MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **78** (2017) 565-580

ISSN 0340 - 6253

Mathematical Function for the Identification of Molecular Fragments as Chemical Groups

Laszlo Tarko

Centre of Organic Chemistry – Romanian Academy, Romania, Bucharest, Sector 6, Spl. Independentei 202B, PO box 35-108, MC 060023

tarko_laszlo@yahoo.com

(Received February 22, 2017)

Abstract

A three variables mathematical function is proposed, useful for the identification of molecular fragments, as "chemical groups", in an analyzed molecule. The non-linear and discontinuous function is a proposed mathematical definition for the term "chemical group". The atoms are included or not in the same fragment according to the value of the atomic numbers, net charges and bond orders, calculated using the semi-empirical quantum-mechanics PM6 method. The proposed algorithm allows automatic virtual fragmentation of the analyzed molecules, organometallics, inorganics and ions and does not need any previously established list of fragments. The non-conjugated fragments coincide with the classical functional groups because of the parameterization of the proposed formula. The neighboring classical groups are included in the same fragment if their conjugation is strong enough or if they are linked by heteroatoms. The aggregate of the classical chemical groups is considered a new chemical group. Two molecules which include the same fragments, regardless of the number of fragments, are considered molecules from the same class, from the point of view of the chemical structure. The text presents the fragments identified in 85 species having a great diversity of chemical structures, including ions, organometallics, inorganics, peptides and keto/enol type equilibrium products. The paper presents a comparison with the resultsobtained using the fragmentation methods SLASH, SDFP (Standard Fragmentation Procedure) and FbSS (Fragment-based Shape Signatures).

1. Introduction

In the last more than 150 years the research of chemical properties of a great number of molecules having a huge diversity of chemical structures imposed the concept of "chemical group", also called "functional group".

The chemical groups are groups made of specific atoms connected by specific chemical bonds. Because of this group of atoms, the molecule participates in certain reactions. During reactions the structure (the number of atoms, the type of atoms and the type of chemical bonds) of the chemical groups changes, while the rest of the molecule remains unchanged.

The empirical knowledge of the chemical structure - chemical properties relationship became gradually more extensive and, consequently, the list of chemical groups has become increasingly large. In order to identify which are the chemical groups of a molecule, one has to rely on the molecular graph and to compare the groups of atoms (fragments) found in the analyzed molecule with the list of the known chemical groups [1]. Sometimes the fragments are defined using fingerprints [2]. There are dozens of fragmentation procedures, as well as lists of fragments [3-5, 7-22, 37-42]. These procedures and lists can be used as the starting point for retro-synthesis [43], for the calculation of some descriptors (calculable molecular characteristics), chemical structure retrieval systems [6, 44], QSPR equations [18, 38], intramolecular synergy [23] and chemical similarity [24, 25].

This paper proposes a three variables mathematical function, useful for the identification, in an analyzed molecule, of the molecular fragments as chemical groups. The proposed algorithm is a modified version of the quoted procedure [26].

2. Methods and formulas

The value of many descriptors strongly depends on the molecular geometry. Identification of the "correct" molecular geometry is called "geometry optimization". After virtually building the molecules, the geometry optimization was performed using the programs PCModel [27] and MOPAC [28], more specifically the included semi-empirical quantum-mechanics PM6 method [29] and MOPAC keywords *pulay charge=n gnorm=0.2 geo-ok mmok bonds*.

After "geometry optimization", one calculates, for all (a_1, a_2) pairs of atoms in the analyzed molecule, the value of the non-linear and discontinuous function F (named so from "fragments").

$$F = 1 - \mathrm{sgn}(a \cdot b \cdot c \cdot d) \qquad (1)$$

where

$$a = (Z_1 - 1)(Z_2 - 1)$$

$$b = (Z_1 - 6)(Z_2 - 6) + \mathrm{int}\left(\frac{k_1}{BO}\right)$$

$$c = (Z_1 - 6)(Z_2 - 6) + \mathrm{int}\left[\frac{k_2}{\max(s_1, s_2)}\right] + \mathrm{int}\left(\frac{k_3}{BO}\right)$$

$$d = \left[\mathrm{int}\left(\frac{6}{Z_1}\right) + \mathrm{int}\left(\frac{6}{Z_2}\right)\right](Z_1 + Z_2 - 10)$$

