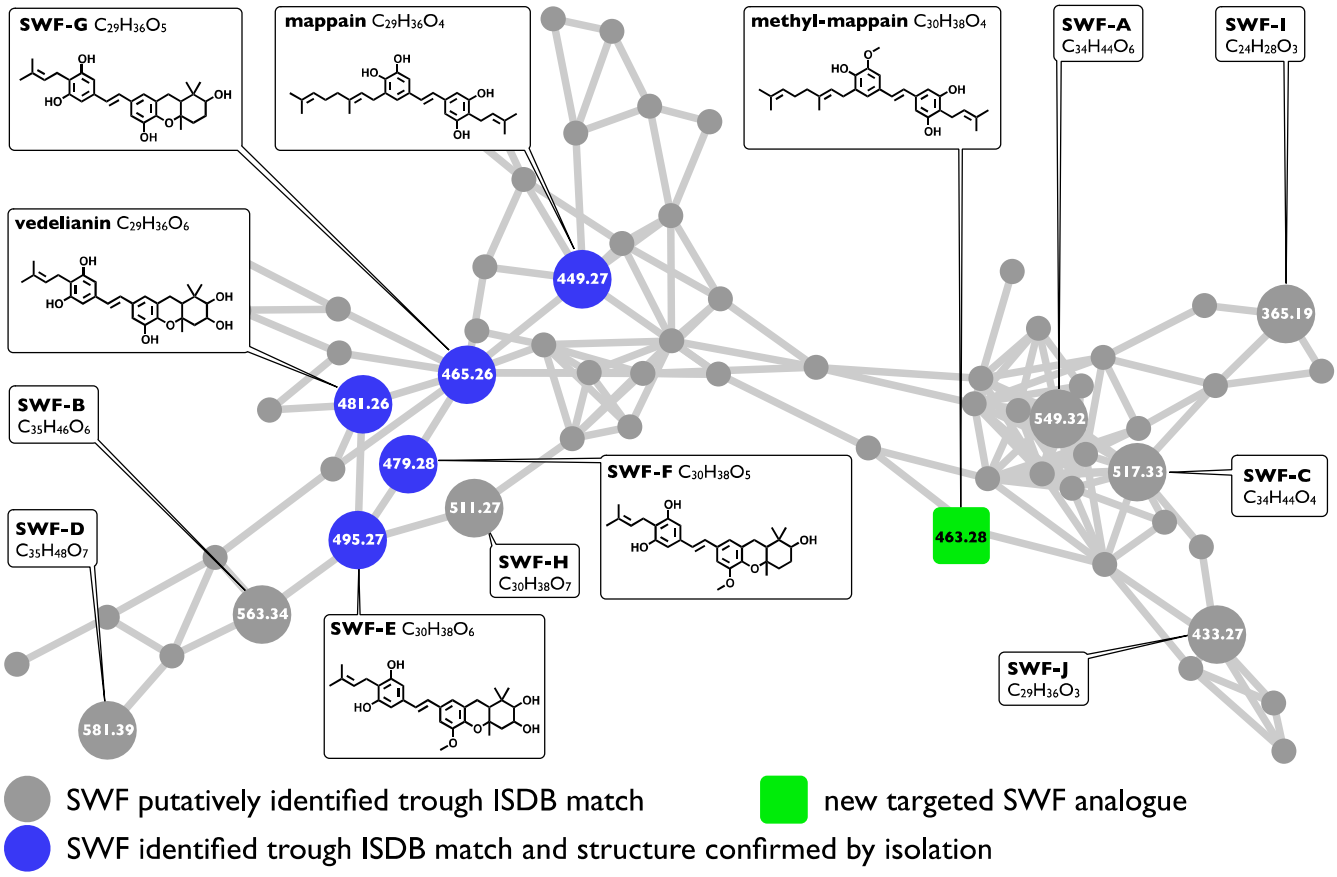


Figure 2: Cluster corresponding to compounds of the SWF (schweinfurthin) family observed in the MN of EtOAc extract of various *Macaranga* species. Match with the ISDB was made using the parent ion mass as filter (**D** in **Fig.1**)

