JOURNAL OF NATURAL PRODUCTS

Article

Cite This: J. Nat. Prod. 2019, 82, 1258−1263

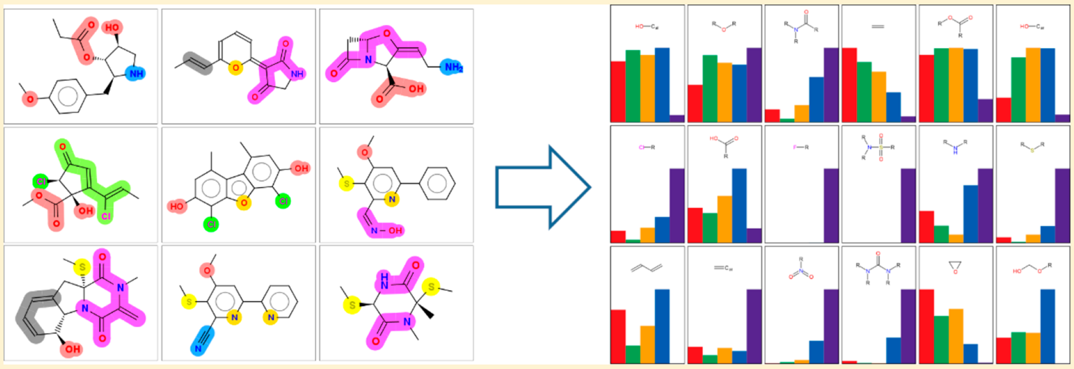
pubs.acs.org/jnp

A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products

Peter Ertl* and Tim Schuhmann

Novartis Institutes for BioMedical Research, CH-4056, Basel, Switzerland

Supporting Information



ABSTRACT: The two most striking features that discriminate natural products from synthetic molecules are their characteristic scaffolds and unique functional groups (FGs). In this study we systematically investigate the distribution of FGs in natural products from a cheminformatics perspective by comparing FG frequencies in natural products with those found in average synthetic molecules. We thereby aim for the identification of FGs that are characteristic for molecules produced by living organisms. In our analysis we also include information about the natural origins of the structures investigated, allowing us to link the occurrence of specific FGs to the individual producing species. Our findings have the potential for being applied in a medicinal chemistry context concerning the synthesis of natural product-like libraries and natural product-inspired fragment collections. The results may be used also to support compound derivatization strategies and the design of "non-natural" natural products.

The concept of functional groups (FGs)—sets of connected atoms that determine properties and reactivity of the parent molecule—forms a cornerstone of organic chemistry, medicinal chemistry, toxicity assessment, and spectroscopy. FGs are key to establish specific interactions with target proteins via the formation of hydrogen bonds and other weaker interactions. FGs also determine the overall bioavailability and metabolic stability of the parent molecules. The presence of unique FGs in natural products (NPs) is one of the most important features that make this class of molecules so distinct. Numerous scientific studies analyzing differences between NPs and synthetic molecules focused so far mostly on physicochemical properties[1,2] and skeletal features such as rings or scaffolds.[3,4] No study has yet been carried out with a focus on investigating the distribution of FGs in NPs. The main reason for this was a lack of available software tools capable of performing such analyses. All currently accessible tools for FG analysis use a predefined list of substructures, a process that works well for standard organic molecules but that is not capable of assessing special classes of molecules with unique structural features such as NPs. In this study we analyze the distribution of FGs in NPs using a newly developed procedure that is generally applicable and able to identify all FGs without the need of having a predefined list of substructures.

■ COMPUTATIONAL METHODOLOGY

The actual extraction of FGs from molecular structures has been done by a procedure developed in-house at Novartis.[5] The method is fully described in ref 5; therefore only a brief summary will be provided here. The algorithm is based on a walk through all nonaromatic atoms in the molecule. Once a heteroatom is found, all neighboring heteroatoms and atoms connected by multiple bonds are recursively added to form a functional group. Additionally, multiple carbon−carbon bonds as well as oxirane, aziridine, and thiirane rings are considered to be functional groups. For a defined list of common FGs, information about the parent carbon is also retained to be able to distinguish between alcohols and phenols or amines and anilines. Aromatic heteroatoms are collected as single atoms, not as part of a larger system. They are extended to a larger FG only when there is an aliphatic functionality connected (for example an acyl group connected to a pyrrole nitrogen).

