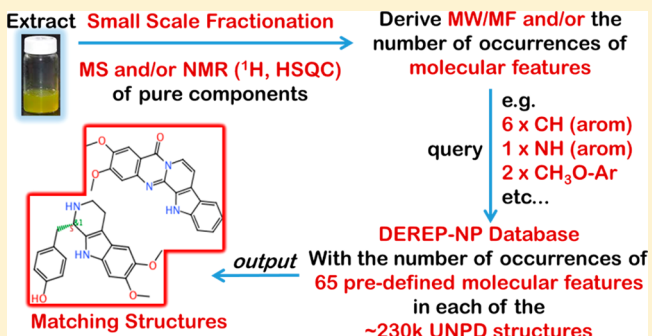


Database for Rapid Dereplication of Known Natural Products Using Data from MS and Fast NMR Experiments

Carlos L. Zani^{*,†} and Anthony R. Carroll[‡][†]Natural Products Chemistry Laboratory, Centro de Pesquisa René Rachou-Fiocruz, Belo Horizonte, 30190-002, MG, Brazil[‡]Griffith School of Environment, Griffith University, Gold Coast Campus, Southport, QLD 4222, Australia

Supporting Information

ABSTRACT: The discovery of novel and/or new bioactive natural products from biota sources is often confounded by the reisolation of known natural products. Dereplication strategies that involve the analysis of NMR and MS spectroscopic data to infer structural features present in purified natural products in combination with database searches of these substructures provide an efficient method to rapidly identify known natural products. Unfortunately this strategy has been hampered by the lack of publicly available and comprehensive natural product databases and open source cheminformatics tools. A new platform, DEREPI-NP, has been developed to help solve this problem. DEREPI-NP uses the open source cheminformatics program DataWarrior to generate a database containing counts of 65 structural fragments present in 229 358 natural product structures derived from plants, animals, and microorganisms, published before 2013 and freely available in the nonproprietary Universal Natural Products Database (UNPD). By counting the number of times one or more of these structural features occurs in an unknown compound, as deduced from the analysis of its NMR (¹H, HSQC, and/or HMBC) and/or MS data, matching structures carrying the same numeric combination of searched structural features can be retrieved from the database. Confirmation that the matching structure is the same compound can then be verified through literature comparison of spectroscopic data. This methodology can be applied to both purified natural products and fractions containing a small number of individual compounds that are often generated as screening libraries. The utility of DEREPI-NP has been verified through the analysis of spectra derived from compounds (and fractions containing two or three compounds) isolated from plant, marine invertebrate, and fungal sources. DEREPI-NP is freely available at <https://github.com/clzani/DEREP-NP> and will help to streamline the natural product discovery process.



A crude biota extract may contain thousands of structurally diverse compounds present in a wide range of concentrations.¹ The separation of these compounds from each other (many of which often occur in low abundance) can be challenging.¹ The majority of these compounds are involved in primary metabolic processes, but a subset, natural products (NPs), are not.² NP research often aims to identify the role of purified NPs in biological processes, and to do this, extracts or semipurified fractions obtained through chromatographic separation of crude extracts are tested in a bioassay.¹ Further purification of the bioactive fractions ultimately leads to the isolation of individual bioactive compounds.¹ The identification of the complete 3D structure of these unknown bioactive NPs is not trivial and requires extensive analysis of data obtained from multiple spectroscopic experiments (mainly MS and 1D and 2D NMR). Often configurational isomers (either enantiomers or diastereomers) are identified, and these molecules provide an added challenge to correctly assign a structure.¹ Even after peer review some structures are incorrectly assigned, and their true identity may subsequently be corrected either through reinterpretation of spectroscopic

data or total synthesis.³ To date, well over 230 000 structurally characterized NPs have been reported from plants, animals, and microorganisms, and the number of new structurally characterized NPs published continues to increase by many thousands each year.⁴ As a consequence, known NPs are frequently reidentified during NP discovery research. The reisolation of known bioactive NPs when the discovery of novel or new ones is the intended outcome of research is an expensive and time-consuming impediment. The concept of dereplication was introduced in 1978 as a methodology to avoid the rediscovery of known anticancer antibiotics from bioactive extracts derived from microorganisms when the discovery of novel anticancer compounds was intended.⁵ Dereplication is now defined as the elimination of known active substances from consideration when a bioactive mixture is being investigated.⁶ Since 1978 a diversity of dereplication strategies have been reported.^{7–12} These strategies generally rely upon the ability to match molecular features present in unknown bioactive NPs (either

Received: November 24, 2016

Published: June 15, 2017

spectroscopic/spectrometric or structural) with data stored in spectroscopic (containing MS and/or ^{13}C NMR data of NPs) and/or structural databases. A dereplication strategy is generally implemented after initial screening of extracts or semipurified fractions, and the most widely used involve the so-called hyphenated techniques, in which a separation device (a chromatograph) is coupled with spectrometers such as MS, UV, IR, and NMR and analysis of the spectra obtained provides structural information on the compounds present in mixtures.^{9,13–15} UV and IR provide the least discriminatory spectroscopic data and can really be used only for dereplication in combination with other spectroscopic data. Mass spectrometry is a powerful tool, but many compounds have identical molecular weights and fragmentation data can be ambiguous. An NMR spectrometer has been described as the universal chemical detector since it generates the most information-rich data to derive structural features present in a compound and molecules with the same molecular weight but different constitution or configuration possess different NMR fingerprints. The intrinsic low sensitivity of this technique has been circumvented by enormous advances in hardware, software, and pulse programs, and nowadays a very small amount of sample is required to generate high-quality NMR data. This frequently allows unambiguous identification of compounds eluting from an HPLC column in submilligram quantities.^{6,16} NMR spectroscopy therefore provides a significant advantage over MS to definitively assign molecular structures. Whatever dereplication approach is used the computational matches (or “hits”) still need to be verified as identical structures through more thorough spectroscopic analysis, literature comparisons with published spectroscopic data, or chromatographic comparison with reference compounds.