This number is however, by far, not comparable to the actual number of described NPs which is estimated to be greater than 200 000.[24] Thus, *in-silico* fragmentation approaches have been envisioned as a possible way to overcome the limit set by the low number of experimental MS/MS data. Various methods for *in-silico* fragmentation exist and their applications in NPs chemistry have been reviewed in 2014.[25] Integration of an *in-silico* fragmentation database of a specific chemical class (lipids) in a dereplication workflow has been successfully developed by Fiehn's lab and is available on an online platform: LipidBlast.[23] Recently, two innovative strategies based on *in-silico* fragmentation approaches have been published. CSI:FingerID computes a fragmentation tree to explain the MS/MS spectrum of an unknown molecule and further predict the molecular fingerprint of this compound. This fingerprint can then be searched within vast chemical databases such a PubChem.[26] Another interesting solution is the MCID MS/MS search program which allows the user to search experimental MS/MS data against an *in-silico* fragmented database of 8021 human metabolites and their predicted metabolic products.[27] CSI:FingerID presented improved scores against other benchmarked tools but has the current limitation of processing MS/MS spectra one at a time.[28] The MCID MS/MS tool offers a batch mode but the searchable database is limited to human metabolites.

In this work we describe an alternative approach allowing to search experimental MS/MS spectra in a batch mode against an extensive database constituted by the *in-silico* data of >220 000 NPs, thus representing a >10 fold increase regarding the actual total number of spectra present in all experimental MS/MS databases.[23] The dereplication process allows annotation of all nodes with positive hits in the database and direct visualization of the results as structures within the MN. By combining MN, an extensive *In-Silico* MS/MS DataBase (ISDB) and chemical structure visualization we propose a comprehensive dereplication pipeline based on Free, open-source tools as well as an accessible database for maximal availability to the community. A proof of principle with two case studies, a plant and a fungal strain extracts, using two complementary dereplication modes is presented.

EXPERIMENTAL SECTION:

Molecular Networking: Molecular networks were created using the online workflow at GNPS (http://gnps.ucsd.edu). The data were then clustered with MS-Cluster with a parent mass tolerance of 0.8 Da and a MS/MS fragment ion tolerance of 0.5 Da to create consensus spectra. Further, consensus spectra that contained less than 2 spectra were discarded. A network was then created, where edges were filtered to have a

