

Properties and Architecture of Drugs and Natural Products Revisited

Kristina Grabowski and Gisbert Schneider*

Johann Wolfgang Goethe-University, Institute of Organic Chemistry and Chemical Biology, Centre for Membrane Proteomics, Siesmayerstr. 70, D-60323, Frankfurt am Main, Germany

Abstract: Computer-based analysis revealed that natural products exhibit a remarkable structural diversity of molecular frameworks and scaffolds that could be systematically exploited for combinatorial synthesis. Natural products offer a rich pool of unique molecular frameworks that complement “drug space”. They possess desirable druglike properties rendering them ideal starting points for molecular design considerations. This review provides an overview of chemotype diversity and molecular properties of collections of drugs and druglike molecules, pure natural products, and natural product-derived compounds. Compared to druglike molecules, pure natural products contain more oxygen atoms and chiral centers, and have less aromatic atoms on average. Among the natural product library we identified more than one thousand scaffolds that were not contained in any other compound set analyzed. This outcome provides a basis for the design of new natural product-derived compound libraries. Our study demonstrates that computational chemical biology can assist in finding suitable molecular entities in collections of natural products for drug discovery.

Keywords: Natural products, drug discovery, combinatorial chemistry, library design, virtual screening.

INTRODUCTION

With the introduction of high throughput screening (HTS) in the 1970s and combinatorial chemistry in the early 90s, a new era in drug discovery had begun. It became possible to effectively generate thousands of compounds at comparably low costs. Despite their appeal these techniques have had surprisingly small impact on the derivation of novel drugs and candidate compounds for lead optimization [1-5]. This limited productivity is partly caused by limited structural diversity and the lack of biological relevance of the underlying structures of screening libraries [6, 7]. In contrast, natural products represent the richest source of inspiration for the identification of novel scaffold structures that can serve as the basis for rational drug design [8]. Among the FDA-approved New Chemical Entities (NCEs) introduced between 1981 and 2002, 49% were of natural product origin or were derived from natural products using computer-based design [4]. As Costantino and Barlocco stated, “.. natural products are considered to contain scaffolds with the potentiality to be privileged structures because in many cases they are synthesized by biological systems to specifically interact with protein targets” [9]. Natural products should thus serve as biologically validated starting points for combinatorial variation, and combinatorial libraries built around these “privileged” structures should facilitate hit and lead finding with increased probability and quality [7], and chemical biology provides a scientific concept to addressing this topic.

One may wonder why so many natural products have proven pharmacological effect on the human organism. The vast majority of microorganisms or plants do not produce compounds that are meant to bind to human proteins. One

answer might lie in the finite structural space of protein folds. Current estimates emanate from approximately 1,700 distinct folds and 4,000 structural superfamilies [10]. As a consequence, human drug targets may consist of the same building blocks or contain similar structural domains to the targets with which natural products coevolved [11, 12]; or as Meinwald stated, “.. natural products have evolved to interact with something, and that something may not be so different from human proteins” (cited in ref. [12]).

Generally, systematic investigations of molecular scaffolds are used as a way of measuring the diversity of a compound library [13-15]. Although we also wanted to investigate the scaffold diversity inherent in synthetic drug and natural product libraries, the main focus of our study is on comparing scaffolds inherent to druglike compounds and natural products. This comparison aims at identifying novel scaffold architectures that might be suitable for combinatorial library design.

DATA COMPILATION AND METHODS FOR PROPERTY AND SCAFFOLD ANALYSIS

Five different sets of molecules were compiled for this study:

1. **Drug molecules** were taken from the Collection of Bioactive Reference Analogues (COBRA) [16] which contains reference molecules for ligand-based library design.
2. **Pure natural products (PNP)** were taken from the natural products database MEGAbolite (Version 050118) by AnalytiCon Discovery which comprises exclusively pure natural products isolated from plants and terrestrial microorganisms [17]. We also added those compounds of the Interbioscreen database (IBS2004N/IBS2005N), which were marked as genuine natural compounds (GNC: 6%) [18].
3. **Natural products (NPs) and derivatives ((Semi) Natural, SNP)** were taken from the following publicly

*Address correspondence to this author at Johann Wolfgang Goethe-University, Institute of Organic Chemistry and Chemical Biology, Centre for Membrane Proteomics, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany; Fax: +49 69 798 24880; E-mail: g.schneider@chemie.uni-frankfurt.de

available collections: BioSpecs natural products library [19], which contains isolated and synthesized natural products and derivatives from natural sources like plants, fungi, bacteria, and sea organisms; Microsource [20] pure natural products and their derivatives; Tulip [21] with 24,694 modified natural compounds listed; GUPPY with 158 natural compounds and their derivatives [22], and the Interbioscreen database IBS2004N/IBS2005N [18] containing 38,821 compounds marked as derivatives and analogs (DNC 88%) and rare derivatives (RAR 6%). 60-65% of the whole Interbioscreen collection (DNC, RAR and GNC) are compounds of plant origin, 5-10% of microorganisms, approximately 5% from marine species and the remaining compounds are from other natural sources.

4. **NP-derived combinatorial compounds** (NatDiv) were taken from the AnalytiCon collection (NatDiverse, Version 050112) [17], where all templates are based on 11 scaffolds from natural products.
5. **Marine natural products** (MNP) were compiled from the literature [23, 26, 27]. The compounds were isolated amongst others from sponges (41%), Coelenterates (21%), marine microorganisms and phytoplankton (10%) (Fig. 1). Marine organisms have to compete with an often inhospitable environment (little light, high salt content) that is very distinct from that of organisms living ashore. This was shown to result in a rich source of broad structural diversity and novel, partially unusual structures, often featuring promising biological activity [30-36]. Marine organisms possess a wide range of bioactive compounds exhibiting anticancer, antiviral, antifungal or antibacterial activity [4, 23-29, 37]. A recent analysis of NP extracts demonstrates that 620 of 9,945 marine organisms tested were selected as "active" against two leukemia cell lines in an NCI human cancer prescreen [37]. Over 60% of them origin from sponges, the phy-

lum which also represents the biggest fraction in our MNP database. Because marine natural products are often said to be distinctly different from other natural products, we did not combine them with the other PNP compounds so that we could distinguish between mainly plant-derived NPs and NPs from marine sources.

Prior to analysis, all datasets were desalted with the CLIFF software (Chemical Library: Interconversion of File Formats) [38] to remove counter ions. In order to obtain consistent entries a script, written in the Scientific Vector Language (SVL) of MOE (Molecular Programming Environment) [39], was applied to protonate and deprotonate selected charged groups respectively (Table 1). Afterwards duplicates were removed with MOE. This resulted in unique datasets containing the following numbers of compounds: 5,648 drug molecules (COBRA), 65,544 (semi)natural products (SNP), 3,784 pure natural products (PNP), 2,247 marine natural products (MNP), and 7,838 NP-derived combinatorial compounds (NatDiv).

Then, various molecular descriptors were calculated with the software-suite MOE: Hydrogen-bond donors (DON) denote the number of OH and NH atoms, hydrogen-bond acceptors (ACC) denote the number of O and N atoms as defined by Lipinski [40]. The number of rotatable bonds (RTB) was calculated using Oprea's definition [41]. Estimation of the octanol/water partition coefficient (SlogP) was done using the Wildman and Crippen approach [42]. A violation of Lipinski's "rule of five" (LipViol) was recorded if $DON > 5$, $ACC > 10$, molecular weight (MW) > 500 and $SlogP > 5$. The topological polar surface area (TPSA) was calculated using group contributions from connection table information to approximate the polar surface as described by Ertl [43]. Globularity is a three-dimensional (3D) shape descriptor. 3D structures were generated with the Corina software [38] and afterwards energy-minimized in MOE (with default parameters). Globularity was calculated by dividing the smallest eigenvalue by the largest eigenvalue of

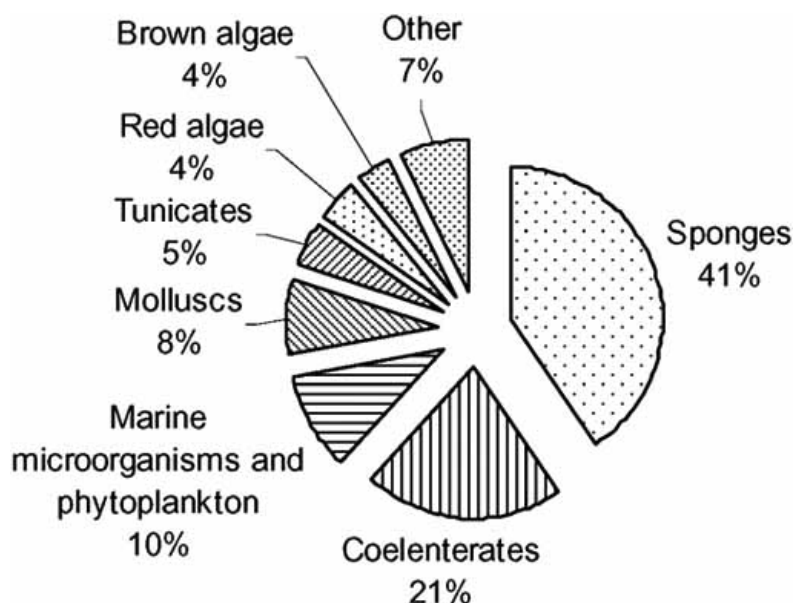


Fig. (1). Distribution of marine natural products origin.

the covariance matrix of atomic coordinates. A value of 0 indicates a two- or one-dimensional object while a value of 1 indicates a perfect sphere.