Received: December 5, 2018
Published: April 1, 2019

DOI: 10.1021/acs.jnatprod.8b01022

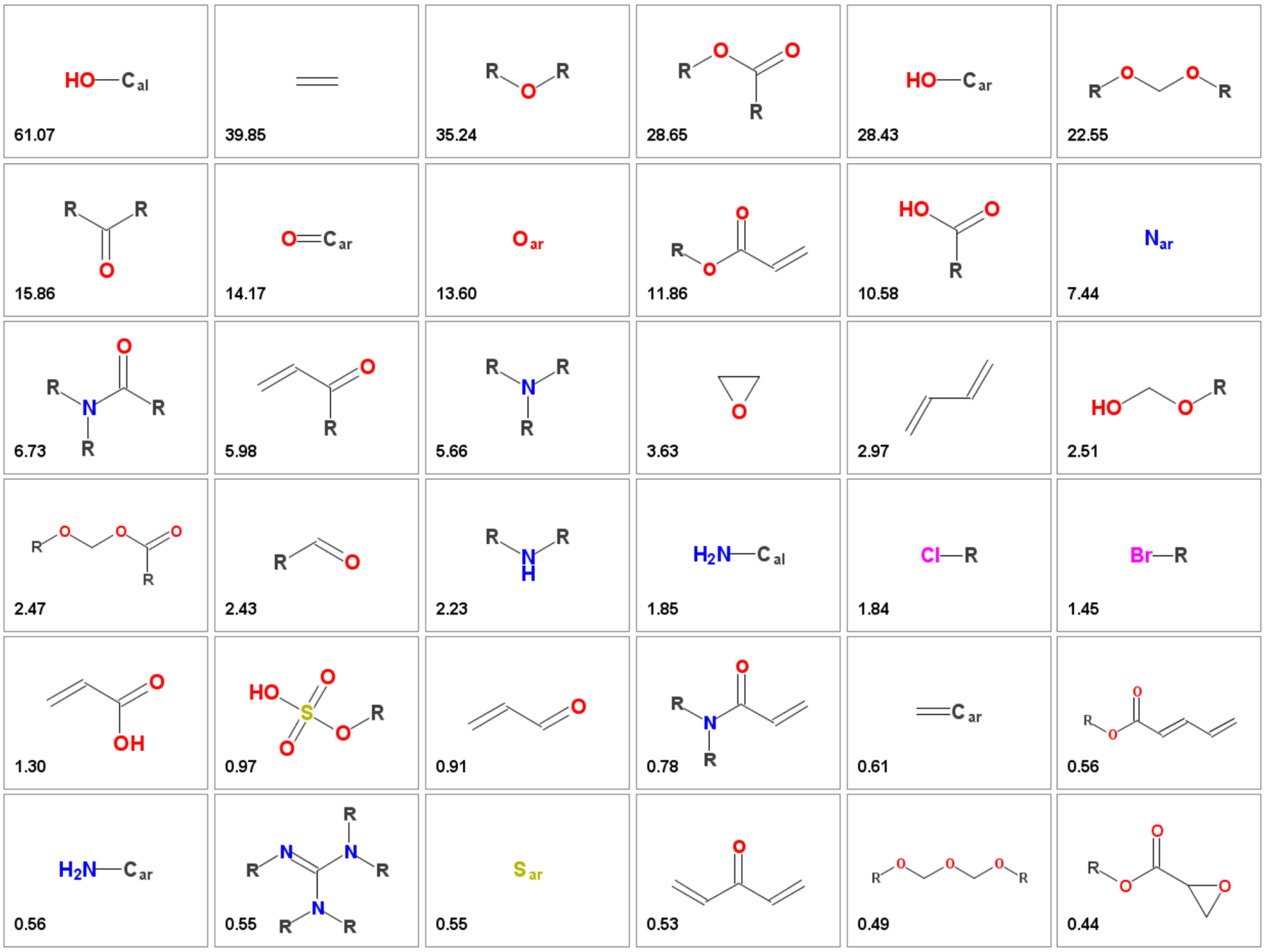


Figure 1. Most frequent functional groups found in natural products. The number indicates percentage of molecules having this group.

Heteroatoms in heterocycles are traditionally not considered to be "classical" FGs by themselves but to be part of the whole heterocyclic system.

