J Comput Aided Mol Des (2013) 27:125–160
DOI 10.1007/s10822-012-9629-z

Oxygen-containing fragments in natural products

Zoya Titarenko · Natalya Vasilevich ·
Vladimir Zernov · Michael Kirpichenok ·
Dmitry Genis

Received: 25 September 2012 / Accepted: 17 December 2012 / Published online: 28 December 2012

Abstract An analysis of the chemical environment of the oxygen atoms in the DNP database compared to the CMC and SCD databases was performed. Some structural clusters were identified which are predominant among the natural products and can be considered as distinctive features of NPs. Fifty-three oxygen-containing structural fragments that are distinctive for the DNP (distinctive set of fragments DSF) in comparison with the SCD have been identified. A new descriptor Mc was introduced for describing the ratio of atoms involved in the DSF to the total number of heavy atoms. A significant difference in the Mc values among the reference databases allowed the use of a specific cluster of the DSF as a tool for performing similarity searches for oxygen-containing NP molecules, or for evaluation or comparison of databases according to their NP-likeness. An example illustrating that the suggested approach could allow not only estimating the NP-likeness, but also serve as a tool for designing new NP-like compounds is provided. The suggested approach for NP-likeness evaluation moves away from the traditional ideas of scaffolds, cycles, linkers and substituents.

Keywords Natural products · Oxygen-containing structural fragments · NP-likeness · NP-likeness score · Distinctive set of fragments

Introduction

In recent years we have seen a renewed and rising interest in natural products (NPs) and NP-like compounds in relation to the drug discovery process [1–8]. In this context a detailed analysis of the structural features of NPs in comparison with the existing drugs database (CMC) and the available commercial screening collections (SCD) seems to be particularly interesting and could provide us with new ideas for compound design. As a source of natural products for analysis the Database of Natural Products (DNP) is often used. Very careful and detailed comparison of global physicochemical properties and some structural features of these model databases have been previously presented in the literature [9, 10]. It is worth mentioning an interesting study devoted to the design of NP-like virtual libraries using a new molecular enumerator [11].

Generally speaking, one of the main distinctive features of NPs, along with a high content of asymmetric and saturated carbon atoms is the enriched amount of oxygen atoms [9, 10]. However, various classes of organic compounds can be found among natural products that stand in stark contrast in terms of oxygen content, for example, sugars versus terpenes and so on. There are several publications that provide a comparative analysis of fragments containing oxygen atoms in the DNP, the CMC, and the SCD databases [12, 13]; however to the best of our knowledge no systematic analysis based on structural fragments has yet been performed.

Materials and methods

Databases

Three basic databases were used: the Dictionary of Natural Products (DNP), release-191-2009 (232,923 molecules, http://www.crcpress.com/product/isbn/9780412491504), the MDL Comprehensive Medicinal Chemistry database (CMC),

Z. Titarenko · N. Vasilevich (✉) · V. Zernov · M. Kirpichenok · D. Genis
ASINEX, 20 Geroev Panfilovtsev Str., Moscow 125480, Russia
e-mail: nvasilevich@asinex.com

release 2008.1 (8987 molecules, http://accelrys.com/products/databases/bioactivity/comprehensive-medicinal-chemistry.html), and the MDL Comprehensive Screening Compounds Directory database (SCD), release 2008.4 (7,237,042 molecules, http://accelrys.com/products/databases/sourcing/screening-compounds-directory.html). These databases were treated as previously published [14, 15] to remove duplicates, normalize charges and remove counterions. Sugar units and glycoside parts of molecules were kept. Final databases contained: the DNP—154,175 entities, the CMC—8530 entities, and three random sets from SCD database including 198,418, 150,000, and 150,000 entities (produced by means of the Microsoft Access Random() function). Since deviation in the data obtained for these selections did not exceed 5 %, in this paper we specify the data for the set of 198,418 compounds.

Only statistically significant structural fragments that encompass not less than 0.1 % of total oxygen atoms in the model databases were considered.

Atom typing

Atom types other than carbon, nitrogen, and oxygen were not additionally differentiated. When necessary C, N and O atoms were split into aromatic (Ca) and aliphatic; aliphatic and unsaturated fragments were divided into atoms in linear chains (Cs) and atoms in non-aromatic cycles (Cr), and a site of any linker or substituent attachment in a cycle was identified (Fig. 1a).

Fragment count

Applied search engine included a count of each fragment in the molecule with possible overlaps. Figure 1b demonstrates an example: the query applied can be met in target molecule six times, so the hit count is equal to 6. Parameter Mc (molecular coverage) that denotes the ratio of C and O atoms belonging to the query to the total number of heavy atoms in the molecule is equal to 0.92 since all atoms except the oxygen belong to the query.

Structure preprocessing

In some cases it was necessary to eliminate from consideration some monotonous repeating fragment (for example, ester, carbohydrate, steroid or any other) and then analyzed remaining structural pieces for hit count. These structural pieces were considered as a single molecule not as two separate ones. In the example presented in Fig. 1c the ester fragment was excluded, transforming the ester molecule into two tetrahydropyran rings. In spite of the absence of any linker between them these two rings were considered as a single molecule and the hit count was performed in its usual way.

Software

Symyx Cheshire 4.1 SDK (including classes and methods for treatment of molecules and substructure search) was used for hit count (Accelrys Cheshire http://accelrys.com/products/informatics/cheminformatics/accelrys-cheshire.html) as a substantial part of our own software written in c# language.

Chemical structures

Chemical structures were drawn and analyzed using MDL® ISIS/Draw 2.5 and MDL® ISIS/Base 2.5 software.

Results and discussion

General approach

Continuing the development of our fragment approach for the analysis of properties of NPs [16], in the present study we attempted to reveal the particular structural fragments that predominantly contain oxygen atoms in the DNP and to find out how the DNP differs from two other model databases, the CMC and the SCD, in this respect. It is worth noting that the analysis of oxygen-containing fragments was performed on the whole DNP database without relating to any particular class of natural products. In this way we tried to consider and classify all oxygen atoms present in the DNP database in the context of the type of their chemical environment.

Three basic databases were chosen for our comparative analysis, namely: the DNP (232,923 molecules), the database of existing drugs—CMC (8987 molecules), and a representative selection including about 200,000 randomly selected compounds from the database of available collections for screening—the SCD (7,237,042 molecules). The later set was employed assuming that such an approximation can be considered as a hypergeometric distribution [14, 17] and it is suggested to be good for subsets of less than 10 % [14]. After removal of duplicates, normalizing charges and removal of counter ions etc. as previously published [15, 18], the DNP database contained 154,175 unique structures, the CMC contained 8,432 unique structures, and the selection from the SCD contained 198,258 entities. However, as opposed to analysis performed by authors of previously published methods [10–12, 15], we deliberately kept sugar units as well as glycoside parts of molecules, considering their presence as important features of natural products. Taking into account

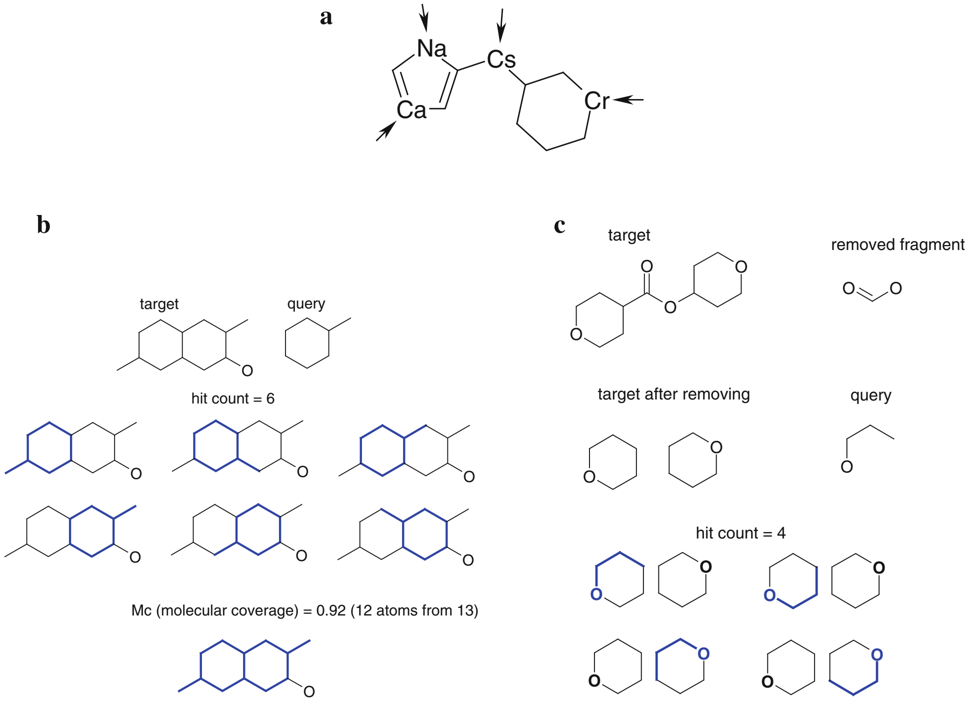


Fig. 1 Illustration of fragment count approach: **a** example of atom typing; **b** example of fragment count strategy; **c** example of fragment count strategy when some fragment is necessary to remove

the high diversity of the DNP we limited our analysis to consideration of statistically significant structural fragments that encompass not less than 0.1 % of total oxygen atoms in the model database.

We used our proprietary software for the analysis. Our approach is conceptually close to that described by Ertl et al. [18] that has developed the method of NP-likeness score evaluation using ideas of Bremser [19] about atom centered fragments. However, in contrast to the described approach we considered molecular structural fragments from the very general point of view as labeled (marked) molecular subgraphs [20] not attaching them to the atom centered fragments only. We were also not attached to any structural units such as cycles, linkers and side chains that are commonly used in generally accepted approaches using Scaffold concept [2, 21, 22] adapted to classical theory of organic chemistry. In general, we considered that a fragment may consist of any arbitrary part of the molecular structural formula. For example it can begin in a cycle and go beyond its scope, or consist of any part of a polycyclic system, or contain any chemical bond, and substituent etc.

Moreover, our program differentiates atomic types in particular details, for instance it allows us to find separate carbon atoms in aromatic (Ca) and aliphatic, unsaturated fragments, to distinguish atoms in linear chains (Cs), and in non-aromatic cycles (Cr), as well as to identify a site of any given linker or substituent attachment in a cycle. Hence in various program settings one can consider any types of bonding simultaneously or, for instance, only cyclic non-aromatic bonds (labeled hereafter as “Rn”), or limit the analysis to only aromatic bonds (“Ar” symbol), or to only single or double or triple bonds in linear chains (“Ch” symbol).

Additionally our software allows eliminating from consideration any monotonous repeating fragment (for example, ester, carbohydrate, steroid or any other) and then analyzing structural features of the remaining pieces of structural formula of organic molecules.

The following symbolism was used in the current paper. Designation of the particular structural fragment[1] by means of single bonds actually denotes any non-aromatic (single, double or triple) bonds in this fragment. Double and triple bonds and aromatic cycles wherever necessary are shown in the usual way. Dashed lines denote any type of bond including aromatic ones. Particular bonds or cyclic systems are marked in a blue colour. Labeling of a double bond with a red colour denotes an aromatic bond. All formulae given in schemes generally describe only part but not the entire molecule and permit any other substituents or cycles. Areas where a particular substituent can be localized are outlined by dashed lines.

Within the present survey our focus has concentrated on a comparison with the patterns found primarily in the DNP database. This database was chosen as a point of departure and the features of this database were systematically analyzed. Specific characters of the CMC and the SCD databases were of secondary significance and were analyzed only for comparative purpose.

[1] Particular structural fragments, presented on schemes, are abbreviated in text as “f” (from ‘fragment’).

To address this problem the total oxygen content in molecular formulae in the model databases was counted. Then we moved level by level, in accordance with continuous complicacy of the closest chemical environment of the examined oxygen atom. We tried to maintain the balance of oxygen atoms in each level and, as was mentioned above, considered only abundant structural clusters that include not less than 0.1 % of the total oxygen atoms in the DNP database.

Our approach can be illustrated in the following way. The DNP database contains 1,065,723 oxygen atoms; this value is taken as 100 %. According to Scheme 1.1, 76 % of total oxygen is bound with carbon by single bonds, at the same time 41 % is found in hydroxy-groups, 22 % in carbonyl groups, and the remaining oxygen is bonded with heteroatoms. This distribution shows the first level of the chemical environment of the oxygen atoms. Even at this level one can observe essential differences between the DNP, the CMC and the SCD databases (Scheme 1.1). Graphical correlation within the first level is shown in Fig. 2. Structural fragments O–H and C–O appeared to dominate in the DNP database. On the contrary, the C=O, N–O, N=O, and S=O fragments are more common in the SCD and the CMC databases than in the DNP database.

Applying this algorithm one should be aware that the O atom could apparently belong to two different structural fragments simultaneously, for instance, to both O–H and C–O moieties. In this case a subtractive procedure can be applied to elaborate the balance of oxygen: the total number of O–H and C–O fragments is counted followed by subtraction of their intersection, i.e. the number of C–O–H fragments. Such a superposition actually means a transition to structural fragments of the next level or an extension of the level of the oxygen chemical environment. Thus, analysis of the structural level shown in Scheme 1.2 reveals that NP hydroxy groups (41.1 %) are almost entirely included in carbon containing fragments C–O–H (40.6 %).

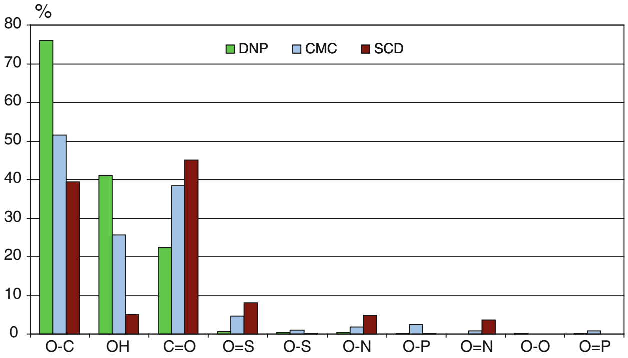
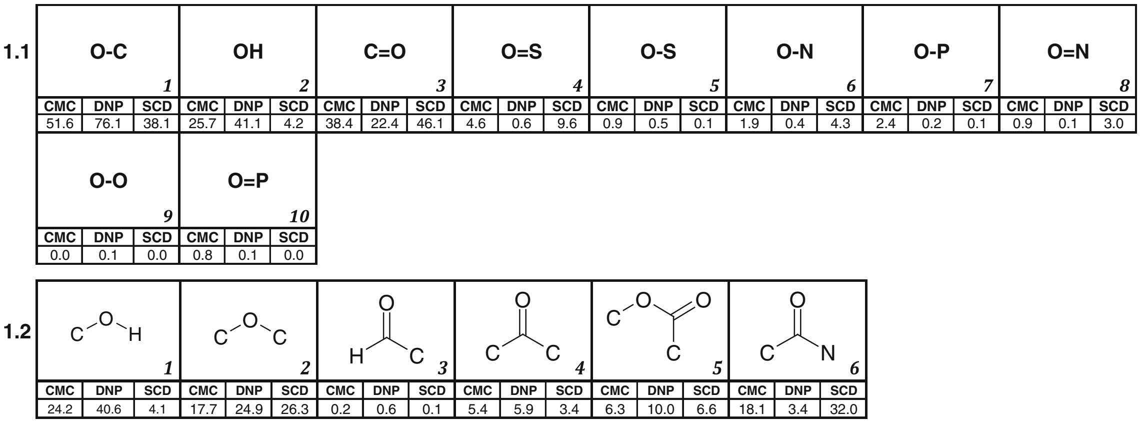


Fig. 2 Abundance of oxygen containing two-atomic structural fragments in the DNP, CMC and SCD databases

Since this approach implies the sharp rise of the number of the levels and structural elements in them, in the course of the analysis we tried to limit ourselves to those familiar to chemists and generally accepted in organic chemistry classes of compounds. Thus we skipped a number of obvious intermediate subgraphs. A general structural fragment discussed below usually constitutes a part of other larger fragments, composing structural formulae of a particular compound or a class of compounds. Thus cholic acid (Fig. 3) includes 3 cyclohexane and 2 decaline structural fragments (one example of each is colored) composing a structural formulae of steroids. The cholic acid contributes 5 oxygen atoms to the total balance of oxygen in the DNP: three of them are included in secondary hydroxy groups that are linked to non-aromatic cycles, one in a carboxylic hydroxy-group, and the last one is in a carbonyl group.

Let us now consider in detail some key oxygen-containing structural fragments significant for the DNP database.



1.1

O-C 1			OH 2			C=O 3			O=S 4			O-S 5			O-N 6			O-P 7			O=N 8		
CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD
51.6	76.1	38.1	25.7	41.1	4.2	38.4	22.4	46.1	4.6	0.6	9.6	0.9	0.5	0.1	1.9	0.4	4.3	2.4	0.2	0.1	0.9	0.1	3.0

O-O 9			O=P 10		
CMC	DNP	SCD	CMC	DNP	SCD
0.0	0.1	0.0	0.8	0.1	0.0

1.2

1			2			3			4			5			6		
CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD	CMC	DNP	SCD
24.2	40.6	4.1	17.7	24.9	26.3	0.2	0.6	0.1	5.4	5.9	3.4	6.3	10.0	6.6	18.1	3.4	32.0

Scheme 1 Overview of oxygen-containing structural fragments. *1.1* Two-atom structural fragments; *1.2* structural fragments corresponding to main classes of organic molecules

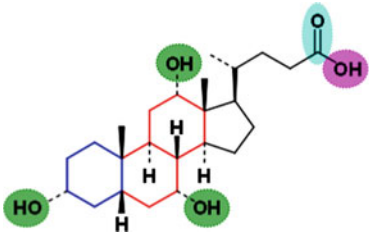
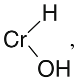


Fig. 3 Some structural fragments of cholic acid: (a) *Blue line* cyclohexane, (b) *orange line* decaline, (c) *dark green circle* secondary hydroxyl-group linked to a cycle  , (d) *rose circle* OH of carboxylic group, (e) *light green circle* C=O of carbonyl group

Oxygen-containing fragments

C–O–H fragments

Hydroxy-groups in the DNP database are mainly found in three structural classes: alcohols (28.0 %), phenols (9.5 %) and carboxylic acids (2.2 %).[2] Specification of alcohols as primary, secondary etc. can be seen in Schemes 2 and 3.

In this section we did not elaborate in detail the distribution of hydroxy groups in common carbohydrates such as particular hexoses, pentoses, disaccharides and so on, which are considered as a one-type cluster. It is much more interesting to reveal other classes of natural products, where the oxygen of hydroxy-groups is distributed. Using a de-glycosylation procedure we performed an estimation of the percentage of oxygen atoms in C–O–H fragments of aglyconic parts (values within the brackets hereafter in Schemes), that seem to be of particular importance for the design of new NP-like compounds.

Acyclic alcohols There are a significant number of primary alcohols in the DNP database (4.5 %, Scheme 2.1, f. 1), many of them belong to structural fragments, where oxy-methyl groups are connected to non-aromatic cycles (f. 3). Alcohols containing acyclic carbon in the β-position (f. 2) are less common, and corresponding aryl derivatives (f. 4) are far less frequent and almost entirely include benzyl alcohols. There is a big cluster of primary alcohols among natural products that contains unsubstituted or substituted oxy-groups in the β- or γ-position (f. 6, 7).