Table 1. Functional Groups that were Protonated or Deprotonated in the Data Sets

Functional Group	Action	Not if in: (*)
$\equiv \text{N}^+$	deprotonate	
$-\text{NH}_3^+$	deprotonate	
H_2^+ N 	deprotonate	
$=\text{NH}_2^+$	deprotonate	
H^+ N 	deprotonate	
$-\text{O}-$	protonate	$\begin{array}{c} \diagup \text{O}^- \diagdown \\ +2\text{S} \\ \diagdown \text{O}^- \diagup \end{array}$ i.e. $\begin{array}{c} \\ -\text{S}=\text{O} \\ \\ \text{O} \end{array}$ $\begin{array}{c} \text{H}^+ \\ \\ \text{S}-\text{O}^- \\ \end{array}$ i.e. $\begin{array}{c} \text{H} \\ \\ \text{S}=\text{O} \\ \end{array}$ $\begin{array}{c} \text{O}^- \\ \\ -\text{P}^+- \\ \end{array}$ i.e. $\begin{array}{c} \text{O} \\ \\ -\text{P}- \\ \end{array}$
		$\begin{array}{c} \\ -\text{N}^+-\text{O}^- \\ \end{array}$ and $\begin{array}{c} \\ \text{N}^+-\text{O}^- \\ / \end{array}$
$-\text{S}-$	protonate	
$-\text{N}^-$	protonate	
C H ⁺	discard	
$=\text{O}^+$ 	discard	

(*) MOE-specific notation of Sulfone, Sulfoxide, Phosphate.

Structural analysis of the molecules was performed on two different levels regarding either atomic properties or graph properties, respectively. Atomic properties contain information like element-type, hybridization and atomic charges. The analogy between a structure diagram and a topological graph is the basis for applying graph theory to chemical structures: A graph consists of vertices and edges, molecular graphs solely describe the linkage of the atoms, at which every non-hydrogen atom is a vertex and every bond is an edge in the molecular graph [44]. Therefore, this provides a way to cluster molecules on a topological level.

To compare the structures of known drugs to NPs and NP derived compounds we dissected molecules into “side-chains” and “frameworks” containing ring systems and linkers as defined by Bemis and Murcko [13]. According to this concept, a molecule is segmented into four units: ring systems, linkers, side-chains and frameworks. In Fig. 2 the different structural units are shown considering the anti-anxiety agent Diazepam as example.

- **Ring systems** are defined as cycles within the molecular graph (rings) or as rings sharing an edge (a bond between two atoms) or vertex (atom) in the molecular graph. A set of rings with fused or spiro connections is defined as a single ring system. Benzene, anthrazene and spiro[5.5]undecane (Fig. 3) are single ring systems; Benzen a monocyclic, Anthrazen a tricyclic and Spiro[5.5]undecane a bicyclic.
- **Linkers** are defined as vertices (atoms) and/or edges (bonds) on the path connecting two different ring systems. Diazepam (Fig. 2) has a zero-atom linker, whereas diphenylmethane (Fig. 3d) has a one-atom linker between two different ring systems.
- **Side-chains** are defined as atoms which are not classified as ring system or linker atoms.
- **Frameworks** are defined as ring systems (if no linker exists) and ring systems connected by linkers, i.e. everything which remains after removing the side-chains. Acyclic molecules do not have any framework by definition. The framework of Diazepam shown in Fig. 2 has two different ring systems (one mono- and one bicyclic) connected by a zero-atom linker.

In addition to Bemis and Murcko we introduced marked frameworks (Fig. 4) where exit-vectors of removed side-chains were kept at ring systems and we expanded the definition of frameworks to scaffolds (Fig. 5).

- **Marked frameworks** additionally contain the exit-vectors of the side-chains as seen in Fig. 4, so that the positions where side-chains come off at ring systems are marked. Side-chains which go off at linker atoms are removed anyway.
- **Scaffolds** are frameworks from which exocyclic double bonds (double bonded single atom side-chains, mostly X=O) are not removed from rings and remain part of the atomic framework (Fig. 5).

This led to eight different levels of abstraction: atomic scaffolds (marked/unmarked), graph scaffolds (marked/unmarked), atomic frameworks (marked/unmarked), and graph frameworks marked/unmarked. Subsequent analysis of the datasets was done with MOE scripts written in SVL. In an earlier related study Xue and Bajorath [15] also implemented Murcko scaffold analysis using SVL. We re-implemented the concept from scratch with the primary aim to provide data formats that can immediately be used for compound *de novo* design and natural-product derived combinatorial library enumeration.

COMPARISON OF MOLECULAR PROPERTIES

To get a first overview, we compared the five datasets considering common descriptors for druglike molecules (Table 2).

The average calculated molecular weight is similar for drugs (MW = 415), pure NPs (MW = 394), semi NPs (MW = 409) and the NP derived combinatorial library (MW = 441), whereas marine NPs have a higher mean value (MW = 504) and show the broadest distribution (std dev = 250) followed by pure NPs (std dev = 196) and drugs (std dev = 143). However, the majority of the natural compounds (63% of MNPs and 82% of PNPs) are inside the normal weight

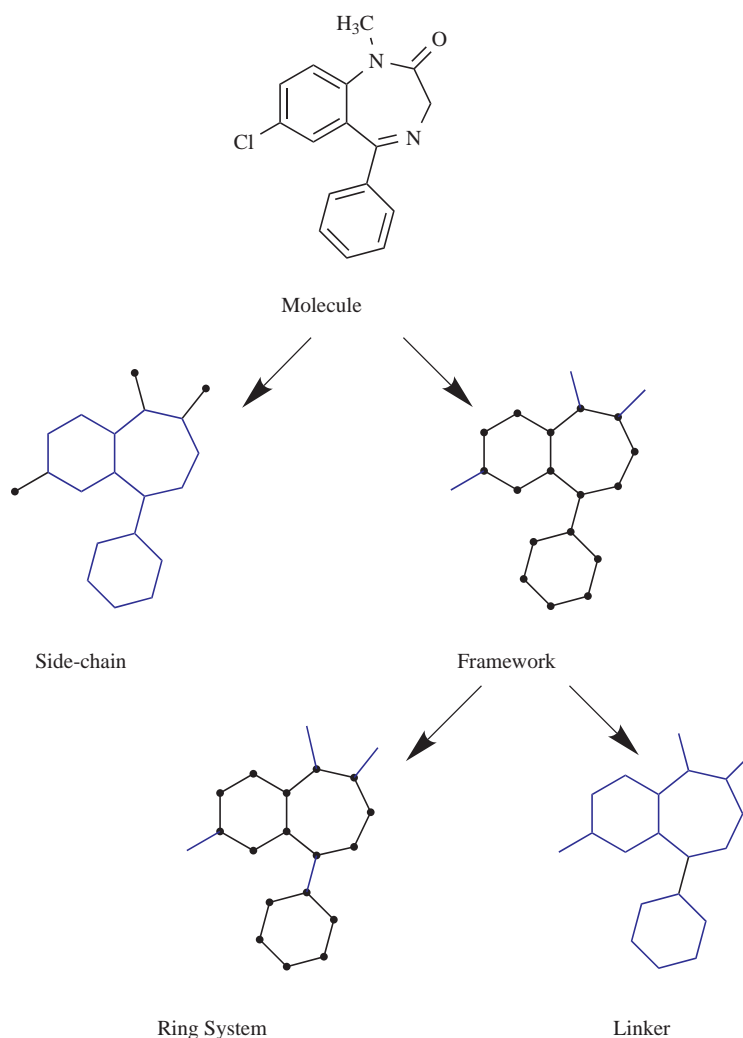


Fig. (2). Dissection of the molecule Diazepam in three side-chains and one framework containing two ring systems and a zero-atom linker (adapted from Bemis and Murcko [13]).

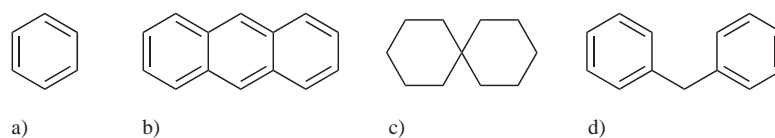


Fig. (3). Benzene (a), Anthracene (b) and spiro[5.5]undecane (c) are three single ring systems; Diphenylmethane (d) contains two ring systems and a one-atom linker.

range of druglike molecules (≤ 500 MW) and it has also been shown that many NP-like compounds with higher molecular weight are able to access intracellular targets [45].