$Z_{1,2}$ are the atomic numbers of atoms a_1 and a_2 (tabulated values)
BO is the bond order of the $a_1 - a_2$ chemical bond (calculated using the PM6 method and read from the MOPAC output file)
$s_{1,2}$ are the net charges of the atoms a_1 and a_2 (calculated using the PM6 method and read from the MOPAC output file)
sgn – means "the sign of the function …"; here sgn( ) has value 0 or +1
int – means "integer part of real number …"; here int( ) is 0 or another natural number [30]
max – means "the greatest value of …"

The values of k_1 = 1.05, k_2 = 0.50, k_3 = 0.24 factors were empirically established, according to the results of the PM6 method, which is parameterized for 70 elements.

The value of the bond order *BO* cannot be measured experimentally, but it is calculated for any pair of atoms in the analyzed molecule. The value of *BO* is high only for "bonded" atoms. In addition, the smaller is the value of *BO* the more "ionic" is the chemical bond; the greater is the value of *BO* the more "multiple" is the chemical bond. The value of *BO* of C-C bonds in alkans is ~ 1, of C-C bonds in benzene is 1.4406 and of "triple" bonds is > 2.8. If $BO < k_3$ the atoms a_1 and a_2 are considered "non-bonded". The value of k_1 is considered a limit between "single" and "aromatic" bonds [31].

The term $(Z_1 + Z_2 - 10)$ in formula of *d* is necessary because the Boron – Boron bond links two heteroatoms, i.e. atoms different from hydrogen and carbon; for other heteroatoms *Z* > 6.

The value of the *F* function is 0 or 1. If *F* = 1 the computer program [32] which uses the formula (1) will include the atoms a_1 and a_2 in the same molecular fragment otherwise the atoms a_1 and a_2 will be included in two different fragments.

The accuracy of grouping the atoms into molecular fragments depends on the correctness of the geometry optimization, which greatly influences the calculated values of

the bond orders and net charges, especially in organometallic compounds. The precision of the PM6 method in calculating the value of net charges and bond orders is not the subject of this paper.

According to formula (1), there are four categories of atoms, included or not in the same fragment:

- H (hydrogen atoms)
- C (carbon atoms with net charge $\leq k_2$)
- C^* (carbon atoms with net charge $> k_2$)
- X (heteroatoms)

The form of functions *a, b, c* and *d* was chosen because the atoms a_1 and a_2 should be included in the same fragment only if they are "bonded" and only if they have a certain value of *Z* and net charge.

The product $a \cdot b \cdot c \cdot d$ has the logic function OR, i.e. the value of the product $a \cdot b \cdot c \cdot d$ is null if $a = 0$ OR $b = 0$ OR $c = 0$ OR $d = 0$.

Consequently, the atoms a_1 and a_2 will be included in the same fragment if:

- a_1 *or* a_2 is H, regardless of the value of the bond order *BO*, because $a = 0$
- a_1 *or* a_2 is C *and* $BO > k_1$, because $b = 0$
- a_1 *and* a_2 are C^* *or* X *and* $BO > k_3$, because $c = 0$ *or* $d = 0$, i.e. $c \cdot d = 0$

According to the proposed algorithm:

- in alkans, iso-alkans and cyclo-alkans the number of fragments CH_x ($x = 0, 1, 2$ or 3) is equal to the number of carbon atoms (C – C bonds and $BO < k_1$)
- the atoms in propadiene, ketene, vinyl groups or (hetero)aromatic cycles will be included in the same fragment (C – C or C – X bonds *and* $BO > k_1$)
- the groups OH_x ($x = 0$ or 1) and NH_x ($x = 0$, 1 or 2) will be included or not in the same fragment with the bonded aromatic cycle, depending on the value of the bond order in the C – X bond
- the atoms in ester, amide, iso-cyanate and carbamate groups will be included in the same fragment (C^* - X or C – X bonds *and* $BO > k_1$)
- the atoms in groups free of carbon, for instance B – B, O – O, S – S, Si – O, HN – NH, SO, SO_2, NO_2, PO_4, will be included in the same fragment (X – X bonds and $BO > k_3$)
- the metal atoms bonded to carbon atoms (C – X bonds) will be included in the same fragment only if $BO > k_3$

In brief, the identified molecular fragments and the classic chemical groups are similar, because of the form of formula (1) and of the value of factors k_1, k_2 and k_3. However,

the neighboring classic groups are included in the same fragment if the conjugation in C – X bond, measured by the value of the bond order, is strong enough or if the bond links two heteroatoms (X – X link).