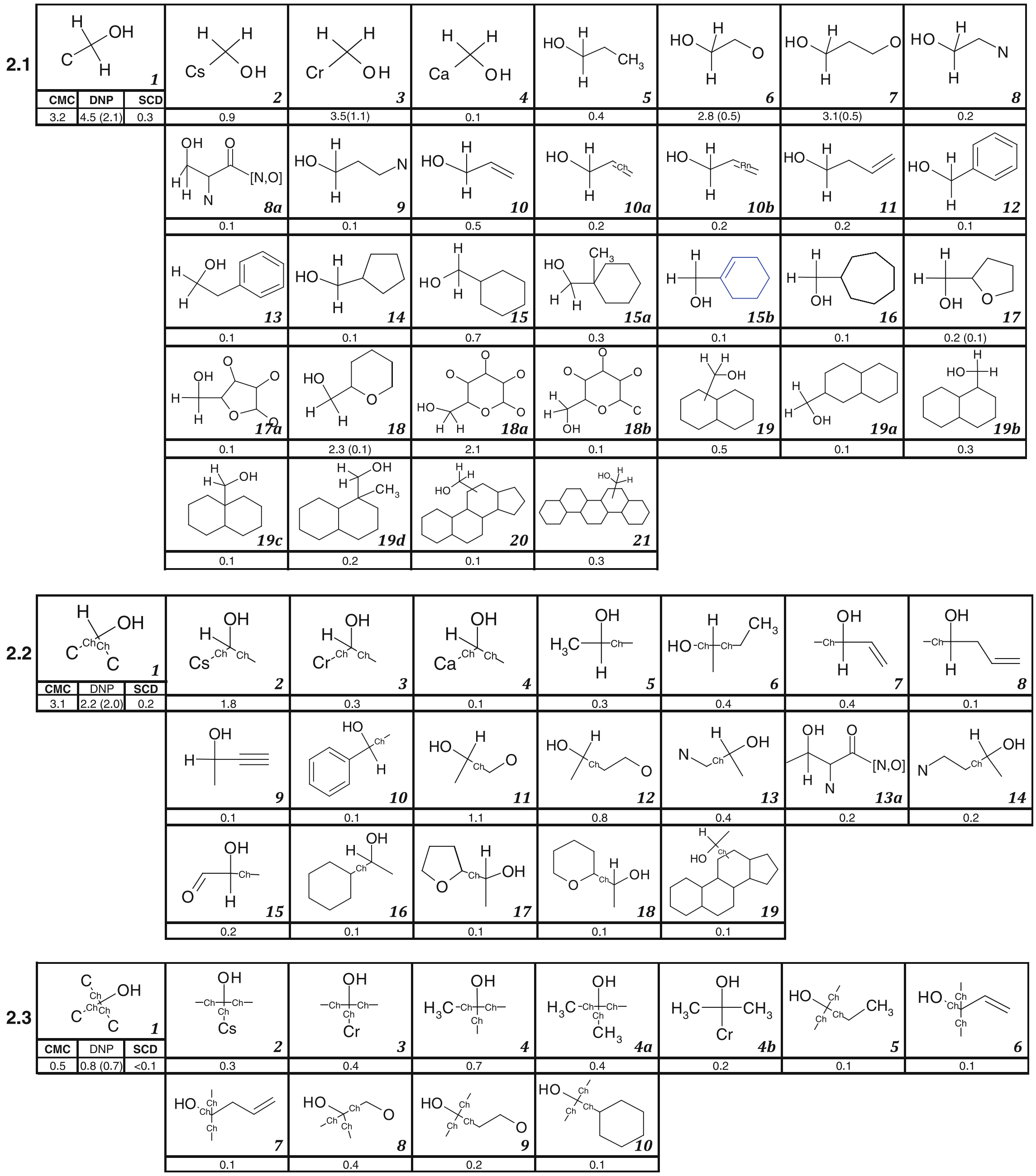
A portion of primary alcohols belonging to aglycones, accounts for approximately half of their total value. The distinctive feature of primary alcohol fragments is the significant fraction of β-methyl derivatives (f. 5), found mainly in cyclohexane derivatives and in methyldecalines. Various oxy- and amino-alcohol fragments are also essential. Detailed consideration of primary amino alcohols (f. 8) reveals that one-third of them are included in aliphatic linear fragments, a little more in serine fragments and a small amount in hydroxy-amino-cyclohexane. Allyl and homoallyl alcohol derivatives can be found among primary alcohols such as diterpenoid Vibsanin F (Table 1, ex. 2.1-1).

Most of the fragments described by subgraph f. 3 belong to cyclohexanes, which evolve into decaline structure (f. 15 → 19), and then into steroids and some other terpenoids. Within the considered cluster one can find some aglycones such as cyclopentane and cycloheptane derivatives. The latter are not usually independent and can be parts of bi- and polycyclic molecules such as Daphlongeranine A—an alkaloid from the fruits of Daphniphyllum Longeracemosum (ex. 2.1-2).

The acyclic secondary alcohols structural class contains twofold fewer compounds than the class of primary alcohols in the DNP database (Scheme 2.2, f. 1). Aglycones are predominant within this class, and the majority of secondary alcohols are aliphatic (f. 2). The portion of alcohols containing cycles in the α-position is significantly less (f. 3, 4). General subgraph pattern of the nearest chemical environment is presented by fragments f. 5–15. Among secondary alcohols of this type one can note the clusters of various oxy- and amino alcohols comprising threonine derivatives. Allyl, homoallyl, propargyl and benzyl alcohols can be also observed in this cluster; α- and β-methyl alcohols are quite common as well. Secondary hydroxy-groups are also observed in α-oxy carboxylic acids and their derivatives for instance, in the molecule of the protease inhibitor Bestatin (ex. 2.2-3). The most valuable monocyclic fragments originated from subgraph 3 belong to cyclohexane, tetrahydrofuran, and tetrahydropyran clusters (f. 16–18). The cyclohexane cluster partly evolves into the steroid class.

The DNP database contains a rather insignificant number of tertiary acyclic alcohols (Scheme 2.3, f. 2–10); almost all of them are aglycones. More than half of aglycones contain a carbinol group connected with the non-aromatic cycle; the remaining ones belong to the aliphatic cluster. Virtually all tertiary acyclic alcohols are α-methyl carbinols such as Thiostreptine (ex. 2.3-2)—a product of degradation of the natural oligopeptide antibiotic Thiostreptone. A great deal of α-methyl carbinols evolve then into α, α-dimethyl derivatives and in turn the dimethyl carbinol group in tertiary alcohols is often connected directly with a cyclic non-aromatic system. Generally all trends in the tertiary alcohol cluster are similar to those observed for the secondary alcohols, for example, one can find unsaturated alcohols, β- and γ-hydroxy carbinols etc. Some examples of natural products that belong to acyclic alcohols are presented in Table 1.

[2] Hereinafter presented values denote the percentage of a particular structural fragment in the DNP database (if attribution to the CMC or the SCD database is not specially mentioned).

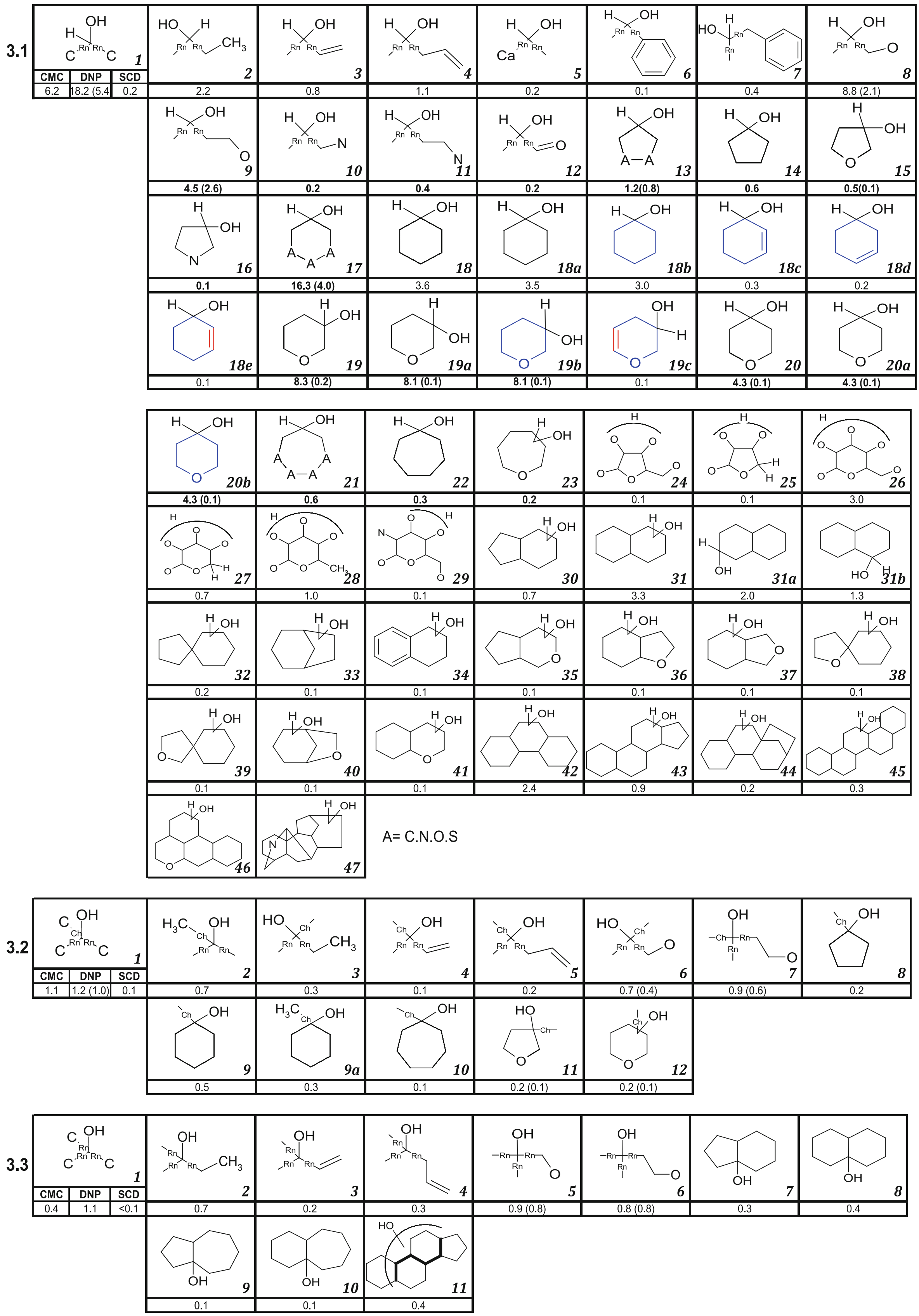


Scheme 2 Acyclic C–O–H structural fragments: *2.1* primary alcohols; *2.2* secondary alcohols; *2.3* tertiary alcohols

Concluding this section, it can be noted that the portion of acyclic alcohols in the DNP database is generally comparable with that in the CMC database and significantly exceeds the level of similar C–O–H fragments in the SCD database (Fig. 4), even if carbohydrates are excluded from consideration.

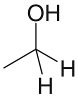
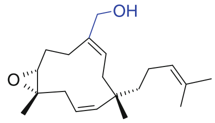
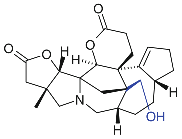
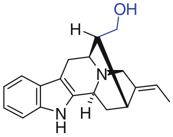
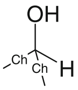
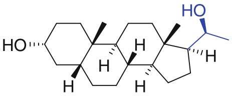
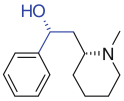
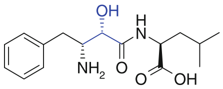
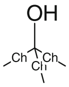
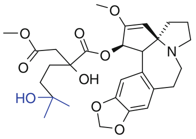
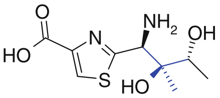
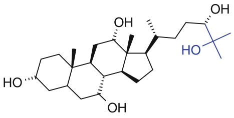
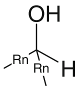
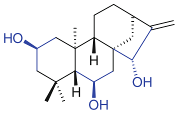
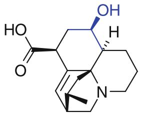
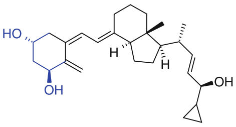
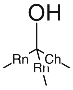
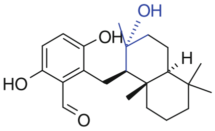
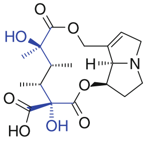
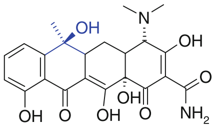
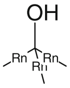
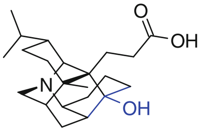
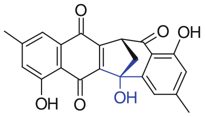
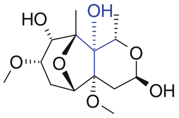
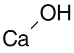
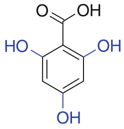
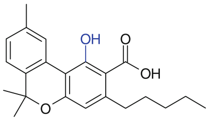
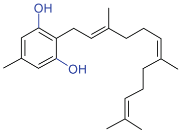
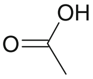
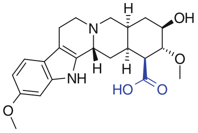
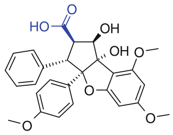
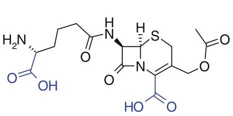
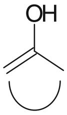
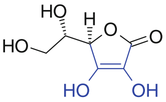
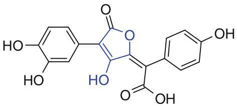
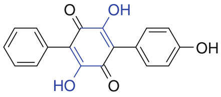
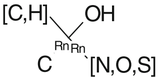
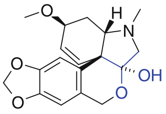
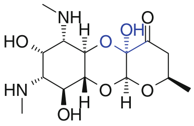
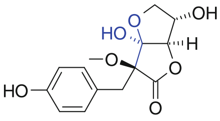
Cyclic secondary and tertiary alcohols The content of cyclic secondary alcohols in the DNP database is essentially higher than the content of acyclic ones (Scheme 3.1, Fig. 3). A high content of cyclic sp3 carbon atoms was continually mentioned in the literature [9, 10] as a distinctive feature of natural products. Obviously, this fact can be explained by the presence of carbohydrate residues. However, some other points can be noted, that are worth considering in more detail.

Considering the nearest environment of carbinol groups in the secondary cyclic alcohols one can easily note clusters containing hydroxy- or alkoxy groups in the β- or γ-position and containing methyl groups in the β-position of the cycle (f. 2). It is also worth mentioning cyclic allyl, homoallyl, benzyl and phenylethyl alcohol moieties as well



Scheme 3 Cyclic C–O–H structural fragments: *3.1* secondary alcohols; *3.2* and *3.3* tertiary alcohols

Table 1 Examples of natural products comprising various C–O–H fragments

Scheme	Fragment	%	Example 1	Example 2	Example 3
2.1		4.5	 Vibsanin F	 Daphlongeranine A	 Tombozine
2.2		2.2	 Pregnane-3,20-diol	 Sedamone	 Bestatin
2.3		0.8	 Harringtonine	 Thiostreptine	 Cholestane-3,7,12,24,25-pentol
3.1		18.2	 16-Kaurene-2,6,15-triol	 Lannotinidine A	 Calcipotriol
3.2		1.2	 Siphonodictyal A	 Sceleratine	 Tetracycline
3.3		1.1	 Dapholdhamine B	 Natalenone	 Urechitol B
4.1		9.5	 Phloroglucinic acid	 Cannabinolic acid	 Grifolin
5		2.2	 Reserpic acid	 Rocaglioic acid	 Cephalosporin C
6.1		0.4	 Ascorbic acid	 Xerocomic acid	 Ascocorynin
6.2		0.6	 Tazettine	 Spectinomycin	 Delesserine

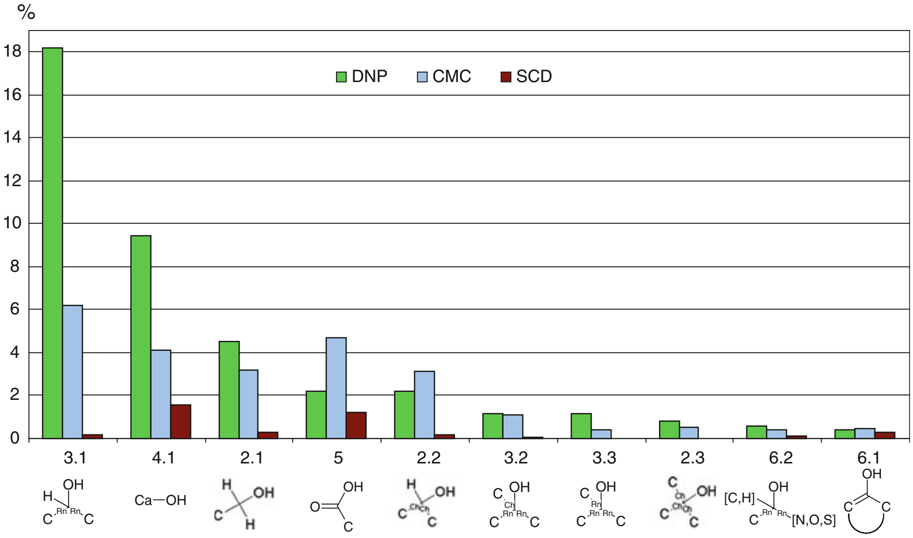


Fig. 4 Abundance of C–O–H fragments in DNP, CMC and SCD databases

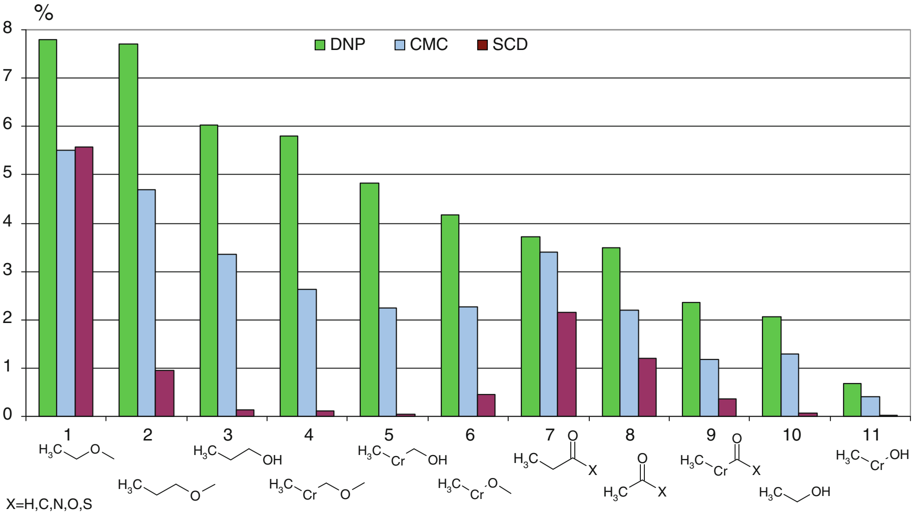


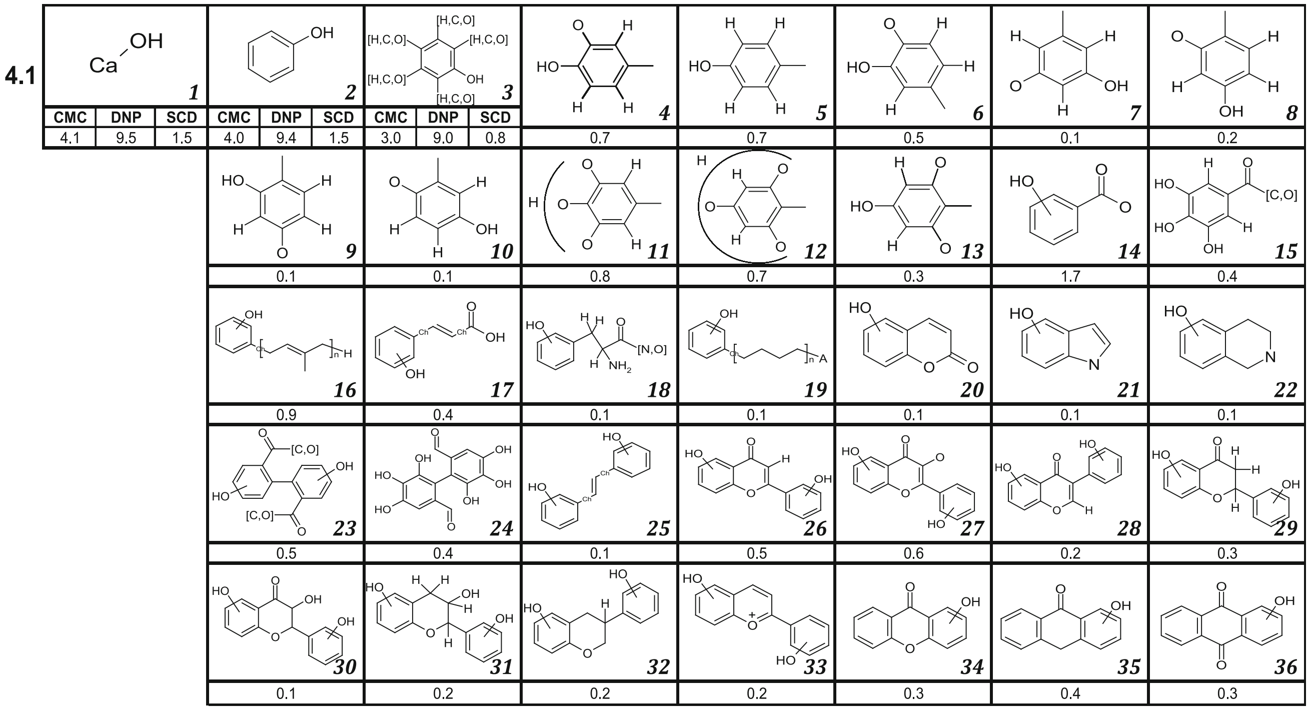
Fig. 5 Abundance of α- and β-methyl substituted oxygen-containing structural fragments in the DNP, CMC, and SCD databases

as α-hydroxy carbonyl derivatives, namely α-hydroxy lactones, lactams and ketones. A fraction of various amino alcohols is also quite essential; among them one can observe β-oxyproline moieties.

