CROATICA CHEMICA ACTA
CCACAA **78** (2) 141–149 (2005)

ISSN-0011-1643
CCA-2990
Original Scientific Paper

Binary Substructure Descriptors for Organic Compounds*

Kurt Varmuza, Wilhelm Demuth, Manfred Karlovits, and Heinz Scsibrany**

Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166-2, A-1060 Vienna, Austria

RECEIVED JUNE 26, 2004; REVISED OCTOBER 28, 2004; ACCEPTED NOVEMBER 2, 2004

Organic chemical structures are represented by binary vectors that contain information about presence or absence of 1365 substructures. The guiding ideas for selecting this set of substructures are described and examples are given. Software SubMat has been developed for a fast and flexible computation of binary substructure descriptors from molecular structures. Examples from structure similarity searches demonstrate the performance of representing organic chemical structures by the described set of substructures.

Keywords
chemical structure similarity
Tanimoto index
software SubMat

INTRODUCTION

Among the many methods used and suggested to represent chemical structures by a set of numerical descriptors,[1–5] substructure descriptors play an important role. Characterization of a molecule by a set of substructures is evident to chemists and directly related to similarity or diversity of chemical structures. Substructure descriptors, however, are less powerful than topological descriptors or other descriptors derived from 2D or 3D structures, for modeling quantitative relationships between chemical structures and properties or activities. If only the presence or absence of substructures is encoded, a binary vector (sometimes called fingerprint) represents a chemical structure. Most methods of multivariate data analysis require vectors of constant length, that means, a fixed set of substructures has to be used for all investigated molecular structures.

A set of substructures for organic molecules is described, and software SubMat is presented which allows an easy and flexible generation of binary substructure descriptors. Applications demonstrate the use of these descriptors for structure similarity searches, characterization of the diversity of chemical structures in databases, and cluster analysis of chemical structures. Clustering of chemical structures by using these substructure descriptors has been described previously;[6,7] furthermore they have been used for examinations and improvements of spectra similarity measures in infrared spectroscopy[8] and in mass spectrometry.[9]

METHODS

Substructures

Aim was the definition of a set of substructures that cover a large diversity of organic molecules. The strategy applied for the creation of substructures was as follows: (i) Restriction to most common elements; (ii) systematic generation of substructures by using an isomer generator; (iii) selection of substructures by chemical experiences; (iv) elimination of very exotic substructures. Finally, a set of 1365 substructures was obtained, divided into eight groups as shown in Table I.

---

* Dedicated to Dr. Edward C. Kirby on the occasion of his 70th birthday.

** Author to whom correspondence should be addressed. (E-mail: kvarmuza@email.tuwien.ac.at)

TABLE I. Substructure groups and number of substructures per group

Group number	Group definition	No. of substructures
1	Elements (single atom substructures)	46
2	Two-atom substructures	78
3	Single, not aromatic rings	404
4	Condensed, not aromatic rings	130
5	Aromatic rings	97
6	Other rings	39
7	Trees (chains and branches)	418
8	Functional groups	153
	Total	1365

The non-hydrogen elements present in the substructures are C, N, O, S, P, F, Cl, Br, I, B, and Si. Furthermore, two pseudo element have been defined; pseudo element A for hetero atoms (any atom except C or H), and pseudo element Q for non-H atoms. In some cases explicit H-atoms are used. Free valences in molecular structures are interpreted as H-substituted; free valences in substructures can be attached to any atom. The bond types used are single, double, triple, and aromatic; additionally the bond type may be not specified (any type). Most substructures are single structures, however, some substructures contain not connected structural parts. No stereoisomeric information or 3D data are used in the substructures, because the applied substructure searches only consider connectivities of atoms. Details about the contents of the eight substructure groups and the creation of the substructures are given below.

(1) Elements (Single Atom Substructures). – Each substructure consists of a single atom or up to six not connected atoms of the same element: $N_{1–6}$, $O_{1–6}$, $S_{1–6}$, $P_{1–6}$, $F_{1–3}$, $Cl_{1–3}$, $Br_{1–3}$, $I_{1–3}$, $B_{1–3}$, Si, $A_{1–6}$. For instance the descriptor for substructure A_3 is 1 if at least three hetero atoms are present in the molecule, otherwise it is 0.

TABLE II. Examples for two-atom substructures

Subgroup	Element		Bond Type[(a)]				
	Atom 1	Atom 2	s	d	t	a	n
C and another	C	C	+	+	+	+	+
	C	N	+	+	+	+	+
	C	O	+	+			+
	C	S	+	+			+
	C	A	+	+	+	+	+
	C	F	+				
	C	Cl	+				
	C	Br	+				
	C	I	+				
N and another	N	N	+	+	+	+	+
	N	O	+	+			+
	N	S	+	+			+
	N	A	+	+	+	+	+
	N	Q	+	+	+	+	+

[(a)]Bond types are s, single; d, double; t, triple; a, aromatic; n, not defined. A plus (+) indicates that this substructure is used.