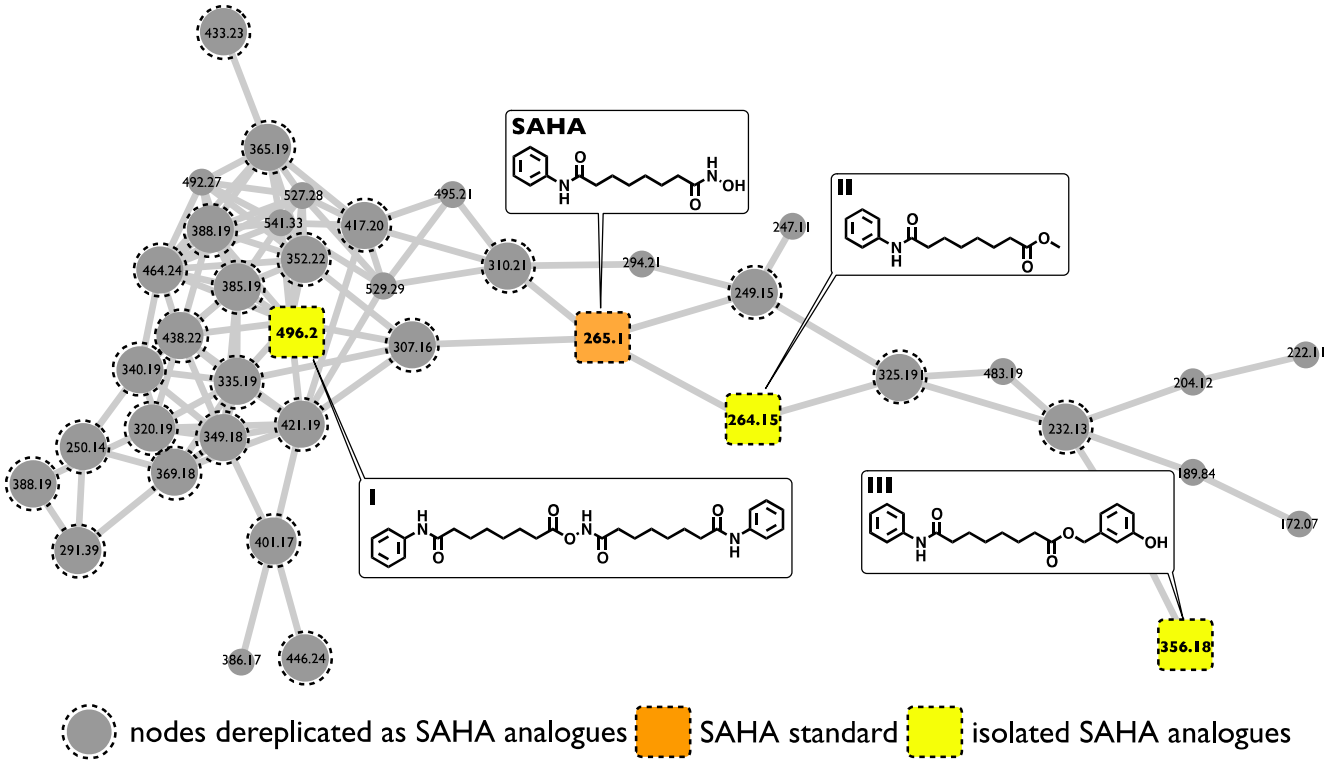


Figure 3: Cluster corresponding to analogues of SuberoylAnilide Hydroxamic Acid (SAHA) observed in the MN of crude extract of a *Penicillium* species treated with the epigenetic modificator. Match with the ISDB was made in variable dereplication mode with a 200 Da shift tolerance

cosine score above 0.7 and more than 6 matched peaks. Further edges between two nodes were kept in the network if, and only if, each of the nodes appeared in each other's respective top 10 most similar nodes. The spectra in the network were then searched against the ISDB spectral library. ChemViz 1.3 plugin (freely available at http://www.cgl.ucsf.edu/cytoscape/chemViz/) was used to display the structure of the dereplication hits directly within Cytoscape 2.8.

Mass Spectrometry Analysis: Chromatographic separation was performed on a Thermo Dionex Ultimate 3000 UHPLC system interfaced to a Q-Exactive Plus mass spectrometer (Thermo Scientific, Bremen, Germany), using a heated electrospray ionization (HESI-II) source. The LC conditions were as follows: column: Waters BEH C18 100x2.1 mm i.d., 1.7 µm; mobile phase: (A) water with 0.1% formic acid; (B) acetonitrile with 0.1% formic acid; flow rate: 600 µL/min; injection volume: 1 µL; gradient: linear gradient of 2%–98% B over 8 min. In positive ion mode, diisooctyl phthalate $C_{24}H_{38}O_4$ $[M+H]^+$ ion (*m/z* 391.28429) was used as internal lock mass. The optimized HESI-II parameters were the following: source voltage: 4.0 kV (pos), sheath gas flow rate (N_2): 50 units; auxiliary gas flow rate: 12 units; spare gas flow rate: 2.5; capillary temperature: 266.25 °C (pos), S-Lens RF Level: 50. The mass analyzer was calibrated using a mixture of caffeine, methionine-arginine-phenylalanine-alanine-acetate (MRFA), sodium dodecyl sulfate, sodium taurocholate and Ultramark 1621 in an acetonitrile/methanol/water solution containing 1% acid by direct injection. The data-dependent MS/MS events were performed on the 5 most intense ions detected in full scan MS (Top5 experiment). The MS/MS isolation window width was 1 *m/z*, and the stepped normalized collision energy (NCE) was set to 20 – 35 – 50 units. In data-dependent MS/MS experiments, full scans were acquired at a resolution of 35 000 FWHM (at *m/z* 200) and MS/MS scans at 17 500 FWHM both with a maximum injection time of 50 ms. After being acquired in MS/MS scan, parent ions were placed in a dynamic exclusion list for 3.0 seconds.

Generation of the in-silico fragmentation databases (ISDB) and library search: The generation of the ISDB was achieved using SMILES input from the commercial Dictionary of Natural Products (DNP v. 24, (http://dnp.chemnetbase.com/)). Permanently charged compounds were removed and the input resulted in a 221 771 entries database. The *in-silico* fragmentation tool used for the generation of the ISDB is CFM-ID v. 1.10 (freely available at http://sourceforge.net/projects/cfm-id/).[29] The cfm-predict module was used with the default pre-trained Combined-Energy model. The probability threshold was set to 0.001 and no post-processing was applied. The generated spectra of protonated species at low, medium and high energy were merged and converted to .mgf files. The input file was split in 455 files of 500 entries and jobs were parallelized on a HPC cluster. The computations were performed at University of Geneva on the Baobab cluster.[30] The library search was achieved using the open-source tool Tremolo[31] freely available at:

http://proteomics.ucsd.edu/Software/Tremolo/. When using parent mass (PM) as filter the PM tolerance was set to 0.005 Da. In variable dereplication mode it was set to 200 Da. The similarity score threshold was set at 0.2.