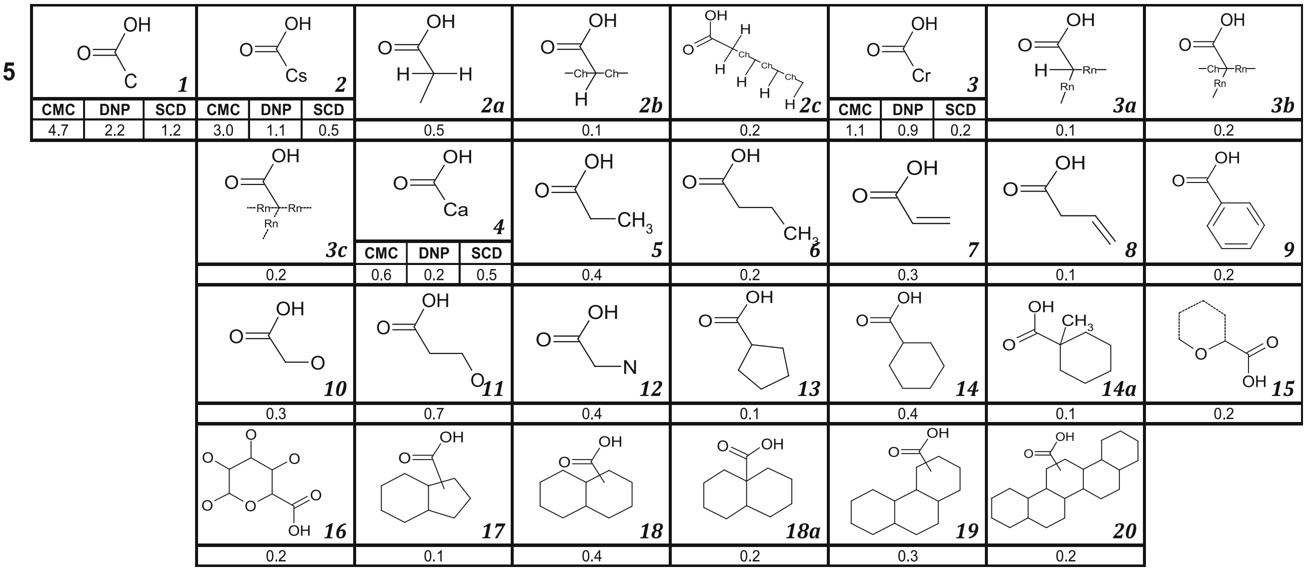
However, expectedly, the biggest clusters of this type are β- and γ-oxy alcohol moieties (f. 8, 9), which compose inter alia parent graphs of carbohydrates. General analysis of the secondary cyclic alcohols shows that the majority of them are 6-member followed by 5-member and 7-member cycles, containing carbon atoms in the β-position to the hydroxy-groups. As expected, a major part of 5- and 6-member cycles are tetrahydrofurans and tetrahydropyrans composing the skeleton of carbohydrate molecules. Examples of common carbohydrate moieties (f. 24–29) are shown in Scheme 3.1. After excluding the common carbohydrate moieties containing glycoside oxygen atoms from consideration, hydroxy derivatives of aglycones are estimated to account for 5.4 % of total oxygen atoms.

As follows from Scheme 3.1 the portion of secondary hydroxy-groups in tetrahydrofuran, tetrahydropyran and oxepane moieties among aglycones is not sufficient. At the same time the fraction of 5- and 6-member rings in aglycones remains relatively high; this fraction mainly consists of carbocycles. Scheme 3.1 demonstrates detail subgraphs of parent cyclohexane and tetrahydropyran (f. 18–20) moieties including subgraph considering any non-aromatic bonds and its particular subtypes. Fragments of 7-member cycles such as cycloheptane and oxepane also serve as an aglycones source. However, these carbocycles are usually included into more complex polycyclic systems in real molecules. Typical representatives of this cluster are Vitamin D derivatives such as Calcipotriol (ex. 3.1-3).

Further analysis of cyclohexane and cycloheptane Cr–O–H fragments reveals that a significant number of them are included in decaline systems, octahydroindene



Scheme 4 Aromatic C–O–H structural fragments: *4.1* phenols

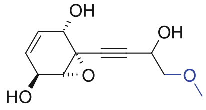
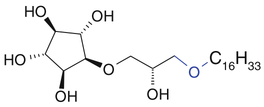
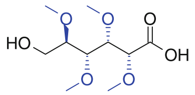
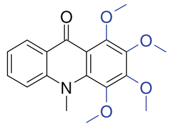
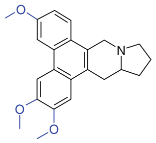
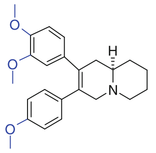
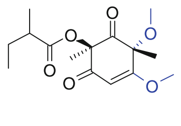
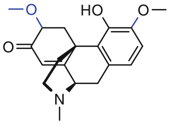
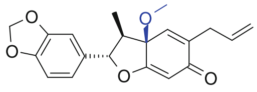
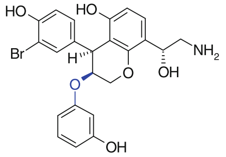
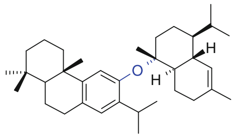
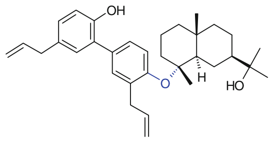
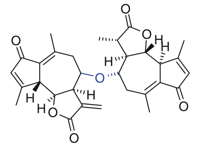
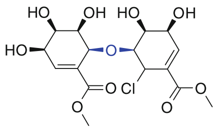
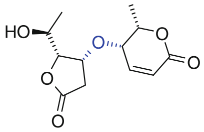
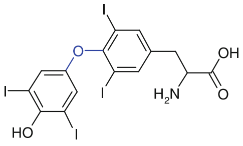
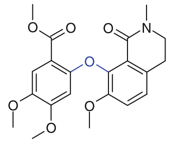
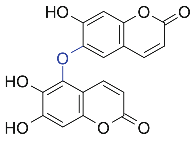
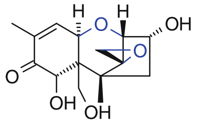
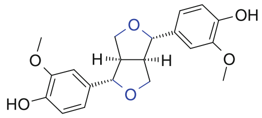
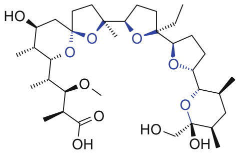
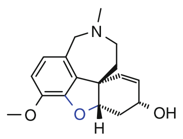
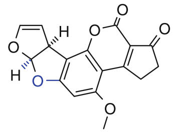
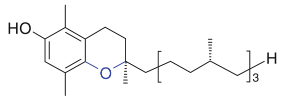
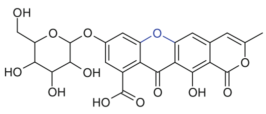
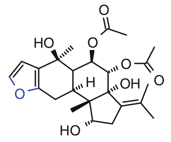
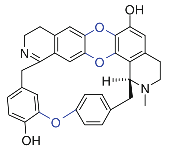


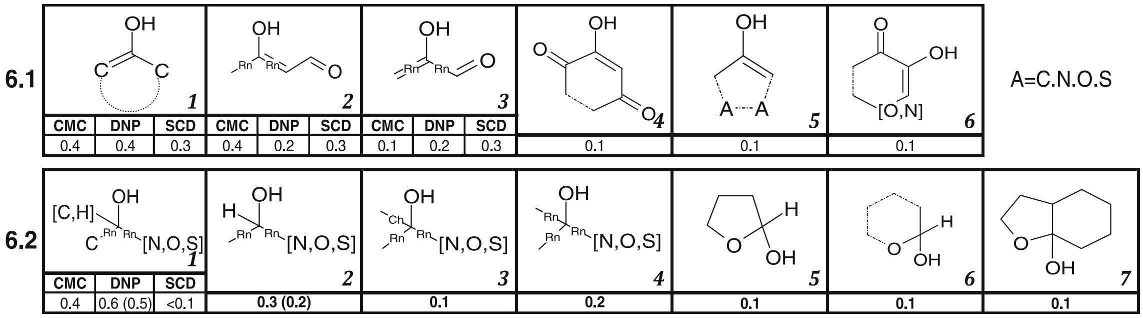
Scheme 5 Carboxylic acids

systems, steroids, polycyclic terpenes and alkaloids (ex. 3.1-2). Besides, at the junction of the parent subgraphs describing 5- and 6-member rings there are many secondary OH-groups in bicylic systems in the DNP database, statistically significant examples of them are shown in Scheme 3.1. In this cluster one can find various fused, spiro and bridged bicyclic fragments, most of them belong to terpenoids, for example Kaurenetriol (ex. 3.1-1). Terpenoid alkaloid Aconitine fragments (f. 47) may be also found among polycyclic systems.

The number of tertiary cyclic alcohols in the DNP database is considerably less than that of secondary ones (totally 2.3 %, Scheme 3.2 and 3.3). Nevertheless, as in the case of acyclic analogues, their number is still higher than that in the CMC and especially in the SCD database. Unsaturated derivatives are seen among both secondary

Table 2 Examples of natural products containing different types of C–O–C fragments

Scheme	Fragment	%	Example 1	Example 2	Example 3
7.2	$\mathrm{Cs}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Cs}$	0.5	 Speciosin A	 Sarcotride C	 2.3.4.5-Tetramethylgluconic acid
7.3	$\mathrm{Ca}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Cs}$	4.4	 Melicopicine	 Antofine	 (-)-Julandine
7.4	$\mathrm{Cr}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Cs}$	2.7	 Phomaligol A	 Salutaridine	 Denudatin A
7.5	$\mathrm{Ca}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Cr}$	0.7	 Ammonificin A	 Calocedimer B	 Eudeshonokiol A
7.6	$\mathrm{Cr}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Cr}$	3.2	 Lactucain A	 Pericosine E	 Angiopterlactone A
7.7	$\mathrm{Ca}^{\mathrm{Ch}}\mathrm{O}^{\mathrm{Ch}}\mathrm{Ca}$	0.2	 Thyroxine	 Noyaine	 Dimeresculetin
8.1	$\mathrm{Cr}^{\mathrm{Rn}}\mathrm{O}^{\mathrm{Rn}}\mathrm{Cr}$	10.6	 Deoxynivalenol	 (+)Pinoresinol	 Monensin A
8.2	$\mathrm{Cr}^{\mathrm{Rn}}\mathrm{O}^{\mathrm{Rn}}\mathrm{Ca}$	1.7	 Galanthamine	 Aflatoxin B1	 Beta-tocopherol
8.3	$\mathrm{Ca}^{\mathrm{Rn}}\mathrm{O}^{\mathrm{Rn}}\mathrm{Ca}$	0.8	 Leprocybin	 Caesalpinin	 Kurramine



Scheme 6 Other C–O–H structural fragments

and tertiary cyclic alcohols. A large part of tertiary alcohols having only two cyclic bonds (Scheme 3.2) contain methyl groups connected with saturated rings in the α- or β-position to hydroxy-groups. More detailed analysis reveals that the largest cluster consists of cyclohexane fragments. The remainder includes cyclopentane, cycloheptane, tetrahydrofuran and tetrahydropyran derivatives. Marcocyclic alcohols such as, for instance, the alkaloid Sceleratine (ex. 3.2-2) are less frequent.

A similar trend is true for tertiary alcohols containing three cyclic C–C bonds (Scheme 3.3). Among them the portion of β-methyl derivatives is surprisingly high (f. 2). In addition there is a rather large fraction of unsaturated and oxy-alcohols. Most of the Cr–O–H fragments are in saturated or partly saturated fused bicyclic systems such as decalin and octahydroindene, which in turn are parts of steroids and various terpenoid molecules. Generally, the distribution of tertiary alcohols in natural products is quite diverse, they can be found in alkaloids, natural quinines and other classes of NPs (ex. 3.3-1-3).

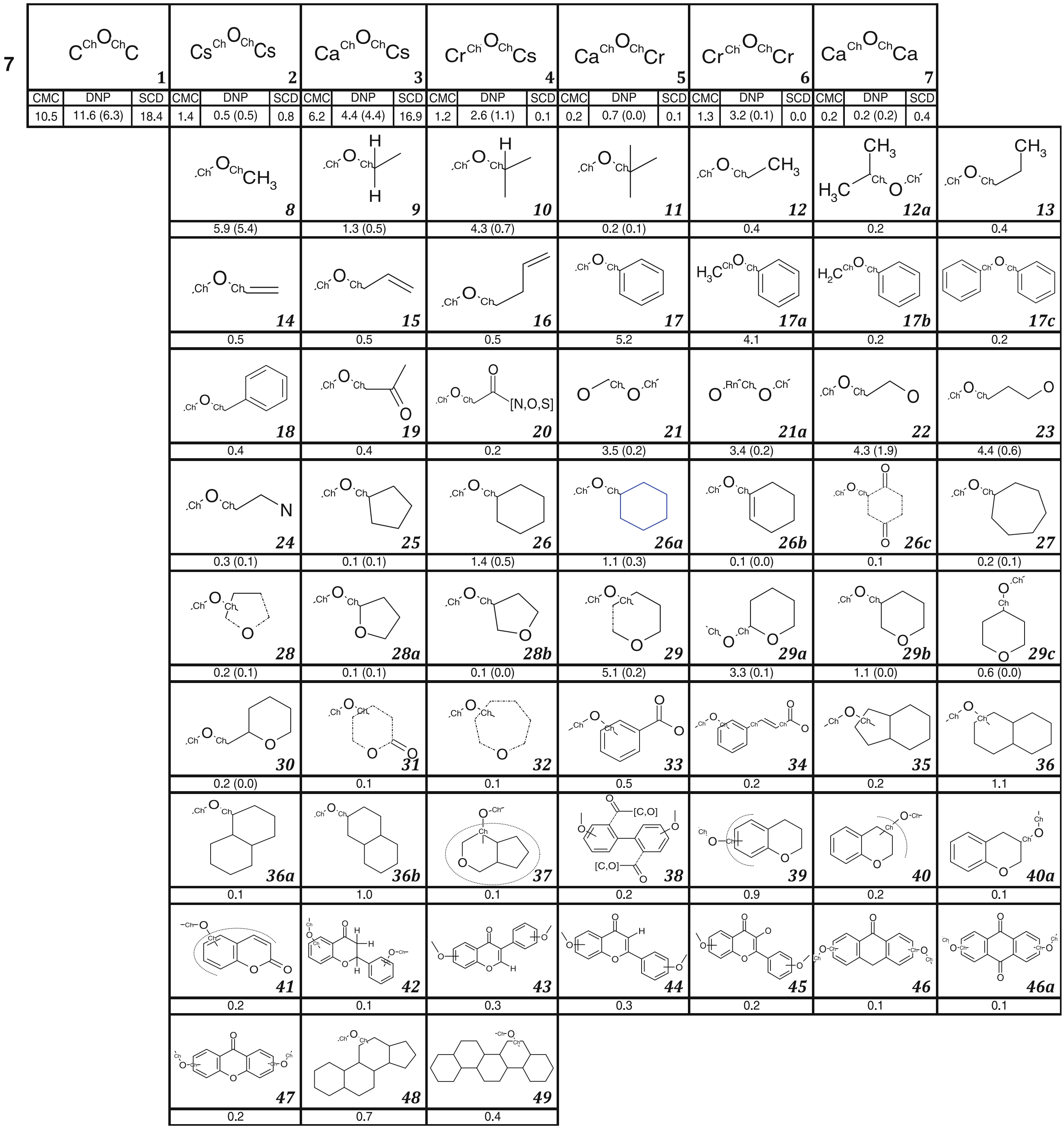
In the previous section we have already discussed the high percentage of α–methyl derivatives among secondary and tertiary acyclic carbinols. This statement seems to be more general: the percentage of all α–and β-methyl alcohols in the DNP database is quite high and significantly exceeds that in the two other model databases. Figure 4 illustrates the estimated number of α–and β-methyl alcohols for all three databases. Data for fragments containing α–and β-methyl groups connected with non-aromatic rings are given separately. Moreover it will be shown below that an increased percentage of α- and β-methyl groups in the nearest environment of oxygen atoms, including methyl groups linked to cycles, is a very characteristic feature of NPs (Fig. 5).

Phenols The next big class of compounds in the DNP database containing the structural fragment C–O–H is the class of aromatic compounds Ca-OH (Scheme 4.1, f. 1, 9.5 %), the great majority of which consists of phenols. Hydroxy-groups linked to other heterocyclic rings (OH-substituted pyridines, quinolines and benzopyrylium salts) account for not more than 0.1 % of total oxygen. The majority of phenols in the DNP database are mono-, di- and polyphenols, containing carbon substituents such as benzoic and Gallic acid derivatives. By contrast, phenol fragments in the CMC and especially in the SCD database contain mainly other types of substituents such as halogens, amino-, sulfo-, nitro-groups etc.

Among mono-substituted phenols fragments with para-allocation of OH groups and carbon substituents are predominant. Disubstituted phenols in most cases contain hydroxy-groups in para- or meta- positions, and one of them may be derivatized. Among these polyphenols a large part consists of fragments with 3,4,5- superposition of oxy-groups as in Gallic acid derivatives (f. 15) and with 2,4,6- superposition as in Phloroglucinic acid (ex. 4.1-1), whereas one or two oxygen atoms can be alkylated.

Derivatives (usually esters) of substituted benzoic acids contain a number of phenolic OH-groups in the DNP database. Of these several belong to derivatives of Gallic acid substituted in the aromatic ring, which fragment constitutes a natural tannin class (f. 23), also containing derivatives of hexahydrodiphenic and Ellagic acids. Prenylated phenols (f. 16) such as Grifolin isolated from Albatrellus confluens (ex. 4.1-3) constitute another large cluster of phenols. One more cluster includes fatty acids and polyketides derived by the acetate biosynthetic pathway (f. 19), for example, Cannabinolic acid (ex. 4.1-2). Another group of natural phenols is Cinnamic acid derivatives (f. 17) occurring in nature mainly as esters. Two-thirds of them belong to Caffeic acid derivatives and one-third to p-Coumaric acid derivatives. An insignificant amount (~0.1 %) of phenol moieties can be found among amino acid derivatives, coumarins, indole and isoquinoline alkaloids and stilbenes.