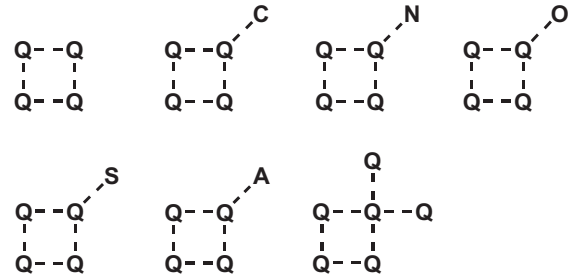

Figure 1. Examples for substructures in group 3 (single, not aromatic rings). Four-membered rings made only by Q-atoms and the used substitutions are shown. All bonds have not-defined type. Such ring substructures have been defined for ring sizes 3 to 8.

(2) Two-atom Substructures. – These substructures contain all two-atom combinations of the elements C, N, O, S, F, Cl, Br, I, A, and Q with all five bond types applied; however, restricted by the chemical valence rules, a minimum of one free valence, and a reasonable chemical meaning. Table II shows as examples the used two-atom substructures containing a C-atom plus another element, and a N-atom plus another element. Combination C+Q, for instance, is not used because already covered by C+C and C+A.

(3) Single, not Aromatic Rings. – These substructures contain a ring with a size between three and eight atoms. Three subgroups have been defined: (a) Rings made only by Q-atoms; (b) rings made only by C-atoms; (c) rings containing hetero atoms.

(a) Rings made only by Q-atoms (size 3 to 8 atoms) have been created with bond type single and bond type not-defined. The presence of any ring with a given size is tested with a corresponding ring substructure consisting of only Q-atoms and with the ring bonds not defined. Figure 1 gives the substituents considered, shown for 4-membered rings. This subgroup contains 42 substructures.

(b) Rings made only by C-atoms (size 3 to 8 atoms) have been more extensively varied resulting in 134 substructures. Figure 2 gives the substituents and unsaturations considered, shown for 4-membered rings.

(c) Heterocyclic rings of size three to five and made by the elements C, N, and O have been exhaustively built by using the isomer generator software Molgen.[10] Selected S-containing rings and 6-membered hetero rings have been created manually to avoid combinatorial explosion. Figure 3 shows all possible 3-membered rings made from elements C, N, and O, using single, double and triple bonds, containing at least one hetero atom, and having at least one free valence. These 17 substructures have been searched in the Beilstein Crossfire database system,[11] containing at this time *ca* 4 million compounds, in an infrared spectral database,[12] containing 13,484 compounds, and in a mass spectral database,[13] containing 106,955 compounds. Only two substructures have not been found in the Beilstein database.

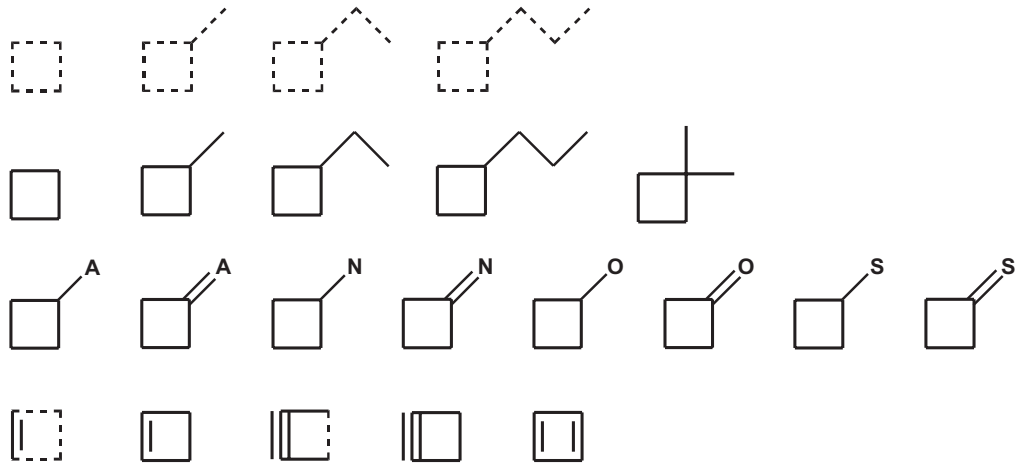

Figure 2. Examples for substructures in group 3 (single, not aromatic rings). Rings made only by C-atoms and used substitutions as well as unsaturations are shown for 4-membered rings. Such ring substructures have been defined for ring sizes 3 to 8. A dotted lined denotes a not defined bond type.