Since DNP is a commercial product, the generated ISDB cannot be distributed here. In order to generate a freely accessible DB, another ISDB using input from the Universal Natural Products Database UNPD (a free DB available at http://pkuxxj.pku.edu.cn/UNPD/) was generated.[32] After removing duplicates and permanently charged compounds a total of 170 602 compounds were fragmented using the same parameters as above. The resulting UNPD-ISDB is available at http://oolonek.github.io/ISDB/.

Isolation and de novo structural elucidation of selected metabolites: Details and protocols used are provided in SI.

RESULTS & DISCUSSION:

A comprehensive DB of MS/MS *in-silico* spectra of NPs was generated using the Dictionary of Natural Products (DNP) as structural input. After filtering permanently charged compounds (for instance, quaternary nitrogen), 221 771 compounds were conserved. The theoretical MS/MS spectrum of each NP structure was then calculated using an open-source tool that proposes a probabilistic generative model for the ESI-MS/MS fragmentation process (CFM-ID). This tool was chosen for various reasons. First, when benchmarked against existing state-of-the-art *in-silico* fragmentation methods (MetFrag and FingerID), using multiple data sets, CFM-ID was previously shown to significantly outperform them.[29] Another advantage of using CFM-ID is that the calculations are done once for all structures and saved in the database thus eliminating the need for further computations during the dereplication process.[26] Finally, the computation process can be parallelized, allowing in our case the full *in-silico* fragmentation of the 221 771 structures in less than 12 hours.

Metabolite profiling of selected crude natural extracts was performed by UHPLC-HRMS with automated acquisition of MS/MS spectra in a data dependent analysis mode. The dereplication strategy then consisted in the combination of two pivotal tools: molecular networks and an extensive *in-silico* fragmentation database of NPs (ISDB). All acquired MS/MS spectra were processed by molecular networking and nodes of the network were dereplicated in an automated batch mode against the ISDB. The spectral library search was done using the parent ion mass as filter or in a modification tolerant manner (variable dereplication). The dereplication results could then be directly visualized as chemical structures within the networks allowing intuitive navigation through the annotated nodes of the analyzed sample.

In order to evaluate the relevance of the dereplication strategy and the pertinence of the results, we applied it to two different complex NPs extracts following complementary approaches.

- In a first case, the ISDB was searched with a strict filter corresponding to the parent ion mass in order to precisely identify nodes of the MN. (See point **C** in **Fig. 1** and **Fig. 2**). Here, the strategy was tested to dereplicate compounds of a specific chemical class of interest (prenylated stilbenes) in various plant species of the *Macaranga* genus and target unknown analogues.

- In a second case, the ISDB was searched in a modification tolerant manner (variable dereplication), allowing spectral match with compounds having different parent ion mass but sharing MS/MS spectral similarities. (See point **E** in **Fig. 1** and **Fig. 3**). Here, a crude extract of a fungal strain of the *Penicillium* genus treated with a synthetic compound, SuberoylAnilide Hydroxamic Acid (SAHA), was dereplicated using this mode in order to rapidly highlight exogenous metabolites related to SAHA.

***Identification of prenylated stilbenes in various Macaranga* spp.**

Euphorbiaceae of the *Macaranga* genus are known to biosynthesize a particular prenylated stilbene class, schweinfurthins (SWF), which display potent anticancer activities.[33–36] In order to discover novel derivatives of SWF, we analyzed crude extracts of 21 *Macaranga* species by UHPLC-HRMS/MS and organized the obtained fragmentation data by molecular networking. In the generated MN, 142 999 individual MS/MS spectra were organized into 2368 nodes forming 361 clusters of connected nodes. (See SI **Fig S-1**). The MN was initially dereplicated against the whole ISDB and afterward against a subset of the ISDB limited to Euphorbiaceae secondary metabolites (3272 compounds) in order to refine the dereplication results. The first step of the chemical exploration consisted in the identification of possible SWF clusters within the full molecular network. By browsing the dereplicated structures through the network using the chemical visualization plugin, a cluster corresponding to SWF analogues was easily highlighted (**Fig.2** and SI **Fig. S-1**). Within this cluster, 12 members of the SWF family (SWF-A, B, C, D, E, F, G, H, I, J, mappain and vedelianin) were dereplicated. Five of these SWF, which were previously isolated by us from *M. tanarius* (SWF-E, SWF-F, SWF-G, mappain and vedelianin), were co-injected and allowed in each case to confirm the identification made by comparison with the *in-silico* database (blue nodes in **Fig. 2**)