Metabolomics is a related cheminformatics methodology that has been developed to study the total metabolic processes within organisms through the identification and quantification of metabolites.¹⁷ A significant distinction between metabolomics and dereplication is that metabolomics research usually aims to identify all metabolites involved in metabolic processes, whereas dereplication is a tool used to identify only known NPs. Metabolomics can however provide insights for NP research since the majority of metabolites present in an organism are involved in primary metabolic processes (the main focus of metabolomics research), but the metabolic pathways of primary metabolism often supply the precursors for NP biosynthesis.² The accuracy of metabolomics analysis can suffer exactly the same limitations as hyphenated dereplication techniques since validation of the true identity of a molecule that matches a database entry requires more rigorous analysis than just a molecular ion match or a fragmentation pattern match, as these give no indication of the constitution or configuration within the molecule. A significant limitation of many metabolomics and hyphenated dereplication strategies is that they ignore configurational isomerism. To this end, databases of ^1H NMR spectra for common primary metabolites have been generated, and these serve as a powerful tool for cross referencing MS data with NMR data to give an orthogonal set of corroborating data to unambiguously assign the true identity of a molecule. Unfortunately NMR data for the vast majority of published NPs are not available in databases and the majority of the metabolome of the world’s biodiversity remains uncharacterized and thus unidentifiable.

There are free Web-based databases available that have been used for fast identification of known compounds. NAPROC,¹³ available at the University of Salamanca (Spain), contains ^{13}C NMR data for about 20 000 natural products,^{18,19} while the CH-NMR-NP database contains ^{13}C NMR chemical shifts for 30 000 natural products published between 2000 and 2014.²⁰ However, besides containing only a limited number of compounds, matching structures in these databases requires researchers to acquire ^{13}C NMR spectra for any isolated compound (an insensitive method). Since a database containing annotated NMR data for the 230 000 structurally characterized NPs does not currently exist, other approaches to utilize the powerful discriminating features of NMR data have been developed. In 2001, Bradshaw and co-workers developed a method to quickly match the planar structures of previously published NPs using MS and NMR data obtained for purified NPs.²¹ They showed the discriminant power of structural information such as molecular weight and the exact counts of the number of methyl, methylene, and methine groups occurring in each structure from a database containing about 126 000 natural product structures. Their approach was further improved by other groups,^{6,22} to include other structural features that also can be easily deduced from ^1H , HSQC, and/or HMBC NMR experiments. Unfortunately, the software and data described by these authors are either proprietary or commercial,²³ restricting their widespread access.

More recently, Williams and co-workers created a subset of the ChemSpider database containing 22 million diverse compounds for the dereplication of natural products using minimal NMR data inputs.²⁴ They concluded that their approach would give better results if the database could be focused on NPs only. This context prompted us to prepare a database that, besides containing the molecular structures, molecular formulas, and exact and nominal molecular weights of known natural products, also included the frequency at which specific substructures, identified from simple and fast NMR experiments (e.g., ^1H NMR and edited HSQC), occur in each structure in the database. For molecules containing methyl groups, HMBC spectra can also be acquired quickly, and this often provides further substructures containing quaternary carbons such as ketones, esters, amides, double bonds, and amino-substituted and oxygenated carbons. The recent publication of the Universal Natural Products Database (UNPD), a compilation of 229 358 natural products from terrestrial and marine macro- and microorganisms that was made publicly available, has made this work possible.⁴ Furthermore, free and open source software necessary to process, store, retrieve, and analyze chemical structures and related data is now available.^{25–27} With these data and tools in hand, a database (DEREP-NP) was generated that provides the number of times that each of the selected 65 structural features (small substructures that can be deduced from NMR experiments) occurs in each NP present in UNPD. This dereplication methodology can be applied to purified natural products or fractions containing a small number of compounds. Fraction libraries that have been developed to contain a small number of individual NPs per fraction through judicious choice of HPLC separation protocols are routinely used in NP drug discovery, and this dereplication methodology is well suited to interrogate these sorts of libraries. The DEREP-NP database is freely available for download at <https://github.com/clzani/DEREP-NP>. It must be emphasized that DEREP-NP is a manual low-throughput dereplication tool that is not useful for

high-throughput metabolomics. It is, however, a powerful tool to rapidly identify “hits”. Hits are defined as published NP structures that are present in DERE-*NP* and that share the queried structural features with the purified unknown compound. Verification of the identity of the unknown compound with one of the “hits” still requires an analysis of the necessary set of experimentally obtained spectroscopic data and comparison with published data of the “hits” for unambiguous identification. Furthermore, if configurational isomers of a structure are present in DERE-*NP*, a search will return “hits” for all published configurational isomers, but again comparison with published data should clarify these stereochemical assignments.

RESULTS AND DISCUSSION

The origin and scope of the data compiled in UNPD (<http://pkuxxj.pku.edu.cn/UNPD>) were detailed by Gu and co-workers.⁴ The DERE-*NP* database was prepared using the procedures outlined in the *Experimental Section*. It contains 1.92 GB of data and takes about 3 min to be processed and displayed by DataWarrior²⁶ using the hardware described. By saving the database in the DataWarrior format (*.dwr), the file size is reduced to 361.6 MB and requires less than 40 s to process and present the data on the screen. Searches usually take less than a second to be completed. The DataWarrior interface is user-friendly and contains a help file that provides detailed instructions to use its functionalities.