Regarding the number of non-hydrogen atoms the higher weight is mostly due to the average larger size of marine compounds (34.6 heavy atoms for marine NPs and between 28.2 and 31.1 for the other databases).

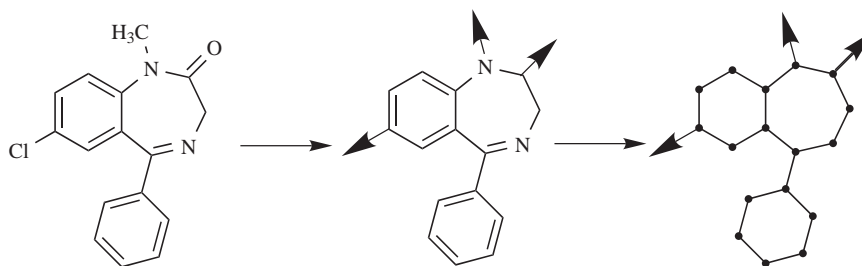


Fig. (4). The “marked framework” of Diazepam. Arrows at the framework mark side-chain positions (exit-vectors).

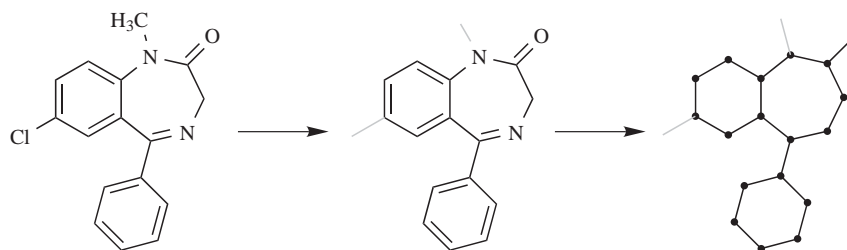


Fig. (5). The scaffold of Diazepam. The keto group remains part of the atomic/graph scaffold.

The average lipophilicity SlogP-value is lower for the mainly plant derived collection PNP (2.3) and NatDiv (2.1) than for drugs (3.5), marine NPs (3.9) and SNPs (3.7). Our database of druglike molecules shows some differences compared to drug datasets used in other studies [46-48], namely slightly higher average MW, more non-hydrogen atoms and higher SlogP values. It has been observed that newer drugs show such a tendency [49]. The COBRA database contains many “newer” drugs; on the other hand it contains a large fraction of comparably large GPCR ligands, which explains the differences between the drug data sets used in other comparative studies between natural products and drugs [46-48].

As reported by other studies [46, 48] one big difference between pure NPs and the other data sets is the average number of nitrogen and oxygen atoms. Marine NPs have nearly three times less nitrogen atoms per molecule than drugs or the NatDiv dataset, plant derived NPs even four times. SNP with 2.1 nitrogen atoms per molecule are in between drugs (3.0) and NPs (0.7, 1.2). In contrast, NPs contain approximately twice as many oxygen atoms than drugs (unlike in the datasets analyzed by Lee and Schneider [48], but consistent with Feher and Henkel [46, 47]), the SNP and

NatDiv datasets are with 4.3 and 4.4 oxygen atoms in between.

The number of hydrogen-bond acceptors therefore is almost the same for drugs, PNPs and SNPs (6.4-6.6), and a bit higher for marine compounds (7.4) and the combinatorial compounds (8.0).

NPs contain slightly higher numbers of hydrogen-bond donor atoms per molecule than drugs and NP derived combinatorial compounds (about 2.7 to 2.2), and less (1.4) for seminatural compounds, as also observed elsewhere [46].

The number of molecules that cause at least two violations of the “rule of five” [40] is low in the drug, SNP and NatDiv dataset (approximately 10%), as also observed by Lee and Schneider [48]. In contrast to their findings our NP datasets contains nearly twice (PNP) and threefold (MNP) as many molecules that cause at least two violations, but still more than 80% of the PNP molecules and 70% of the MNP compounds account for less than 2 violations.

Because the “rule of five” depends on molecular size, and NPs are bigger molecules in general (larger compounds spawn more functional groups on average) we applied a recently developed size-independent “Chemistry Space Fil-

Table 2. Comparison of Molecular Properties Between Drugs (COBRA), Pure Natural Products (PNP, MNP), NPs and Derivatives/Analogues (SNP), and NP-Based Combinatorial Compounds (NatDiv)

	COBRA	PNP	MNP	SNP	NatDiv
Molecular mass	414.5 (142.7)	393.9 (196.3)	503.6 (250.1)	409.2 (102.4)	441.3 (74.2)
# heavy atoms	29.1	28.2	34.6	29.1	31.1
SlogP	3.5 (2.2)	2.3 (2.7)	3.9 (2.6)	3.7 (1.7)	2.1 (1.8)
# nitrogen atoms	3.0	0.7	1.2	2.1	3.6
# oxygen atoms	3.4	5.9	6.1	4.3	4.4
ACC	6.4	6.6	7.4	6.4	8.00
DON	2.1	2.7	2.6	1.4	2.26
LipViol ≥ 2	10%	18%	30%	10%	8%
TPSA	90.5 (55.4)	98.9 (82.1)	108.9 (88.4)	83.2 (35.1)	104.7 (35.9)
# aromatic atoms	12.4	5.1	3.5	11.8	9.5
# chiral atoms	1.4	5.5	6.3	1.4	3.3
RGB	18.8	19.5	18.6	19.4	21.4
Rings	3.3	3.6	2.9	3.5	4.0
RTB	6.7	5.2	11.5	6.1	5.3
Globularity	0.10	0.12	0.14	0.08	0.08

Numbers in parentheses are standard deviations.

ACC: hydrogen-bond acceptors, DON: hydrogen-bond donors, LipViol: fraction of compounds with more than one violations of Lipinski’s rule of five, TPSA: topological polar surface area, RGB: number of rigid bonds, RTB: number of rotatable bonds.

ter” [50]. It combines two descriptors, a molecular saturation related one and a hetero-atom proportion descriptor, which yielded good results in distinguishing druglike molecules (MACCS-II Drug Data Report, MDDR [51], and Comprehensive Medicinal Chemistry database, CMC [51]) from nondruglike molecules (Available Chemicals Directory, ACD [51]). This filter also classified 73% of the Chinese Natural Products Database (CNPD) [52] as druglike. When we applied it to our data sets (Table 3), the score was high for PNP, NatDiv and MNP (69-83%) suggesting that these natural product collections contain many druglike molecules. Only 48% of the COBRA set were classified as druglike indicating that this descriptor is not database-independent and failed to classify the drug collection correctly. For comparison, we employed a different druglikeness prediction, which is based on a Support-Vector-machine approach [53]. Here all data sets are scored as druglike (68-87%), but again the overall score for the COBRA drug set was lower than for genuine NPs (PNP and MNP). Summarizing, a general finding is that both “druglikeness” predictions scored NP databases even more druglike than actual drugs indicating a general “biophoric” character of NPs [54].

Table 3. “Druglikeness” Scores of the Data Sets Computed by the Chemistry Space Filter [44] and Druglikeness SVM [47]

	COBRA	PNP	MNP	SNP	NatDiv
Chemistry Space Filter	48%	69%	83%	53%	78%
Druglikeness SVM	73%	82%	87%	68%	73%

With respect to the charge at physiological pH, TPSA can be a critical factor for oral bioavailability [55-57] and is correlated to the number of hydrogen-bond donors and acceptors. Typically compounds are considered to be orally bioavailable with a TPSA value between 75 and 150 Å² [57]. Marine NPs (109 Å²) and the NP derived combinatorial dataset (105 Å²) have the highest average TPSA values, followed by pure NPs (99 Å²). Drugs (91 Å²) and natural product derivatives and analogs (83 Å²) yield lower values on average. Genuine natural products have often higher TPSA values and a broader distribution (std.dev. > 80) than drugs (std.dev. = 55), where most of the molecules have a TPSA below 120 Å².

The average number of aromatic atoms shows a pronounced difference between NPs (PNP: 5.1, MNP: 3.5) and drugs, NP derivatives, analogs and the NP-based combinatorial NatDiv (COBRA: 12.4, SNP: 11.8, NatDiv: 9.5). In drugs and SNPs approximately every second atom per molecule is aromatic, every third in NatDiv, but only every fifth and tenth in PNP and MNP, respectively.

In contrast, PNPs contain on average approximately four times more chiral centers than drugs or the SNP compounds. As Feher and Schmidt stated [46] this difference might be due to the inconvenience of chiral separation in synthetic chemistry (drugs, NP derivatives and analogs). On the other hand, nature with many stereospecific reactions often produces compounds with high numbers of chiral centers which favors selective binding to mostly stereospecific binding

sites [11]. This elusion of chiral centers in synthetic chemistry causes the bias of aromatic rings in synthetics, which is not found in natural products. Consequently, the NP-based combinatorial set (NatDiv) yields values between synthetics and NPs.