Several molecular databases were analyzed in this study. The NPs were represented by the molecules from The Dictionary of Natural Products (DNP).[6] Only the NPs where the producing organism could be identified were included in the analysis (altogether over 186 000 molecules). The NP-producing organisms were identified with the help of the Taxonomy Database of the National Center for Biotechnology Information.[7] A Python script was used to analyze the information from the biological source field (BSRC) of DNP, to identify the scientific name if present and to match it to the entry in the taxonomy database. This allowed assigning the source organism to one of four classes (animals, plants, fungi, or bacteria). If no scientific name could be found, another program written in Python examined the words given in the source description field and compared them with the keywords typical for particular classes. For example "extracted from leaves" identified the origin as a plant, or "isolated from marine sponge" as an animal (see also the Supporting Information for details of this procedure and the list of keywords). Particularly the class of "animals" is highly heterogeneous, including species ranging from mammals to marine invertebrates, and the results need to be interpreted with this in mind.

In addition to the commercial DNP database, a set of open NP structures also has been analyzed. This set consisted of 21 000 fungal and microbial metabolites from the Natural Products Atlas[8] and 53 000 metabolites from the TCM Database@Taiwan[9] containing mostly plant metabolites. After cleaning and normalization (as described below) this collection that we call the Open Natural Products collection (OpenNP) contained 67 000 molecules.

The synthetic molecules were represented by the commercially available samples from the ZINC database[10] containing over 13 million structures. These molecules represent well the chemical space of "common" synthetic molecules.

Before the actual extraction of FGs all molecules were normalized, generating the standard SMILES as molecule representation. The normalization included removal of structures with valence errors, neutralization of atomic charges, and removal of counterions. At this stage duplicate structures were also removed. Such normalization may be performed by using the open source cheminformatics toolkit RDKit.[11] The actual extraction of FGs has been done by an in-house toolkit written in Java.[5] In the meantime, open source versions of this protocol have been made available in Python[12] and also in Java;[34] therefore the whole computational procedure can now be performed using open source software.