DERE-*NP* searches can be performed using the different formats available in DataWarrior: textual/numeric, with the options “starts with”, “contains”, and “equals”; the slider format is useful to search ranges of values. If partial structures can be derived from NMR data, they can be searched using the substructure filter format. All structures retrieved by a query are shown in a single window pane, making it easier for the user to inspect and decide which feature to query next or what signals to check in the NMR spectra, thus facilitating the iterative nature of the process. As the data file also contains 3D coordinates, the 3D viewer of the software can be used to facilitate the visualization of structures that may be difficult to inspect in a 2D representation.

The distribution of the MW of the compounds in the database is shown in *Figure 1*. Overall, the database contains 33 339 different MW values, with the most frequent value being 264.13615, representing 640 compounds with the molecular formula $C_{15}H_{20}O_4$.

Interpreting edited HSQC NMR data in combination with analysis of 1H NMR integrals to distinguish CH's from CH_3 's

provides the most discriminant feature since the numbers of CH_3 's, CH_2 's, and CH's can be rapidly identified. The addition of specific types of quaternary carbons deduced from HMBC NMR correlations provides an added level of discrimination. Although the number of structures with singular values for each of these features can be as high as 35 000 (*Figure 2*), if combined, the discriminant effect is increased and can lead, in many cases, to very few “hit” structures, as demonstrated by Bradshaw et al.²¹ Indeed, the number of different $CH_3-CH_2-CH-Cq$ combinations in the database is 70 138, which is almost twice the number of distinct MW combinations. Thus, the combination of these four features also has a high discriminating power.

The most frequent combination in the database, for example, is for structures with 4 CH_3 's, 4 CH_2 's, 4 CH's, and 3 Cq's, which occurs only 454 times. If MW is included as a criterion, the number of structures retrieved can be drastically reduced. However, while some MW values return just one structure, others can retrieve as many as 152 (*Table 1*).

These results emphasize that although being valuable, the criteria used by Bradshaw et al.²¹ are not always enough to reduce the number of hits to a manageable number. This is the reason that Lang et al.⁶ and Bitzer et al.²² included more features that can be extracted from simple NMR experiments. For the same reason, DERE-*NP* includes the ability to search for 65 features (*Table 2*), which can help the researcher to reduce the number of candidate structures to be analyzed and increase the chances for the rapid identification of the isolated compound under investigation if it is present in the database.

The papers from Bitzer et al.,²² Bradshaw et al.,²¹ and Lang et al.⁶ provide good advice and strategies to make the best use of computational databases using NMR data for dereplication. The small-scale fractionation procedure described by Lang and co-workers is especially noteworthy.⁶ It is important to emphasize that queries with incorrect input values will undoubtedly result in wrong structures being retrieved. Even if wrong “hits” are retrieved, they will ultimately be eliminated because their published spectroscopic data will not match those obtained experimentally. To avoid this situation and if one is not sure about the frequency of a specific feature from the spectra, either this feature should not be used or the number of these features should be estimated using a larger value range. Although the latter approach increases the number of “hits”, at least the correct structure, if present in the database, will be among those retrieved. It must be stressed that the absence of a feature can be as important as its presence to reduce the number of candidate structures. So, if a given output shows structures containing features that cannot be corroborated by NMR data, a zero value for that feature should be entered to indicate its absence and hence exclude those structures containing that feature. Dereplication searches should aim to extract the most information from the 1H NMR and edited-HSQC experiments, as the number of possible structures can be dramatically reduced if one can use, for example, the number (or absence) of aromatic protons, the aromatic substitution pattern, number of oxygenated carbons (methoxy, carbinol, anomeric, and methylenedioxy), sp^3 , sp^2 , and sp -hybridized carbon atoms, terminal methylene groups, the multiplicity of the methyl signals, number of acidic OH and formyl groups, and so on (*Table 2*). For those with access to accurate mass spectrometric data the use of the text filter with the option “starting with” to enter the MW values, starting with an integer value and then inserting, one by one, the dot and the other

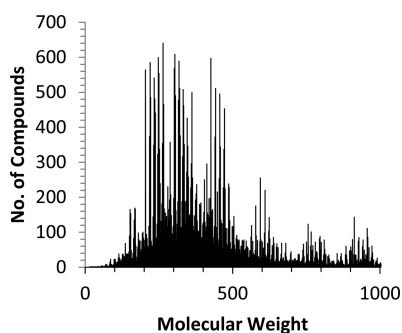


Figure 1. Graphic showing the number of compounds with MW values in the range 16–1000 present in the DERE-*NP* database.

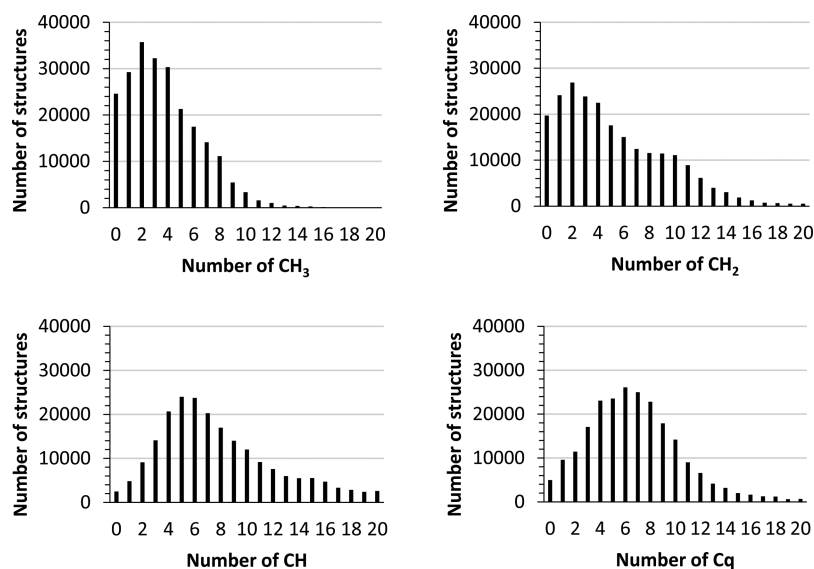


Figure 2. Histograms showing the number of structures vs the number of CH₃, CH₂, CH, and Cq (quaternary) groups in the DEREPP-NP database.