The number of rings (3-4) and rigid bonds (approx. 19) are almost identical for all datasets, albeit slightly higher for the NatDiv collection because it does not contain any acyclic molecules.

The average number of rotatable bonds (RTB) per molecule is lowest for PNP and NatDiv (approx. 5), followed by SNP (6.1) and COBRA (6.7). Similar numbers were obtained in a related study by Feher and Schmidt [46]. The marine natural products contain nearly twice as many rotatable bonds on average. MNP and PNP molecules contain on average twice as many side-chains per molecule than mostly synthetic compounds. The marine compounds also tend to have larger side-chains (Fig. 7) than plant-derived NPs. This could explain the huge discrepancy between MNPs and the other datasets regarding RTBs.

Globularity was the only 3D descriptor analyzed in our study. A value of 0 indicates a one- or two-dimensional object while a value of 1 indicates a perfect sphere. The hypothesis was that NPs might be more globular or “sterically complex” than drugs, because the targets they evolved to interact with are inherently three-dimensional and chiral [11]. Almost all molecules yielded a globularity value below 0.5 (approximately 99% for all data sets). This means that all the analyzed small compounds are rather planar than spheroidal. However, genuine NPs (PNP, MNP) are slightly more spheroidal than NP derivatives, analogs, combinatorial or druglike compounds. Among the genuine NPs more than 32% of the molecules have a globularity value greater than 0.14, whereas we counted only 26, 18 and 15% for the COBRA, SNP and NatDiv data sets, respectively.

Finally, we screened the databases for “unsuitable” natural products, that is, derivatives known to interfere with common assay procedures [58]. As expected, the genuine NP data sets PNP and MNP contain the most unsuitable NPs (7% and 3.8% respectively, Table 4), followed by SNP (1.1%) and COBRA (0.8%). The NatDiv does not contain any of those compounds as it represents a specifically tailored combinatorial library build around 11 NP scaffolds. This example demonstrates that NP-derived combinatorial libraries can be well-suited for screening. However, the percentage of unsuitable NPs even in the PNP and MNP data sets is generally low and although saponins, for example, might be unsuitable for common assay procedures they are known as potent immunological adjuvants and cancer vaccines [59-61].

MOLECULAR FRAMEWORK AND SCAFFOLD ANALYSIS

With the aim to identify potentially novel scaffolds in natural products, which might be suitable for synthetic chemistry, we then virtually dissected all molecules into frameworks, corresponding scaffolds and side-chains. We compared the structural diversity of the individual data sets on different levels of abstraction. We performed a solely

Table 4. Absolute Numbers and Cumulative Percentages of “Unsuitable” Natural Products in All Data Sets Investigated

Unsuitable NPs	COBRA	PNP	MNP	SNP	NatDiv
Quinones	14	60	35	319	0
Polyenes	7	6	32	20	0
Saponin derivatives	2	183	7	139	0
Cytochalasin derivatives	2	1	0	178	0
Cycloheximide derivatives	6	3	0	14	0
Monensin derivatives	0	4	10	1	0
Cyanidin derivatives	6	3	0	14	0
Squalestatin derivatives	6	3	0	14	0
Sum/percent	0.8%	7.0%	3.8%	1.1%	0%

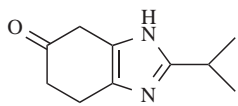
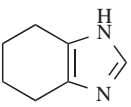
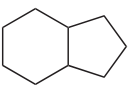
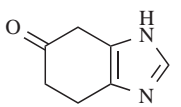
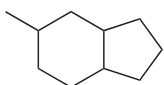
graph-based analysis, which is more general, and an atom-based analysis including atom and bond types that further divide framework classes. On both levels of abstraction, the study was performed with and without considering side-chain positions (“exit-vectors”) on the frameworks and with (scaffold analysis) and without (framework analysis) treating double-bonded single atom side-chains (exocyclic double bonds like X=O) as part of the framework. The results are summarized in Table 5. The set of druglike molecules (COBRA) turned out to be the most diverse on every level of

abstraction, followed by marine NPs and plant-derived NPs (PNP). As expected, AnalytiCon’s combinatorial NatDiv compounds and the NP derivatives and analogs built around single natural products are according to our definitions least diverse. Here, acyclic molecules do not have a framework. The number of acyclic compounds in the different sets is low, for the analyzed data sets their percentage ranges from 0–6.9% of all compounds in a particular collection (Table 5).

Graph-Based Frameworks and Scaffolds

We will first focus on graph-based frameworks and scaffolds. With 1,971 graph-based frameworks (35%) among the 5,562 cyclic COBRA compounds, the COBRA set exhibits the greatest overall chemotype diversity. For MNP, PNP, NatDiv, and SNP we found 619 (30%), 823 (22%), 925 (12%) and 6,017 (9%) different frameworks, respectively. When Bemis and Murcko published their analysis of the Comprehensive Medicinal Chemistry database in the year 1996 (CMC, v. 94.1) they found only 32 graph-based frameworks that account for the “shape” of 50% of the database compounds [13]. Rognan and coworkers [14] recently introduced the term “NC50C” for this metric to measure scaffold diversity. In Fig. 6A Bemis’ and Murcko’s results for the CMC are plotted together with our results. Obviously all databases in our study are more diverse than the CMC data set from 1994. Noteworthy, the COBRA set was manually compiled to cover drug space most broadly with respect to “scaffold architecture” [16]. In contrast to Bemis’ and Murcko’s observation that only 32 frameworks are the basis for 50% of all analyzed druglike molecules, 160 graph-based frameworks are needed to describe the chemotype of 50% of the COBRA compounds. For the (semi) NPs 113 frameworks account for 50% of the database molecules, 76 frameworks for the combinatorial NatDiv, followed by 64 for the marine NPs and 37 for the PNP data set.

Table 5. Reduction of Molecular Diversity on Different Levels of Abstraction, Expressed as Percent of Cyclic Molecules

		COBRA	Marked*	PNP	Marked*	MNP	Marked*	SNP	Marked*	NatDiv	Marked*
	# Molecules, acyclic	5648, 1.5%		3784, 3.1%		2234, 6.9%		65544, 0.4%		7838, 0%	
	Atomic Framework	64% (57%)	78% (73%)	42% (34%)	62% (42%)	47% (41%)	65% (60%)	20% (19%)	44% (42%)	32% (31%)	51% (51%)
	Graph Framework	35% (24%)	65% (54%)	22% (11%)	52% (31%)	30% (20%)	58% (50%)	9% (8%)	34% (32%)	12% (9%)	32% (31%)
	Atomic Scaffold	64% (58%)	78% (74%)	42% (25%)	65% (45%)	52% (45%)	67% (63%)	21% (20%)	44% (43%)	32% (32%)	52% (51%)
	Graph Scaffold	41% (30%)	67% (57%)	30% (15%)	57% (36%)	40% (30%)	62% (55%)	12% (10%)	35% (34%)	12% (11%)	32% (32%)

Numbers in parentheses give the fraction of unique frameworks/scaffolds found in a data set. These were computed as the number of unique scaffolds divided by the total number of scaffolds found in each data set.

*Marked frameworks/scaffolds contain the exit-vectors of the side-chains at the frameworks/scaffolds, different substitution types lead to different frameworks/scaffolds.

Since exocyclic double bonds are rigid and can participate in ligand-receptor interaction, we extended the framework definition to “scaffolds”, where these exocyclic double bonds remain part of the framework (Fig. 5). Marine NPs contain most of those groups (e.g. C=O), the SNPs and NatDiv compounds the least (Table 5). The COBRA compounds contain 2,284 different graph-based scaffolds (41% of all scaffolds), directly followed by marine NPs with 823 (40%) scaffolds. The structural diversity reduces to 30% (1,103 scaffolds) for the PNP and to 12% for the SNP (7,815 scaffolds) and the NatDiv set (961 scaffolds). When the numbers of graph-based scaffolds are plotted against the percent of compounds for which they represent the basic chemotype (Fig. 6B, Table 6) the order of the data sets slightly changes compared to graph-based frameworks: With 242 scaffolds accounting for 50% of the molecules COBRA is still most diverse, followed by SNPs (198), MNPs (128) and PNPs (86). The number of scaffolds for the genuine NPs (MNP and PNP) doubled compared to frameworks, which demonstrates that many of the natural frameworks are subdivided into different classes of scaffolds. With 198 scaffolds the SNP database appears to be “diverse”. Note that 50% of the database refers to more than 32,000 molecules, which are reduced to 198 scaffolds. On the other hand, 129 scaffolds account for about 1,100 MNP compounds. Krier and co-workers [14] introduced the PC50C value as a new metric for scaffold diversity: The PC50C is the percentage of all scaffolds accounting for 50% of the classified compounds of each collection. To build the compounds for half of the SNP database only 3% of all scaffolds are needed, whereas up to 16% are needed in case of marine NPs (Table 6). Here one has to keep in mind the reduction of the whole data set: For example, the SNP data set is already reduced to 12% on this level. From these scaffolds less than 3% represent the chemotype of 50% of all molecules. Hence the genuine natural products, particularly the compounds of our marine set, may be considered as remarkably diverse.