Formula (1) is a first proposed mathematical definition for the term "chemical group". The formula is an attempt to explain, at atomic and sub-atomic level, *why* the atoms should be or should not be included in the same fragment.

3. Results and comments

For example, in Figure 1, Table 1 and Table 2 are presented the structure and the calculated values of the net charge and bond orders in Niclosamide. In Figure 1 the numbers are conventional indices of the heavy atoms (different from hydrogen) in the MOPAC output file.

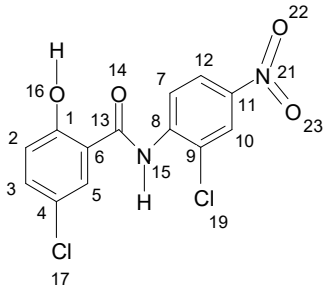


Figure 1 The structure and indices of atoms in niclosamide.

Table 1 Calculated net charges of the heavy atoms in Niclosamide

Atom's index	Net charge	Atom's index	Net charge	Atom's index	Net charge	Atom's index	Net charge
1	0.458	**7**	-0.292	**12**	0.032	**17**	-0.062
2	-0.317	**8**	0.296	**13**	0.660	**19**	-0.048
3	0.009	**9**	-0.161	**14**	-0.570	**21**	0.845
4	-0.101	**10**	0.012	**15**	-0.463	**22**	-0.465
5	0.008	**11**	-0.274	**16**	-0.453	**23**	-0.466
6	-0.440						

Table 2 Calculated bond orders in Niclosamide

Bonded heavy atoms (bond)	Bond order	Bonded heavy atoms (bond)	Bond order	Bonded heavy atoms (bond)	Bond order	Bonded heavy atoms (bond)	Bond order
1-2	1.250	**1-16**	1.164	**7-8**	1.362	**7-12**	1.471
2-3	1.573	**4-17**	1.005	**8-9**	1.308	**9-19**	1.004
3-4	1.276	**6-13**	1.005	**9-10**	1.457	**11-21**	0.904
4-5	1.543	**13-14**	1.638	**10-11**	1.371	**21-22**	1.484
5-6	1.285	**13-15**	1.063	**11-12**	1.363	**21-23**	1.482
1-6	1.315	**8-15**	1.057				

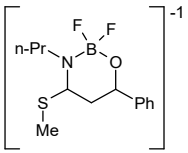
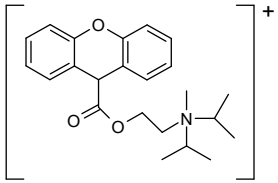
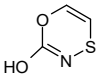
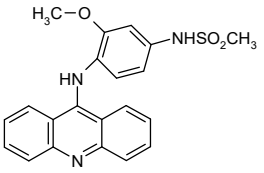
For exemplification, Table 3 includes many examples of fragments, identified in species (molecules, organometallics, inorganics, ions) with very diverse chemical structures. The structures in Table 3 are, in our opinion, a broad and good illustration of the proposed algorithm. The species can be identified in ChemIDplus Advanced database [33], according to the Registry Number RN or name.

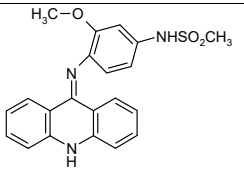
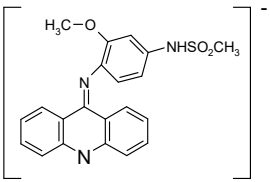
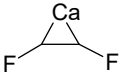
The anion **66** [34] was described as an "organic complex" [35]. The structures **70**, **71** and **79**, not yet synthesized, were imagined by author. In solution there is, probably, equilibrium between the three species **74 – 76** of the same acridine derivative.