In general, the systematic generation of this type of heterocyclic rings can be defined as follows. All possible rings have been created by the Molgen software, using molecular formulae $C_cN_nO_oH_h$ with three, four or five atoms of the elements C, N, O (with at least one C-atom), and at least one H-atom.

$$c + n + o = r \text{ for } r = 3, 4, 5 \quad (1)$$

$$h = h_{max},\ h_{max} - 2,\ h_{max} - 4,\ldots \text{ with } h > 0 \quad (2)$$

$$h_{max} = 2c + n \quad (3)$$

The H-depleted ring structures have been used as substructures. Substructures generated by this approach that have been found in less than 20 entries of the Beilstein Crossfire database have been eliminated. From 17 created 3-membered rings 12 survived, from 56 created 4-membered rings 18 survived, and from 189 created 5-membered rings 86 survived.

(4) Condensed, not Aromatic Rings. – These substructures have been created manually. Combining two rings of size three to six resulted in 93 substructures. As an example, Figure 4 shows the nine substructures obtained by a combination of a 5-membered ring with a 3-, 4- or 5-membered ring.

	N—O	N—O	N—N	N—N	N=N	O
BCF	1,603	3	1,023	12	912	75,463
IR	0	0	4	0	0	93
MS	15	0	13	0	3	1,232

	O	N	N	N	O—O	N
BCF	35	16,425	24	625	54	0
IR	0	2	0	0	0	0
MS	0	229	0	11	0	0

	N	N—O	N—N	N—N	N=N
BCF	4	0	31	185	41
IR	0	0	0	0	0
MS	0	0	0	4	0

Figure 3. Exhaustive set of 3-membered hetero cyclic rings made from elements C, N, and O, containing at least one hetero atom and having at least one free valence. Number of occurrences in the Beilstein Crossfire Database (BCF; 4 million compounds), in an infrared spectral database (IR; 13,484 compounds), and in a mass spectral database (MS; 106,955 compounds) are given.

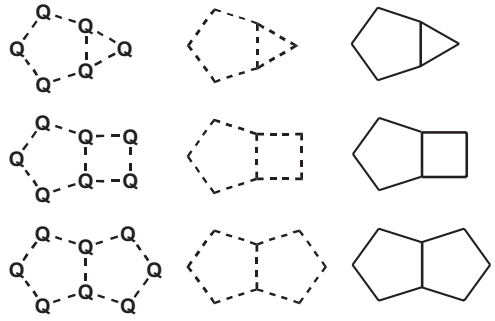

Figure 4. Selected substructures with condensed rings (group 4) obtained by the combination of a 5-membered ring with a 3-, 4- or 5-membered ring.

Further 37 substructures contain three to four condensed rings, as for example present in polycyclic hydrocarbons or steroids.

(5) Aromatic Rings. – The four subgroups are: (a) Benzene ring(s) including for instance biphenyl, and benzene rings substituted by C- or Q-atom(s) (27 substructures); (b) A single benzene ring directly substituted by a hetero atom (N, O, S, F, Cl, Br, A) or substituted by a chain C-atom connected to a hetero atom (33 substructures); (c) Benzene ring(s) condensed to another ring including for instance dibenzothiophene (17 substructures); (d) Substructures with N-aromatic rings (6-membered rings with one, two or three N-atoms) including for instance adenine; see Figure 5 (20 substructures). For substructure searches, the bond type aromatic is used instead of alternating double and single bonds.

(6) Other Rings. – Bridged ring systems as for instance present in terpenes or similar compounds have been created manually. In one subgroup all atoms are C, connected by single bonds; in another subgroup all atoms are C, connected by bonds with a not defined type, and in the third subgroup all atoms are Q, connected by bonds with a not defined type.

(7) Trees (Chains and Branches).– This group contains non-cyclic substructures with elements C, N, O, and Q. Most substructures have been generated systematically by using the isomer generator software Molgen. Three subgroups have been defined: (a) Alkyl and similar; (b) alkenyl, alkinyl and similar; (c) 3- and 4-atom substructures with C, N, O. Details of isomer generation are as follows.

TABLE III. Number of tree substructures (isomers) with three to six C-atoms and one or two double bond equivalents (DBE)

C-atoms	Number of isomers		
	DBE = 1 one double bond	DBE = 2 two double bonds	DBE = 2 one triple bond
3	1	1	1
4	3	2	2
5	5	6	3
6	13	16	7
sum	22	25	13

(a) The saturated substructures only containing C-atoms comprise the n-alkyl groups with 3–12, 15, 20, 25, 30 atoms either with single bonds or with the not-defined bond type. The branched substructures have been obtained by generating all isomers for C_4H_{10} to C_6H_{18}, and eliminating the non-branched ones. The resulting 32 substructures are either defined with single bonds or with not-defined bond type.

(b) The unsaturated substructures only containing C-atoms have been restricted to 3 to 6 atoms, and to 1 or 2 double bond equivalents (DBE); the numbers of isomers are listed in Table III.