Table 6. Numbers of Scaffolds (NC50C) and Percentages of Scaffolds that Account for 50% of the Molecules (PC50C)

	Total Number of Molecules	Number of Cyclic Molecules	Number of Graph Scaffolds	Number of Scaffolds (NC50C)	PC50C
COBRA	5648	5562	2284	242	11%
PNP	3784	3667	1103	87	8%
MNP	2234	2078	823	129	16%
SNP	65544	65253	7815	198	3%
NatDiv	7838	7838	961	81	8%

Diversity in databases is also generated by the number of singletons, that is, scaffolds that occur in only one molecule of the particular database. The number of such unique scaffolds strongly differs for the considered data sets. With 66% of all scaffolds the COBRA database contains more than twice as many singletons than the NatDiv and SNP data sets (Table 7). Only 11% and 14% of the COBRA and MNP scaffolds occur at least four times, whereas it is more than 50% in the case of the NatDiv set. This, of course, arises from the fact that the NatDiv is a combinatorial database, and all structures can be ascribed to 11 natural chemotypes. In the SNP data set, which contains many analogs and synthetics, 30% of all scaffolds are found at least four times.

thetics, 30% of all scaffolds are found at least four times. This also explains why the overall shape diversity (Table 5) is low for the SNP and NatDiv compounds.

The reduction of database diversity on the different levels can be high, for example the 65,544 molecules of the SNP set are reduced to only 7,815 different graph scaffolds (12%). But of these scaffolds, the vast majority (6,841 scaffolds or 88%) are unique to this data set meaning they do not have identical counterparts in any other collection analyzed (Table 5). The percentage of unique scaffolds for the NatDiv, MNP, COBRA and PNP collections is 86%, 75%, 74% and 51%, respectively. Hence the majority of the scaffolds is exclusively found in the particular database as it was previously shown with different scaffold definitions and data sets [14, 48].

The five most common graph-based scaffolds of each data set are shown in Fig. 8. Clearly the six-member ring is the most common scaffold for drugs and pure natural products. Interestingly, the six-member ring is just at rank two for the SNP database and at position six for the NatDiv data set. The most common scaffolds-with the exception of NatDiv scaffolds-are in accordance with the set found by Bemis and Murcko in the CMC data set of 1994. Still, there are some different graph-based scaffolds among the top-ranked scaffolds, especially for marine NPs (for example, number 5 from MNP in Fig. 8). The NatDiv collection shows many markedly different scaffolds among the most common ones, which again is due to the fact that the NatDiv set is build around eleven NP derived scaffolds.

Atom-Based Frameworks and Scaffolds

When considering atomic properties (atom and bond-types) of the frameworks or scaffolds a higher “diversity” is observed for all data sets, which simply reflects that a graph-based scaffold class is subdivided into further classes based

Table 7. Number of Unique Scaffolds (Singletons) in the Different Datasets

	Graph-Based Singletons*	Atom-Based Singletons*
COBRA	1515 (66%)	2835 (79%)
PNP	652 (59%)	985 (64%)
MNP	473 (57%)	697 (65%)
SNP	3575 (46%)	7671 (56%)
NatDiv	277 (29%)	1385 (55%)

*Expressed as percentage of all scaffolds of the database.

on the atomic properties of the scaffolds. As expected, the number of singletons per database on the atomic level increases compared to graph-based scaffolds. The number of singletons ranges from 55% of all atom-based scaffolds for the combinatorial NatDiv up to 79% for the druglike COBRA database (Table 7). The majority of all atomic scaffolds of the analyzed databases are present in just a single molecule of this data set. Considering such an atom-based scaffold definition for diversity estimation might thus not be appropriate.

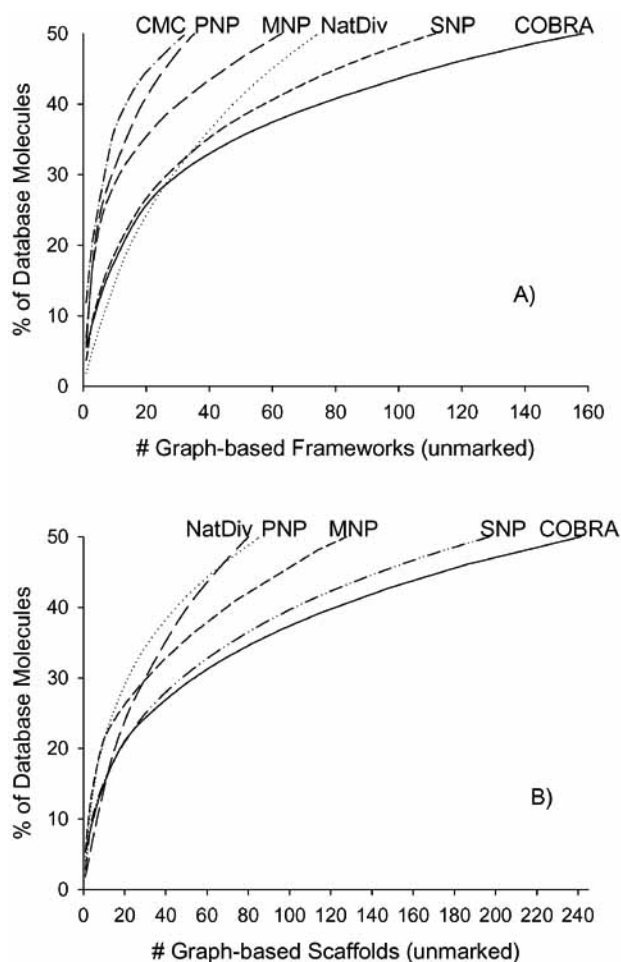


Fig. (6). Number of graph-based frameworks (A) and scaffolds (B) that account for 50% of the drugs (CMC and COBRA), pure natural products (PNP and MNP), NPs and derivatives (SNP), and NP-based combinatorial compounds (NatDiv). In A) the results for the CMC compounds according to Bemis and Murcko [13] are shown.

Frameworks and Scaffolds With Marked Side-Chain Positions

We finally extended the framework and scaffold definitions to include the exit-vectors of the side-chains. Side-chain positions in NPs might have evolved to optimally present functional groups for the binding to three-dimensional binding sites. The information about observed

side-chain attachment points could be useful for designing natural product-derived combinatorial libraries. As a general observation we found that one scaffold can come in different variations of exit-vector points. This is expressed in the higher diversity of the data sets when comparing marked scaffolds to unmarked ones (Table 5). Generally, we found many scaffolds that are present in both drug molecules and NPs. This study also revealed that substitution patterns of the same scaffolds often differ between NPs and druglike compounds or NP derivatives (see, for example, scaffolds 3, 4, 5 of SNPs in Fig. 9; scaffold 2 of the MNP set in Fig. 11). This might be a consequence of different synthetic tractability of substitution patterns.

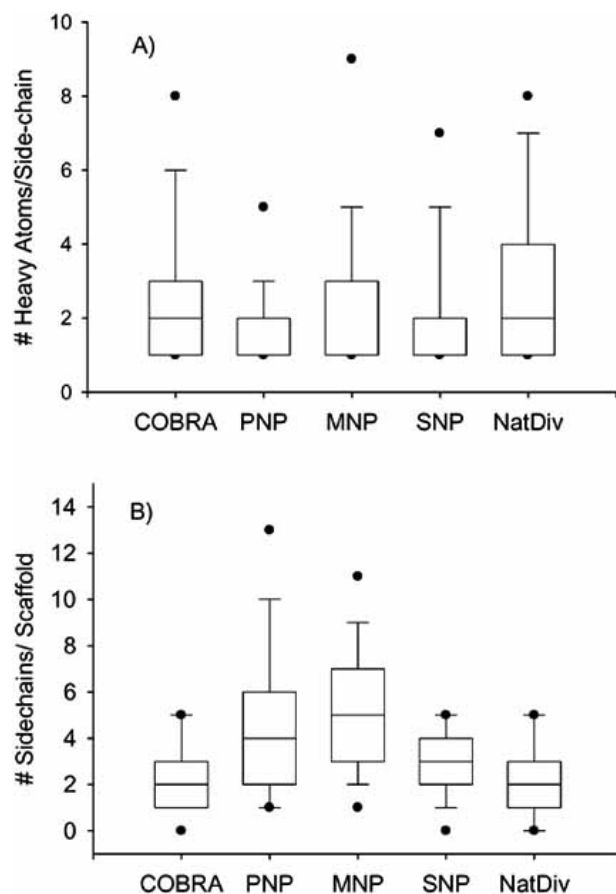


Fig. (7). Distribution of A) heavy atoms per side-chain and B) side-chains per scaffold in drugs (COBRA), pure natural products (PNP, MNP), NPs and derivatives/analogs (SNP) and NP based combinatorial compounds (NatDiv). Exocyclic double bonds (e.g. X=O) are not considered as side-chains. Box plots give the median with the 25% and 75% percentiles, whiskers indicate the 5% and 95% percentiles, and the dots represent minimal and maximal values.