Table 1. MW Distribution of the 454 Natural Products with 4 CH₃'s, 4 CH₂'s, 4 CH's, and 3 Cq's

entry	MW	number of structures
1	204	152
2	220	77
3	236	42
4	237	1
5	238	105
6	252	8
7	254	31
8	268	1
9	272	10
10	301	9
11	302	3
12	317	1
13	318	2
14	333	1
15	348	1
16	354	1
17	380	2
18	396	1
19	417	1
20	433	5

digits, can be beneficial. Some caution should be made here since many molecules produce adducts by electrospray MS, and therefore blindly using the accurate mass data without paying due attention to potential adducts being observed could lead to erroneous results.

To verify DEREPP-NP outputs, we revisited the six examples given by Bitzer et al.,²² Bradshaw et al.,²¹ and Lang et al.,⁶ querying the same features in the same order used in their work. The results are summarized in Table 3. Since DEREPP-NP includes 229 358 structures of natural products published in the literature up to 2013,⁴ the number of “hits” is usually higher than those reported in previous studies with smaller or focused databases.

The following examples were selected from our own research. A compound was purified from the endophytic fungus *Cochliobolus* sp., and its HRMS indicated a molecular

Table 2. 65 Structural Features Counted for the 229 358 Structures in the DEREPP-NP Database

CH ₃ (all)	CH ₂ (all)	CH (all)	Cq ^a (all)
CH ₃ (singlet)	CH ₂ (sp ³)	CH (sp ³)	Cq (sp ³)
CH ₃ (doublet)	CH ₂ (sp ²)	CH (sp ² all)	Cq (sp ² all)
CH ₃ (triplet)	CH ₂ -O (all)	CH (sp ² olefinic)	Cq (sp ² olefinic)
CH ₃ (isopropyl)	CH ₂ (dioxy)	CH (sp ² arom)	Cq (sp ² arom)
CH ₃ -Ar ^b	CH ₂ -N (all)	CH (arom-singlet)	Cq (sp)
CH ₃ (vinyl)		CH (sp)	
CH ₃ (acetyl)		CH-O (all)	
CH ₃ -O (all)		CH/CO (aldehyde)	
CH ₃ O-Ar		CH (anomeric)	
CH ₃ -N (all)		CH (peptide)	
CO (all)	OH (all)	CH-N (all)	CH-X (arom) ^d
CO (ester/lactone)	OH (alcohol)		CH-X ₂ (arom) ^e
COOH	OH (phenol)	benz ^c 1-monosubst	CH ₃ NCH ₂ ^f
	OH (acidic)	benz 1,2-disubst	CH ₃ NCH ^f
		benz 1,3-disubst	CH ₂ NCH ₂ ^f
		benz 1,4-disubst	CH ₂ NCH ^f
		benz 1,2,3-trisubst	CH ₃ NC=O ^f
		benz 1,3,5-trisubst	CH ₃ OC=O ^f
		benz 1,2,4-trisubst	CH ₃ C-C=O ^f
		benz 1,2,3,4-tetrasubst	NH (all)
		benz 1,2,3,5-tetrasubst	NH ₂ (all)
		benz 1,2,4,5-tetrasubst	NH (arom)
		benz 1,2,3,4,5-pentasubst	

^aCq = quaternary carbon. ^bAr = aromatic ring. ^cbenz = benzene ring. ^d¹J_{CH} < 200 Hz. ^e¹J_{CH} > 200 Hz. ^fHMBC experiments.

formula of C₃₀H₄₂O₇. Analysis of the integrals obtained from a ¹H NMR spectrum in combination with correlations observed in an edited HSQC spectrum (Figure 3) highlighted the presence of eight methyl groups, five methylenes, and seven methines.

Table 3. Comparison of Results Obtained with DEREPP-NP Using the Examples Described by Bitzer et al.,²² Bradshaw et al.,²¹ and Lang et al.⁶

example	step	criteria	number of hits	
			literature ^{a-c}	DEREP-NP
1 ^a	1	MW 333; 3 CH ₃ ; 2 CH ₂ ; 8 CH	narcissidin	narcissidin, <i>epi</i> -narcissidin, 11-hydroxy galanthine
2 ^a	1	MW 250; 3 CH ₃ ; 4 CH ₂ ; 4 CH	40	72
	2	1 CH sp ²	15	29
	3	1 CH ₂ -O	9	12
	4	3 CH ₃ singlets	3 (1 compatible with signals below 1.1 ppm)	9 (4 compatible with signals below 1.1 ppm)
3 ^b	1	MW 266; 1 CH ₃ ; 1 CH ₃ O; 2 CH ₂ sp ³ ; 3 CH sp ²	diaportinol	diaportinol and pestalasin D
4 ^b	1	MW 493; 4 CH ₃ ; 1 CH ₂ sp ²	cytochalasin H	cytochalasin H
5 ^c	1	4 CH ₃ ; 1 OCH ₃ ; 1 H-C≡O	23	106
	2	0 CH ₂ , 0 CH (sp ³)	11	25
	3	one 1,2,3,5-tetrasubst benzene ring	phomosine	phomosine A and thamnoliadepside C
6 ^c	1	7 CH ₃	327	14 104
	2	1 CH ₃ (acetyl)	34	1562
	3	2 CH ₃ (vinyl)	5	180
	4	MW 502	NF-00659-A3	NF-00659-A3 and calophinone-6-O-acetate

^aNatural Products Database²¹ with 126 000 natural products. ^bPrivate database²² with 15 000 unique natural products. ^cAntiMarin database⁶ with 47 000 unique compounds.