(c) All tree structures containing three (respectively four) atoms of the elements C, N, O with at least one C-atom have been generated; the number of double bond equivalents has been varied between zero and the maximum possible value. The molecular formulae used for isomer generation and the number of isomers are listed in Table IV.

(8) Functional Groups. – Functional groups not present in other substructure groups have been defined manually and collected in seven subgroups according to the presence of

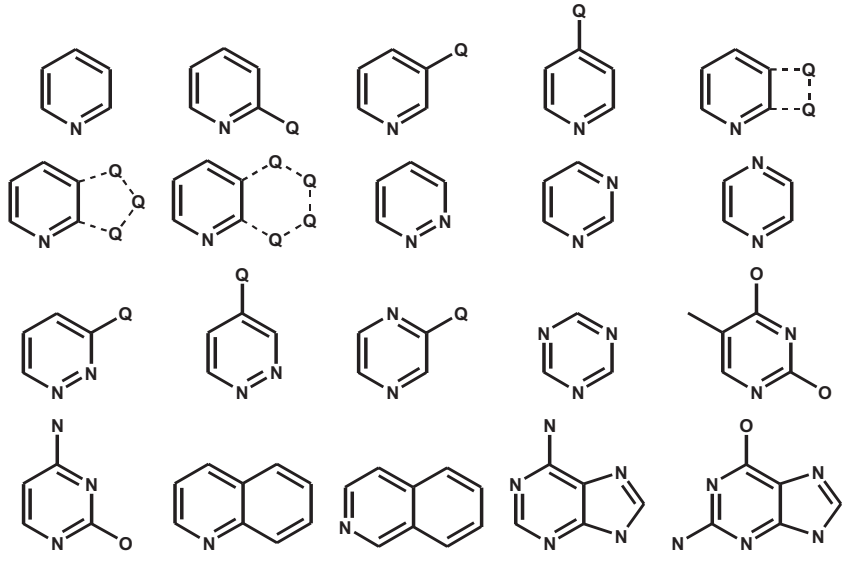

Figure 5. Substructures used containing a 6-membered N-aromatic ring (group 5).

TABLE IV. Number of substructures (isomers) with all combinations of three or four atoms from C, N, O containing at least one C-atom and one H-atom. All possible values for the number of double bond equivalents (DBE) have been used.

No. of atoms	Formula	No. of H-atoms, no. of isomers					
		DBE = 0	DBE = 1	DBE = 2	DBE = 3	DBE = 4	Row sum
3	C2 N	H7 2	H5 3	H3 3			8
	C2 O	H6 2	H4 2	H2 2			6
	C N2	H6 2	H4 3	H2 2			7
	C N O	H5 3	H3 4	H1 2			9
	C O2	H4 2	H2 1				3
	Sum						33
4	C3 N	H9 4	H7 8	H5 9	H3 4	H1 1	26
	C3 O	H8 3	H6 6	H4 6	H2 2		17
	C2 N2	H8 6	H6 12	H4 11	H2 4		33
	C2 N O	H7 8	H5 15	H3 12	H1 3		38
	C2 O2	H6 5	H4 6	H2 4			15
	C N3	H7 4	H5 7	H3 4	H1 1		16
	C N2 O	H6 8	H4 14	H2 7			29
	C N O2	H5 8	H3 9	H1 3			20
	C O3	H4 3	H2 2				5
	Sum						199