Examining the number of side-chains per scaffold shows that NPs tend to have more side-chains than druglike and NP-derived compounds (Fig. 7B) (note that for this study exocyclic double bonds were considered as part of the scaffold). The side-chain distribution of NPs peaks at four side-chains per scaffold and has a wider distribution than drugs

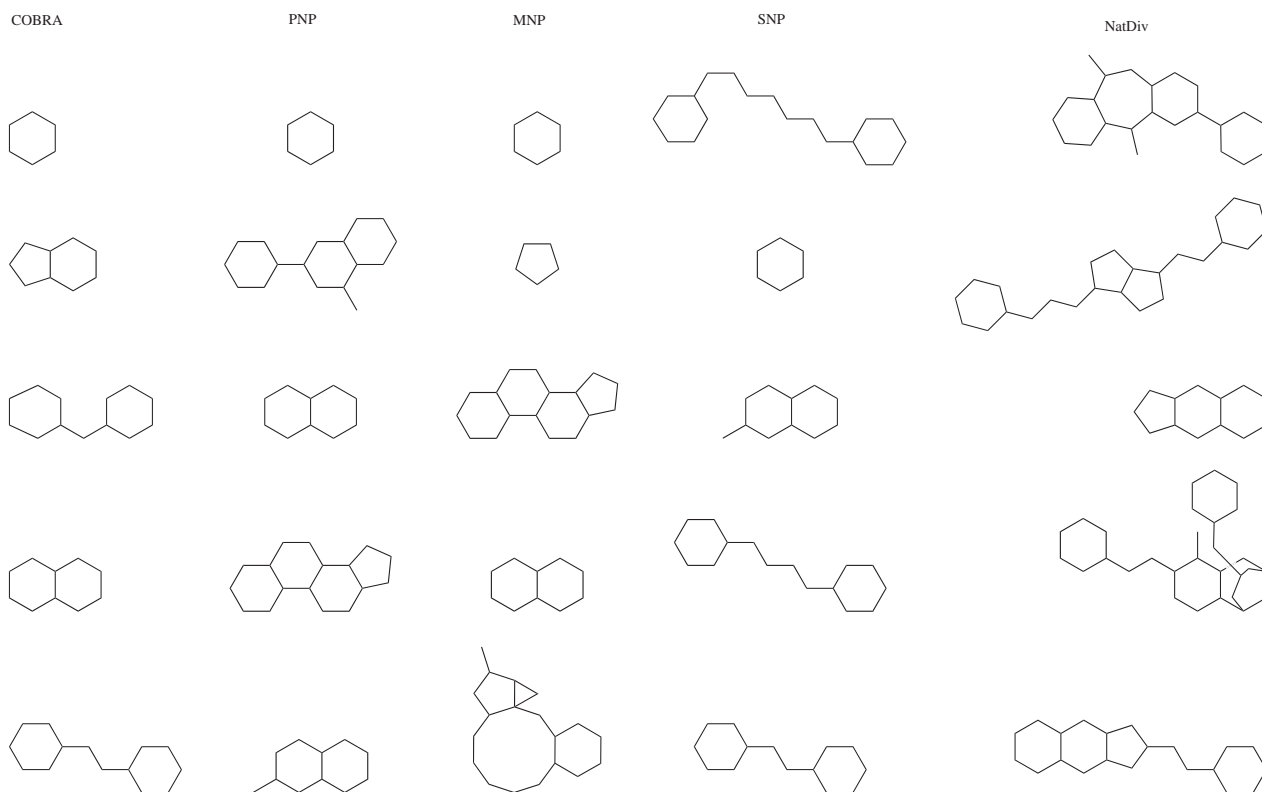


Fig. (8). Most common graph-based scaffolds of druglike compounds (COBRA), genuine NPs (PNP and MNP), NP-derivatives and -analogs (SNP), and NP-derived combinatorial compounds (NatDiv).

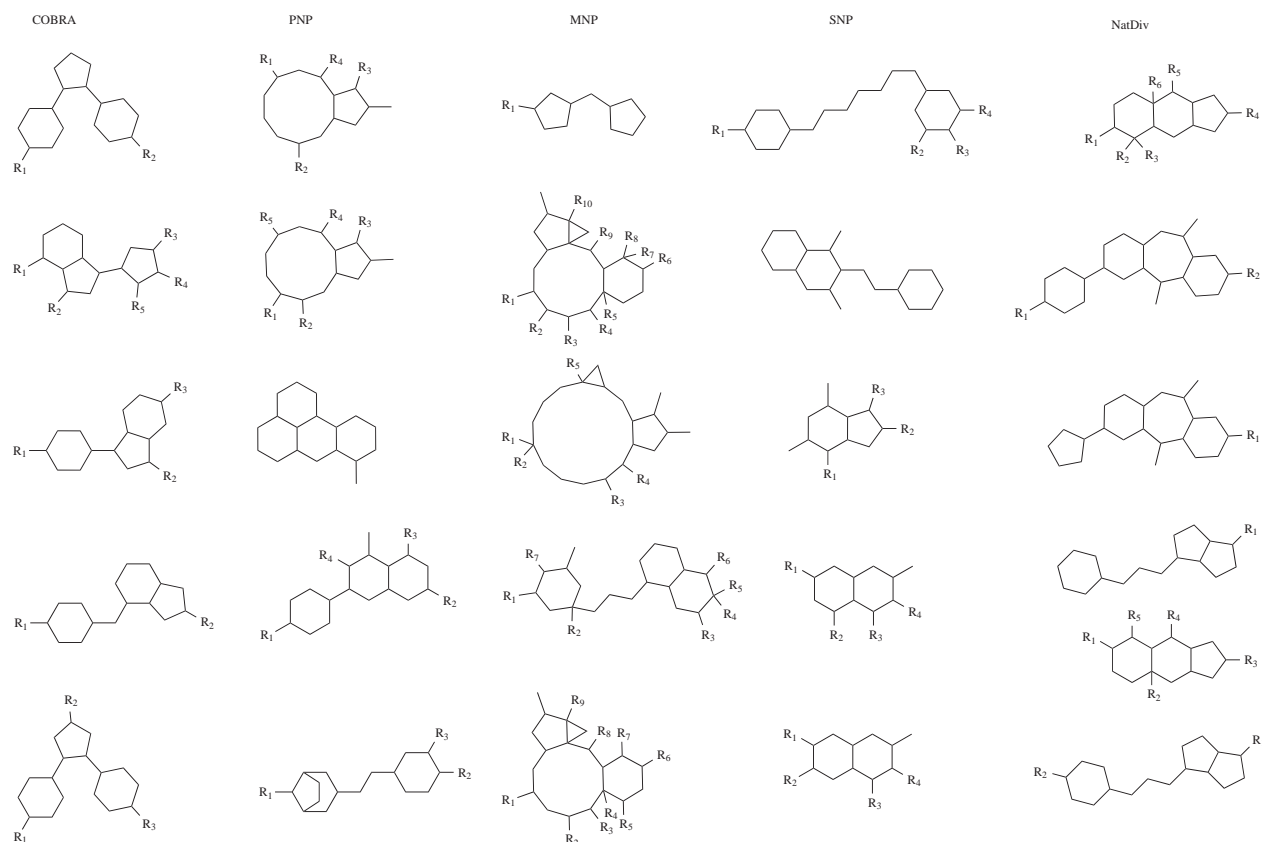


Fig. (9). Most common graph-based scaffolds with marked side-chain positions that are exclusively found in druglike molecules (COBRA), pure NPs (PNP), marine NPs (MNP), NP-derivatives and -analogs (SNP), and NP-derived combinatorial compounds (NatDiv).