In parallel, dereplication of the metabolites detected by UHPLC-HRMS (positive ion mode) was made based on HRMS1 data only in order to compare both approaches. A peak list was built by reconstructing chromatographic peaks for each detected features (specific m/z at a specific retention time). All exact masses of the peak list were then compared to the exact masses of the

compounds reported in the DNP with a mass tolerance of ± 3 ppm. In this case, the superiority of tandem MS/MS over HRMS1 based dereplication was clearly highlighted. For example, for vedelianin ($C_{29}H_{36}O_6$) or SWF G ($C_{29}H_{36}O_5$), 10 different isomers exist for each molecular formula only within the Euphorbiaceae family, and 27 and 28, respectively, within all described NPs (according to the DNP). Here, since structural information related to fragmentation is taken into account, MS/MS based dereplication allowed ranking of the correct structure first for both metabolites. Mappain ($C_{29}H_{36}O_4$) was ranked 1st out of the 4 described isomers when using the subset database limited to Euphorbiaceae metabolites. When searched against the whole ISDB, it was ranked 2nd out of the 15 isomers present in the DNP. These examples demonstrate the efficiency of the proposed dereplication strategy and also highlight the importance of using orthogonal information such as chemotaxonomy to take into account plausible biosynthetic routes.[37]

Various nodes (smaller grey nodes in **Fig. 2**) could not be identified as known members of the SWF family, but their presence within the cluster indicated potential novel analogues. One of these metabolites, $[M+H]^+$ ion at *m/z* 463.28 (green node in **Fig. 2**), could be localized in the MS profiles of two Vietnamese *Macaranga* species: *M. lowii* and *M. tanarius*. It corresponded to the molecular formula $C_{30}H_{38}O_4$ for which 21 isomers are present within the whole DNP. None of the *in-silico* MS/MS spectra of these isomers matched with the experimental spectra of this feature indicating a potential novel metabolite. The compound was finally isolated from the crude extract of the fruits of *M. tanarius*. As expected, the structural elucidation indicated that the molecule is a novel prenylated stilbene, wich corresponded to an O-methylated mappain derivative (structural elucidation is detailed in SI). According to the obtained network, various novel compounds are likely to be present in the analyzed *Macaranga* spp. extracts. Their phytochemical study is ongoing and the use of the annotated MN and the systematic extraction of their corresponding ion traces will help guiding their MS-targeted isolation.

Identification of SAHA metabolites in the culture broth of a Penicillium sp. *strain.*

Epigenetic modifiers have been used as chemical triggers of fungal cryptic biosynthetic pathways with success in the past few years.[38–40] One of the most commonly used epigenetic modifiers is the Histone DeACetylase (HDAC) inhibitor Suberoyl Anilide Hydroxamic Acid (SAHA). In order to explore the biosynthetic potential of *Penicillium vulpinum* it was grown for 15 days in SAHA supplemented Potato Dextrose Broth (PDB) media versus a standard PDB media. The culture broths of 6 replicates were extracted and analyzed by UHPLC-HRMS/MS. HRMS1 data of the control and SAHA-treated crude extracts were compared by statistical analysis. Several features were found to be only present in the SAHA-treated cultures indicating the possible activation of a previously silent biosynthetic pathway. In order to gain insight on the chemical identity of these features, MN were generated on SAHA treated and control fungal culture extracts. The features differentially detected by the statistical analysis were found to be present in a common MN cluster (**Fig. 3** and SI **Fig. S-2**). An ion at *m/z* 265.1 corresponding to the protonated form of SAHA was observed within this cluster. In order to check if all features in this cluster were structurally related to SAHA, its *in-silico* fragmentation data was added to the ISDB. In this case, since the objective was to identify all possible analogues of a precise metabolite (SAHA), the library search was run in the variable dereplication mode allowing for a ±200 Da mass shift between the parent ion mass and the database mass. More than 60% of the nodes of this particular cluster were annotated as structural analogues of SAHA. Mass-targeted isolation of three ions at *m/z* 496.2, 264.15 and 356.18 (yellow squares in **Fig. 3**) indeed resulted in the isolation of SAHA derivatives. Compound **I** (octanedioic acid phenylamide(7-phenylcarbamoyl-heptanoyloxy)-amide) and **II** (methyl suberanilate) are synthetic compounds that have been previously described.[41,42] Compound **I** is described as a prodrug of SAHA presenting HDAC inhibiting activity as well as increased aqueous solubility and cellular permeability compared to SAHA.[41] Compound **III** (3-hydroxybenzyl 8-oxo-8-(phenylamino)octanoate) is a novel molecule and is likely generated by the condensation of SAHA with 3-hydroxybenzyl alcohol, a metabolite described in *Penicillium vulpinum*.[43] Whether this condensation is enzymatically catalyzed or merely results from the intrinsic reactivity of the SAHA hydroxamate moiety with 3-hydroxybenzyl alcohol remains to be elucidated. The observation of numerous nodes related to SAHA (**Fig. 3**) indicated its important reactivity in the culture media. The HDAC inhibition activity described[41] for compounds such as **I** also indicates that the epigenetic modifying effects observed in SAHA-treated fungal cultures [44–47] might be due to a mixture of HDAC inhibiting compounds rather than attributed to SAHA only. In our case, the use of the variable dereplication mode against the ISDB incremented with *in-silico* fragmentation data of SAHA allowed to annotate a large range of features structurally related to SAHA. In the search for novel induced compounds in response to epigenetic modifications such a tool can thus be efficiently used to differentiate compounds related to the modifier itself from those that might be induced.