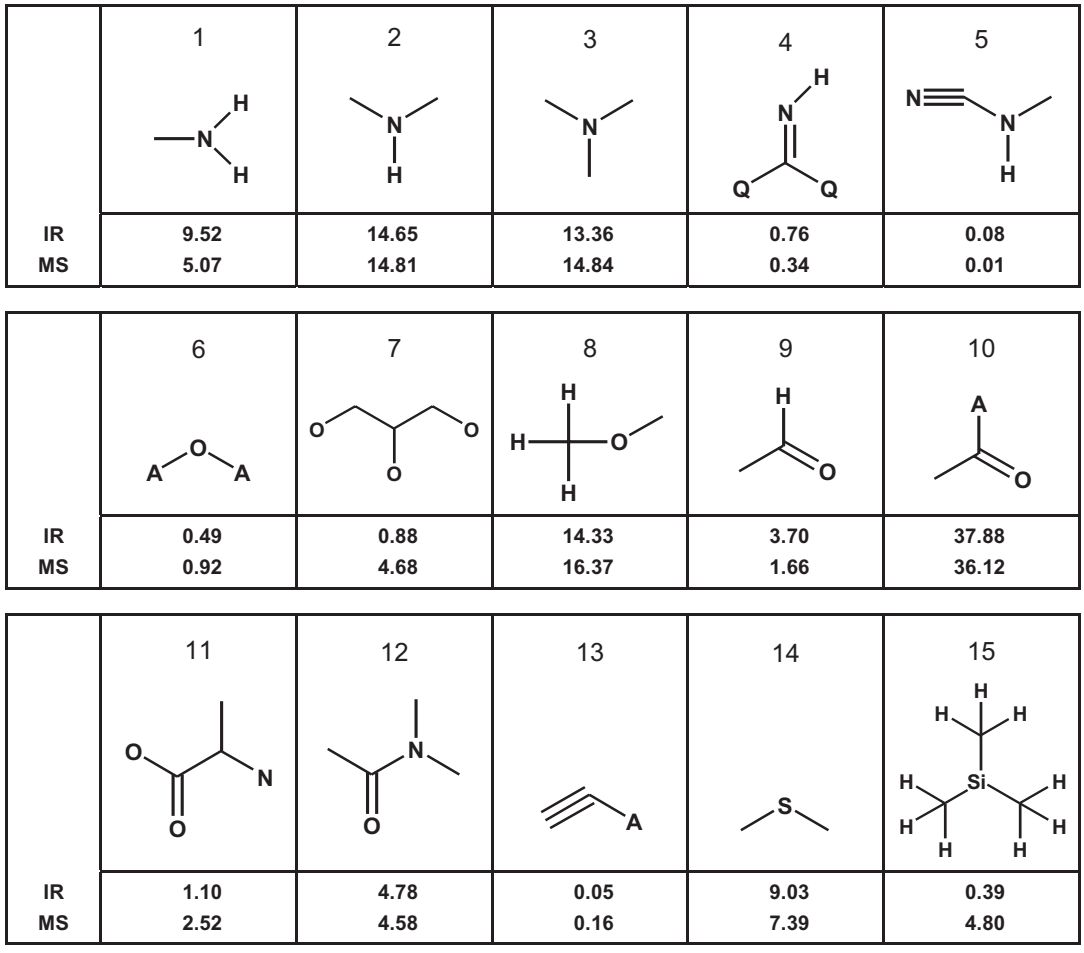

Figure 6. Examples of substructures from group 8 (functional groups). The percent of compounds containing the substructure is given for two spectral databases, one with 13,484 infrared spectra, the other with 106,955 mass spectra; see Figure 9.

hetero atoms. (a) Only N (36 substructures); (b) only O (48 substructures); (c) only S (15 substructures); (d) N and O (29 substructures); (e) N and S (4 substructures); (f) O and S (10 substructures); (g) others (11 substructures). Examples are shown in Figure 6 together with the percent of compounds containing the substructure for two spectral databases.

Software SubMat

Software SubMat has been developed for a fast and flexible generation of binary substructure descriptors.[14] SubMat requires two input files, one for the molecular structures, the other for the substructures; all structures have to be in Molfile format.[15,16] A few parameters control the output format of the generated descriptors to a text file (Figure 7). The result file contains a row for each molecular structure and a column for each substructure; value »1« means that the substructure is present in the molecular structure, value »0« means that it is absent. All descriptors together, arranged in the so called substructure isomorphism matrix, can be easily imported from the result file into other software, for instance Excel or multivariate data analysis programs. Computing time on personal computers with Pentium IV and 2.6 GHz is typically one second for 1000 molecular structures and 200 substructures, that is 5 μs per descriptor value. SubMat has been written in Borland Delphi 6.0 and is running under operating systems Windows 95/98/ME/NT/2000/XP.

The substructure search algorithm[17] contains complete atom-atom and bond-bond matching based on the backtracking principle;[18] no 3D information or stereochemistry are considered. The current limits are 127 explicitly defined atoms and 255 bonds per structure with a maximum of 11 bonds at an atom. The maximum number of molecular structures and substructures is limited by the available memory; with 256 MB RAM *ca* 20,000 molecular structures and 1,000 substructures have been processed successfully within one run.

SubMat can be used in two different operating modes. One is the interactive mode with a typical Windows graphical user interface. The other is the remote mode by calling SubMat from another software. A command file in text format is used to transfer parameters to SubMat, such as file names, and parameters for the output format. So-called semaphore files are used for a simple and safe communication between the calling program and SubMat. Main information obtained from a semaphore file is its existence (the presence of a semaphore file means »on«). For instance, an end semaphore is an empty file created by SubMat immediately after the result file has been closed. Regularly checking the existence of the end semaphore file by the calling program allows to recognize the termination of the result file, and thereby a safe continuation with re-opening the result file and reading it. Furthermore, a progress semaphore is regularly produced by SubMat; it includes an integer number that gives the percent of already computed descriptors; in the case an error is detected by SubMat, the progress file contains an error code as a negative number. A stop semaphore can be used by the calling program; its existence is regularly checked by SubMat and SubMat is automatically terminated if this semaphore file is found. The filenames of the semaphores are defined by the user in the command file. Examples for calling SubMat with a command file »command.txt« are as follows (for simplicity no paths are used here).