DOI: 10.1021/acs.jnatprod.8b01022
J. Nat. Prod. 2019, 82, 1258−1263

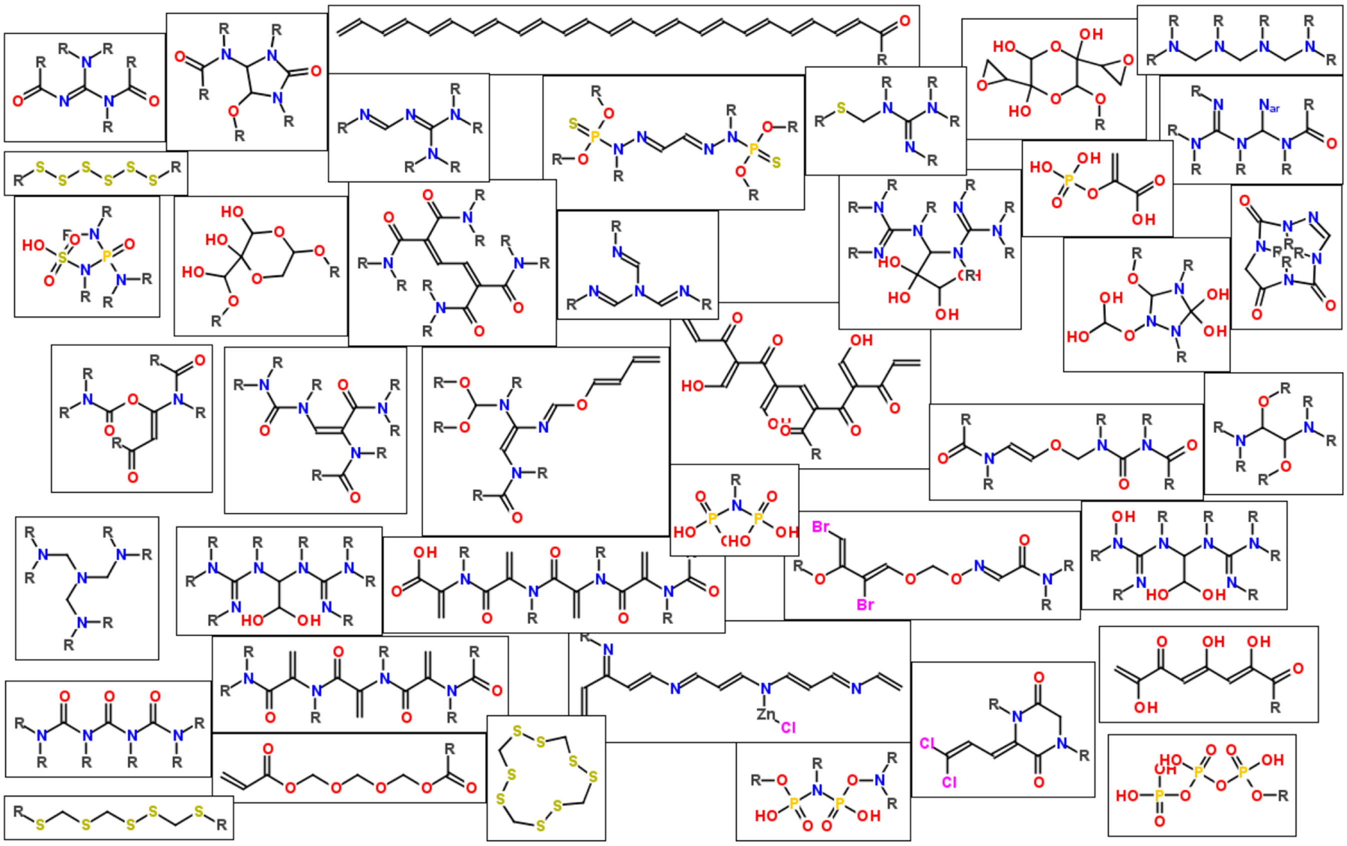


Figure 2. Examples of exotic functional groups found in natural products (randomly ordered).

RESULTS AND DISCUSSION

Cheminformatics analysis of over 186 000 NPs from the DNP identified 2785 unique FGs. They show a typical power law (or "long tail") distribution with few common groups and a large number of infrequent groups. Twenty-five FGs are present in more than 1% of the molecules and 93 in more than 0.1% of molecules. Over 1000 groups (1214, 43.6%) are singletons (present in only one molecule). The most frequent FG in NPs is the alcohol hydroxy group, which is present in 61.1% of molecules, followed by alkene (39.9%), ether (35.2%), ester (28.7%), and the phenolic hydroxy group (28.4%). The 36 most frequent FGs from NPs are shown in Figure 1.

On the lower end of the frequency distribution one can find several quite exotic FGs (Figure 2). These examples nicely document the enormous structural diversity that biosynthetic machineries involved in the synthesis of these molecules can produce (and also illustrates the challenges natural product chemists have to cope with).

Analysis of the OpenNP collection (67 000 molecules) provided very similar results to those obtained for DNP besides differences in the distribution of a few FGs, particularly alkynes and secondary amines. The fact that the TCM Database@Taiwan contains molecules that are not present in other NP databases has already been reported by Kramer et al.[13] Frequencies of specific FG occurrences in DNP, the OpenNP, and synthetic molecule collections are provided in the Supporting Information.

One of the major goals of our study was to analyze the differences between FGs present in NPs vs those occurring in synthetic molecules. To achieve this, we compared the FG distribution of 186 000 molecules present in the DNP with 13 million synthetic molecules from catalogues of commercial compound providers. The differences for the 50 most frequent FGs from each set (69 unique FGs totally) are shown in Figure 3. The horizontal axis in the diagram represents FG frequency (the most common groups are on the left side, less common on the right), whereas the vertical axis indicates the propensity of FGs for NPs (green area at the top) or synthetic molecules (red area at the bottom). The figure nicely visualizes the differences between the two sets. The FGs typical for NPs contain mostly oxygen atoms (hydroxy, ester, peroxide, polyglycol, epoxide rings), ethylene-derived groups and various *α*,*β*-unsaturated systems. In synthetic molecules nitrogen-containing and chemically more easily accessible FGs are over-represented, such as amide, urea, sulfonamide, sulfone, or imide functionalities and substituents such as fluoro or nitro.

The fact that oxygen atoms are part of most FGs in NPs can be attributed to several reasons. Oxygen is the most abundant element in nature and has, since the onset of aerobic life hundreds of millions of years ago, been easily accessible to organisms from the atmosphere and finally to the evolution of their biosynthetic machineries.[14] Molecular oxygen requires—in contrast to nitrogen fixation—far less energy for chemical activation and can thus be incorporated into organic molecules in an efficient manner. Polyketide synthases, the source for a variety of structures produced by bacteria and fungi, introduce oxygen as acetyl- and malonyl-CoA via catalyzing a decarboxylative Claisen condensation to incorporate these extender units into growing polyketide chains. The thus introduced keto functionalities can subsequently be converted into alcohols stereospecifically or can be further reduced to alkenes or finally even fully saturated systems.[15] Plants, and to