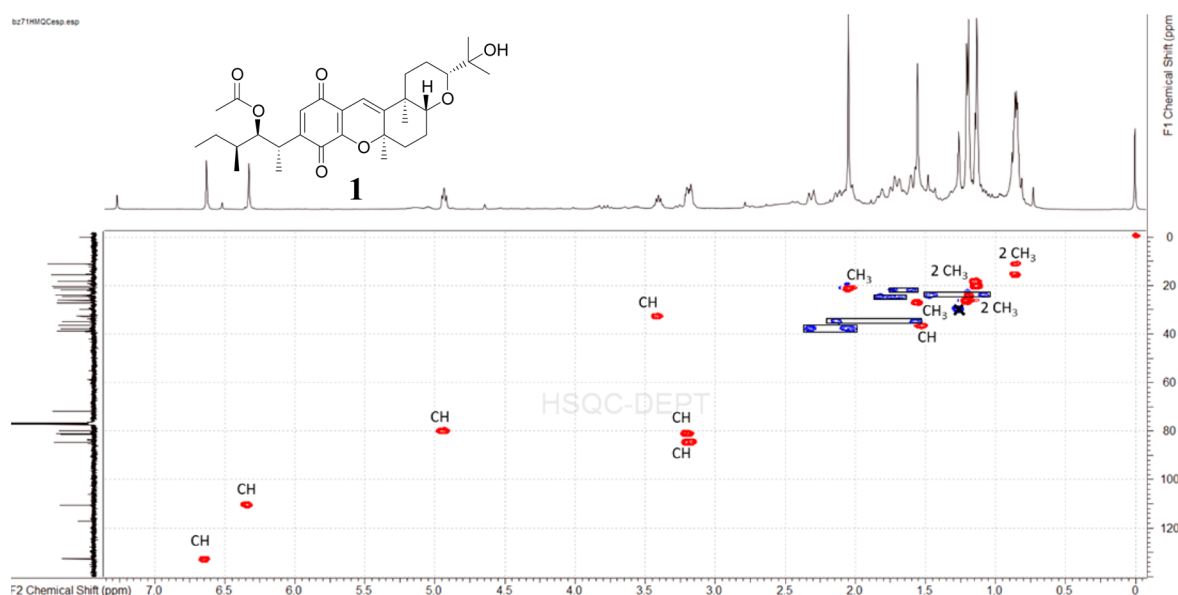


Figure 3. HSQC contour map of **1** (400 MHz, CDCl₃). CH₂ cross-peak pairs are enclosed in rectangles. (See Figure S1, Supporting Information, for an expansion of the high-field region.)

A molecular formula search in DEREP-NP resulted in 90 matches, and restricting the search to structures with the above-mentioned number of methyl, methylene and methine groups reduced the matches to just one compound, anhydrococlinoquinone A (**1**). Comparison of the complete set of NMR and mass spectra with those published for this compound confirmed its identity.²⁸

For those researchers who lack ready access to mass spectrometry facilities, ¹H/HSQC NMR data alone can provide sufficient evidence to obtain hits to structures present in the database. Extracts from a plant in the genus *Corymbia* (Myrtaceae) yielded NMR data from semipurified fractions that contained signals for eight methyl singlets, an additional six methyls, and a phenolic proton (Figure S2, Supporting Information). A search of DEREP-NP using the three criteria (8 CH₃ (singlet), 14 CH₃ (all), and 1 OH (phenol)) yielded nine hits. Since the HSQC data also suggested the molecule

contained three methylenes, supplementing the above search with the addition of CH₂ (all) = 3 reduced the hits to one compound, rhodomertosone C (**2**).²⁹ Comparison of the ¹H and HSQC NMR data with those reported in the literature confirmed the compound to be rhodomertosone C.

An extract from the leaves of a *Triunia* species (Proteaceae) yielded a fraction containing a compound with three methyls, four methines, three methylenes, and no aromatic protons (Figure S3, Supporting Information). A search of DEREP-NP using these four criteria yielded 458 hits. The ¹H NMR spectrum indicated that two of the methyls are doublets, the HSQC spectrum (Figure S4, Supporting Information) suggested that one of the methyls is attached to a nitrogen and one of the methines is oxygenated, while the HMBC spectrum (Figure S5, Supporting Information) indicated the molecule also contains a ketone carbonyl adjacent to one of the methyl groups. Adding these criteria (CH₃ (doublet) = 2,