The usefulness of the dereplication strategy could thus be illustrated by these two cases. A step-by-step detailed workflow describing how to annotate MN nodes by spectral matching against the ISDB and visualize the results as chemical structures within the network is provided in the Supporting Information. A version of the ISDB based on the freely available Universal Natural Products Database[32] (UNPD-ISDB) as well as a script,

allowing to easily launch the spectral search and generate the Cytoscape output files for results visualization, are also provided as additional material.

CONCLUSION

The proposed dereplication workflow is generic and based on a combination of Free, open-source tools allowing availability and further improvements. By increasing the number of individual MS/MS spectra (> 10 fold compared to all currently available experimental MS/MS spectra), the generated ISDB drastically expands the chemical space searchable by tandem MS/MS based dereplication.

The two complementary case studies described in this work highlight the interest of the ISDB to efficiently annotate nodes within a MN. Nevertheless, it is important to keep in mind that matches against *in-silico* fragmentation data have to be taken with caution and will not replace comparison against spectra of standards acquired under the same ionization and fragmentation conditions. However, the combination with MN provides increased confidence in the annotation process by efficiently displaying the structural relationship between metabolites. Thus, combined with chemical structure visualization within the network, the *in-silico* fragmentation of an extensive NPs database offers a powerful tool to explore and proceed to the early annotation of complex mixtures of secondary metabolites. It is expected that integration of such ISDBs in a specific dereplication workflow of online MN platforms such as GNPS[22] would be helpful for automated annotation of generated networks and usefully complement the existing experimental databases.

Various aspects of this combined MN-ISDB dereplication approach could be improved in the future to further increase the confidence in metabolite identification. Besides the development of novel *in-silico* fragmentation tools, experimental MS/MS data acquisition parameters should be optimized to maximize structural information. All efforts in this direction as well as improvements in the score ranking of candidates would be welcomed. Usage of orthogonal information such as integration of phylogenetic data in this ranking process could be an interesting development.

Another exciting possibility is the creation of *in-silico* fragmentation libraries of hypothetical metabolites[8,48], for which, by definition, no experimental spectra can be acquired. By combining *in-silico* metabolization[37,49] and *in-silico* fragmentation, the searchable chemical space of actual databases could be exponentially increased. For example, generation of MS/MS fragmentation libraries of *in-silico* predicted polyketide-synthase metabolites[50,51] could be an effective analytical approach to target novel NPs scaffolds in microorganisms and effectively complement genome-mining strategies. The availability of such bioinformatics tools should open brand new perspectives in NPs research.

ASSOCIATED CONTENT

Supporting Information

The full molecular networks of *Macaranga* spp. and *Penicillium* sp. extracts, cultivation conditions, isolation details and NMR data of all isolated compounds are presented in the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org. The full UNPD-ISDB, scripts to launch spectral search and generate ouput files for results visualization and a detailed step-by-step workflow are available at http://oolonek.github.io/ISDB/.