A class of natural products containing the largest portion of phenolic groups, which accounts for approximately 2 % of total oxygen is flavonoids (f. 26-33). Distribution of Ca–OH groups in various types of flavonoids, namely flavones, isoflavones, flavans, isoflavans and anthacyanines is demonstrated in the Scheme. A considerable number of



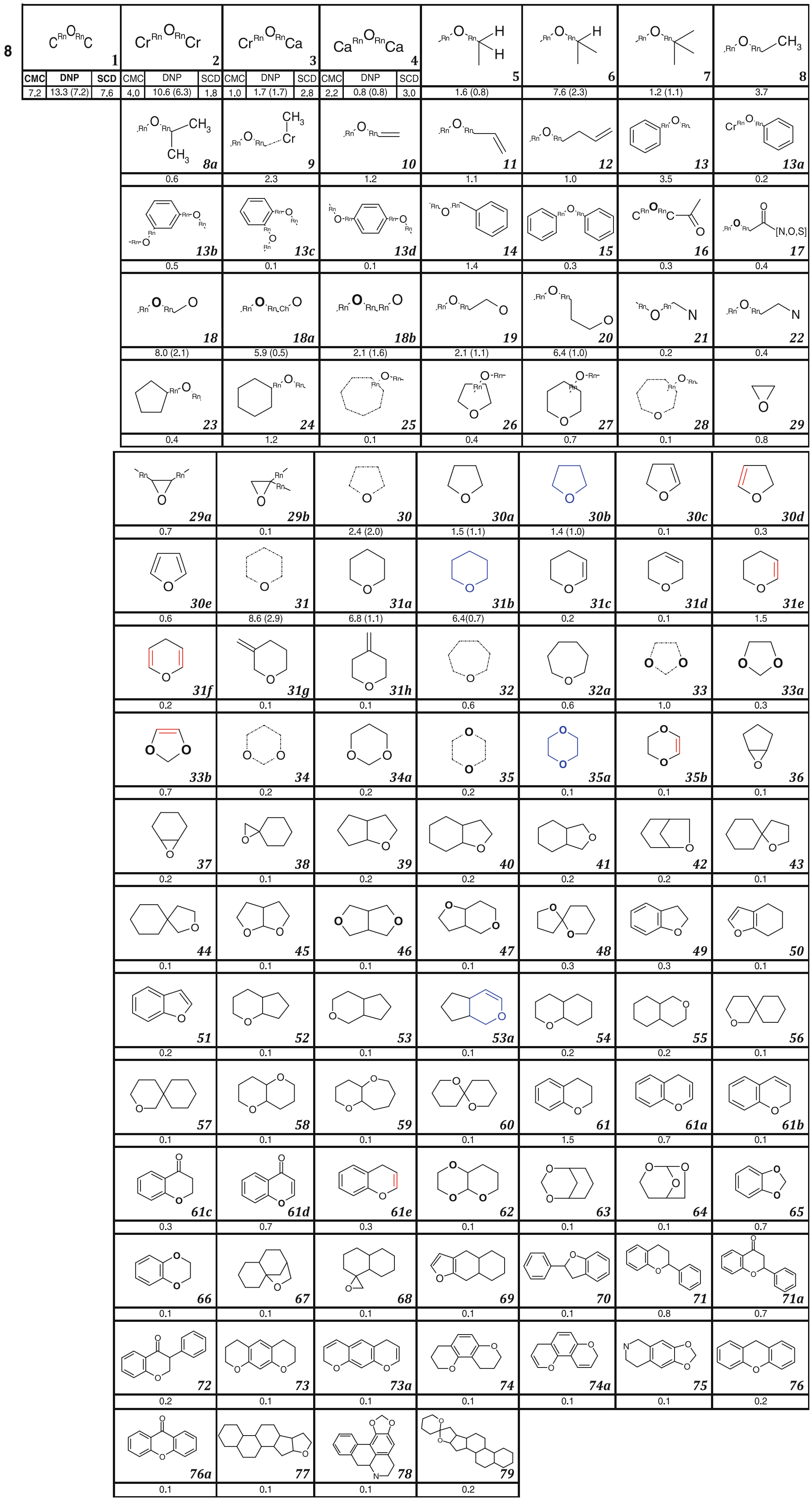
Scheme 7 Acyclic ether C–O–C structural fragments

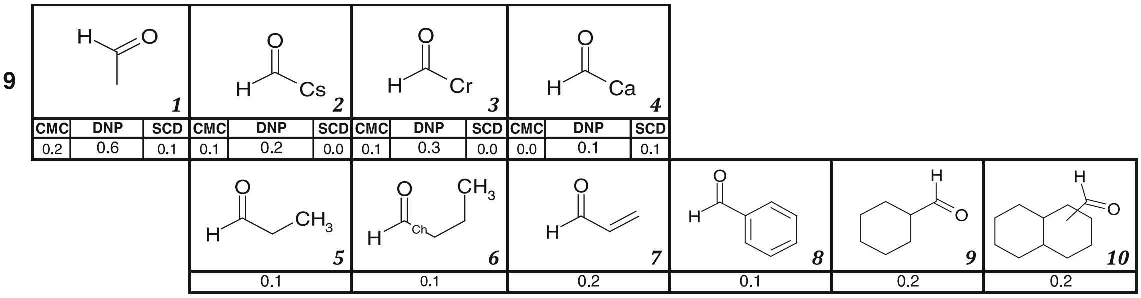
phenolic OH-groups are present in xanthonoids, dianthrones and anthraquinones.

Carboxylic acids The portion of free carboxylic acids[3] (Scheme 5, f. 1) in the DNP is not very significant (2.2 %) compared to the CMC database but it is still higher than that in the SCD database. The spectrum of carboxylic acids available in the DNP database is highly diversified. A fraction of acids where the carboxylic group is linked to acyclic carbon atoms (f. 2), in particular, to unsubstituted methylene groups, is the most essential. This structural type is a parent for fatty aliphatic acids, in some cases partly unsaturated with the number of carbon atoms more than 5. The majority of such acids contain linear carbon

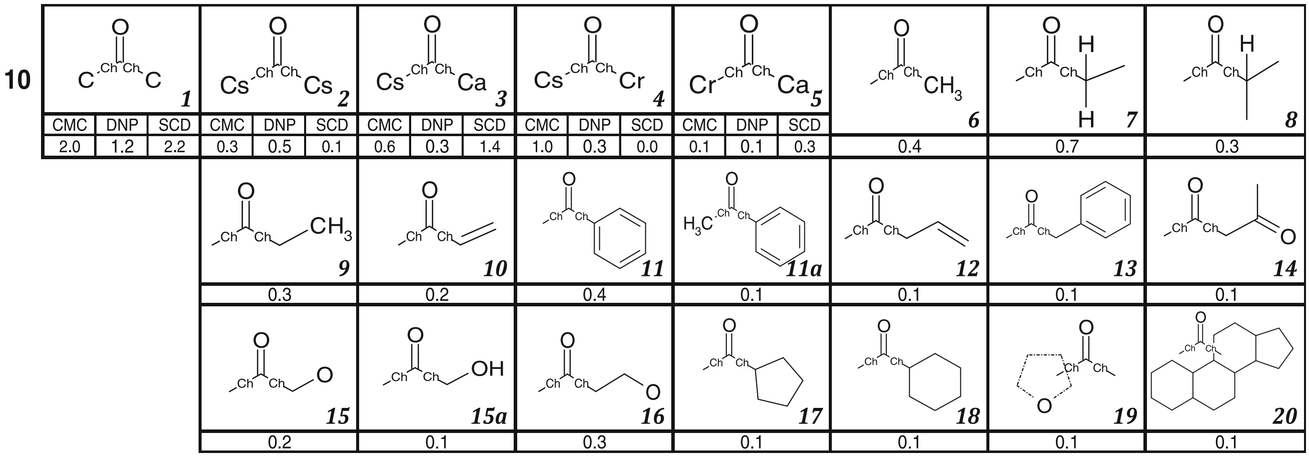
[3] Since carboxylic acid, ester, and lactone groups contain two oxygen atoms, without logical contradiction we reduce the predetermined statistical threshold to 0.05%.

Scheme 8 Cyclic ether C–O–C structural fragments




9
1 CMC 0.2 DNP 0.6 SCD 0.1
2 CMC 0.1 DNP 0.2 SCD 0.0
3 CMC 0.1 DNP 0.3 SCD 0.0
4 CMC 0.0 DNP 0.1 SCD 0.1
5 0.1
6 0.1
7 0.2
8 0.1
9 0.2
10 0.2

Scheme 9 Aldehyde structural fragments


10
1 CMC 2.0 DNP 1.2 SCD 2.2
2 CMC 0.3 DNP 0.5 SCD 0.1
3 CMC 0.6 DNP 0.3 SCD 1.4
4 CMC 1.0 DNP 0.3 SCD 0.0
5 CMC 0.1 DNP 0.1 SCD 0.3
6 0.4
7 0.7
8 0.3
9 0.3
10 0.2
11 0.4
11a 0.1
12 0.1
13 0.1
14 0.1
15 0.2
15a 0.1
16 0.3
17 0.1
18 0.1
19 0.1
20 0.1

Scheme 10 Acyclic ketone structural fragments

chains. Sometimes such a structural fragment can be found in terpenes and steroids, where the COOH group is connected with cyclohexane moiety not directly, but through a C_1–C_3 alkyl tether.

Fragments where the carboxylic group is connected with methine carbon atoms provide an origin to a small cluster of branched acids, the great majority of which belong to α-methyl derivatives, and to important classes of free α-oxy- and α-amino-acids, though not very abundant. It is interesting to note the rather high level of α-methyl- and particularly β-methyl acids of different types.

A large cluster is composed by subgraph f. 3 derivatives, where the carboxylic group is connected directly with non-aromatic cycles. This structural class then evolves into derivatives of cyclohexanecarbonic, cyclopentanecarbonic, and tetrahydropyrancarbonic acids (f. 13–15). Cyclohexanecarbonic structural fragments are parent structures for decaline derivatives and various terpenes with tri- and penta-cyclic carbon skeletons being the most frequent. The proportion of carboxylic groups linked to steroid moieties is surprisingly low and does not exceed the threshold value (~0.04 %). Interestingly, a cluster of acids with carboxylic groups linked to quaternary cyclic carbon atoms (f. 3b, 3c) is quite substantial, α-methyl cyclohexane and decaline fragments belong to this cluster. A fragment of 2-tetrahydropyrancarboxylic acid (f. 15) almost entirely evolves into uronic acids in carbohydrates. There are some α, β- unsaturated acids including acrylic, cinnamic, maleic, and fumaric acids, in the DNP database, Clusters containing β, γ- unsaturated and aromatic acids, mainly benzoic, are rather small.

The remaining part of the carboxylic group found in the DNP database is much diversified among the classes of organic compounds. One can observe here derivatives of succinic, pipecolic acids, proline, as well as phenylacetic acids, penicillines and cephalosporins (ex. 5-3), acids containing porphirinic ring, alkaloid derivatives and so on.

Other classes C–OH fragments in the DNP database can be observed in other less frequently occurring classes of compounds (Scheme 6).

One of these classes is enols (Scheme 6.1), the typical example of which is ascorbic acid (Table 1, ex. 6.1-1). Double bonds in natural enols are predominantly cyclic and usually stabilized by additional carbonyl groups. In other words, natural enols are usually cyclic α, β- unsaturated compounds that can be described by one of the following subgraphs f. 2 and f. 3 in approximately equal proportions.

Hydroxy-derivatives of benzoquinones and naphtoquinones suit both of these subgraphs. Natural fungal pigment Ascocorynin (ex. 6.1–3) belongs to this class of compounds. The DNP database also contains flavonols (f. 6) and 5-member heterocyclic moieties that are fragments of tetronic and tetramic acids (f. 5); an example molecule of this cluster is the mushroom pigment Xerocomic acid (ex. 6.1–2). One more enolic system in the DNP database contains the 2H-pyran moiety (<0.1 %) comprising triacetic acid lactones and hydroxy-coumarins. Steroids, meroterpenoids, tetracyclines and alkaloids containing hydroxy-cycloheptatrienone occasionally occur in DNP database.

It is worth mentioning another class of NPs containing the C–O–H moiety, namely hemiacetal and hemiketal derivatives (Scheme 6.2, f. 1) that can be divided into several subgraphs, however their occurrence in the DNP is not significant. The majority of cyclic hemiacetals (f. 2) consists of carbohydrate derivatives. Detailed analysis of carbohydrate molecules shows that the portion of free glycoside OH-groups is less than 0.3 %. The aglyconic part in this structural type is highly diversified. One can find here molecules containing reduced moieties of γ-butyrolactone or corresponding butenolide, maleinimide, partly reduced moieties of tetronic acids, phtalic anhydrides, amino acetals, indol alkaloids, terpenoids and other .

Hemiketals containing one linear C–C bond are even less frequent (f. 3), but are also very diversified. Cyclic polyethers where the C–C bond is a bond with a methyl group or a bond tethering two cyclic ethers are quite common in this cluster.

Finally, hemiketal systems with three cyclic bonds (f. 4) contain ketal OH-groups mainly in the bridge carbon atom that joins two saturated rings. Usually these system are various combinations of tetrahydropyran, tetrahydrofuran and γ-butyrolactone moieties. Examples of such hemiketals are the alkaloid Tazettine, the antibiotic Spectinomycin, and Delesserine an ascorbic acid derivative isolated from red alga (ex. 6.2-1-3).

In conclusion of section "Introduction", the distinctive features of the DNP database in terms of hydroxyl groups are essentially a larger proportion of cyclic structural fragments such as cyclic alcohols and phenols and a considerably higher degree of branching of carbon skeletons, supported by a larger portion of tertiary alcohols in comparison with the CMC and especially, with the SCD databases (Fig. 3).

C–O–C fragments

A widespread structural class of oxygen-containing compounds is ethers (Table 2), their percentage in the DNP database accounts for 24.9 %. In this section we will also consider C–O–C fragments in acetals and ketals, which we consider can be analyzed together with relative ether subgraphs. C–O–C fragments localized in esters groups will be considered in section "Results and discussion" below.

Ethers can be divided into acyclic and cyclic ones that are present in the DNP database in approximately equal amounts (Schemes 7 and 8, 9, 10, and 11.6 and 13.3 % respectively).

Acyclic ethers The presence of two carbon substituents leads to more complexity of the chemical environment, so acyclic ethers will be considered independently, from two different points of view (Scheme 7). Firstly, they will be analyzed in terms of the nearest chemical environment of the oxygen atom depending on the neighboring α-carbon atom types (f. 2–7).

The DNP database contains a quite small cluster of ethers where the oxygen atoms are connected with two acyclic

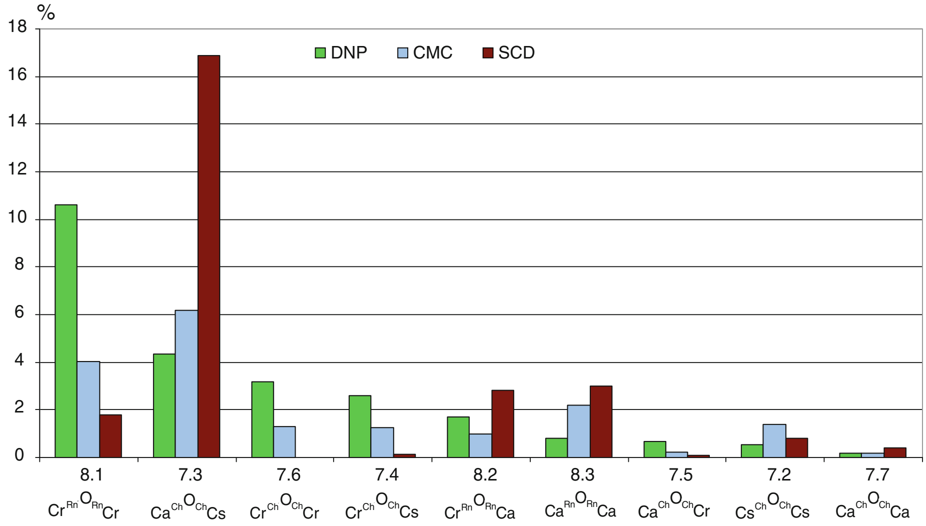
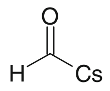
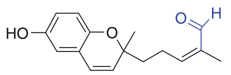
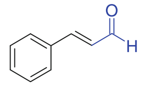
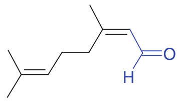
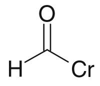
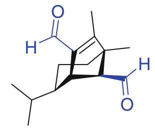
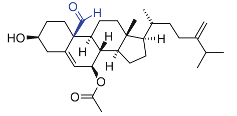
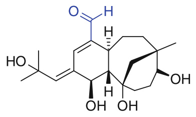
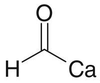
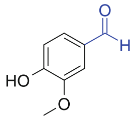
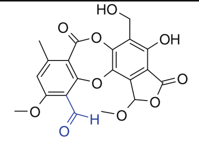
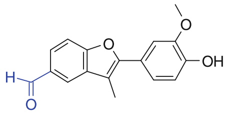
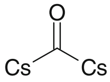
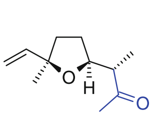
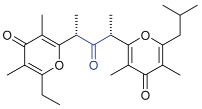
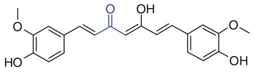
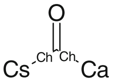
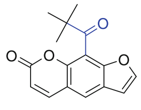
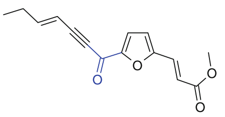
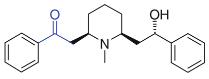
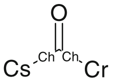
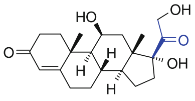
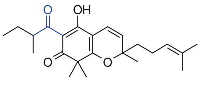
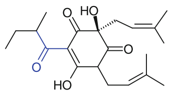
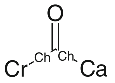
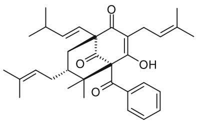
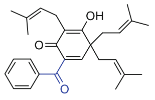
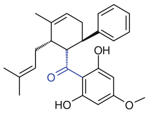
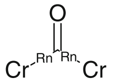
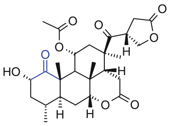
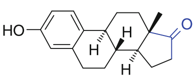
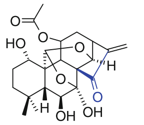
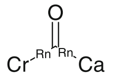
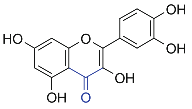
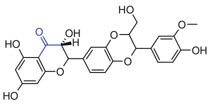
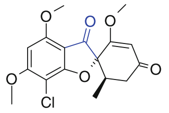
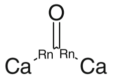
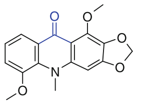
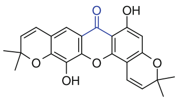
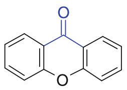
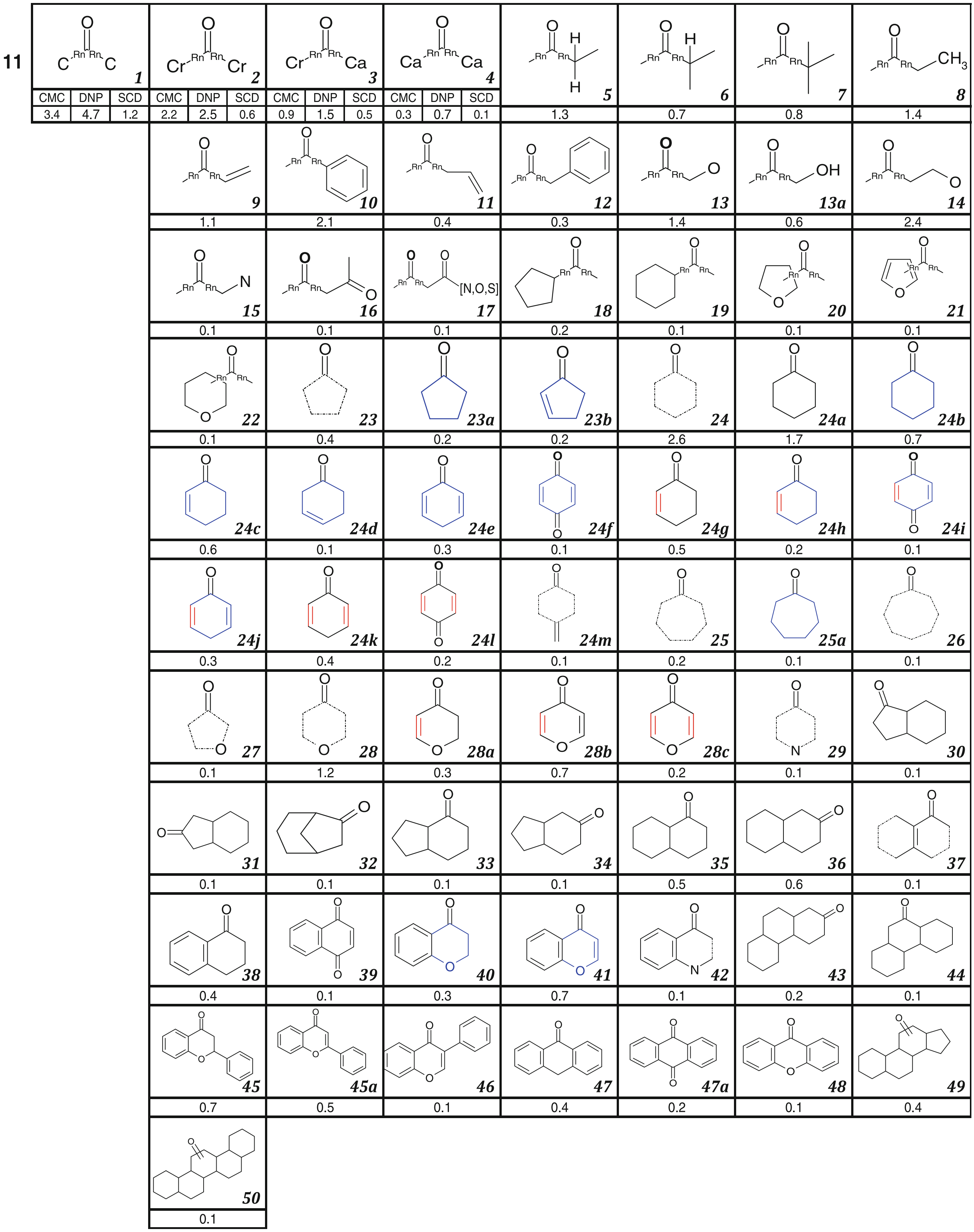


Fig. 6 Abundance of ether C–O–C fragments in the DNP, CMC, and SCD databases

Table 3 Examples of natural products containing different types of aldehyde and ketone fragments

Scheme	Fragment	%	Example 1	Example 2	Example 3
9.2		0.2	 Elaeagin	 Cinnamaldehyde	 Citral
9.3		0.3	 Helminthosporal	 Ergosta-5.24(28)-diene-3.7.19-triol	 Floridicin B
9.4		0.1	 Vanillin	 Salazinic acid	 Eupomatenoid 10
10.2		0.5	 Nordavanone	 Maurapyrone C	 Curcumin
10.3		0.3	 Swietenone	 Wyeron	 Lobeline
10.4		0.3	 Hydrocortisone	 Petiolin B	 Adhumulone
10.5		0.1	 Nemorosone	 Grandone	 Panduratin A
11.2		2.5	 Simarolide	 Estrone	 Ponicidin
11.3		1.5	 Quercetin	 Isosilybin	 Griseofulvin
11.4		0.8	 Tecleanthine	 Rheediaxanthone A	 Xanthone



Scheme 11 Cyclic ketone structural fragments

carbon atoms (f. 2); almost all oxygen of this cluster belongs to aglycones. Such ethers can include long aliphatic substituents as found in Sarcotide—a cyclitol derivative isolated from a marine sponge (Table 2, ex. 7.2-2).