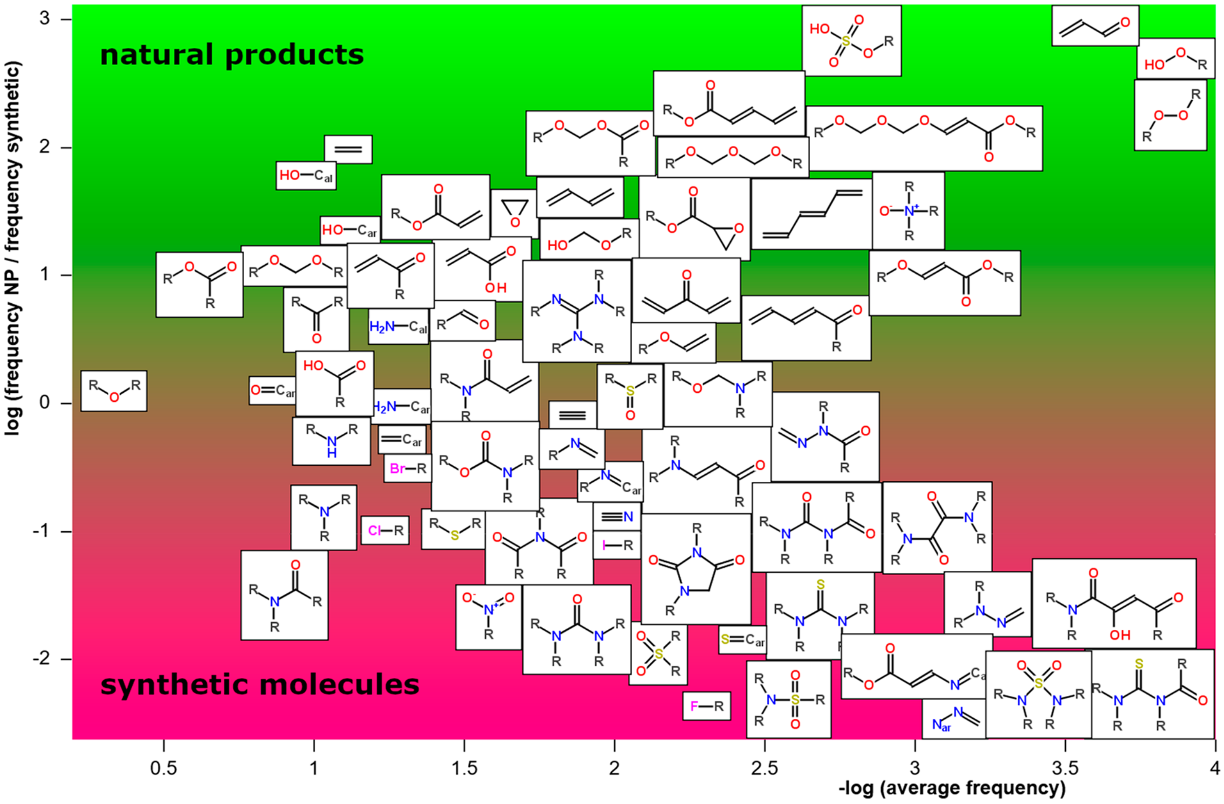


Figure 3. Plot of common functional groups displaying their preference for natural products (green area) and synthetic molecules (red area) expressed as logarithm of the ratio between their frequencies in NPs and synthetic molecules. Position on the horizontal axis is proportional to the frequency of functional groups expressed as negative logarithm of their average frequencies. The most common groups are on the left and less common on the right.