AUTHOR INFORMATION

Corresponding Author

jean-luc.wolfender@unige.ch

ACKNOWLEDGMENT

PMA is grateful to FNS for fellowship on Subside 200020_146200. We would like to thank Felicity Allen (University of Alberta, Canada) for discussions and precious help with the CFM-ID tool. We would like to thank Mingxun Wang, Pieter Dorrestein and Nuno Bandeira (University of California, San Diego, USA) for developing and making public the GNPS platform. We thank the NCCR Chemical Biology for use of their facilities. This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01). Part of this work was carried out within the framework of an International Associated Laboratory (LIA) between the Centre National de la Recherche Scientifique (CNRS, France) and the Vietnam Academy of Science and Technology (VAST, Vietnam). Figure 1 and TOC contain modified icons by Fredrik Edfors, Edward Battistini, Nimal Raj, Sergey Demushkin, Jakob Vogel and Michael Wohlwend from thenounproject.com

REFERENCES

(1) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129.

(2) Wishart, D. S. *Bioanalysis* **2011**, *3*, 1769–1782.

(3) Gaudêncio, S. P.; Pereira, F. *Nat. Prod. Rep.* **2015**, *32*, 779–810.

(4) Wolfender, J.-L.; Marti, G.; Thomas, A.; Bertrand, S. *J. Chromatogr. A* **2015**, *1382*, 136–164.

(5) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2007**, *8*, 105.

(6) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234.

(7) Nielsen, K. F.; Larsen, T. O. *Front. Microbiol.* **2015**, *6*, 1–15.

(8) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Vervoort, J.; Bino, R. J. *Anal. Chem.* **2014**, *86*, 4767–4774.

(9) van der Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; de Vos, R. C. H. *Metabolomics* **2012**, *8*, 691–703.

(10) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.;

Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743–E1752.

(11) Traxler, M. F.; Kolter, R. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10128–10129.

(12) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686–1699.

(13) Kleigrewe, K.; Almaliti, J.; Tian, I. Y.; Kinnel, R. B.; Korobeynikov, A.; Monroe, E. a.; Duggan, B. M.; Di Marzo, V.; Sherman, D. H.; Dorrestein, P. C.; Gerwick, L.; Gerwick, W. H. *J. Nat. Prod.* **2015**, *78*, 1671–1682.

(14) Brito, Â.; Gaifem, J.; Ramos, V.; Glukhov, E.; Dorrestein, P. C.; Gerwick, W. H.; Vasconcelos, V. M.; Mendes, M. V.; Tamagnini, P. *Algal Res.* **2015**, *9*, 218–226.

(15) Duncan, K. R.; Crüsemann, M.; Lechner, A.; Sarkar, A.; Li, J.; Ziemert, N.; Wang, M.; Bandeira, N.; Moore, B. S.; Dorrestein, P. C.; Jensen, P. R. *Chem. Biol.* **2015**, *22*, 460–471.

(16) Klitgaard, A.; Nielsen, J. B.; Frandsen, R. J. N.; Andersen, M. R.; Nielsen, K. F. *Anal. Chem.* **2015**, *87*, 6520–6526.

(17) Mohimani, H.; Pevzner, P. A. *Nat. Prod. Rep.* **2016**, *00*, 1–14.

(18) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703–714.

(19) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther Drug Monit* **2005**, *27*, 747–751.

(20) Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T.; Hirai, M. Y.; Saito, K. *Phytochemistry* **2012**, *82*, 38–45.

(21) The National Institute of Standards and Technology. NIST. http://www.nist.gov/srd/nist1a.cfm (accessed Dec 16, 2015).

(22) GNPS http://gnps.ucsd.edu/ (accessed Dec 16, 2015).

(23) Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755–758.

(24) Mander, L. N.; Liu, H.-W. *Comprehensive Natural Products II: Chemistry and Biology*; Amsterdam: Elsevier, 2010.

(25) Hufsky, F.; Scheubert, K.; Böcker, S. *Nat. Prod. Rep.* **2014**, *31*, 807.

(26) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. *Proc. Natl. Acad. Sci.* **2015**, *112*, 12580–12585.

(27) Huan, T.; Tang, C.; Li, R.; Shi, Y.; Lin, G.; Li, L. *Anal. Chem.* **2015**, *87*, 10619–10626.

(28) da Silva, R. R.; Dorrestein, P. C.; Quinn, R. a. *Proc. Natl. Acad. Sci.* **2015**, *112*, 12549–12550.

(29) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98–110.

(30) Baobab HPC cluster https://plone.unige.ch/distic/pub/hpc/baobab_en (accessed Dec 16, 2015).