A rather large proportion of ether moieties occur in alkyl aryl ethers (f. 3). The majority of these ethers belong to methoxy-groups connected with benzene rings. Most of alkoxy-derivatives are benzoic and cinnamic acid derivatives. They are also common in coumarins, flavonoids, tannins, anthrone, anthraquinone derivatives, xanthanoids (f. 38, 46–47), and to a lesser extent alkaloids (ex. 7.3-1, 2, 3).

Structural moieties described by subgraphs f. 4–6 belong to ethers with one α-carbon atom being a part of non-aromatic rings, and a second one being acyclic, aromatic or also a part of non-aromatic rings. In case of ethers encompassed by subgraph f. 6 the main portion of oxygen atoms is concentrated in carbohydrates, particularly in glycoside linkers in di- and polysaccharides. On the contrary, subgraph f. 4 comprises a lot of aglycone moieties. A cluster of acyclic diaryl ethers is the least distributed in the DNP database.

Comparing subgraphs f. 2–7 one can note distinctive features of the DNP database compared with the two model ones: a small number of alkyl aryl ethers and a high content of ethers with at least one α-carbon in non-aromatic ring. Some examples of acyclic ethers are given in Table 2.

Independent classification of ether fragments concentrates on one carbon atom in C–O–C fragments (f. 8–23) but moves further along their chemical environments.

As follows from Scheme 7 half of all ethers contain a methyl group at least in one α-position to the oxygen atoms (f. 8). α-Methylene containing ethers occur considerably less frequently. However, the number of moieties containing oxygen neighboring to secondary carbon atoms increases again and most of them are carbohydrates. This can be explained by the high percentage of ethers connected with saturated cycles that usually contain methine moieties in the DNP database.

α-Oxyethers as well as β- and γ-oxy-derivatives (f. 21–23) constitute big clusters; almost all of them are carbohydrates. A small cluster of ethers of β-aminoalcohols can be also observed in the DNP database.

Aromatic ethers compose a substantial segment of the DNP database; most of them are phenyl derivatives, to large extent methoxy phenyl ethers. There are also measurable amounts of vinylic, allylic, homoallylic and benzylic fragments, as well as α-keto- and α-carboxy-derivatives (f. 14–20).

Subsequent development of fragments f. 2–24 is demonstrated in Scheme 7. A distinct branch consists of carbohydrate ethers; among them the main part includes tetrahydropyran derivatives (f. 29) and tetrahydrofuran derivatives are present in considerably fewer numbers (f. 28). The tetrahydropyran cluster also contains bicyclic fused systems such as for example benzopyran (f. 40, Table 2, ex. 7.5-1) and structural subtype f. 39 including coumarins and flavonoids (f. 41–45). Other examples of tetrahydropyran containing systems are iridoids (f. 37) and ethers containing the hydroxymethylene moiety (f. 30).

The aglyconic segment among acyclic ethers containing aromatic (f. 16) and non-aromatic (f. 25–27) carbocycles is quite appreciable in the DNP database. Among them cyclohexanol derivatives evidently dominate over cyclopentanol and cycloheptanol structural fragments (f. 25–27), these moieties mainly belong to fused or spiro-systems, including steroids, triterpenes, occasionally other terpenoids and iridoids.

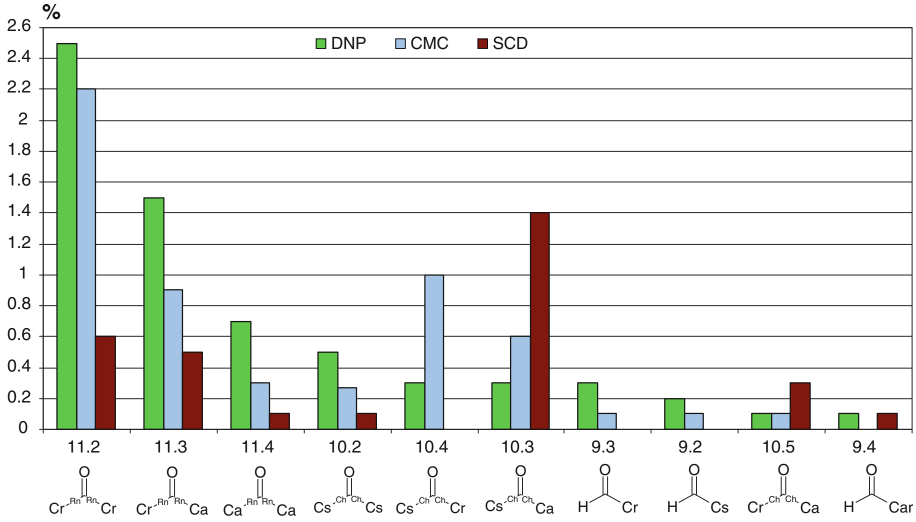
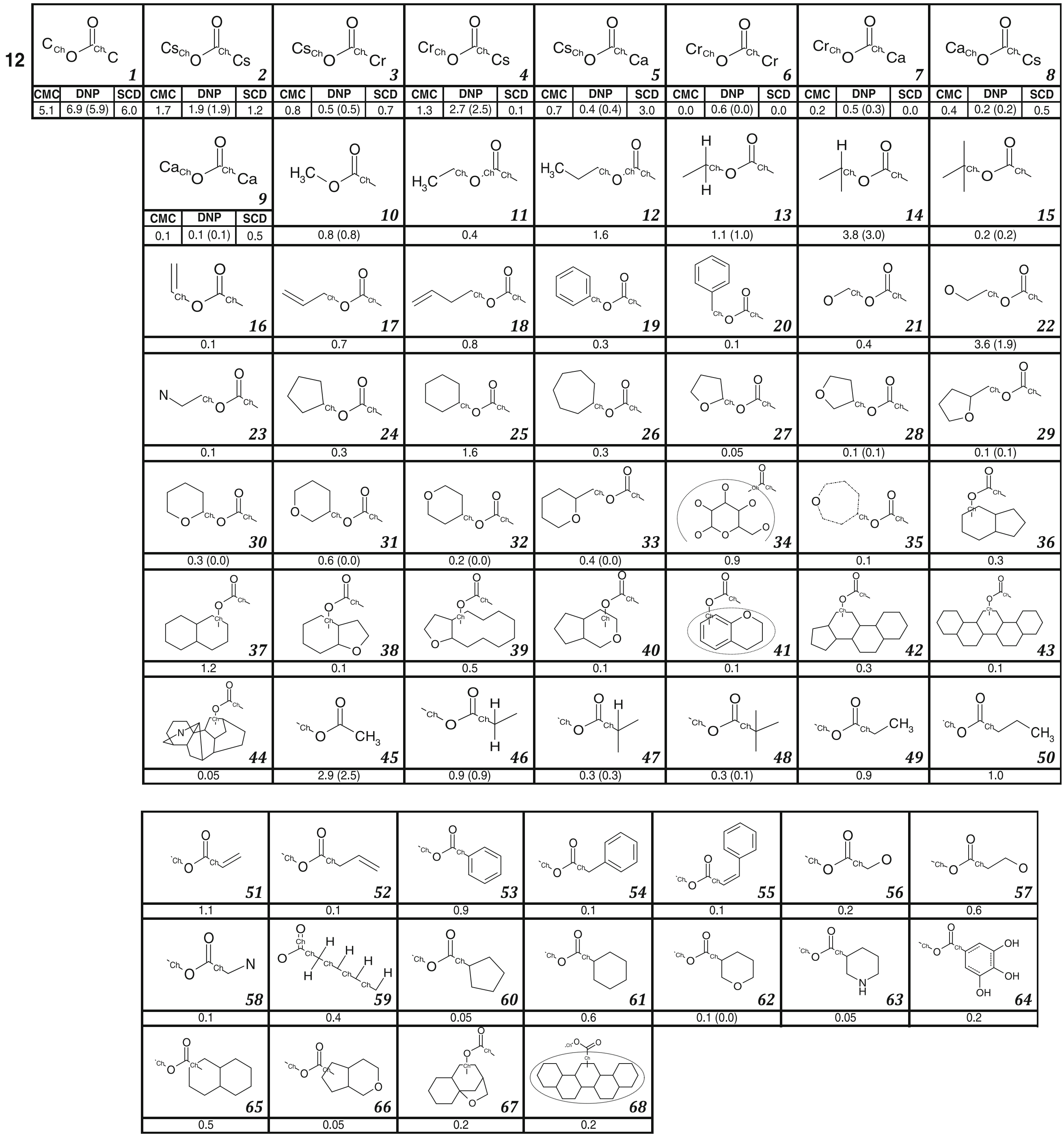


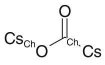
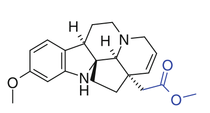
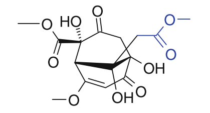
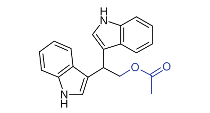
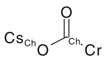
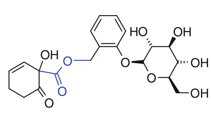
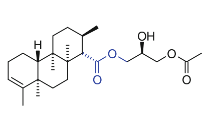
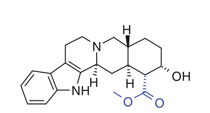
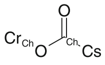
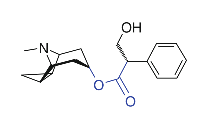
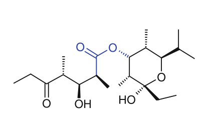
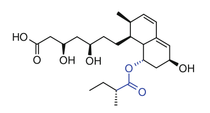
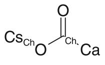
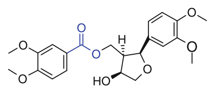
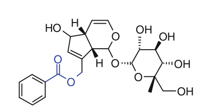
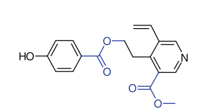
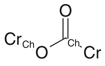
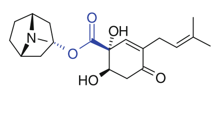
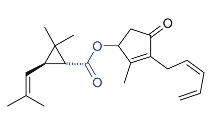
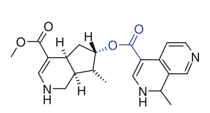
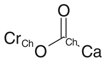
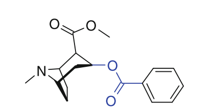
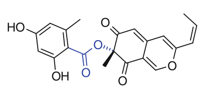
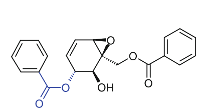
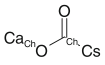
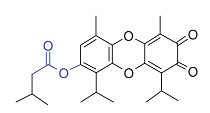
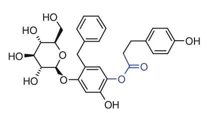
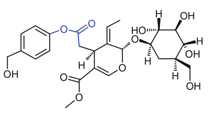
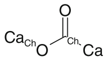
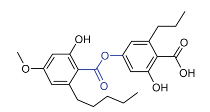
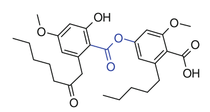
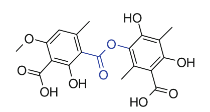
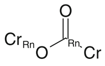
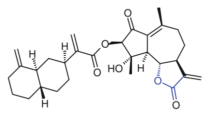
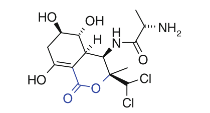
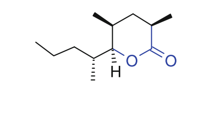
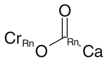
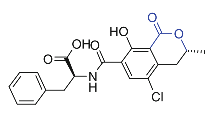
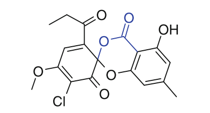
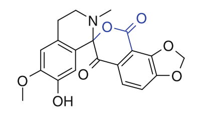
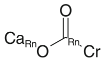
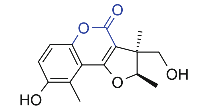
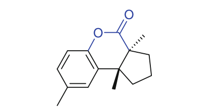
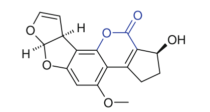
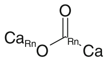
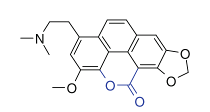
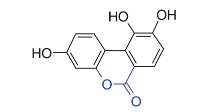
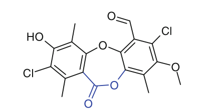
Fig. 7 Abundance of aldehyde and ketone fragments in the DNP, CMC, and SCD databases

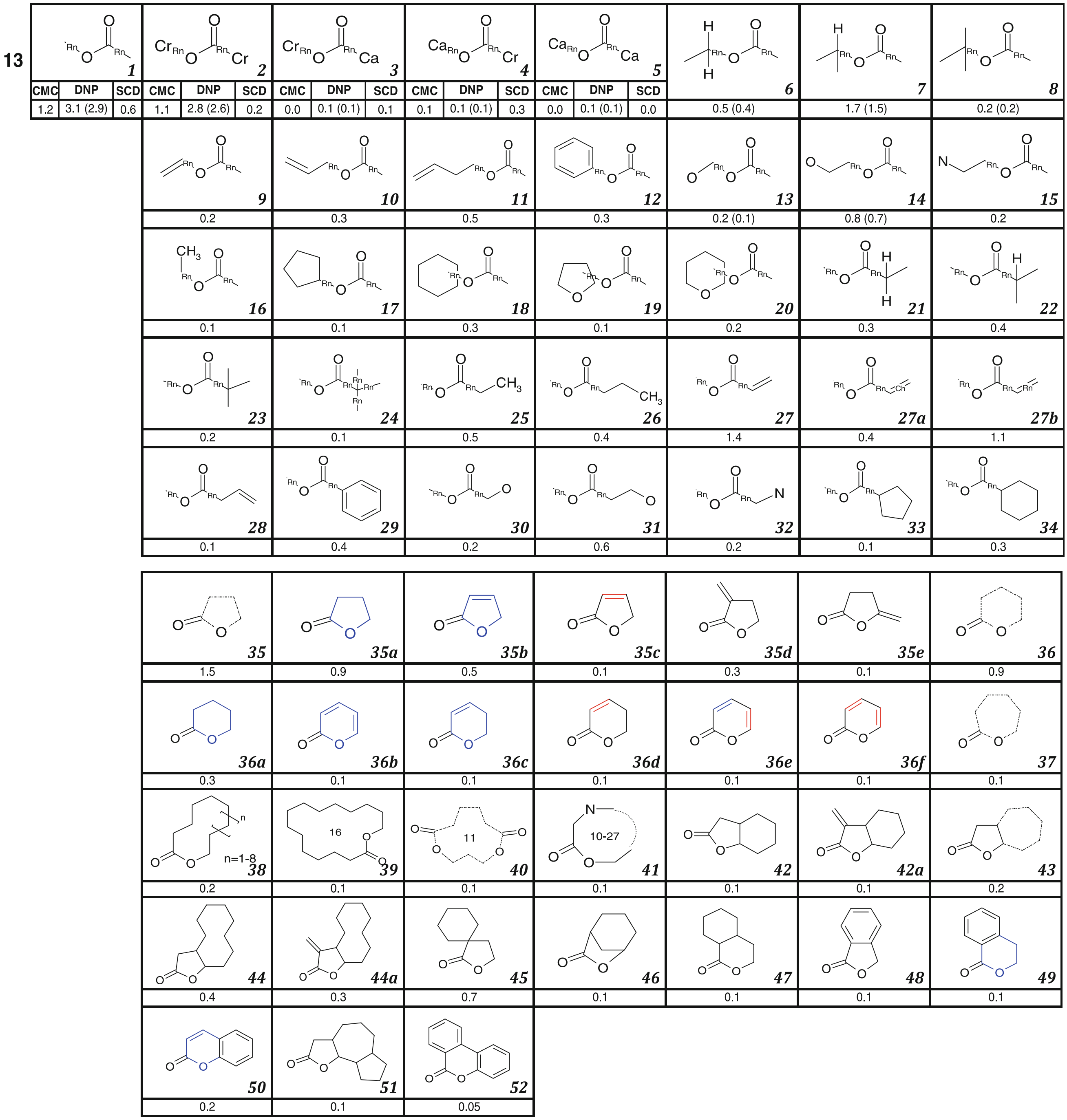


Scheme 12 Ester structural fragments

Cyclic ethers The number of structural fragments containing cyclic ether groups in the DNP database is almost twofold higher than those in the CMC and SCD databases (f. 1, Scheme 8). In accordance with the types of the nearest carbon atoms they can be divided into three main clusters (f. 2–4). The largest group includes the ether fragments containing oxygen atoms connected with two non-aromatic carbon atoms (f. 2). In the DNP quite a large part of them belong to carbohydrates existing in furanose and pyranose forms, and more than half belong to aglycones. Diversification of aglycone forms within this cluster is very high: mono-, bi-, polycyclic fused and spiro-cyclic system containing oxygen, including cyclic acetals and so on (f. 23–79). The alkyl aromatic cluster (f. 3) including only aglycones, is much lower in abundance. This structural type is characteristic for partly hydrogenated benzocyclic

Table 4 Examples of natural products containing different types of ester and lactone moieties

Scheme	Fragment	%	Example 1	Example 2	Example 3
12.2		1.9	 Schizophylline	 Sargassumketone	 Streptindole
12.3		0.5	 Salicortin	 Verrucosin 2	 Yohimbine
12.4		2.7	 Scopolamine	 Dolabriferol	 Pravastatin
12.5		0.4	 Magnostellin B	 Melampyroside	 Fontaphilline
12.6		0.6	 Consiculine	 Pyrethrin I	 Scaevodimerine A
12.7		0.5	 Cocaine	 Mitorubrin	 Pipoxide
12.8		0.2	 Ecklonoquinone A	 Punarnavoside	 Excelsioside
12.9		0.1	 Imbricaric acid	 Confluentic acid	 Hypothamnolic acid
13.2		2.8	 Podachaenin	 Bactobolin A	 (-)-Invictolide
13.3		0.1	 Ochratoxin A	 Maldoxin	 Procumbine
13.4		0.1	 Glaupadiol	 (–)-Herbertenolide	 Aflatoxicol
13.5		0.1	 Thaliglucinone	 Fasciculiferol	 Allorhizin



Scheme 13 Lactone structural fragments

derivatives. Several accessory subgraphs (f. 13a–d) give some ideas about further transformations of this cluster; they lead to various hydrogenated furo- and pyranochromens (for example, f. 73, 74) and macrocyclic ethers including those containing aromatic rings in this macrocycle.

The least abundant cluster contains an O atom between two aromatic carbon atoms (f. 4). This cluster mainly involves furan derivatives (f. 30e) including its fused analogues (f. 50, 51), and a cyclic diphenyl moiety (f. 15) being a part of xanthanoids (f. 76).