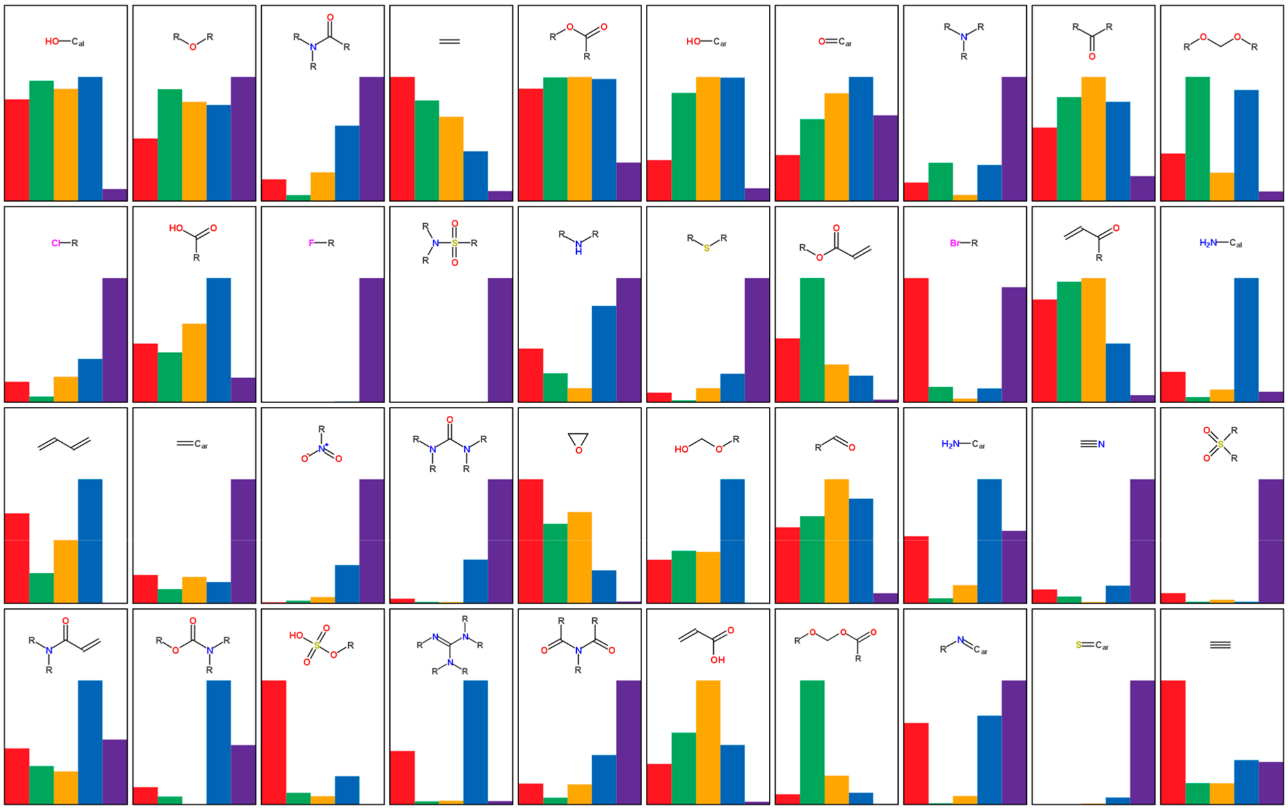


Figure 4. Relative frequencies of common functional groups from natural products produced by animals (red), plants (green), fungi (orange), and bacteria (blue) and those present in synthetic molecules (magenta).

some extent also fungi and bacteria, harness the mevalonate and methylerythritol pathways to produce terpenoids, another very diverse class of secondary metabolites. Tailoring enzymes such as mono- and dioxygenases as well as epoxidases are used to introduce oxygen into these scaffolds, mainly resulting in hydroxy, keto, and/or aldehyde functionalities. Alkaloids, on the other side, are usually biosynthesized from amino acids, which serve as a direct source of oxygen and nitrogen atoms for these plant metabolites. In addition, some hydroxy moieties present in secondary metabolites can also have their origin in sugars that are transferred onto terpene, polyketide, and alkaloid scaffolds via dedicated glycosyl transferases.

We were also interested in comparing the distribution of FGs among NPs in relation to their producing organism classes: animals, plants, fungi, or bacteria. For all of these four classes we calculated frequencies of common functional groups and compared them to synthetic molecules. The results of this analysis are shown in Figure 4. Heights of the bars in this graph are relative to the class a particular FG most frequently occurs in. Whereas for some FGs (alcohols, esters, keto groups, aldehydes) their distribution is very similar for all types of source organisms, in most cases one can see clear differences. The distribution of FGs in synthetic molecules differs significantly from NPs, as noted already in the previous section. This holds in particular true for amide, fluorine, cyano, nitro, sulfonamide, and sulfone groups. Detailed information about differences in FG occurrence between different types of source organism is available in the Supporting Information.

One can use the distribution of FGs in NPs produced by different classes of organisms also for an estimation of the similarity between these four sets. Results of this analysis are shown in Table 1. The relationship between different groups was calculated as a cosine similarity between vectors representing the percentage of FG occurrences in the different classes (details in the Supporting Information). According to this measure, NPs produced by plants and fungi are most closely related, followed by bacterial and fungal NPs. This goes along with the taxonomic difference of these species, with fungi and plants being eukaryotes and bacteria belonging to the prokaryotic world.

Table 1. Similarity between Different Classes of Natural Products and Synthetic Molecules Calculated Based on the Occurrence of Functional Groups[a]

source	plants	fungi	bacteria	synthetic
animals (21 007)	56.9	57.4	48.6	20.7
plants (133 480)		68.0	55.3	20.1
fungi (18 412)			58.9	21.9
bacteria (13 575)				28.9

[a] The figures in parentheses indicate number of molecules included in the analysis.