Table 3 The fragments in the analyzed species

No.	Specie	Number	Fragment(s)
1	2,2,4-trimethyl-pentane	8	C, CH, CH_2 and CH_3
2	Methyl-cyclohexane	7	CH, CH_2 and CH_3
3	2-Methyl-1,3-butadiene	3	CH_3, HC=CH and $HC=CH_2$
4	Cyclohexene	5	CH_2 and HC=CH
5	1,3-dimethyl-benzene	3	CH_3 and C_6H_4
6	Styrene	2	$HC=CH_2$ and C_6H_5
7	Naphthalene	1	$C_{10}H_8$
8	Hexafluoroethane	8	C and F
9	Hexaiodoethane	7	C-C and I
10	(Trifluoromethyl)benzene	2	C_6H_5 and CF_3
11	Diethyl ether	5	CH_2, CH_3 and O
12	Diphenyl ether	3	C_6H_5 and O
13	1,1'-dioxodibenzene	3	C_6H_5 and O-O
14	1,1'-dithiodibenzene	3	C_6H_5 and S-S
15	Ethanol	3	CH_2, CH_3 and OH
16	Phenol	2	C_6H_5 and OH
17	2,4-Difluorophenol	3	F and C_6H_3-OH
18	Aniline	1	C_6H_5-NH_2
19	3-Nitroaniline	2	C_6H_5-NH_2 and NO_2
20	Acetic acid	2	CH_3 and COOH
21	Peracetic acid	2	CH_3 and COO-OH
22	Methyl acetate	3	CH_3 and COO
23	γ-Butirolactone RN 96-48-0	4	CH_2 and COO
24	Dicyanoacetylene	1	NC-C≡C-CN
25	Guanidine	1	H_2N-C(=NH)-NH_2
26	Pimagedine RN 79-17-4	1	H_2N-C(=NH)-NH-NH_2
27	Acetamide	2	CH_3 and $CONH_2$
28	(N-methyl)methylcarbamate	3	CH_3 and NHCOO
29	Alloxane RN 50-71-5	4	HN-CO-NH and CO
30	Maleic hydrazide RN 123-33-1	2	HC=CH and CO-NH-NH-CO
31	Maleimide	2	HC=CH and OC-NH-CO
32	N-Bromo-maleimide	2	HC=CH and OC-N(Br)-CO
33	Allyl isothiocyanate	3	CH_2, H_2C=CH and N=C=S

34	3,4-Dichlorophenyl isocyanate	4	C_6H_3, Cl and N=C=O
35	Benzoxazolone RN 59-49-4	2	C_6H_4 and NHCOO
36	3,5-Dimethyl isoxazole RN 300-87-8	3	CH_3 and C_3HNO
37	Thiotepa RN 52-24-4	7	CH_2 and N_3P=S
38	4-Toluenesulfonyl chloride RN 98-59-9	3	CH_3, C_6H_4 and SO_2Cl
39	Isoniazide RN 54-85-3	2	CO-NH-NH_2 and C_5H_4N
40	Dithizone RN 60-10-6	4	C_6H_5, N=N and NH-NH-C=S
41	Niclosamide RN 50-65-7	5	Cl, NO_2, C_6H_3-OH and C_6H_3-NHCO
42	Atropine RN 51-55-8	14	CH, CH_2, CH_3, C_6H_5, N, OH and COO
43	Diobutil RN 51-38-7	8	CH_2, CH_3, I, C_6H_2-OH and COO
44	Picadex RN 99-00-3	7	CH_2, NH, N-C=S and SH
45	Famphur RN 52-85-7	7	CH_3, C_6H_4, N-SO_2 and O_3P=S
46	Spironolactone RN 52-01-7	23	C, CH, CH_2, CH_3, C=CH, COO and S-CO
47	Verazide RN 93-47-0	7	CH_3, C_6H_3, O, C_5H_4N and CO-NH-N=CH
48	Cyclohexanone peroxide RN 78-18-2	13	C, CH_2, O-OH and O-O-C-OH
49	Mitomycine RN 50-07-7	14	C, CH, CH_2, CH_3, O, NH, CO, COO-NH_2, C=C-NH_2 and N-C=C-CO
50	Sulfallate RN 95-06-7	9	CH_2, CH_3, C=CH_2, Cl, S and N-C=S
51	Mefenamic acid RN 61-68-7	5	CH_3, C_6H_3, COOH and C_6H_4-NH
52	3,4,5-Trimethoxyanthranilic acid RN 61948-85-4	7	CH_3, O, COOH and $C_6H(O)(NH_2)$
53	Strychnine RN 57-24-9	17	C, CH, CH_2, C=CH, C_6H_4, N, O and N-CO
54	Bromcresol green RN 76-60-8	11	C, CH_3, C_6H_4, Br, C_6H-OH and SO_3
55	Bisphenol S	3	C_6H_4-OH and SO_2
56	Tetradifon RN 116-29-0	7	C_6H_2, C_6H_4, Cl and SO_2
57	D-Lysergic acid diethylamide RN 50-37-3	13	CH, CH_2, CH_3, C=CH, N, N-CO and C_8H_5N
58	Glucose RN 50-99-7	11	CH, CH_2, OH and CHO
59	Paracetamol RN 103-90-2	4	CH_3, C_6H_4, OH and NHCO
60	Pyramidon RN 58-15-1	7	CH_3, C_6H_5, N and C_3N_2O
61	Aspirin RN 50-78-2	4	CH_3, C_6H_4, COO and COOH
62	Novobiocin RN 303-81-1	22	C, CH, CH_2, CH_3, C=CH, C_6H_3, O, OH, NHCO, OCONH_2 and $C_6H_2(C_3O)(OH)(O)$
63	Tabun RN 77-81-6	6	CH_2, CH_3, CN and NPO_2
64	Aspartame RN 22839-47-0	10	CH, CH_2, CH_3, C_6H_5, NH_2, COO, COOH and NHCO
65	Cyclochlorotine RN 12663-46-6	23	CH, CH_2, CH_3, C_6H_5, Cl, OH, N-CO and NHCO