Now let us consider the more distant environment of one of the carbon atoms in the C–O–C moiety (f. 5–21). Comparing subgraphs (f. 5–7) one can conclude that fragments containing oxygen atoms linked to secondary carbon atoms are predominant among cyclic ethers compared with oxygen atoms linked with primary and tertiary

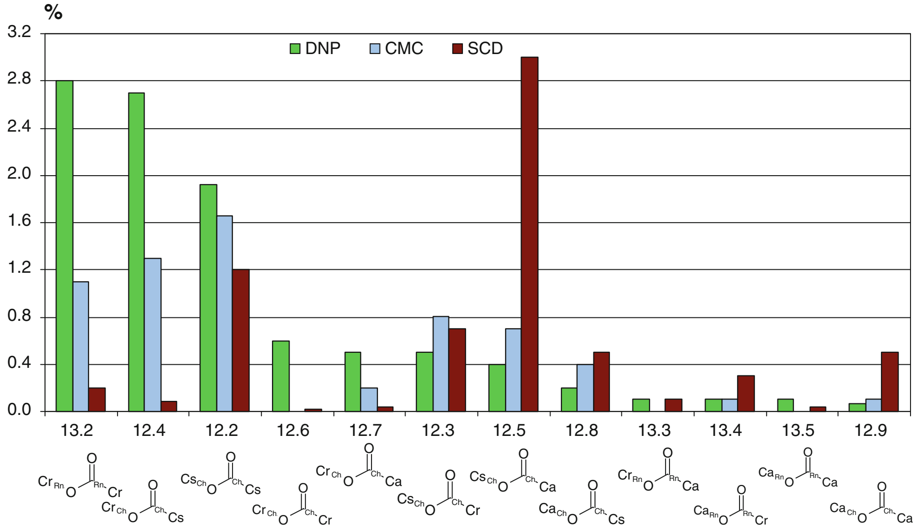
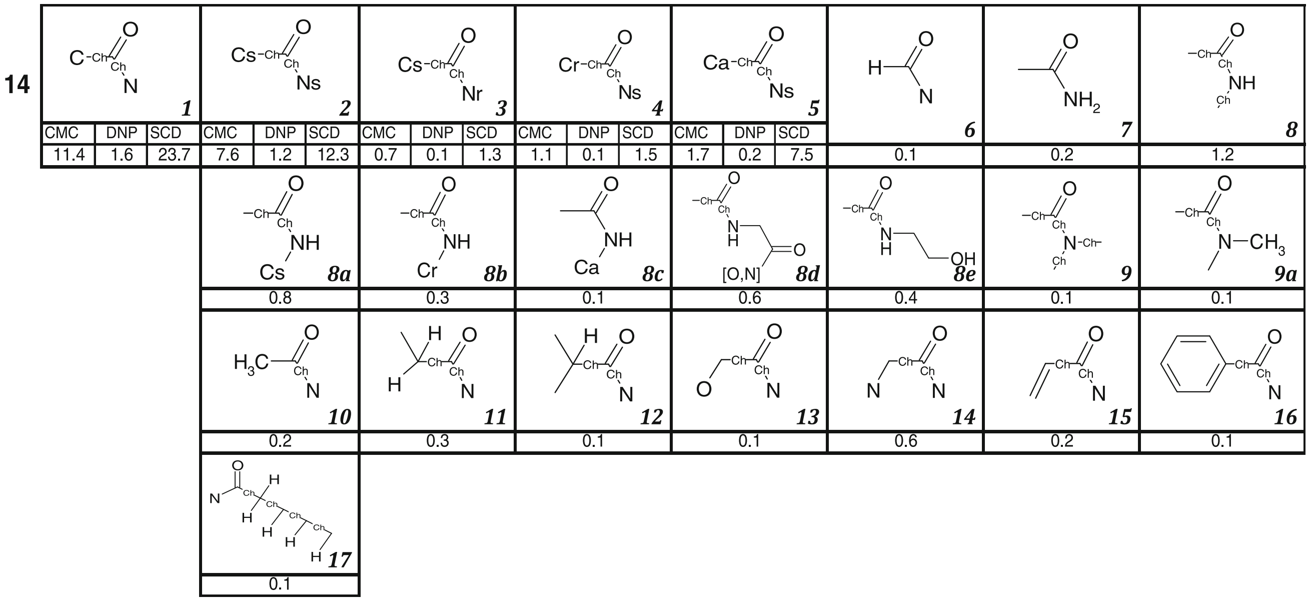


Fig. 8 Abundance of ester and lactone fragments in the DNP, CMC, and SCD databases



Scheme 14 Amide structural fragments

ones. α-Methyl- and α,α-dimethyl-substituted cyclic ethers as well as β-methyl derivatives are widespread among natural products (see Fig. 4).

α-Oxyethers (f. 18) constitute a large cluster within the considered ether fragments, most of which belong to carbohydrates. One can see that while almost three quarters of these acetals contain an acyclic bond C–O being inter alia a link in di- and polysaccharides, an entirely cyclic moiety C–O–C–O occurs considerably less frequently though includes more aglycones. Clusters of β- and γ-oxyethers (f. 19 and 20) are observed and consist mainly of carbohydrates.

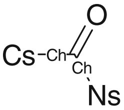
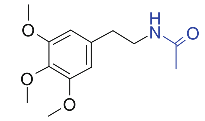
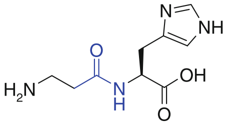
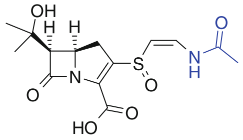
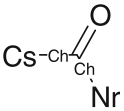
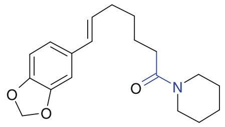
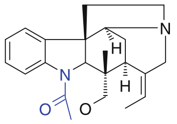
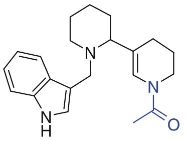
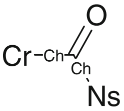
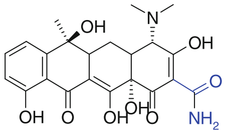
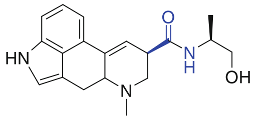
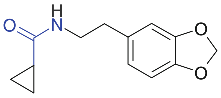
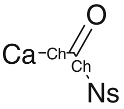
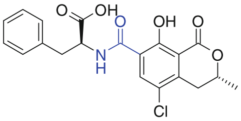
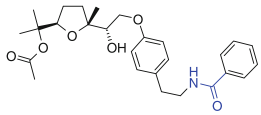
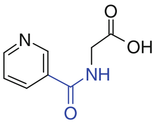
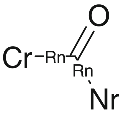
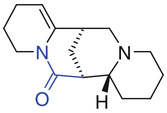
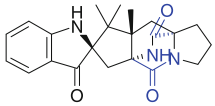
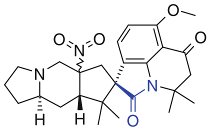
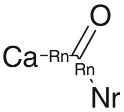
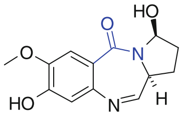
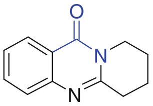
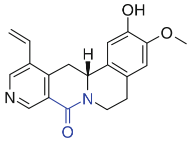
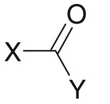
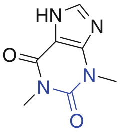
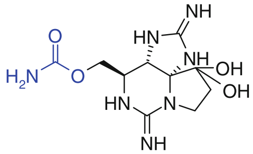
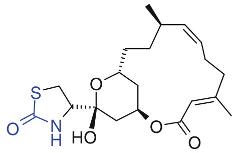
Allylic, homoallylic and benzylic cyclic ethers occur in the DNP database in approximately equal proportions. Cyclic aminals, β-aminoethers as well as α-keto- and α-carboxy derivatives can be found in lesser amounts.

Analysis of the next level of complexity in structural moieties results in ether derivatives adjoining to non-aromatic rings (f. 23–28), in this way fused bridge and spiro systems can be formed.

According to ring size several big monocyclic clusters of ethers can be selected (f. 29–35). The DNP database contains a surprisingly high percentage of epoxides (f. 29); almost all of them belong to two subtypes: fused epoxides based on non-aromatic, mainly cyclopentane and cyclohexane rings and spiro-epoxides (ex. 8.1-1).

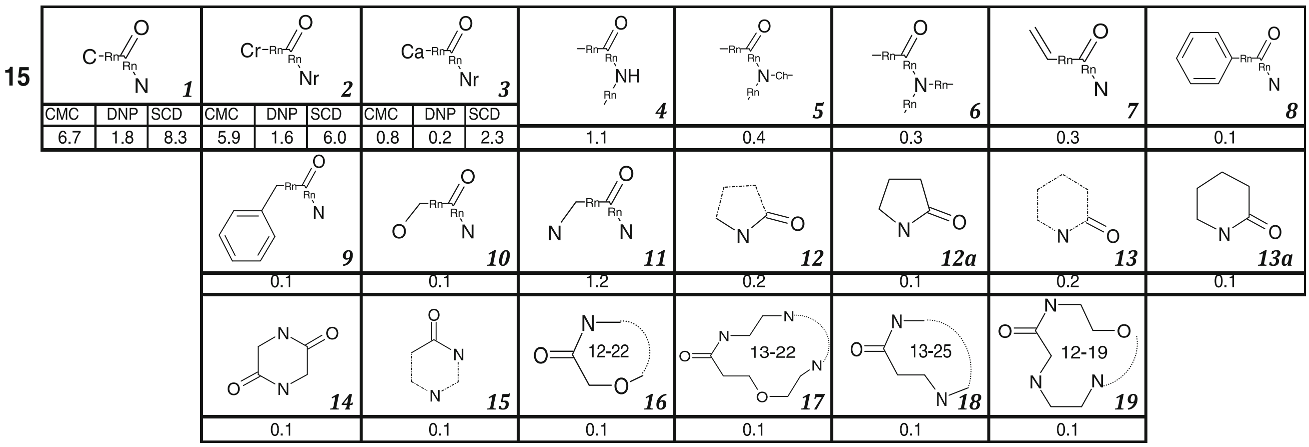
As should be expected the largest monocyclic ether clusters are derivatives of tetrahydrofuran and tetrahydropyran

Table 5 Examples of natural products containing different types of amide and lactam fragments

Scheme	Fragment	%	Example 1	Example 2	Example 3
14.2		1.2	 Mescaline	 Canosine	 Capetimycin A
14.3		0.1	 Piperolein A	 Retuline	 Gramodendrine
14.4		0.1	 Tetracycline	 Ergometrine	 Rubesamide
14.5		0.2	 Ochratoxin A	 Tuberine	 Nicotinuric acid
15.2		1.6	 Aphyllidine	 Brevianamide A	 Cyclopiamine A
15.3		0.2	 Neothramycin A	 Mackinazolinone	 Alangimaridine
16.1	 X,Y = N,O,S	0.2	 Theophylline	 Saxitoxin	 Latrunculin B

(f. 30, 31). Though they are often structural units of carbohydrates, the proportions of aglycones among these clusters are still significant. General structural types f. 30 and f. 31 give various bicyclic systems; some of them that are statistically significant are shown in the Scheme (f. 39–65). Although the majority of carbohydrates are encompassed by furanose and pyranose forms containing only single C–C bonds within the cycle, it seems more interesting to consider the further development of the aglycone portions of these moieties.

Aglycones of the tetrahydrofuran parent graph (f. 30) include non-aromatic fused, bridge and spiro bicycles (f. 39–46), derivatives of dihydrofyran and furan (f. 49–51), as well as bicycle (f. 47) and spiro-ketal (f. 48).



Scheme 15 Lactam structural fragments

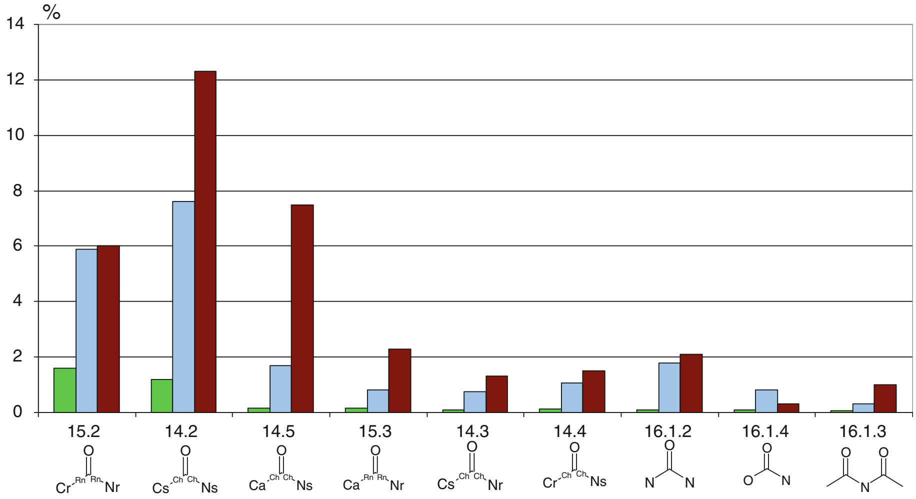


Fig. 9 Abundance of amide, lactam and some related fragments in the DNP, CMC, and SCD databases

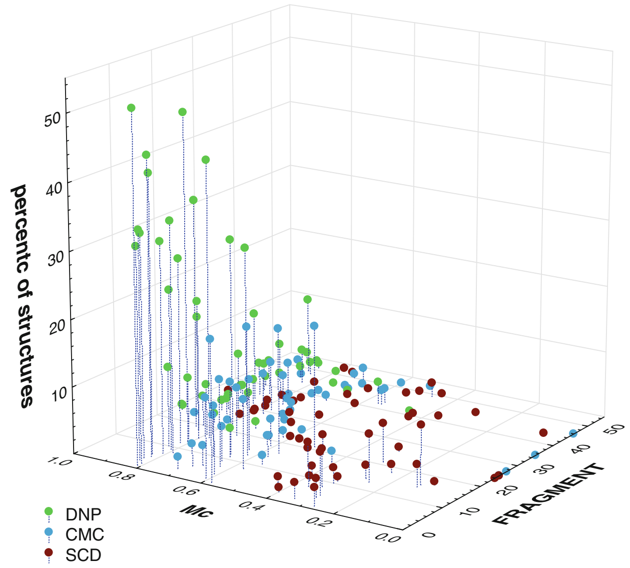
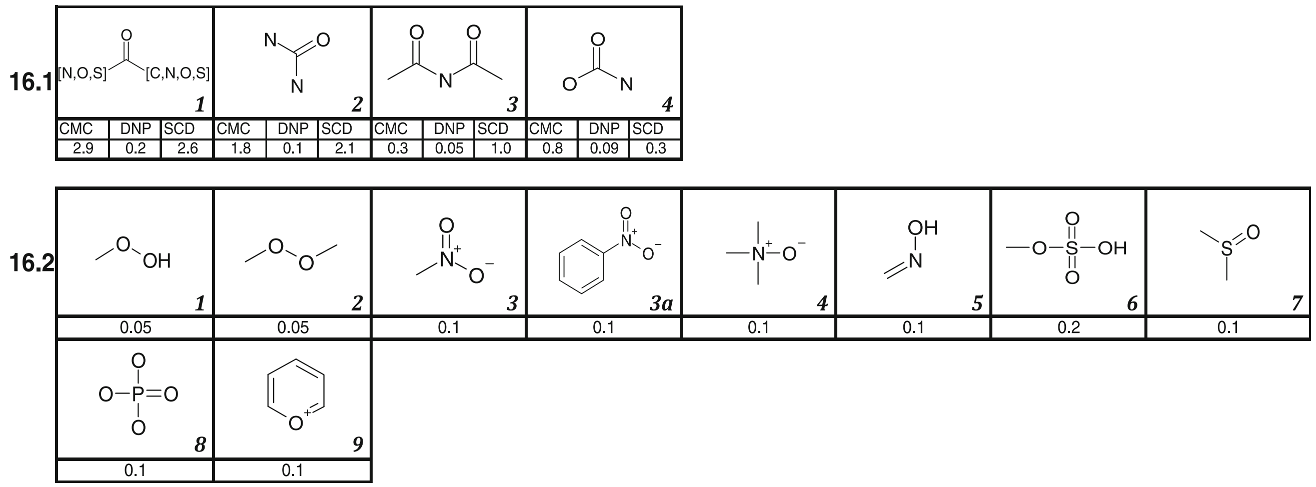


Fig. 10 Scatter diagram for DSF in the DNP, the CMC, and the SCD databases

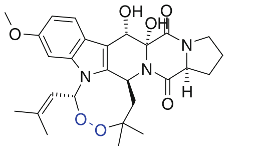
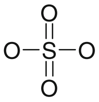
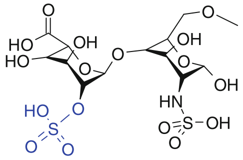
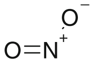
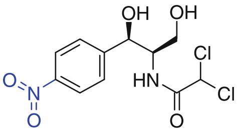
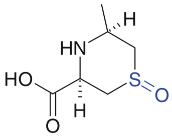
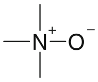
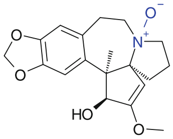
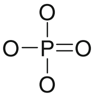
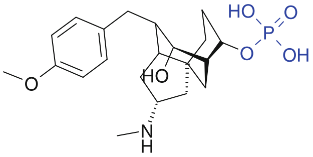
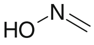
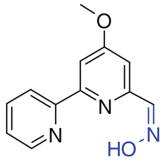
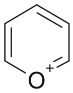
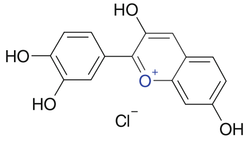
Further development of the previously mentioned bicyclic fragments is highly diversified. Thus a fragment of bridged bicycle (f. 42) can be found in tricyclic dihydroagarofuran system (f. 67), which in turn can be a part of other terpenoid systems. Furan moieties can be found in tricyclic systems like furanoeremophilane (f. 69), for example, in the terpenoid Caesalpinin (ex. 8.3-2). Tetrahydrofuran moieties occur for instance in terpenoids furostane (f. 77) and spirostane (f. 79). Tetrahydrofuran derivatives that belong to furocoumarins, morphine alkaloids, lignans like Pinoresinol (ex. 7.1.1-2) and other classes of natural products can be found in the DNP. Other examples include Galanthamine (ex. 8.2-1), used for the treatment of mild to moderate Alzheimer's disease and various other memory impairments and Aflatoxin B1 (ex. 8.2-2), a widely spread mycotoxin possessing a strong carcinogenic effect.

The evolution of the tetrahydropyran graph (f. 52–62) is also very diverse. One can find there a number of various bicyclic fragments such as non-aromatic fused and spiro systems including cyclic ketals. Bicyclic subgraph (f. 53)



Scheme 16 Other oxygen-containing structural fragments

Table 6 Examples of natural products containing other types of oxygen containing fragments

Scheme	Fragment	Example	Scheme	Fragment	Example
16.2-2	O—O	 Verruculogen	**16.2-6**		 Heparin
16.2-3		 Chloramphenicol	**16.2-7**	S=O	 Cycloalliin
16.2-4		 Cephalotaxine alpha-N-oxide	**16.2-8**		 FR 901483
16.2-5		 Caerulomycin - antibiotic	**16.2-9**		 Fisetinidin

almost entirely belongs to iridoids. The dihydropyran moiety (f. 61) serves as a base for an important class of vitamins, the Tocopherols (ex. 8.2-3). Spiro-cyclic and fused tetrahydropyran ether fragments together with monocyclic tetrahydrofuran and tetrahydropyran fragments linked by C–C bonds in α-position to cyclic oxygen can be found in polyether antibiotics, such as Monensin A (ex. 8.1-3), and in toxins like marine toxins Brevetoxin and Okadaic acid. Unsaturated derivatives (f. 31 c-f) of the parent tetrahydropyran graph evolve into fragments (f. 61a-e), that are parent structures for flavonoids, pyranochromens, and xanthnoids. The fluorescent pigment Leprocybin (ex. 8.3-1) illustrates this cluster.

Another cluster of monocyclic ether moieties includes oxepane derivatives (f. 32) that occur in polycyclic molecules, for instance, in brevetoxins mentioned above. Some monocyclic ether clusters contain two oxygen atoms in the same cycle. The acetal structural fragment (f. 33) can be found in carbohydrates and some aglycones, mostly 5-member dioxymethylene fragments found in some tetrahydroisoquinoline alkaloids (f. 75), flavonoids and lignans. A small 1, 3-dioxane cluster (f. 34) evolves into bicyclic systems (f. 63, 64). Approximately half of 1, 4-dioxane fragments (f. 35) in the DNP database can be found in carbohydrates. Another half is observed mainly among benzodioxine derivatives (f. 66) and to a lesser degree (<0.1 %) in dibenzo-1, 4-dioxine fragments like polyphenols and isoquinoline alkaloids like Kurramine (ex. 8.3-3).

Concluding this section in Fig. 6 we compare the percentage of oxygen in C–O–C subgraphs for all three databases. The figure demonstrates that non-aromatic ring systems containing Cr–O–Cr ether moieties are a distinctive feature of natural products.