Despite basic similarity in biosynthetic logic between fungi and bacteria, one can see several clear differences in their FGs (Figure 4). Bacteria metabolites contain more nitrogen-containing FGs (primary and secondary amines, guanidines, and urethanes).

The large majority of the roughly 140 000 plant-derived secondary metabolites published to date either belong to the isoprenoid family or are derived from the shikimic acid biosynthesis pathway, a pattern that shows some similarity also to fungal metabolites. Fungi, however, produce a variety of peptidic structures via the NRPS pathway not present in plants and thus bridge the gap to the bacterial world. From the latter, far fewer terpene-derived secondary metabolites are reported, whereas especially actinomycetes possess a rich polyketide (PKS) metabolism. With PKS II-derived metabolites showing similar structural features for both bacteria and fungi, the relatively small PKS I enzymes in fungi process precursors iteratively in contrast to bacterial PKS I, where extender units are incorporated by individual modules.[16] As a result, one can find larger and thus more diversely functionalized PKS I products such as the desertomycins and amphotericins in bacterial metabolism, whereas fungi tend to produce smaller, less functionalized products. In addition, although underlying enzymatic processes follow a quite similar fundamental biosynthetic logic, fungi and bacteria use slightly different strategies to assemble nonribosomal peptides.[17] To exemplify this, fungi often use amino-isobutyric acid (AIB) as a building block, resulting in the production of a large variety of petaibols and other AIB-containing peptides.[18] In bacterial-derived NRPS products, however, AIB is not a common component, and higher functionalized building blocks such as serine, threonine, and asparagine become thus more likely part of the resulting peptides, a factor that further contributes to the difference in FG distribution. At this point it needs to be clarified that our conclusions are based on information about isolated and fully characterized natural products only and do not take into account any genome-derived evidence about the biosynthetic potential of secondary-metabolite-producing organisms. The synthetic molecules are clear outliers; they show only very low similarity to all four NP classes.

The analysis of FG frequencies can also help to provide information about the structural diversity NPs exhibit with respect to their producing organisms. As a measure of diversity we used the number of unique FGs present in at least 0.1% of molecules from different classes. The largest number of FGs is found in bacteria (148), followed by animals (121), fungi (106), and plants (85). It is remarkable that NPs produced by plants, in spite of being the largest and most thoroughly investigated group, are the least diverse according to this measure. The most diverse are NPs produced by bacteria,[19,20] followed by those produced by animals. The high diversity of NPs extracted from animals is probably due to the fact that many molecules, although isolated from marine animals, are actually produced by symbiotic bacteria.[21] The NPs isolated from marine animals, in particular corals and sponges, represent about one-half of the molecules in our "animal-originated" set.

Since the earliest days of rational drug discovery NPs have been used as a source of inspiration for the design of novel bioactive molecules.[22,23] At the beginning, this effort was limited to slight modifications of the original NPs with the goal to improve the biological activity profile and/or physicochemical properties. More recently, it has become possible to analyze large NP databases, which allows for a more comprehensive understanding of structural features that enable specific target interactions, knowledge that is being applied in modern molecular design. In this way, several NP-like combinatorial libraries have been generated.[24−26] By comparing structural features between NPs and synthetic molecules, a method to calculate the NP-likeness score, a measure of how a molecule is fitting to the chemical space covered by NPs, was developed.[27] This score has been used in virtual screening,[28] the design of NP-like virtual libraries,[29] selection of NP-like fragment collections for fragment-based screening,[30,31] and the design of novel NPs by deep neural networks.[32] The information about FG distribution in NPs obtained by the present analysis can provide additional help for *in silico* design of novel NP-like molecules. The list of FGs characteristic for NPs can be used as another measure to assess NP-likeness of existing libraries and to design "non-natural" NPs by machine learning or evolutionary algorithms. Knowing about FG frequencies in NPs can provide guidance for the derivatization of NP mixtures or extracts, as it allows for tailoring reaction conditions accordingly.[33] Last but not least, knowing FGs and the enzymatic processes involved in their generation enables the search for FG-encoding genes in large bacterial, fungal, and plant genomes. As secondary metabolite genes are often clustered, this could help in the identification of novel biosynthetic gene clusters.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jnatprod.8b01022.

> Technical details about the methodology; Table S1, showing comparison of occurrence (in percentage) of functional groups identified in the DNP, the OpenNP, and the synthetic molecules, and Table S2, showing comparison of occurrence (in percentage) of functional groups in natural products from the DNP originated from animals, plants, fungi, bacteria, and synthetic molecules (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: peter.ertl@novartis.com. Web: peter-ertl.com.