Chart 1

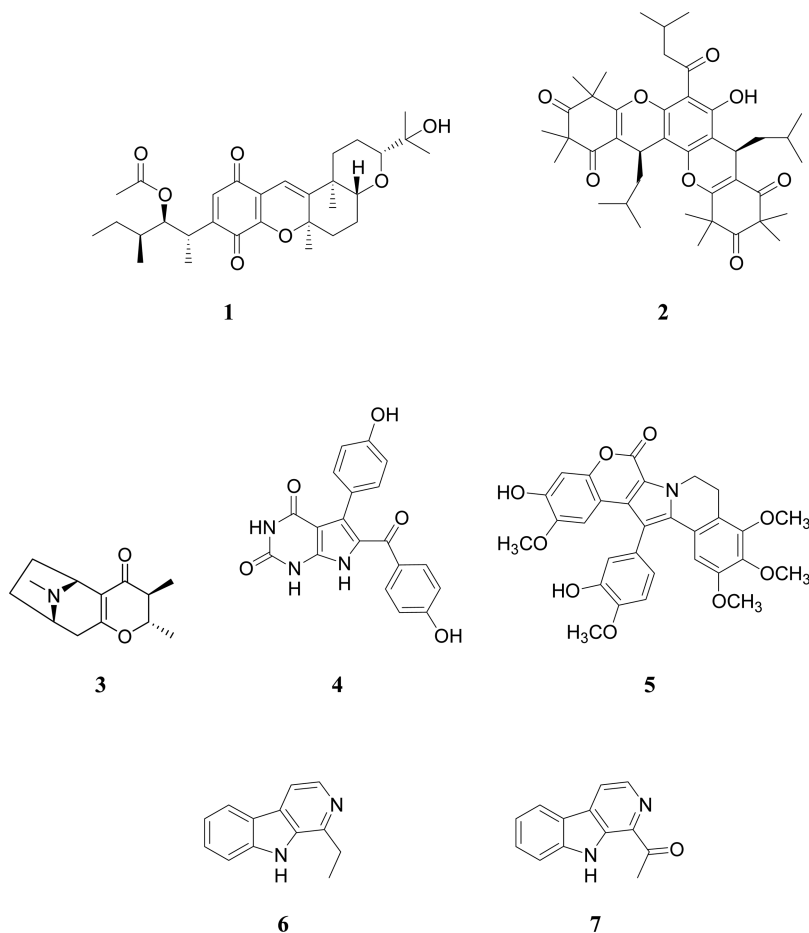


Figure 4. ¹H NMR spectrum of a 3:1 mixture of 6 and 7 (600 MHz, DMSO-d₆).

CH₃N = 1, CH–O (all) = 1, and CO (all) = 1) to the search reduced the hits to one compound, 2,3-dihydrodarlingine (3), and comparison of its NMR data with the literature confirmed its structure.³⁰

Many marine natural products are proton poor and can often contain isolated proton spin systems. However, even these

groups of molecules can be identified with a limited number of search criteria. A blue-colored fraction obtained from purification of an extract from the marine ascidian *Eudistoma glaucum* gave a ¹H NMR spectrum that contained a mixture of three compounds in a ratio of 1:2:12 (Figure S6, [Supporting Information](#)). The negative-mode ESIMS exhibited a major [M

– H]⁺ ion at m/z 362 consistent with a MW of 363. The ¹H and HSQC NMR data indicated the major compound to have two 1,4-disubstituted aromatic systems and five downfield protons not attached to carbons. A DERE- NP search using benz-1,4-disubst = 2 and MW = 363 yielded 10 hits. Adding CH₂ (all) = 0, CH₃ (all) = 0, and CH (all) = 8 retrieved only the structure of rigidin (4),³¹ and comparison with literature data confirmed the major compound to have this identity. There were not enough distinguishing features in the ¹H NMR spectrum to allow for database searching of the two minor compounds in the mixture.

Another ascidian produced several fractions containing aromatic compounds with few ¹H NMR signals. One compound was found to contain a 1,2,4-trisubstituted aromatic system and three aromatic singlets (Figure S7, [Supporting Information](#)). The molecule also contained five methoxy group proton singlets and two methylenes. A DERE- NP search using benz-1,2,4-trisubst = 1, CH arom (singlet) = 3, CH₂ (sp³) = 2, and CH₃-O-Ar (methoxy) = 5 yielded eight hits, all of which were lamellarin-type alkaloids.³² Comparison of the ¹H NMR data obtained with that reported in the literature for the eight hits established that the compound isolated was lamellarin T (5).³³

There has been a trend over recent years to produce libraries of natural product fractions for biological screening. This approach provides advantages over screening crude extracts since the libraries can be tailored to comply with “lead-like” properties such as logP, and each fraction is likely to contain only a small number of individual compounds.^{34,35} These fractions are therefore also well suited for dereplication using the above-described technique. A fraction from a bryozoan extract contained two compounds in a ratio of ~3:1 (Figure 4). MS analysis indicated the two compounds had MWs of 196 and 210. Both compounds contained aromatic NH, a 1,2-disubstituted aromatic, and two additional aromatic protons. The major compound contained an ethyl group and the minor compound contained a methyl singlet, which showed a correlation to a ketone carbon in the HMBC spectrum. A search of DERE- NP with the above criteria for each compound yielded only two compounds, 1-ethyl-9H-beta-carboline (6) and 1-(9H-beta-carboline-1-yl)ethanone (7).^{36,37}

The examples given above demonstrate the utility of the DERE- NP database and search tools to identify known NPs rapidly using a limited set of spectroscopic data. This approach is limited to the evaluation of spectra of purified compounds or spectra containing a mixture of a small number of compounds. Although many organisms can contain many thousands of individual small molecules in concentrations varying from >1% to less than 0.0001% dry weight, the use of a small number of separation steps can generally yield fractions containing a small number of components in quantities visible by NMR spectroscopy, as demonstrated by Lang and co-workers.⁶ Since modern NMR spectrometers can be used to generate ¹H and HSQC NMR data on submilligram quantities of material in minutes, this technique should be able to provide a quick assessment of the uniqueness of purified or semipurified natural product fractions and thus provide a useful starting point for any natural product discovery process. By limiting the search criteria to only definitive features (and spreading the net wide where ambiguous data occur, especially in crowded regions of NMR spectra) a manageable number of “hits” can still be obtained. Once “hits” have been obtained, however, the identity of the compound still requires validation though