C=O fragments

Approximately one-fifth part of all oxygen belongs to carbonyl groups (22.4 %). As follows from Scheme 1.1 and Fig. 2 the portion of C=O structural fragment in the DNP database is 1.7-fold fewer than that in the CMC database and twofold fewer than that in the SCD database (Scheme 1.1, f. 3). However C=O distribution across different classes of organic compounds is very variable. Thus abundance of aldehydes, ketones, esters and lactones in the DNP is not less but sometimes even higher than that in the other two databases, in particular the abundance of cyclic fragments (see sections “Ketones”, “Esters and lactones”, “Amides and lactams” below). On the contrary, the CMC and especially the SCD database include a higher percentage of amides and lactams (Scheme 1.2).

Aldehydes In spite of a rather small proportion of aldehyde moieties in the DNP database (Scheme 1.2, 0.6 %) it is still higher than in the CMC and SCD (0.2 and 0.1 %, respectively). Some of the aldehydes occurring in nature are shown in Table 3.

Natural aldehydes belong mainly to three clusters (Scheme 9, f. 1–3). The most abundant of them contains non-aromatic cyclic aldehydes (f. 2), approximately two-third of them contain saturated or partly unsaturated fragments of cyclohexanecarbaldehyde (f. 8) that are usually included into decaline fragments (f. 9) and can be found among steroids and terpenoids.

The second cluster contains acyclic aldehydes with linear or branched chains (f. 1). Interestingly, like alcohols and ethers, aldehydes include a considerable fraction of derivatives comprising methyl groups in the β- or γ- positions to the oxygen atom (f. 4, 5), for example, in the molecule of the fungal crop destroying toxin Helminthosporal (ex. 9.3-1). Among aldehydes of the first and second clusters one can select a noticeable group of unsaturated aldehydes (f. 6), including cyclohexene derivatives (ex. 9.3-3) as well as acyclic compounds: acrylic, methacrylic, cinnamic aldehydes (ex. 9.2-2), polyunsaturated aldehydes and terpenoids containing the isoprene moiety (ex. 9.2-3).

Lastly, the third cluster contains aromatic aldehydes (f. 3) mainly benzaldehyde derivatives (f. 7) such as, for instance, Vanillin (ex. 9.4-1).

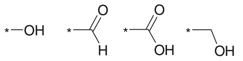
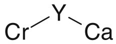
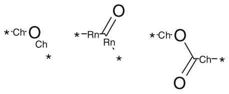
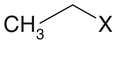
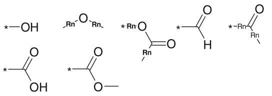
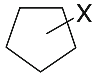
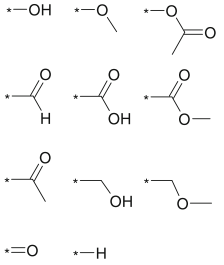
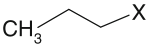
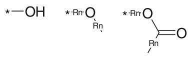
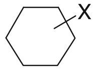
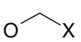
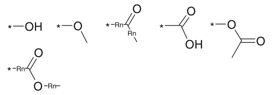
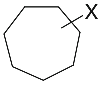
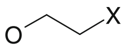
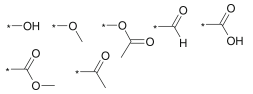
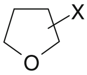
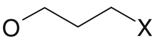
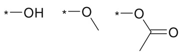
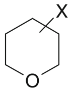
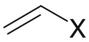
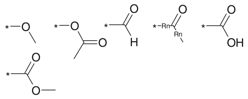
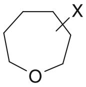
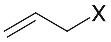
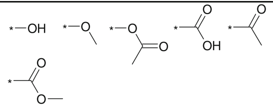
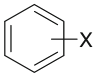
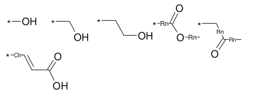
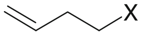
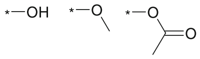
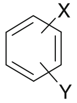
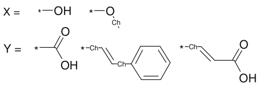
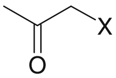
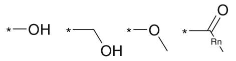
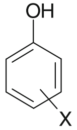
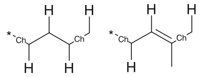
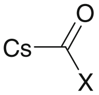
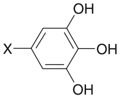
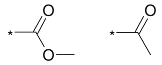
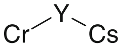
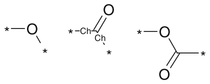
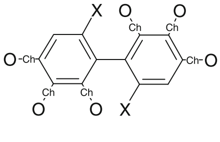
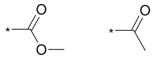
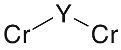
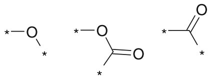
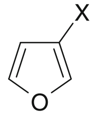
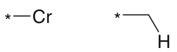
Ketones The proportion of ketone structural fragments in the DNP (Scheme 1.2, 5.9 %) is comparable with that in the CMC database (5.4 %) and exceeds that in the SCD database (3.4 %). However, detailed analysis reveals that the picture is completely different for cyclic and acyclic derivatives (compare Schemes 10 and 11). Thus while the proportion of acyclic ketones among NPs is twofold fewer than in the two other databases (Scheme 10, f. 1), the proportion of cyclic ketones in the DNP (4.7 %) is considerably higher than in the other two databases and fourfold higher than the portion of acyclic ketones in the DNP.

Acyclic ketones. Fragments f. 2–5 (Scheme 10) give an idea about the specific character of acyclic ketones in the DNP database. The major part belongs to ketones with both α-carbon atoms being acyclic, mostly secondary (f. 2), an example is the molecule of the principal component of davana oil, Nordavanone (Table 3, ex. 10.2-1). Two other big structural clusters contain ketone moieties where one of the α-carbon atoms is aromatic (f. 3) or non-aromatic cyclic (f. 0.4).

Fragments f. 6–20 demonstrate modifications of the distant chemical environment of one α-carbon atom. One can note that the content of the analyzed carbonyl fragments decreases in the course of proceeding from methylene to methyl and then to methine derivatives (f. 7 → 6 → 8), and ketones with the carbonyl group connected with a quarternary carbon atom are virtually absent. As previously pointed out for alcohols and ethers there is a big cluster of methyl substituted fragments (f. 9) among alkyl ketones.

Natural products include quite a significant amount,of α, β-unsaturated ketones (f. 10), of them about a third belong to linear vinyl ketone derivatives, and the remaining two-thirds possess double C=C bonds in non-aromatic rings. Phenyl ketones constitute the majority of

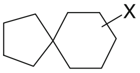
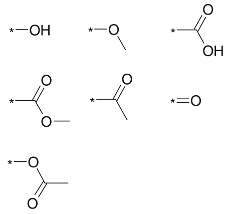
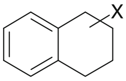
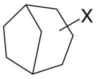
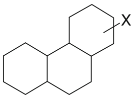
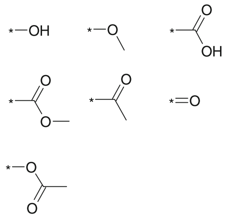
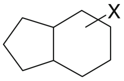
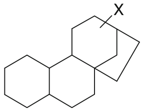
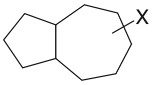
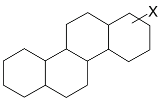
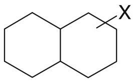
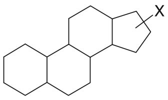
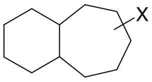
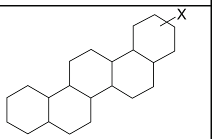
Table 7 Discriminatory set of oxygen containing structural fragments (DSF)

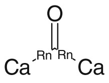
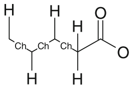
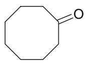
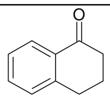
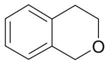
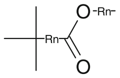
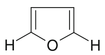
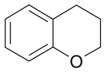
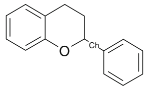
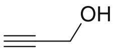
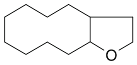
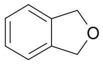
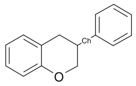
№	Fragment	X, Y	№	Fragment	X, Y
1	Cr—X		14		
2			15		
3			16		
4			17		
5			18		
6			19		
7			20		
8			21		
9			22		
10			23		
11		*—H *—Cs	24		
12			25		
13			26		

X and Y moieties present in the Table should be generally considered as a fragment but not a specific substituent

aryl ketones; about a third part of them are acetophenones. There are small fractions of cyclopentyl- and cyclohexyl-ketones and furan derivatives with different degrees of saturation in the cycle (f. 17–19). Steroids (f. 20) and other terpenes also include a small number of ketone substituents.

Table 7 continued

№	Fragment	X	№	Fragment	X
27			33		*—OH
28			34		
29			35		
30			36		
31			37		
32			38		

№	Fragment	№	Fragment	№	Fragment	№	Fragment	№	Fragment
39		42		45		48		51	
40		43	O—O	46		49		52	
41		44		47		50		53	

X and Y moieties present in the Table should be generally considered as a fragment but not a specific substituent

While the proportion of complexing α-oxy- and β–oxy-ketones (f. 15, 16) is quite appreciable, the percentage of allyl- and benzylketone and 1, 3-diketone fragments does not exceed 0.1 %.

Acyclic ketone moieties are present in very different classes of natural products (Table 3); among them Wyeron possessing antifungal activity (ex. 10.3-2), the alkaloid Lobeline (ex. 10.3-3), the corticosteroid Hydrocortisone (ex. 10.4-1), curcurma component Curcumin (ex. 10.2-3), aroma component of hops Adhumulone (ex. 10.4-3), and a naturally-occurring polycyclic polyprenylated acylphloroglucinol with strong in vitro anti-cancer action Nemorosone (ex. 10.5-1).

Cyclic ketones. Cyclic ketone fragments constitute a representative cluster in the DNP database (Scheme 11, f. 1, 4.7 %). In terms of α-carbon atoms the main cluster contains non-aromatic cyclic ketones (f. 2).

There are 1.5-fold fewer mixed aryl alkyl ketones (f. 3), and even fewer diaryl ketones (f. 4).

A cluster described by f. 2 is highly diversified. This cluster contains a significant amount of fragments having methylene groups in the α-position to carbonyl groups (f. 5) and a reduced amount of corresponding methine derivatives (f. 6). However, a surprisingly high proportion of ketones containing at least one quaternary α-carbon atom (f. 7) seems unexpected. Perhaps such sterically

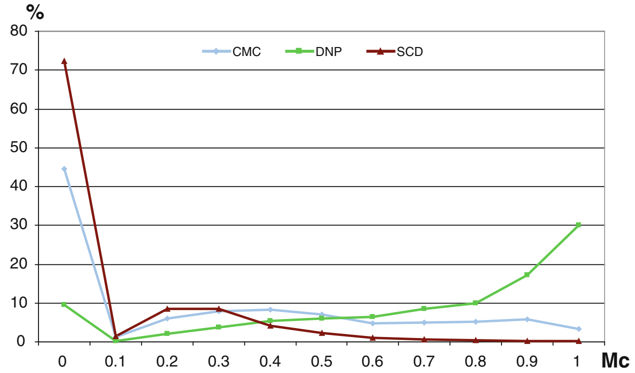


Fig. 11 The proportion of structures in the DNP, CMC, and SCD databases depending on the Mc parameter

hindered and not enolizable ketone fragments are generally more stable and are common for NPs. Another characteristic feature of the DNP is considerable fraction of α-methyl ketones (f. 8, Fig. 5).

Looking at the distant chemical environment of one of the α-carbon atoms adjoining carbonyl groups, various groups of ketone fragments can be identified (f. 9-17). Two large clusters contain phenyl ketones and vinyl ketones. Among other fragments characteristic for NPs the α-oxyketone structural cluster (f. 13) stands out. A half of this cluster is composed of keto-alcohols and the other half by their methoxy-, acyl-, and glycosyl- derivatives. An example of this cluster is the quassinoid Simarolide (ex. 11.2-1). Another remarkable cluster contains cyclic β-oxyketones (f. 14). Allyl- and benzyl- ketone moieties (f. 11, 12) as well as α-amino ketones (f. 15) and 1, 3-dicarbonyl compounds (f. 16, 17) are less common.

Other approaches to cyclic ketone classification are given by subgraphs f. 18–22 that describe ketone fragments connected with monocycles characteristic for NPs, and by subgraphs f. 23–29 describing general monocyclic fragments containing keto-group.

It can be seen that most cyclic ketones belong to cyclohexanone (f. 24) and pyranones (f. 28) clusters. Cyclopentanone, cycloheptanone, cyclooctanone (f. 23, 25, 26) clusters as well as furanone and piperidone (f. 27, 29) clusters contain fewer numbers of ketones. In addition, quite a large group of cyclic ketones (~0.2 %) consists of macrocyclic ketones with ring sizes of more than 9, which belong to carbocyclic and heterocyclic systems but can be hardly formalized due to their diversity, and break up into small clusters.

Cyclopentanone moieties (f. 23) are found in saturated and partly unsaturated cycles in approximately equal proportions. Their transformation leads inter alia to bicyclic fragments (f. 30–32) that can be observed in steroids such as Estrone (ex. 11.2-2) and some terpenoids. It is worth mentioning a bicyclic fragment (f. 32) occurring usually in kaurane derivatives, for example, in the herbal diterpenoid, Ponicidin (ex. 11.2-3).

Cyclohexanone structural fragments (f. 24) evolve mainly into non-aromatic subgraphs with different degrees of saturation. These structural units often lead to hydrogenated moieties of indanones and decalinones (f. 33–38), occurring in steroids and pentacyclic triterpenes. Cyclohexanone fragments including elements of aromatic bonds are parents of bi- and tricyclic systems including benzocyclohexanone, anthrone, naphtoquinone and anthraquinone. Small cycloheptanone and cyclooctanone structural clusters (f. 25, 26), like the macrocycles mentioned above, are mainly involved in complex polycyclic systems.

Ketone heterocyclic moieties of (f. 27–29) mainly include a large pyranone cluster, occurring in flavonoids (ex. 11.3-1) and xanthanoids. The piperidone fragment is almost completely transformed into the quinolone cluster (f. 42) found for example in acridone molecules (ex. 11.4-1).

Concluding the analysis of aldehyde and ketone moieties it is worth noting that the distinctive feature of natural products of these types is the occurrence of non-aromatic rings participating in the construction of the nearest chemical environment of carbonyl groups. Similar tendencies were observed for alcohols and ethers considered above. In the case when the carbonyl group is included in non-aromatic cycles the contrast between the DNP and CMC and especially SCD databases becomes even more crucial (see Scheme 11, f. 2). This trend is illustrated by Fig. 7, showing a graphical comparison of general structural fragments for all three considered databases.

Esters and lactones Oxygen-containing acyclic and cyclic (lactone) ester fragments (see footnote 3) constitute the

Table 8 Percentage coverage by DSF and Mc parameter for the DNP, CMC and SCD and their oxygen containing fractions—$\mathrm{DNP}^{(O)}$, $\mathrm{CMC}^{(O)}$ and $\mathrm{SCD}^{(O)}$

DB	Test DB size	% of covering by DSF	Average Mc for DB	$\mathrm{DB}^{(o)}$ size	$\mathrm{DB}^{(o)}$/DB (%)	% of covering $\mathrm{DB}^{(o)}$ by DSF	Average Mc for $\mathrm{DB}^{(o)}$
DNP	154,175	90.4	0.67	110,550	71.7	97.6	0.78
CMC	8,432	55.3	0.28	1,174	13.1	94.8	0.69
SCD	198,258	27.6	0.08	130,450[a]	1.8[a]	77.9	0.43

[a] From initial DB of 7,237,042 structures

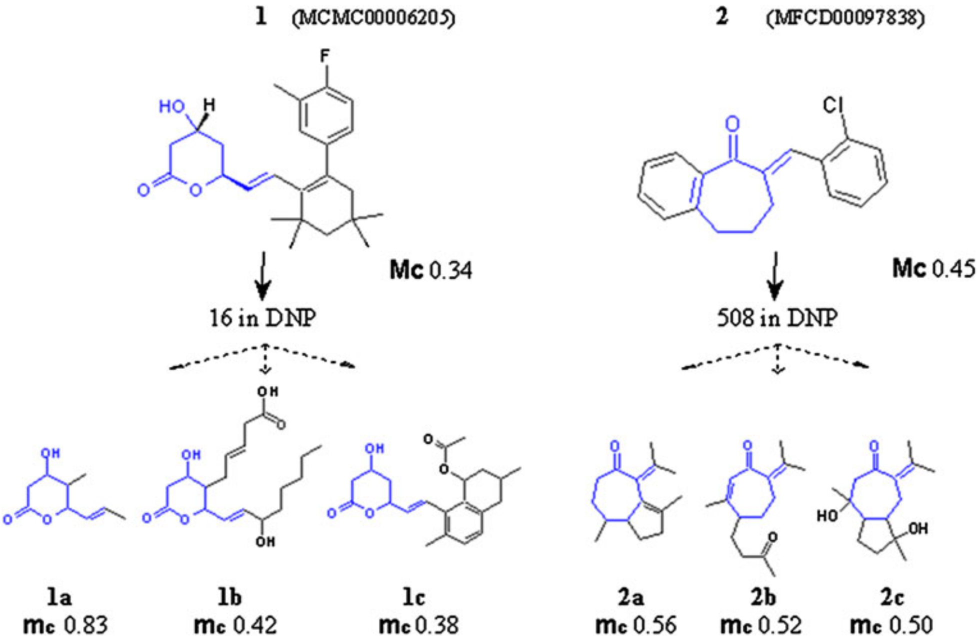


Fig. 12 Search for oxygen-containing molecules in DNP with colored subgraphs

third most abundant cluster in the DNP database (totally 10 %). As follows from Scheme 6.0 the DNP database contains 1.5-fold more ester and lactone groups than the CMC and SCD databases.

Esters. The abundance of acyclic ester fragments among NPs (Scheme 12, f. 1, 6.9 %) is comparable with that in the two other model databases (CMC-5.1 %; SCD-6.0 %), with most of them belonging to aglycones (f. 1). However, detailed analysis of particular clusters (f. 2–9) reveals some nuances.

The two largest ester clusters comprise esters derived from acyclic carboxylic acids and aliphatic (f. 2) and alicyclic (f. 4) alcohols. As expected, in the DNP database esters derived from cyclic non-aromatic alcohols considerably exceed similar esters in the CMC and SCD databases (compare f. 4, 6, and 7). However, the esters derived from cyclic non-aromatic acids (f. 3) are quite rare among NPs.

Consistently, there are some portions of carbohydrate molecules in subgraphs f. 4 and f. 7, and virtually all ester moieties in cluster f. 6 belong to carbohydrates.

In terms of alcohol component the largest cluster in the DNP is that containing secondary alkyl or cycloalkyl groups (f. 14). The abundance drops while proceeding to primary alcohol carbon atom, methyl esters and lastly to the tertiary alcohols. There are small numbers of esters of vinyl alcohols among NPs but appreciable numbers of allylic and homoallylic derivatives, as well as esters derived from phenols and acetals. Esters of carbohydrates participate in the formation of a remarkable cluster of β-oxy alcohol-containing esters (f. 22).

Alcohol oxygen of acyclic esters is often connected with non-aromatic monocycles (f. 24–35) forming cyclohexanol (f. 25) and pyranose (f. 30–34) derivatives and rarer cyclopentanol, cycloheptanol and furanose derivatives.

The DNP database also contains oxepane moieties (f. 35) like the similar cycloheptane fragments (f. 26) and are usually parts of other bi- and polycycles, such as a molecule of the plant metabolite Pipoxide (ex. 12.7-3).

Development of monocyclic systems leads to bicyclic saturated, partly unsaturated and aromatic moieties (f. 36–41). As should be expected the proportion of decaline derivatives is quite noticeable (f. 37), these fragment along with octahydroindene derivatives (f. 36) can be seen in steroids and polycyclic terpenes. Tetrahydropyran moieties as a part of the alcohol component of esters can be found in iridoids (f. 40) and flavanoids (f. 41).

Considering an acid component of acyclic esters one can note that acetates (f. 45) constitute the most remarkable cluster; most of them are in aglycones. An example of an acetate molecule is Streptindole, a genotoxic metabolite isolated from intestinal bacteria (ex. 12.2-3). The proportion of α- and β-methyl derivatives is quite significant (f. 49, 50). There are appreciable numbers of esters formed by α, β-unsaturated acids and aromatic, mainly benzoic acids among NPs. In addition, the DNP database contains derivatives of allylic and phenylacetic acids as well as α-oxy-, β-oxy- and α-amino acids (ex. 12.6-1). Further development of the mentioned subgraphs demonstrates that there are many esters formed by fatty linear acids with the number of carbon atoms more than 5 (f. 59) among NPs, in some cases these esters contain a double bond in the aliphatic chain and are a basis of lipid molecules.