From a DOS batch file: submat.exe command.txt

From a Basic program: shell »submat.exe command.txt«

From a C++ program: system(»submat.exe command.txt«)

From a Matlab program: dos('submat.exe command.txt &')

Finally, a special application of SubMat is mentioned. If the same chemical structures are used as molecular structures and as substructures, the result is a square matrix with binary descriptors. It has been shown that this matrix contains full information about the topological hierarchy of the used structures,[6] and a graph can be derived that shows all substructure relationships between the used structures.

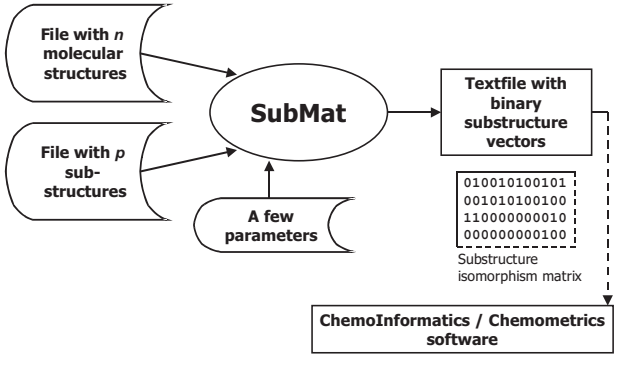

Figure 7. Software SubMat for the generation of binary substructure descriptors from a file with molecular structures and a file with substructures (both in Molfile format). The result file in text format can be easily imported into other software.

APPLICATIONS

Structure Similarity Searches

In the context of this work, chemical structures are represented by vectors of equal length, with the vector components being 0 or 1 and having identical meaning in all represented structures. The similarity of chemical structures can be measured by the similarity of such binary vectors. Among the many similarity criteria applicable to vectors, the Tanimoto index,[5] also called Jaccard similarity coefficient,[19] is successfully used for chemical structures. For two binary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ (each characterizing a chemical structure) the Tanimoto index t is given by

$$t = \sum \mathrm{AND}[x(i), y(i)] \,/\, \sum \mathrm{OR}[x(i), y(i)] \qquad (4)$$

with AND[$x(i)$, $y(i)$] being the result of the logical AND of $x(i)$ and $y(i)$, and OR[$x(i)$, $y(i)$] the result of the logical OR. AND[$x(i)$, $y(i)$] is 1 if $x(i) = y(i) = 1$, otherwise 0;

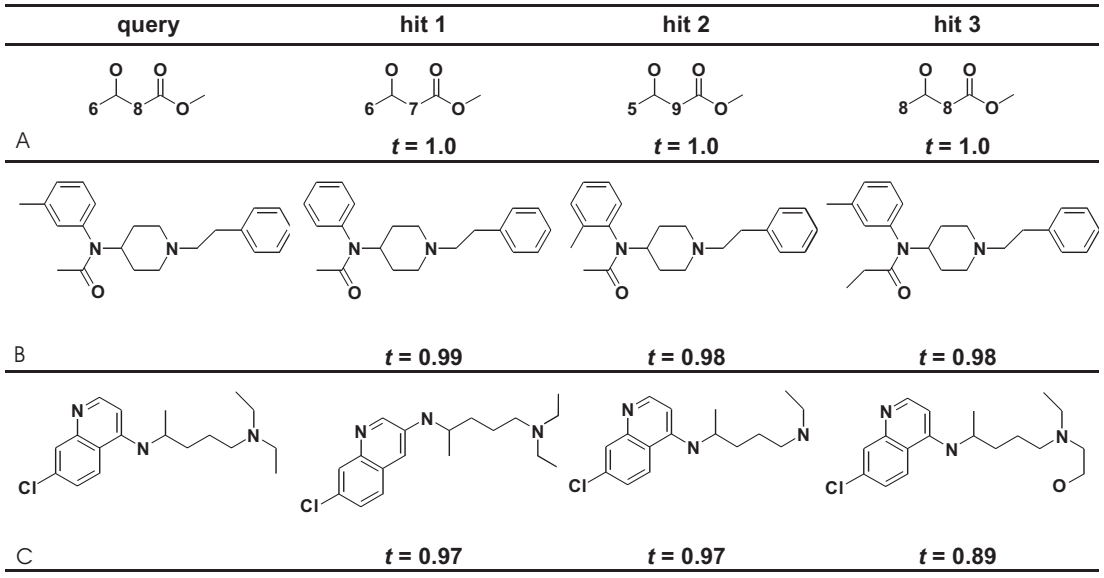

Figure 8. Examples for structure similarity searches. The query structures have been searched in a spectral database containing 106,955 compounds. *t*, Tanimoto index (structure similarity) between query and hit. Numbers within a structure denote a chain of C-atoms of the given length. Query structures are (A) 10-hydroxypalmitic acid methylester; (B) fentanyl; (C) resochine.