ORCID

Peter Ertl: 0000-0001-6496-4448

Notes

The authors declare no competing financial interest.

Supporting tables may be obtained from the corresponding author upon request as text files.

■ REFERENCES

(1) Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 218−227.

(2) Wetzel, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. *Chimia* **2007**, *61*, 355−360.

(3) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272−17277.

(4) Grabowski, K.; Baringhaus, K. H.; Schneider, G. *Nat. Prod. Rep.* **2008**, *25*, 892−904.

(5) Ertl, P. *J. Cheminf.* **2017**, *9*, 36.

(6) *Dictionary of Natural Products 27.1*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2018.

(7) Taxonomy, https://www.ncbi.nlm.nih.gov/taxonomy.

(8) Natural Products Atlas, https://www.npatlas.org/joomla/index.php.

(9) Chen, C. Y. C. *PLoS One* **2011**, *6* (2011), No. e15939.

(10) Sterling, T.; Irwin, J. J. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.

(11) Landrum, G. RDKit: Open-source cheminformatics, http://www.rdkit.org.

(12) Hall, R. https://github.com/rdkit/rdkit/tree/master/Contrib/IFG.

(13) Kramer, C.; Podewitz, M.; Ertl, P.; Liedl, K. *Planta Med.* **2015**, *81*, 459−466.

(14) Jiang, Y.-Y.; Kong, D.-X.; Qin, T.; Li, X.; Caetano-Anollés, G.; Zhang, H.-Y. *PLoS Comput. Biol.* **2012**, *8*, 1−8.

(15) Robbins, T.; Liu, Y. C.; Cane, D. E.; Khosla, C. *Curr. Opin. Struct. Biol.* **2016**, *41*, 10−18.

(16) Keller, N. P.; Turner, G.; Bennett, W. *Nat. Rev. Microbiol.* **2005**, *12*, 937−47.

(17) Yu, D.; Xu, F.; Zhang, S.; Zhan, J. *Nat. Commun.* **2017**, *8*, 15349.

(18) Degenkolb, T.; Karimi Aghcheh, R.; Dieckmann, R.; Neuhof, T.; Baker, S. E.; Druzhinina, I. S.; Kubicek, C. P.; Brückner, H.; von Döhren, H. *Chem. Biodiversity* **2012**, *9*, 499−535.

(19) Demain, A. L. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 185−201.

(20) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *22*, 5601−5606.

(21) Blunt, J. W.; Copp, B. R.; Keyzers, R. A.; Munro, M. H.; Prinsep, M. R. *Nat. Prod. Rep.* **2016**, *33*, 382−431.

(22) Cragg, G. M.; Newman, D. J.; Snader, K. M. *J. Nat. Prod.* **1997**, *60*, 52−60.

(23) Harvey, A. L. *Drug Discovery Today* **2008**, *13*, 894−901.

(24) Nicolaou, K. C.; Pfefferkorn, J. A.; Roecker, A. J.; Cao, G. Q.; Barluenga, S.; Mitchell, H. J. *J. Am. Chem. Soc.* **2000**, *122*, 9939−9953.

(25) Boldi, A. M. *Curr. Opin. Chem. Biol.* **2004**, *8*, 281−286.

(26) Shang, S.; Tan, D. S. *Curr. Opin. Chem. Biol.* **2005**, *9*, 248−258.

(27) Ertl, P.; Roggo, S.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.

(28) Schuster, D.; Wolber, G. *Curr. Pharm. Des.* **2010**, *16*, 1666−1681.

(29) Yu, M. J. *J. Chem. Inf. Model.* **2011**, *51*, 541−557.

(30) Prescher, H.; Koch, G.; Schuhmann, T.; Ertl, P.; Bussenault, A.; Glick, M.; Dix, I.; Petersen, F.; Lizos, D. E. *Bioorg. Med. Chem.* **2017**, *25*, 921−925.

(31) Pahl, A.; Waldmann, H.; Kumar, K. *Chimia* **2017**, *71*, 653−660.

(32) Li, Y.; Zhou, X.; Liu, Z.; Zhang, L. *J. Chin. Phar. Sci.* **2018**, *27*, 451−459.

(33) Ramallo, I. A.; Salazar, M. O.; Mendez, L.; Furlan, R. L. *Acc. Chem. Res.* **2011**, *44*, 241−250.

(34) Fritsch, S.; Neumann, S.; Schaub, J.; Steinbeck, Ch.; Zielesny, A. *J. Cheminformatics* **2019**, submitted.