(31) Wang, M.; Bandeira, N. *J. Proteome Res.* **2013**, *12*, 3944–3951.

(32) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. *PLoS One* **2013**, *8*, e62839.

(33) Thoison, O.; Hnawia, E.; Guiéritte-Voegelein, F.; Sévenet, T. *Phytochemistry* **1992**, *31*, 1439–1442.

(34) Beutler, J. a.; Shoemaker, R. H.; Johnson, T.; Boyd, M. R. *J. Nat. Prod.* **1998**, *61*, 1509–1512.

(35) Beutler, J.; Jato, J.; Cragg, G.; Boyd, M. *Nat. Prod. Res.* **2000**, *14*, 399–404.

(36) Turbyville, T. J.; Gürsel, D. B.; Tuskan, R. G.; Walrath, J. C.; Lipschultz, C. a; Lockett, S. J.; Wiemer, D. F.; Beutler, J. a; Reilly, K. M. *Mol. Cancer Ther.* **2010**, *9*, 1234–1243.

(37) Audoin, C.; Cocandeau, V.; Thomas, O.; Bruschini, A.; Holderith, S.; Genta-Jouve, G. *Metabolites* **2014**, *4*, 421–432.

(38) Du, L.; Robles, A. J.; King, J. B.; Powell, D. R.; Miller, A. N.; Mooberry, S. L.; Cichewicz, R. H. *Angew. Chem. Int. Ed. Engl.* **2013**, *53*, 804–809.

(39) Wang, X.; Sena Filho, J. G.; Hoover, A. R.; King, J. B.; Ellis, T. K.; Powell, D. R.; Cichewicz, R. H. *J. Nat. Prod.* **2010**, *73*, 942–948.

(40) Fisch, K. M.; Gillaspy, a F.; Gipson, M.; Henrikson, J. C.; Hoover, a R.; Jackson, L.; Najar, F. Z.; Wägele, H.; Cichewicz, R. H. *J. Ind. Microbiol. Biotechnol.* **2009**, *36*, 1199–1213.

(41) Miller, T. A.; Witter, D. J.; Belvedere, S. HISTONE DEACETYLASE INHIBITOR PRODRUGS. US 2009/0023786 A1, 2009.

(42) Gediya, L. K.; Chopra, P.; Purushottamachar, P.; Maheshwari, N.; Njar, V. C. O. *J. Med. Chem.* **2005**, *48*, 5047–5051.

(43) Makino, M.; Endoh, T.; Ogawa, Y.; Watanabe, K.; Fujimoto, Y. *Heterocycles* **1998**, *9*, 1931–1934.

(44) Chung, Y.; El-Shazly, M.; Chuang, D.-W.; Hwang, T.; Asai, T.; Oshima, Y.; Ashour, M. L.; Wu, Y.; Chang, F. *J. Nat. Prod.* **2013**, *76*, 1260–1266.

(45) Miao, F.; Liang, X.; Liu, X.; Ji, N. *J. Nat. Prod.* **2014**, *77*, 429–432.

(46) Du, L.; King, J. B.; Cichewicz, R. H. *J. Nat. Prod.* **2014**, *77*, 2454–2458.

(47) Henrikson, J. C.; Ellis, T. K.; King, J. B.; Cichewicz, R. H. *J. Nat. Prod.* **2011**, *74*, 1959–1964.

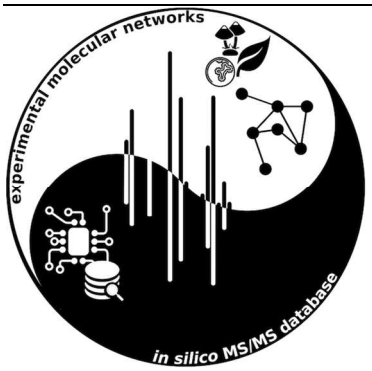
(48) Menikarachchi, L. C.; Hill, D. W.; Hamdalla, M. a.; Mandoiu, I. I.; Grant, D. F. *J. Chem. Inf. Model.* **2013**, *53*, 2483–2492.

(49) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E. J.; Henry, C. S. *J. Cheminform.* **2015**, *7*, 44.

(50) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. a.; Muller, R.; Wohlleben, W.; Breitling, R.; Takano, E.; Medema, M. H.

Nucleic Acids Res. **2015**, *43*, 237–243.

(51) Yadav, G.; Gokhale, R. S.; Mohanty, D. *PLoS Comput. Biol.* **2009**, 5, e1000351.



For TOC only