comparison with published spectroscopic data. The complexity of extracts obtained from diverse organisms still provides researchers with a challenge to distinguish known compounds from new and novel metabolites. This is particularly the case in the study of herbal medicines, where the bioactive ingredients are often unknown. The approach outlined in the study is not intended to be used to identify new and novel chemistry, but to reduce the time and resources spent on reidentifying known chemistry. Researchers in the field of metabolomics have tried to address this discovery bottleneck through the generation of MS- and NMR-based databases. These databases however have generally targeted primary metabolites of interest as biomarkers of disease and environmental disturbance or essential oils present in plants. When applied to natural product research, where compound diversity is high and generally selective for specific biota, the generation of spectroscopic databases for metabolomics research is a major challenge. As mentioned earlier, ¹³C NMR spectroscopic natural product databases have been generated, but these are of limited utility due to the small number of compounds present and a reliance on the acquisition of ¹³C NMR data for searching purposes. There are many limitations to this approach including standardization of a specific NMR solvent used to acquire all data and the challenge of reisolating all published NPs to generate data for inclusion into such a database. An alternative approach would be to generate NMR spectroscopic databases from the published literature and to make these freely available. Searching such databases would still be problematic if spectra were acquired in different solvents than those used to generate the database. On the other hand, structural features that can be derived from NMR analysis, but which do not rely on a specific search of chemical shift data, provide a more powerful tool to identify potential matches to known structures. The protocols outlined in this study have demonstrated that this approach can be an effective tool to identify published NPs, and by including 65 searchable structural (small substructures) features, this open access database is a powerful tool for dereplication. Its usefulness was compared with similar approaches previously described and found to yield similar results, but, since DERE- NP contains 229 358 natural product structures, its added benefit is that it includes a much larger number of compounds isolated from terrestrial and marine macro- and microorganisms published before 2013 than previously described databases. The methodology described above will encourage researchers to find structural alternatives when no verifiable match (through literature comparisons of spectroscopic data with database “hits”) is obtained. The methodology therefore provides a powerful tool to aid and encourage new and novel bioactive natural product discovery. DERE- NP is freely available and can be used on desktop computers running open source software. It can be used as is or improved by the users to fit their needs. To this end, DERE- NP is deposited at <https://github.com/clzani/DEREP-NP>, where new versions will be included as the database evolves and interested researchers can upload their own versions and make requests for corrections and additions.

■ EXPERIMENTAL SECTION

General Experimental Procedures. The UNPD files containing structures (SDF format) and associated data (csv file format) of 229 358 NP structures were downloaded from the Web site indicated by Gu et al.⁴ (<http://pkuxj.pku.edu.cn/UNPD>). These files were processed using KNIME²⁵ (see [Supporting Information](#) for details) in

order to merge and select only the desired information, namely, 2D and 3D structures, available names and CAS numbers, InChI, InChI Key, SMILES, canonical SMILES, molecular formula (MF), and exact molecular weight (MW) calculated from the monoisotopic masses using the mass of the most abundant isotope of each element. The 65 structural features (Table 2) were chosen because they can be easily recognized using ^1H NMR and HSQC spectra and correlations from intense proton signals such as methyl singlets in HMBC spectra. These features were translated into SMARTS queries using the SMARTSEditor²⁶ and processed in a KNIME workflow (Figure S8, Supporting Information) to determine the number of occurrences of each feature in each structure of the UNPD data set. An SDF file containing the original information (structure, MF, etc.) plus the count number of each of the NMR features was generated. This SDF file, named DEREP-NP, can be read by any chemistry-aware software. Adhering to the principle of bringing a freely accessible platform, we used DataWarrior, an open-source software for chemical data visualization and analysis.²⁷ All procedures were run on a MacMini 6.2 computer equipped with an Intel Core i7 processor working at 2.6 GHz, 16 GB RAM and a GPU Intel HD Graphics 4000.

Since correlations between protons and carbons can be distinguished even when the resolution in the carbon dimension in HSQC and HMBC spectra is low, these spectra can be acquired successfully for dereplication purposes even with a moderately small number of experiments (or increments, usually less than 100). Furthermore, most protons have a relatively short spin–lattice relaxation time (T_1), and therefore the delay between scans (or transients) can be reduced to less than a second (0.8 s is a good compromise). By applying both approaches, HSQC and HMBC spectra can successfully be obtained in less than 5 min even for compounds isolated in 1 mg quantities on a high-field NMR spectrometer.

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jnatprod.6b01093.

^1H and HSQC NMR spectra for compounds 2–5, tips for using the database as well the KNIME workflow and node settings used to count the substructures (PDF) DEREP-NP-v1 file (RAR)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +55 31 3349 7791. Fax: +55 31 3295 3115. E-mail: zani@cpqrr.fiocruz.br.

ORCID

Carlos L. Zani: 0000-0003-1859-177X

Notes

The authors declare no competing financial interest. Updated versions of the DEREP-NP file in DataWarrior format (dwr) can be downloaded at <https://github.com/clzani/DEREP-NP>; Osiris DataWarrior can be downloaded for free at <http://openmolecules.org/datawarrior/download.html>

■ ACKNOWLEDGMENTS

We are grateful to G. A. Landrum for help with substructure counting using KNIME, to Oswaldo Cruz Foundation–FIOCRUZ for financial support, and to CNPq for a Science without Borders Program Senior Training Fellowship for C.L.Z. in the laboratory of A.C.