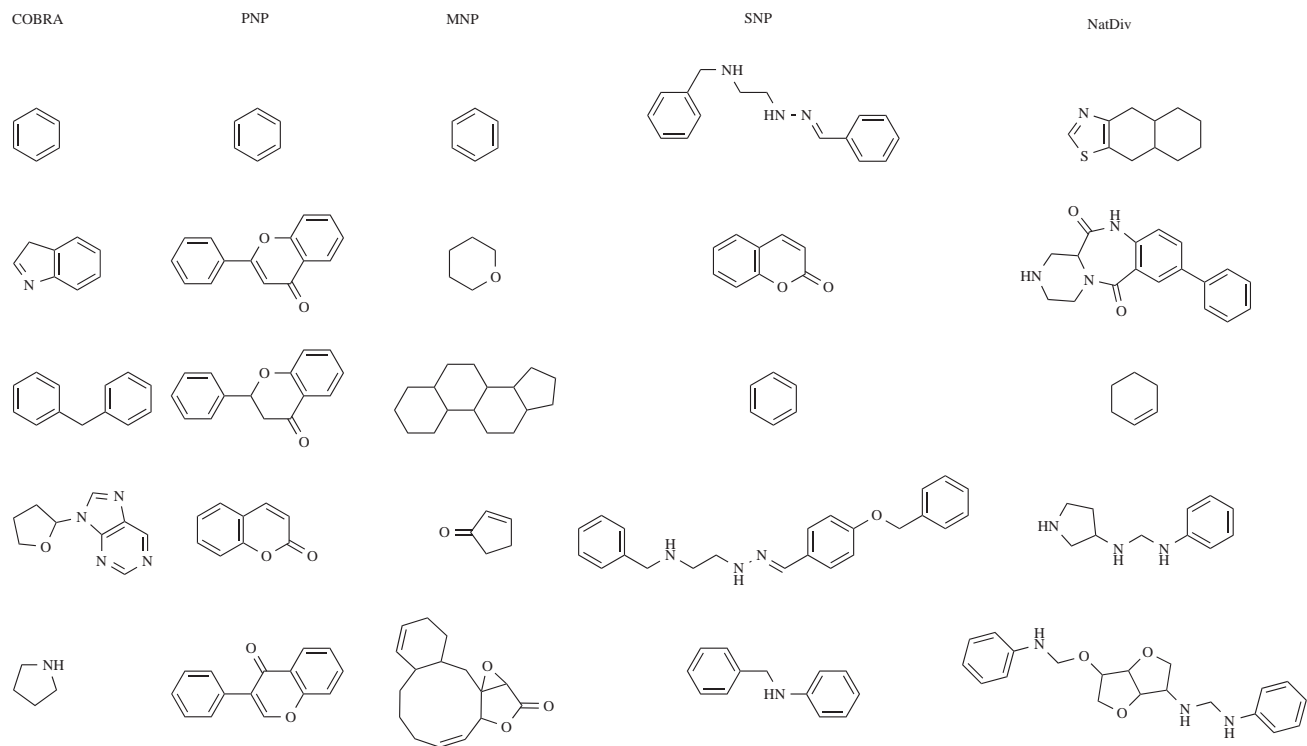


Fig. (10). Most common atom-based scaffolds of druglike molecules (COBRA), genuine NPs (PNP and MNP), NP-derivatives and -analog (SNP), and NP-derived combinatorial compounds (NatDiv).

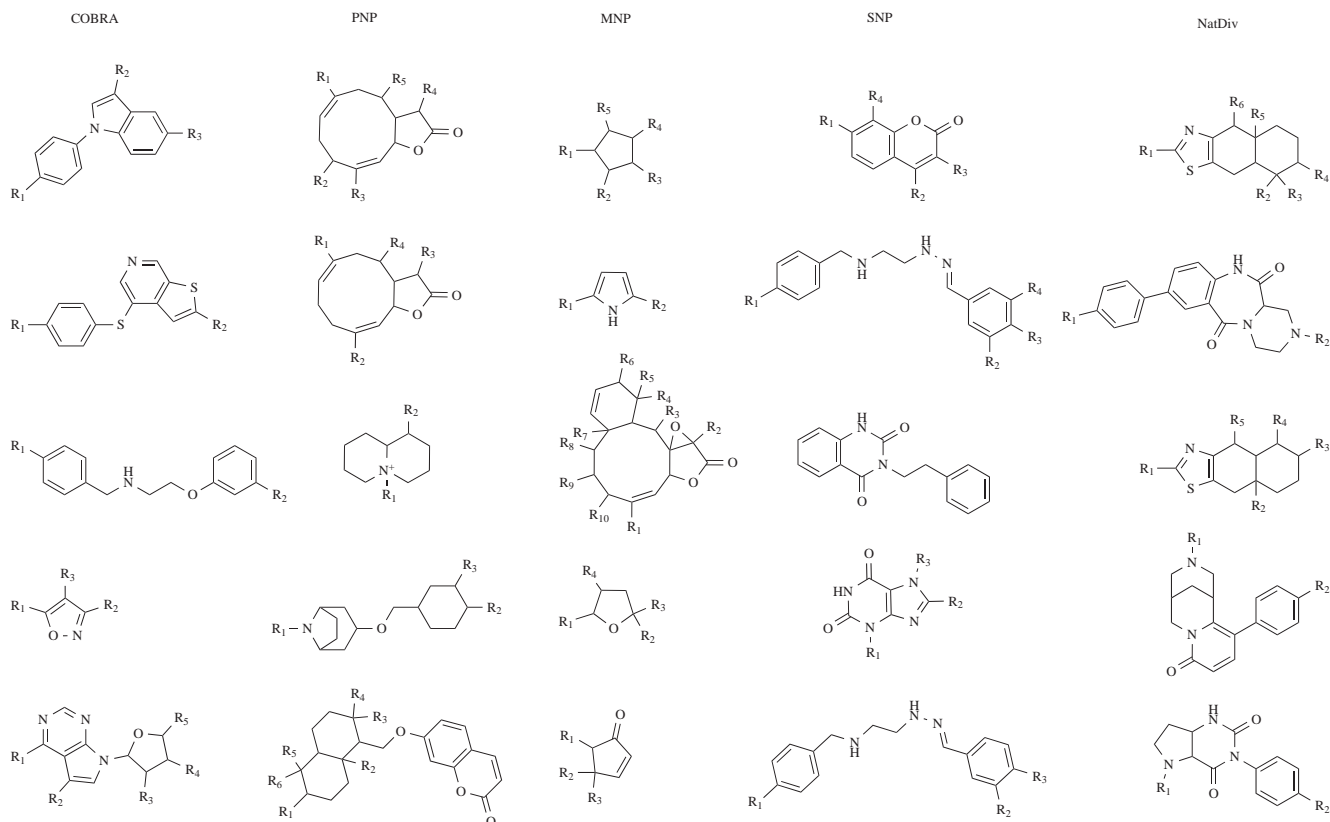


Fig. (11). Most common atom-based scaffolds with marked side-chain positions that are exclusively found in druglike compounds (COBRA), pure NPs (PNP), marine NPs (MNP), NP-derivatives and -analog (SNP), and NP-derived combinatorial compounds (NatDiv).

and NP-derivatives or -analogs which peak at one and two side-chains per scaffold. The relatively high number of natural compounds with more than seven side-chains is mainly due to glycosylation. 95% of the MNP and 88% of the PNP scaffolds contain at least two side-chains, thereby also entailing high structural diversity. Still 78% of the (semi) natural products, 70% of the COBRA, and 58% of the NatDiv scaffolds contain at least two side-chains. The distribution of heavy atoms per side-chain did not show significant differences between the individual data sets. As shown previously [13] the majority of all side-chain contains just one heavy atom, more than 60% in the case of NPs which might also be due to glycosylation (Fig. 7A). The most common side-chains are almost identical to those published by Bemis and Murcko although the order slightly changes. PNP is the only data set that does not contain halogens in the 10 most common side-chains (data not shown).

CONCLUSION

Computer-based analysis revealed that natural products exhibit a remarkable structural diversity of molecular frameworks and scaffolds that could be systematically exploited for combinatorial synthesis. Natural products offer a rich pool of unique molecular frameworks that complement "drug space". They possess desirable druglike properties rendering them ideal starting points for molecular design considerations [62-65]. Recently, the term "BIOS" (biology-oriented synthesis) was introduced by Waldmann and co-workers to highlight this concept as a "... selection of library scaffolds based on relevance to and pre-validation by Nature" [66]. Certainly, not all natural product-derived scaffolds will be directly accessible to synthesis due to their inherent structural complexity or side-chain substitution pattern. Still, systematic scaffold analysis can provide ideas about chemotype variations to the medicinal chemists. We wish to point out that the results of our study might be valid only for the time being-many more natural products will be discovered in the future. We are aware that our data set compilation and the way we defined the compound classes could be a matter of dispute. It is evident that different applications and projects will require modified definitions. Irrespective of these considerations, our study demonstrated that computational chemical biology can assist in finding suitable molecular entities in collections of natural products for drug discovery.

ACKNOWLEDGEMENTS

The authors are grateful to AnalytiCon Discovery (Potsdam, Germany) for providing their databases for analysis, and Dr. Hajo Schiewe and Dr. Lutz Müller-Kuhr for valuable discussion. Michael Meissner is thanked for proof-reading the manuscript. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften (Frankfurt am Main, Germany), and the Centre for Membrane Proteomics (Goethe-University Frankfurt am Main, Germany).