OR[*x(i), y(i)*] is 1 if *x(i)* = 1 or/and *y(i)* =1, otherwise 0. The Tanimoto index is a similarity measure in the range 0 to 1; it can be described as »the number of descriptors with a 1 in both structures (AND relationship), normalized by the number of descriptors with a 1 in at least one of the structures (OR relationship)«. The value of *t* is 1 if $\boldsymbol{x} = \boldsymbol{y}$, however, the corresponding structures are not necessarily identical.

For a structure similarity search all structures of a database have to be encoded by using the same substructures; the query structure has to be encoded in the same way. In a sequential search the Tanimoto indices between query structure and all reference structures are calculated and the reference structures with highest Tanimoto indices are collected in a so-called hitlist.

Figure 8 shows the first three hits for some query structures, using the 1365 substructures described before, and searching in a mass spectral database[13] containing 106,955 compounds. Reference structures which are identical to the query structure have been excluded from the search. Software used was written in Matlab 6.0; typical search time for one query is 40 s. Query A is 10--hydroxy palmitic acid methylester, CAS registry number 56247-30-4. For this query ten reference compounds have been found all with a Tanimoto index of 1.0; all are monohydroxy fatty acid methylesters without C-branches; the three shown compounds are arbitrary selected from these first ten hits. Query B is fentanyl, CAS registry number 437-38-7, a highly toxic compound that has been used in chemical weapons. The found hits have the same connected ring system as the query, although such a substructure has not been used. The highly specific combination of substructures for the query yielded this good result. Query C is resochine, CAS registry number 54-05-7, a drug against malaria. Also in this case the first hits show the same skeleton as the query, only differently substituted. In summary, the results demonstrate a satisfying overall structure similarity between the queries and the found hits.

Structural Diversity of Databases

The molecular structures of a database can be characterized by the frequencies of selected substructures occurring in the molecular structures. Figure 6 contains these frequencies for a set of 15 substructures, comparing an infrared spectral database,[12] containing 13,484 compounds, and a mass spectral database,[13] containing 106,955 compounds. For some substructures the frequencies are very different in the two databases. For instance the trimethylsilyl substructure occurs in 4.8 % of the compounds in the MS database but only in 0.4 % of the compounds in the IR database. This reflects the different origins of the databases and different relevance of compounds, in this case the prominent use of silylated compounds in GC-MS. Most substructures have similar frequencies in both databases as shown in Figure 9. A detailed investigation[7] showed that most of the 1365 substructures have frequencies below 5 %; only a few small substructures are present in more than 80 % of the database structures, demonstrating the high structural diversity of the databases.

The diversity of a structure database (in combinatorial chemistry often called a library) can for instance be characterized by the distribution of Tanimoto indices calculated for a random sample of structure pairs. The

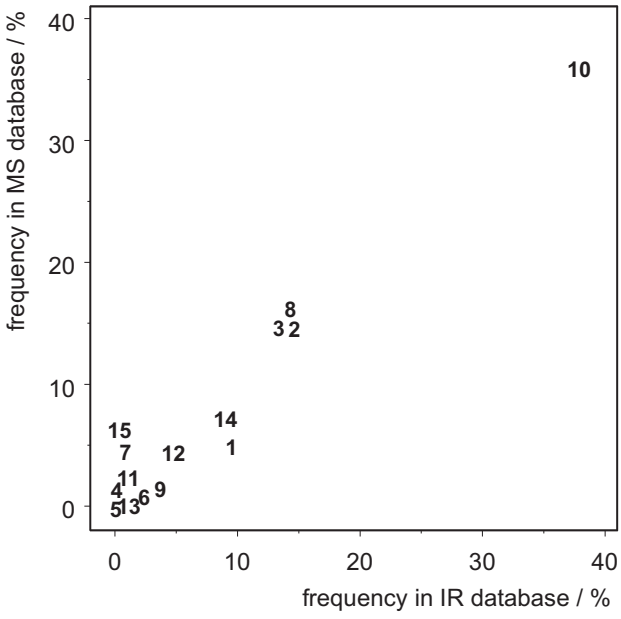

Figure 9. Frequencies of 15 selected substructures from group 8 (functional groups) in two spectroscopic databases. Numbers denote the substructures used in Figure 6.

distributions obtained for the two spectral databases – calculated from one million random pairs – are skewed bell shaped curves. Both distributions are similar but that for the MS database is shifted to lower Tanimoto indices, indicating a higher structural diversity of the MS database than of the IR database.[7]

Spectral Versus Structural Similarity

Spectral library search methods are very common in spectroscopy and routinely used to identify compounds and to obtain structure information if the unknown is not in the database. For the latter case the spectra similarity search must have a high interpretative power, that means the found hitlist structures should be very similar to the query structure. This structural similarity can been measured by the average Tanimoto index between query structure and for instance the first five hitlist structures. Using some hundred to some thousand test queries, several spectra similarity criteria for IR[8] and MS[9] have been tested and improved by this approach.