Among the esters of cyclic acids (f. 60–64) the cyclohexanecarboxylic acid derivatives are the most abundant; they often evolve into decalinecarboxylic acid derivatives and then dihydroagarofuran derivatives (f. 67), steroids and triterpenes (f. 68). Some examples of natural esters are given in Table 4.

Lactones. Though the propotion of cyclic ester (lactone) fragments among NPs is not very significant (Scheme 13, f. 1, 3.1 %) a contrast between the databases is really sharp in this cluster. The total occurrence of lactones in the DNP database is twofold higher than this value for drugs and more than fourfold higher than for commercial libraries. Most lactones among NPs are aglycones (f. 1, 2). A major structural cluster where both carbon atoms of the ester moiety are parts of non-aromatic cycles (f. 2) is distinctive for the DNP.

Analysis of the alcohol part of lactones present in the DNP database reveals that the proportions of lactone moieties decreases from secondary to primary and then to tertiary alcohol carbon atoms. Fragments derived from vinylic, allylic, homoallylic alcohols and phenols (f. 9–12) are quite appreciable as well as β-oxyalcohols (f. 14).

The later subgraph and acetal subgraph f. 13 encompass all carbohydrates (~0.3 %). The DNP database also contains lactones formed by β-amino alcohols (f. 15) and lactones formed by cyclic non-aromatic alcohols (f. 17–20).

Consideration of lactones in terms of acid component shows that their content in the DNP database drops while proceeding from secondary to primary and tertiary carbon atom in the α-position to the carbonyl group (compare f. 22 → 21 → 23). However, the main portion of lactones is derived from α, β-unsaturated acids (f. 27). A small but interesting cluster containing a quaternary carbon atom in the α-position to the carbonyl group (f. 24) provides an origin to various fused, spiro and bridged polycyclic derivatives. Subgraphs f. 25–26 and f. 30–31 describe characteristic clusters of α- and β-methyl and α- and β-oxy derivatives.

A series of monocyclic fragments f. 35–41 contains cyclic lactones of various ring size, the largest are the 5-member cyclic subgraph (f. 35) including γ-butyrolactones and butenolides and the 6-member parent subgraph (f. 36), comprising δ- valerolactone and its unsaturated analogues. Notably, an essential part of the γ-butyrolactone cluster (f. 35a) evolves then into subgraph f. 35d containing exocyclic double bonds, an example of which is the molecule Podachaenin (ex. 13.2-1). An appreciable number of macrolide lactones with ring size 10–20 atoms (f. 38) occur in the DNP database, however, only clusters containing 16-member macrolides and dilactone 11-member ring (f. 39, 40) are statistically significant. Cyclic polypeptides having ring size 20–27 also contain lactone bond (f. 41).

Monocyclic systems (f. 35, 36) then evolve into bicyclic fragments (f. 42–50) giving rise to derivatives of benzofuranone pyranone and coumarins (f. 48–50, ex. 13.5-2), and other fused, spiro and bridged systems (f. 42–47). Usually non-aromatic bicyclic fragments including clusters f. 43, 44 are not independent and are parts of more complex tricyclic (for instance, f. 51, ex. 13.2-1 -sesquiterpene lactone Podachaenin) and polycyclic terpenoid systems.

The abundance of oxygen atoms in ester and lactone fragments in the three databases is illustrated by Fig. 8. Our data supports a characteristic feature of the DNP database which is a prevalence of cyclic non-aromatic structural fragments in comparison with the CMC and SCD databases.

Amides and lactams The DNP database contains a rather small number of peptides, so the content of oxygen in amide structural fragments (Scheme 1.2, f. 6, 3.4 %) is considerably lower than that in the database of drugs (18.1 %) and even less than in the SCD database (32.0 %). Almost a half of total N–C=O moieties in the DNP belong to linear amides (1.6 %), another half can be found in lactams (1.8 %).

Amides. Distribution of amide structural fragments (Scheme 14, f. 2–5) according to types of the nearest environment of the oxygen atom reveals that most of them are found in aliphatic systems (f. 2). This trend is also true for the two other databases. There are considerably fewer amides derived from aromatic acids (f. 5). The remaining statistically significant clusters (f. 3, 4) are even less abundant. The DNP database contains virtually no amides derived from aromatic heterocyclic amines (<0.01 %) nor amides formed by cyclic amines and cyclic (both aromatic and non-aromatic) acids (totally < 0.02 %). Common chemical classification also demonstrates that secondary amides constitute the majority (f. 8) in this group. The fractions of primary (f. 7) and tertiary amides (f. 9) are considerably less; almost all tertiary amides are N-methyl derivatives (f. 9a). Secondary amides (f. 8) evolve into subgraphs (f. 8a–e) where nonaromatic fragments are predominant. The last fragments are mostly those having a residue of carboxylic acid, other amides (f. 8d, ex. 14.6-1-the widespread food-contaminating mycotoxins—Ochratoxin A) or a hydroxy group (f. 8e, ex. 14.4-2—a derivative of lysergic acid Ergometrine) in the β-position to the nitrogen atom.

Detailed analysis of the nearest chemical environment of the carbonyl group reveals that the number of amides decreases during proceeding from α-methylene- to N-acetyl- and then to α-methine derivatives (f. 11 → 10 → 12). There are virtually no amides having a quaternary carbon atom in α–position to the carbonyl group (29 only!). As should be expected a subgraph of peptide fragments (f. 14) is the most abundant amide cluster in the DNP database. Corresponding amides formed by α-oxy acids are considerably less frequent

(f. 13). There are a few more appreciable clusters such as α, β-unsaturated acids derivatives (mainly acrylic and cinnamic), benzamides, and fatty aliphatic acid derivatives (f. 15–17). Some examples of characteristic molecules containing amide moieties are presented in Table 5.

Lactams. The contrast in the proportions of lactam fragments among the DNP, CMC, and SCD databases (Scheme 15, f. 1, 1.8 %) is not as marked as in the case of linear amides. Two statistically significant clusters in the DNP are formed by non-aromatic amines and non-aromatic or aromatic acids (f. 2, 3).

Analysis of the amine component of lactams shows that most of them do not have any substituent at the nitrogen atom (f. 4).

Detailed consideration of the chemical environment of the carbonyl group reveals that the largest cluster contains a cyclic peptide fragment (f. 11). One can also observe lactams formed by α, β-unsaturated acids (f. 7), as well as lactams based on benzoic, phenylacetic and α-oxy acids (f. 8–10). Approximately 0.1 % of lactams in the DNP are fragments of hetaryl substituted lactams—purine, thiazole, oxazole etc. derivatives and can be described by subgraph f. 3.

Further development of analyzed fragments leads to the most valuable monocyclic types of lactams: pirrolidone, piperidone, pyrimidones, and diketopiperazine derivatives (f. 12–15, ex. 15.2-1-3 and 15.3-2-3). It is worth mentioning an important group of penicillin and cephalosporin β-lactams, though the total content of β–lactam moieties is not large (~0.02 %). A distinctive feature of lactams is the presence of quite a big (totally ~ 0.3 %) cluster of cyclic peptides with ring size 10–27 described by parent peptide subgraph (f. 11), which is difficult to formalize.

Graphical comparison of the abundance of amide and lactam fragments in the model databases is given in Fig. 9.

Other classes of carbonyl compounds Scheme 16.1 and Fig. 8 outline some rather rare structural fragments containing a carbonyl group. Urea moieties (f. 2) appear to be the most essential and most of them are parts of pyrimidone type heterocyclic systems, in particular they can be found in barbituric acid derivatives and purine bases. A molecule of the alkaloid Theophylline (ex. 16.1-1) is an example of urea. The related cluster of carbamate fragments (f. 4) includes mainly acyclic moieties. Natural imides (f. 3) are mainly cyclic, the largest fraction contains succinimide derivatives. Derivatives of thio-acids, thio-carbamates, anhydrides and hydrazides of carboxylic acids are rare. Examples of these classes of compounds are given in Table 5.

Other oxygen-containing structural fragments Some other oxygen containing fragments having an oxygen-heteroatom bond are present in the DNP database in insignificant amounts (Scheme 16.2).

A relatively big cluster of natural products contains organic peroxides (Scheme 1.1, f. 9, 0.1 %), a half of them are cyclic fragments (ex. 16.2-2—the indole alkaloid Verruculogen), the other half are hydroperoxides (Scheme 16.2, f. 1, 2).

Natural products containing a nitro-group (f. 3) do not occur frequently and are mainly present as nitrophenyl moieties (f. 3a). N-Oxides (f. 4) and oximes (f. 5) are also rare among natural products.

Some examples are shown in Table 6 (ex. 16.2-3, 4, 5).

Oxygen containing sulfur derivatives can be found mainly as fragments of sulfonic acids (f. 6, 7); the most valuable of them are sulfonic esters (f. 6). Most of these esters can be observed among steroids and tetrahydropyrans. There are small amounts of sulfoxides (f. 7) and fragments of benzene sulfonic acids. In contrast to the CMC and SCD the DNP database contains only a few sulfamides.

Phosphorus in the DNP is mainly included in phosphoric acid residues (f. 8) that form esters with one, two or three alcohol residues or diphosphate or triphosphate moieties. These moieties can be found in nucleotides.

Pyrillium salts forming quite an appreciable cluster (f. 9, ex. 16.2-9) are among other oxygen containing fragments in the DNP database.

Small amounts of fragments of other oxygen containing compounds such as sulfones, nitroso-compounds, as well as arsenic or silicium derivatives can be found among NPs. However, the content of these fragments does not exceed the predetermined statistical threshold.

Some applications and perspective

Our classification of oxygen containing structural fragments found in the DNP database in statistically significant numbers enables us to suggest some generalizations. First of all, some structural clusters that are predominant among natural products can be found. They can be considered as distinctive for the description of NPs and NP-like structures. These fragments are usually non-aromatic and bear the properties of cyclicity and branching, moreover, they often contain an enhanced amount of methyl groups. On the other hand, aromatic oxygen-containing natural products also have their own distinctive features: they usually include oxy- and polyoxyphenyl groups as well as furan, benzofuran, benzopyran and benzoquinone moieties. We consider that the two key questions that arise from our analysis are firstly, how can this knowledge be applied for the design of new NP-like molecules and secondly, where is the cut-off when oversimplification results in the loss of useful biological properties?

Though the design of new NP-like molecules is beyond the scope of this paper, a number of short examples for illustration of possible applications of the current study are given below.

For this purpose we selected a number of subgraphs that are distinctive for the DNP in comparison with the database of commercial chemicals SCD (see Table 7) based on the content values given in Schemes. Subgraphs describing the nearest chemical environment of oxygen atoms and statistically significant mono- bi- and polycylic fragments related to these parent subgraphs were taken. Then for illustration purposes we restricted ourselves to subgraphs containing only C, H, and O atoms. Altogether 53 types of structural fragments were chosen and are schematically shown in Table 7. In real molecules these fragments are not necessarily independent but can overlap with each other, additionally confirming the discriminatory capacity of these fragments. Compounds found simultaneously in two or three model databases were omitted.

These fifty-three fragments occur in the DNP database with different frequencies; however in all cases their relative contents among natural products are considerably higher than among drugs and especially synthetic compounds. On the other hand, any of DSF can contribute to the molecule to different degrees. To evaluate this factor we introduced a parameter Mc which denotes a ratio of C and O atoms belonging to a particular fragment, to the total number of heavy atoms in a molecule. The bigger this parameter suggests that the larger part of a molecule belongs to a considered fragment. Figure 10 demonstrates a scatter diagram for the distinctive set of fragments in the three model databases. Axis Z denotes the percentage of structures in the database at least once containing some fragment, which order number is shown on axis X and corresponds to the number in Table 7 above. The Mc parameter is plotted on the ordinate. One can see that points corresponding to the three model databases occupy clearly different areas of space, confirming selected fragments to be characteristic.

Moreover, in total approximately 90 % of all structural formulae in the DNP, 55 % in the CMC and 28 % in the SCD databases contain at least one fragment from DSF. The mean value of the Mc parameter is 0.67 for the DNP, 0.28 for the CMC and 0.08 for the SCD database. This means that the selected DSF encompasses approximately 67 % of all heavy atoms in the molecules contained in the DNP database and only 8 % contained in the SCD database.

Figure 11 reflects the Mc parameters depending on the portion of encompassed molecules in the databases. It is evident from the figure how different the model databases are in terms of the proportion of NP-like oxygen containing fragments. Moreover, the detailed analysis shows that when $Mc > 0.7$–0.8, the DSF picks out of the CMC and the SCD databases structural formulae virtually indistinguishable from NPs. Without any analysis regarding the source of such molecules in these databases one can suggest that most of them are indeed NPs but for some reason are not included in the current version of the DNP.

Next, to maintain the experimental integrity, we restricted the model databases to only C, H, O and halogen containing molecules having at least one oxygen atom in their structural formula. The size of the model databases was reduced to 72 % of the original size for the DNP (hereafter $DNP^{(O)}$) and to 13 % for the CMC. Due to very strong shrinkage of the SCD database its initial version (7,237,042 molecules) was taken. After standard purification and described criteria application the size of the $SCD^{(O)}$ accounted for only 130,450 structures (1.8 % of the original size), and this database was taken for further calculations (See Table 8).

Let us point out that such a strong shrinkage of the original CMC and especially SCD databases compared to the DNP database can be considered as evidence of “underestimation” of oxygen containing NP-like fragments in these databases.

The results of superposition of the DSF and oxygen containing fractions present in the analyzed databases are shown in Table 8. In this case the DSF encompassed 98 % of molecules from the $DNP^{(O)}$, 95 % from the $CMC^{(O)}$ and 78 % from the $SCD^{(O)}$. The mean value of the Mc parameter became 0.78 for the $DNP^{(O)}$, 0.69 for the $CMC^{(O)}$ and 0.43 for the $SCD^{(O)}$.

Thus, even for oxygen containing fractions of the model databases one can observe a considerable excess of the Mc parameter for the database of natural products. We believe that such a considerable difference in Mc among the model databases suggests that the provided distinctive set of fragments (DSF) can be used as a tool for searches for NPs and NP-like molecules including oxygen containing structural fragments.

The independent experiment described below can be considered as a confirmation of such a suggestion.

Two random molecules were taken from the CMC (1) and the SCD (2) databases; both of them had nonzero values for Mc (Fig. 12). All C and O atoms encompassed by DSF and bonds connecting them are colored blue, the values of Mc are also provided. Both molecules are not included in the DNP database. In order to find out whether the revealed colored subgraphs are NP-like fragments we used them as a structural query in the DNP database. We wished to know if real clusters of NPs described by these subgraphs exist in the DNP.

For the molecule 1 the colored subgraph appeared to find a structural cluster in the DNP database that contains 16 molecules of natural products with various substituents and other structural elements. For the molecule 2 a cluster comprising 508 compounds was found. Therefore we can state that indeed the colored portions of the molecules 1 and 2 are NP-like oxygen containing fragments. Examples of some analogues from the DNP for both these molecules are presented in Fig. 10. Moreover, the selected molecules

(1 a–c… and 2 a–c…, Fig. 12) in turn can be sorted in accordance with correspondence of the mc parameter to particular colored subgraph in order to estimate their resemblance to parent molecules 1 and 2.

Thus we believe that the DSF suggested by us can be used as a tool for searches for NP-like molecules having oxygen containing moieties.

From our point of view the suggested approach can be useful not only for searching for existing NP-like molecules but for the design of new NP-like molecules within the framework of one or another cluster. Obviously to get a complete picture it would be helpful to consider N-, S-, and Hal-containing structural fragments as we have done for O-containing fragments. However, this task lies beyond the scope of this paper. We wanted only demonstrate the tool for classification and design of new NP-like oxygen containing molecules. Every skilled medicinal chemist can use and further develop this tool.

Additionally, the suggested approach could allow the design of hybrid molecules combining NP-like and drug-like properties. During the design of such molecules it is possible to use bioisosteric replacement, scaffold-hopping [22, 23] of some elements not including (or partly including) into the part of molecules encompassed by distinctive fragments. For example, altering of fragments is possible beyond the distinctive area or even partial replacement of oxygen containing NP-like fragment for the corresponding nitrogen containing drug-like version and so on.

Conclusions

A new analytical approach is suggested that is distinctive from the numerous existing approaches using scaffold concept [13, 22, 24] or some other related concepts (scaffold tree approach, scaffold network [2, 8, 25]). In contrast to them our approach moves away from the traditional ideas of scaffolds, cycles, linkers and substituents and suggests relying on any structural elements (connected and disconnected subgraphs) of the molecules. It is only important for these elements to be predominant among the molecules that possess some particular properties. In our study these are oxygen containing NP-like fragments.

An analysis of the chemical environment of the oxygen atoms in the DNP database compared to the CMC and SCD databases was performed. Fifty-three oxygen-containing structural fragments distinctive for the DNP (distinctive set of fragments DSF) in comparison with the SCD have been identified. A new descriptor Mc was introduced for describing the ratio of atoms involved in the DSF to the total number of heavy atoms. A significant difference in the Mc values among the reference databases allowed using a specific cluster of DSF as a tool for performing similarity searches for oxygen-containing NP molecules or for evaluation or comparison of databases accordingly to their NP-likeness. We provide an example illustrating that the suggested approach could allow not only estimating the NP-likeness but serve as a tool for designing new NP-like compounds.

Acknowledgments The authors thank Dr. David Langley for his valuable comments and suggestions.

References

1. Grabowski K, Schneider G (2007) Properties and architecture of drugs and natural products revisited. Curr Chemcial Biol 1:115–127
2. Grabowski K, Baringhausb KH, Schneider G (2008) Scaffold diversity of natural products: inspiration for combinatorial library design. Nat Prod Rep 25:892–904
3. Quinn RJ, Carroll AR, Pham NB, Baron P, Palframan ME, Suraweera L, Pierens GK, Muresan S (2008) Developing a drug-like natural product library. J Nat Prod 71:464–468
4. Camp D, Davis RA, Campitelli M, Ebdon J, Quinn RJ (2012) Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod 75:72–81
5. Cragg GM, Grothaus PG, Newman DJ (2009) Impact of natural products on developing new anti-cancer agents. Chem Rev 109:3012–3043
6. Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI (2009) Novel chemical space exploration via natural products. J Med Chem 52:1953–1962
7. Carlson EE (2010) Natural products as chemical probes. ACS Chem Biol 5:639–653
8. Lachance H, Wetzel S, Kumar K, Waldmann H (2012) Charting, navigating, and populating natural product chemical space for drug discovery. J Med Chem 55:5989–6001
9. Feher M, Schmidt MS (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J Chem Inf Comput Sci 43:218–227
10. Ertl P, Schuffenhauer A (2007) Chemoinformatic analysis of natural products: lessons from nature inspiring the design of new drugs. In: Petersen F, Amstutz R (eds) Natural products as drugs. Birkhaeuser, Basel, Switzerland, pp 217–236
11. Camp D, Davis RA, Campitelli M, Ebdon J, Quinn RJ (2012) Drug-like properties: guiding principles for the design of natural product libraries. J Nat Prod 75:72–81
12. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 47:47–58
13. Bon RS, Waldmann H (2010) Bioactivity-guided navigation of chemical space. Acc Cnem Res 43:1103–1114
14. Yeap SK, Walley RJ, Snarey M, van Hoorn WP, Mason JS (2007) Designing compound subsets: comparison of random and rational approaches using statistical simulation. J Chem Inf Model 47:2149–2158
15. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci USA 102:17272–17277
16. Genis D, Kirpichenok M, Kombarov R (2012) A minimalist fragment approach for the design of natural-product-like synthetic scaffolds. Drug Discovery Today 17:1170–1174

17. Young SS, Farmen M, Rusinko A III (1996) Random versus rational: which is better for general compound screening? Netw Sci [Online] 2, Article 9. http://netsci.org/Science/Screening/feature09.html
18. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 48:68–74
19. Bremser W (1978) HOSE—a novel substructure code. Anal Chim Acta 103:355–365
20. De Grave K, Costa F (2010) Molecular graph augmentation with rings and functional groups. J Chem Inf Model 50:1660–1668
21. Bemis W, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893
22. Hu Y, Stumpfe D, Bajorath J (2011) Lessons learned from molecular scaffold analysis. J Chem Inf Model 51:1742–1753
23. Böhm H-J, Flohr A, Stahl M (2004) Scaffold hopping. Drug Discovery Today Technol 1:217–224
24. Agrafiotis DK, Wiener JM (2010) Scaffold explorer: an interactive tool for organizing and mining structure-activity data spanning multiple chemotypes. J Med Chem 53:5002–5011
25. Varin T, Schuffenhauer A, Ertl P, Renner S (2011) Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. J Chem Inf Model 51:1528–1538