■ REFERENCES

- (1) Molinski, T. F. *Org. Lett.* **2014**, *16*, 3849–3855.
- (2) Drew, S. W.; Demain, A. L. *Annu. Rev. Microbiol.* **1977**, *31*, 343–356.
- (3) Nicolaou, K. C.; Snyder, S. A. *Angew. Chem., Int. Ed.* **2005**, *44*, 1012–1044.
- (4) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. *PLoS One* **2013**, *8*, e62839.
- (5) Hanka, L. J.; Kuentzel, S. L.; Martin, D. G.; Wiley, P. F.; Neil, G. L. *Recent Results Cancer Res.* **1978**, *63*, 69–76.
- (6) Lang, G.; Mayhudin, N. A.; Mitova, M. I.; Sun, L.; Van Der Sar, S.; Blunt, J. W.; Cole, A. L. J.; Ellis, G.; Laatsch, H.; Munro, M. H. G. *J. Nat. Prod.* **2008**, *71*, 1595–1599.
- (7) Mohimani, H.; Pevzner, P. A. *Nat. Prod. Rep.* **2016**, *33*, 73–86.
- (8) Gaudencio, S.; Pereira, F. *Nat. Prod. Rep.* **2015**, *32*, 779–810.
- (9) Hubert, J.; Nuzillard, J.-M.; Renault, J.-H. *Phytochem. Rev.* **2017**, *16*, 55–95.
- (10) Tawfik, A.; Tawfik, N.; Edrada-Ebel, R. *Planta Med.* **2015**, *81*, 1403–1404.
- (11) Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. *Planta Med.* **2014**, *80*, 831–832.
- (12) Ito, T.; Masubuchi, M. *J. Antibiot.* **2014**, *67*, 353–360.
- (13) Miao, Z.; Jin, M.; Liu, X.; Guo, W.; Jin, X.; Liu, H.; Wang, Y. *Anal. Bioanal. Chem.* **2015**, *407*, 3405–3416.
- (14) Kokkotou, K.; Ioannou, E.; Nomikou, M.; Pitterl, F.; Vonaparti, A.; Siapi, E.; Zervou, M.; Roussis, V. *Phytochemistry* **2014**, *108*, 208–219.
- (15) Staerk, D.; Kesting, J. R.; Sairafianpour, M.; Witt, M.; Asili, J.; Emami, S. A.; Jaroszewski, J. W. *Phytochemistry* **2009**, *70*, 1055–1061.
- (16) Guo, W.; Jin, M.; Miao, Z.; Qu, K.; Liu, X.; Zhang, P.; Qin, H.; Zhu, H.; Wang, Y. *PLoS One* **2015**, *10*, e0127583.
- (17) Rochfort, S. J. *Nat. Prod.* **2005**, *68*, 1813–1820.
- (18) Luis Lopez-Perez, J.; Theron, R.; del Olmo, E.; Santos-Buitrago, B.; Francisco Adserias, J.; Estevez, C.; Garcia Cuadrado, C.; Eguiluz Lopez, D.; Santos-Garcia, G. *8th Int. Conf. Pract. Appl. Comput. Biol. Bioinf* **2014**, *294*, 9–19.
- (19) Lopez-Perez, J. L.; Theron, R.; del Olmo, E.; Diaz, D. *Bioinformatics* **2007**, *23*, 3256–3257.
- (20) Asakura, K. *Yuki Gosei Kagaku Kyokaiishi* **2015**, *73*, 1247–1252.
- (21) Bradshaw, J.; Butina, D.; Dunn, A. J.; Green, R. H.; Hajek, M.; Jones, M. M.; Lindon, J. C.; Sidebottom, P. J. *J. Nat. Prod.* **2001**, *64*, 1541–1544.
- (22) Bitzer, J.; Koepcke, B.; Stadler, M.; Heilwig, V.; Ju, Y.-M.; Seip, S.; Henkel, T. *Chimia* **2007**, *61*, 332–338.
- (23) Blunt, J.; Munro, M.; Upjohn, M. In *Handbook of Marine Natural Products*; Springer: Dordrecht, The Netherlands, 2012; Chapter 6, pp 389–421.
- (24) Williams, R.; O’Neil-Johnson, M.; Williams, A.; Wheeler, P.; Pol, R.; Moser, A. *Org. Biomol. Chem.* **2015**, *13*, 9957–9962.
- (25) Berthold, M. R.; Hebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinel, T.; Ohm, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner*; Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation; Albert-Ludwigs-Universität, Freiburg, March 7–9, 2007; Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer-Verlag: Berlin Heidelberg, 2008; pp 319–326.
- (26) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. *J. Chem. Inf. Model.* **2015**, *55*, 460–473.
- (27) Schomburg, K. T.; Wetzler, L.; Rarey, M. *Drug Discovery Today* **2013**, *18*, 651–658.
- (28) Campos, F. F.; Rosa, L. H.; Cota, B. B.; Caligiorno, R. B.; Rabello, A. L. T.; Alves, T. M. A.; Rosa, C. A.; Zani, C. L. *PLoS Neglected Trop. Dis.* **2008**, *2*, e348.
- (29) Hiranrat, A.; Mahabusarakam, W. *Tetrahedron* **2008**, *64*, 11193–11193.
- (30) Yang, F.; Zhao, H.; Carroll, A. R. *Tetrahedron Lett.* **2017**, *58*, 736–689.
- (31) Kobayashi, J.; Jie-Fei Cheng, J.; Kikuchi, Y.; Ishibashi, M.; Yamamura, S.; Ohizumi, Y.; Ohtac, T.; Nozoe, S. *Tetrahedron Lett.* **1990**, *31*, 4617–4620.

- (32) Carroll, A. R.; Bowden, B. F.; Coll, J. C. *Aust. J. Chem.* **1993**, *46*, 489–501.
- (33) Ploypradith, P.; Petchmanee, T.; Sahakitpichan, P.; Litvinas, N. D.; Ruchirawat, S. *J. Org. Chem.* **2006**, *71*, 9440–9448.
- (34) Camp, D.; Campitelli, M.; Carroll, A. R.; Davis, R. A.; Ebdon, J.; Quinn, R. J. *Chem. Biodiversity* **2013**, *10*, 524–537.
- (35) Camp, D.; Davis, R. A.; Campitelli, M.; Ebdon, J.; Quinn, R. J. *J. Nat. Prod.* **2012**, *75*, 72–81.
- (36) Prinsep, M. R.; Blunt, J. W.; Munro, M. H. *J. Nat. Prod.* **1991**, *54*, 1068–1076.
- (37) Huang, H.; Yao, Y.; He, Z.; Yang, T.; Ma, J.; Tian, X.; Li, Y.; Huang, C.; Chen, X.; Li, W.; Zhang, S.; Zhang, C.; Ju, J. *J. Nat. Prod.* **2011**, *74*, 2122–2127.