Cluster Analysis of Structures

From searches in chemistry databases, sets of compounds are obtained with typically 20 to 200 chemical structures. A cluster analysis of these structure may be helpful for further interpretations of the search result. The binary substructure descriptors can be used in multivariate data analyses, such as principal component analysis (PCA), Kohonen-mapping or hierarchical cluster analysis. A feature (descriptor) selection is essential because in small sets of structures many descriptors may be zero in all compounds. Applications have been demonstrated for instance for sets of isomers,[6] and for hitlist data from mass spectra similarity searches.[7,9]

CONCLUSIONS

A set of 1365 substructures has been defined for a representation of organic chemical structures by binary substructure descriptors. Software SubMat has been developed for an easy and flexible generation of such descriptors. Structure similarity searches using these substructures, and the Tanimoto index as similarity measure, yielded highly relevant hitlists from spectra/structure databases. A further application is the evaluation of spectra similarity search methods with the aim to yield hitlists which contain high structure information about the unknown. Binary substructure descriptors have been successfully applied for cluster analysis of structures by multivariate mapping methods.

Acknowledgments. – The authors thank A. Kerber and R. Laue for the isomer generator software Molgen. The work was supported by the Austrian Science Fund, project P14792-CHE.

REFERENCES

1. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
2. A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, M. Karelson, A. E. Visser, and R. D. Rogers, *J. Chem. Inf. Comput. Sci.* **42** (2002) 225–231.
3. H. Kubinyi, QSAR: *Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993.
4. M. Randić, *J. Chem. Inf. Comput. Sci.* **37** (1997) 672–687.
5. P. Willet, *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth, 1987.
6. K. Varmuza and H. Scsibrany, *J. Chem. Inf. Comput. Sci.* **40** (2000) 308–313.
7. H. Scsibrany, M. Karlovits, W. Demuth, F. Müller, and K. Varmuza, *Chemom. Intell. Lab. Syst*. **67** (2003) 95–108.
8. K. Varmuza, M. Karlovits, and W. Demuth, *Anal. Chim. Acta* **490** (2003) 313–324.
9. W. Demuth, M. Karlovits, and K. Varmuza, *Anal. Chim. Acta* **516** (2004) 75–85.
10. A. Kerber and R. Laue, *Software Molgen (Isomer Generator Software)*, Institute for Mathematics II, University of Bayreuth, http://www.mathe2.uni-bayreuth.de/molgen4/, Bayreuth, 2000.
11. MDL-Information-Systems-Inc., *Database Crossfire Beilstein*, MDL Information Systems Inc., http://www.beilstein.com/products/xfire/, Frankfurt am Main, 1997.
12. Chemical-Concepts, *Spectroscopic Database System SpecInfo*, Chemical Concepts, http://www.chemicalconcepts.com/, Weinheim, 1996.
13. NIST, *Mass Spectral Database* 98, National Institute of Standards and Technology, http://www.nist.gov/srd/nist1a.htm, Gaithersburg, MD, 1998.
14. H. Scsibrany and K. Varmuza, *Software SubMat (Generation of Binary Substructure Descriptors)*, Laboratory for Che-

mometrics, Institute of Chemical Engineering, Vienna University of Technology, http://www.lcm.tuwien.ac.at, Vienna, 2004.
15. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, *J. Chem. Inf. Comput. Sci.* **32** (1992) 244–255.
16. MDL-Information-Systems-Inc., *CT File Format*, MDL Information Systems Inc., http://www.mdli.com/downloads/public/ctfile/ctfile.jsp, San Leandro, CA, 2002.
17. H. Scsibrany and K. Varmuza, in: C. Jochum (Ed.), *Software Development in Chemistry*, Vol. 8, Gesellschaft Deutscher Chemiker, Frankfurt am Main, 1994, pp. 235–249.
18. G. A. Hopkinson, in: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner (Ed.), *The Encyclopedia of Computational Chemistry*, Vol. 4, Wiley, Chichester, 1998, pp. 2764–2771.
19. B. G. M. Vandeginste, D. L. Massart, L. C. M. Buydens, S. De Jong, and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier, Amsterdam, 1998.

SAŽETAK

Deskriptori binarnih podstruktura organskih molekula

Kurt Varmuza, Wilhelm Demuth, Manfred Karlovits i Heinz Scsibrany

Strukture organskih molekula prikazane su binarnim vektorima koji sadrže informacije o prisustnosti ili odsutnosti 1365 podstruktura. Razmotrene su ideje koje su dovele do izbora baš toga skupa podstruktura i navedeno nekoliko je primjera. Razvijen je program *SubMat* za brzo i fleksibilno računanje deskriptora binarnih podstruktura iz molekularnih struktura. Demonstrirana je uporaba opisanoga skupa podstruktura u prikazivanju struktura organskih molekula na primjerima pronalaženja strukturno sličnih molekula.