REFERENCES

- [1] Böhm HJ, Schneider G Eds, Virtual Screening for Bioactive Molecules. Weinheim, Wiley-VCH 2000.
- [2] Leach AR, Hann MM. The *in silico* world of virtual libraries. *Drug Discov Today* 2000; 5: 326-36.
- [3] Cragg GM, Newman DJ, Snader KM. Natural products in drug discovery and development. *J Nat Prod* 1997; 60: 52-60.
- [4] Newman DJ, Cragg GM, Snader KM. Natural products as sources of new drugs over the period 1981-2002. *J Nat Prod* 2003; 66: 1022-37.
- [5] Proudfoot JR. Drugs, leads, and druglikeness: an analysis of some recently launched drugs. *Bioorg Med Chem Lett* 2002; 12: 1647-50.
- [6] Fox S, Farr-Jones S, Sopchak L, Boggs A, Comley J. High-throughput screening: searching for higher productivity. *J Biomol Screen* 2004; 9: 354-8.
- [7] Breinbauer R, Manger M, Scheck M, Waldmann H. Natural product guided compound library development. *Curr Med Chem* 2002; 9: 2129-45.
- [8] Koehn FE, Carter GT. The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* 2005; 4: 206-20.
- [9] Costantino L, Barlocco D. Privileged structures as leads in medicinal chemistry. *Curr Med Chem* 2006; 13: 65-85.
- [10] Sadreyev RI, Grishin NV. Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC Struct Biol* 2006; 6: 6-20.
- [11] Ortholand JY, Ganesan A. Natural products and combinatorial chemistry: back to the future. *Curr Opin Chem Biol* 2004; 8: 271-80.
- [12] Rouhi M. Rediscovering natural products. *Chem Eng News* 2003; 81: 77-8, *ibid.* 82-3, *ibid.* 88-91.
- [13] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996; 39: 2887-93.
- [14] Krier M, Bret G, Rognan D. Assessing the scaffold diversity of screening libraries. *J Chem Inf Model* 2006; 46: 512-24.
- [15] Xue L, Bajorath J. Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases. *J Mol Model* 1999; 5: 97-102.
- [16] Schneider P, Schneider G. Collection of bioactive reference compounds for focused library design. *QSAR Comb Sci* 2003; 22: 713-8.
- [17] AnalytiCon Discovery GmbH, Hermannswerder Haus 17, D-14473 Potsdam, Germany. <http://www.ac-discovery.com>
- [18] InterBioScreen Ltd., 121019 Moscow, P.O. Box 218, Russia; <http://www.ibscreen.com/>.
- [19] Specs, Delftechpark 30, 2628 XH Delft, The Netherlands; <http://www.specs.net>
- [20] MicroSource, Discovery Systems, Inc., 21 George Washington Plaza, Gaylordsville, CT 06755 USA, <http://www.msdiscovery.com>.
- [21] Vitas-M Laboratory, Ltd., Center of Molecular Medicine, Vorob'evi Gori, 119829 Moscow, Russia. <http://www.vitasmmlab.com>
- [22] Moscow Medchemlabs, 1812 year str. 7 apt.6, 121170 Moscow, Russia, <http://www.mosmedchemlabs.com>
- [23] Faulkner DJ. Marine natural products. *Nat Prod Rep* 2001; 18: 1-49.
- [24] Faulkner DJ. Marine natural products. *Nat Prod Rep* 2002; 19: 1-48.
- [25] Faulkner DJ. Marine natural products. *Nat Prod Rep* 2000; 17: 7-55.
- [26] Blunt JW, Copp BR, Munro MH, Northcote PT, Prinsep MR. Marine natural products. *Nat Prod Rep* 2004; 21: 1-49.
- [27] Blunt JW, Copp BR, Munro MH, Northcote PT, Prinsep MR. Marine natural products. *Nat Prod Rep* 2005; 22: 15-61.
- [28] Blunt JW, Copp BR, Munro MH, Northcote PT, Prinsep MR. Marine natural products. *Nat Prod Rep* 2006; 23: 26-78.
- [29] Blunt JW, Copp BR, Munro MH, Northcote PT, Prinsep MR. Marine natural products. *Nat Prod Rep* 2003; 20: 1-48.
- [30] Faulkner DJ. Highlights of marine natural products chemistry (1972-1999). *Nat Prod Rep* 2000; 17: 1-6.
- [31] Bhadury P, Mohammad BT, Wright PC. The current status of natural products from marine fungi and their potential as anti-infective agents. *J Ind Microbiol Biotechnol* 2006; 33: 325-37.
- [32] König GM, Kehraus S, Seibert SF, Abdel-Lateff A, Müller D. Natural products from marine organisms and their associated microbes. *Chembiochem* 2006; 7: 229-38.
- [33] Costantino V, Fattorusso E, Menna M, Tagliatalata-Scafati O. Chemical diversity of bioactive marine natural products: an illustrative case study. *Curr Med Chem* 2004; 11: 1671-92.
- [34] Liu B, Zhou J. SARS-CoV protease inhibitors design using virtual screening method from natural products libraries. *Comput Chem* 2005; 26: 484-90.

- [35] Lei J, Zhou J. A marine natural product database. *J Chem Inf Comput Sci* 2002; 42: 742-8.
- [36] Alonso D, Khalil Z, Satkunanathan N, Livett B. G. Drugs from the sea: conotoxins as drug leads for neuropathic pain and other neurological conditions. *Mini Rev Med Chem* 2003; 3: 785-7.
- [37] Cragg GM, Newman DJ, Yang SS. Natural product extracts of plant and marine origin having antileukemia potential. The NCI experience. *J Nat Prod* 2006; 69: 488-98.
- [38] Molecular Networks GmbH, Computerchemie, Nägelsbachstraße 25, D-91052 Erlangen, Germany. <http://www.mol-net.de>
- [39] Chemical Computing Group, 1010 Sherbrooke St. West, #910, Montreal, Canada H3A. <http://www.chemcomp.com>
- [40] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997; 23: 3-25.
- [41] Oprea TI. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* 2000; 14: 251-64.
- [42] Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci* 1999; 39: 868-73.
- [43] Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem* 2000; 43: 3714-7.
- [44] Engel, T. In: Gasteiger J, Engel T Eds, *Chemoinformatics*. Weinheim, Wiley-VCH 2003; 31-2.
- [45] Tan DS. Current progress in natural product-like libraries for discovery screening. *Comb Chem High Throughput Screen* 2004; 7: 631-43.
- [46] Feher M, Schmidt JM. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci* 2003; 43: 218-27.
- [47] Henkel T, Brunne R, Reichel F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew Chem Int Ed* 1999; 38: 647-9.
- [48] Schneider G, Lee ML, Stahl M, Schneider P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J Comput Aided Mol Des* 2000; 14: 487-94.
- [49] Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Meth* 2000; 44: 235-49.
- [50] Zheng S, Luo X, Chen G, *et al.* A new rapid and effective chemistry space filter in recognizing a druglike database. *J Chem Inf Model* 2005; 45: 856-62.
- [51] MDL Information Systems Inc., San Leandro, CA, 94577. <http://www.mdli.com/>
- [52] Chinese Natural Product Database is available from the Shanghai Institute of Material Medica and contains the Chinese natural compounds published.
- [53] Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J Chem Inf Comput Sci* 2003; 43: 1882-9.
- [54] Pickett S. In: Boehm HJ, Schneider G Ed, *The biophore concept, in: Protein-Ligand Interactions*. Weinheim, Wiley-VCH. 2003; 73-106.
- [55] van de Waterbeemd H, Gifford E. ADMET *in silico* modelling: towards prediction paradise? *Nat Rev Drug Discov* 2003; 2: 192-204.
- [56] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 2002; 12: 615-23.
- [57] Martin Y. A bioavailability score. *J Med Chem* 2005; 48: 3164-70.
- [58] Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N. Strategic pooling of compounds for high-throughput screening. *J Chem Inf Comput Sci* 1999; 39: 897-902.
- [59] Kensil CR, Patel U, Lennick M, Marciani D. Separation and characterization of saponins with adjuvant activity from *Quillaja saponaria* Molina cortex. *J Immunol* 1991; 146: 431-7.
- [60] Schaed SG, Klimek VM, Panageas KS, *et al.* T-cell responses against tyrosinase 368-376(370D) peptide in HLA*A0201+ melanoma patients: randomized trial comparing incomplete Freund's adjuvant, granulocyte macrophage colony-stimulating factor, and QS-21 as immunological adjuvants. *Clin Cancer Res* 2002; 8: 967-72.
- [61] Zheng L, Zheng J, Zhao Y, Wang B, Wu L, Liang H. Three anti-tumor saponins from *Albizia julibrissin*. *Bioorg Med Chem Lett* 2006; 16: 2765-8.
- [62] Kulkarni BA, Roth GP, Lobkovsky E, Porco JA Jr. Combinatorial synthesis of natural product-like molecules using a first-generation spiroketal scaffold. *J Comb Chem* 2002; 4: 56-72.
- [63] Liao Y, Hu Y, Wu J, *et al.* Diversity oriented synthesis and branching reaction pathway to generate natural product-like compounds. *Curr Med Chem* 2003; 10: 2285-316.
- [64] Mang C, Jakupovic S, Schunk S, Ambrosi HD, Schwarz O, Jakupovic J. Natural products in combinatorial chemistry: an andrographolide-based library. *J Comb Chem* 2006; 8: 268-74.
- [65] Koch MA, Waldmann H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov Today* 2005; 10: 471-82.
- [66] Nören-Müller A, Reis Corrêa Jr I, Prinz H, *et al.* Discovery of protein phosphatase inhibitor classes by biology-oriented synthesis. *Proc Natl Acad Sci USA* 2006; 103: 10606-11.