66		9	CH, CH_2, CH_3, C_6H_5, S and CH-N-BF_2-O
67		15	CH, CH_2, CH_3, C_6H_4, O, N and COO
68	2,3-dihydronaphthalene-1,4-dione	5	CH_2, C_6H_4 and CO
69	naphthalene-1,4-diol	3	$C_{10}H_6$ and OH
70		2	HC=CH and S-NH-COO
71		2	HC=CH and S-N=C(OH)O
72	Phenolphthalein (phenol-lactone) RN 77-09-8	6	C, C_6H_4, OH, COO and C_6H_4-OH
73	Phenolphthalein (phenol-acid-one) RN 5768-87-6	8	C=C, C_6H_4, OH, CO and COOH
74		4	CH_3, O-C_6H_3-$NHSO_2$ and NH-$C_{13}H_8N$

75		6	CH_3, C=N, C_6H_3-O, $(C_6H_4)_2NH$ and $NHSO_2$
76		4	CH_3, $(C_6H_4)_2N$-C=N-C_6H_3-O and $NHSO_2$
77	Tetraethyl lead	9	CH_2, CH_3 and Pb
78	Phenyl magnesium chloride	2	C_6H_5 and MgCl
79		5	CH, F and Ca
80	Auranofin RN 34031-32-8	22	CH, CH_2, CH_3, O, COO and S-Au=P
81	NCl_3	1	NCl_3
82	$AlCl_3$	1	$AlCl_3$
83	$CaCO_3$	1	$CaCO_3$
84	Calcium carbide	1	CaC_2
85	Cyclooctasulfur	1	S_8

For exemplification, Figures 2 - 6 present the fragments identified in molecules **29, 41**, **57**, **60** and **73**.

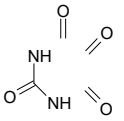


Figure 2 The four fragments in Alloxane (molecule **29**).

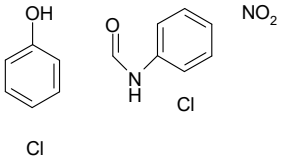


Figure 3 The five fragments in Niclosamide (molecule **41**).

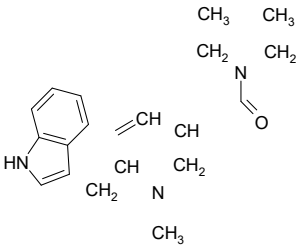


Figure 4 The thirteen fragments in D-Lysergic acid diethylamide (molecule **57**).

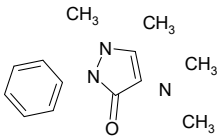


Figure 5 The seven fragments in Pyramidon (molecule **60**).

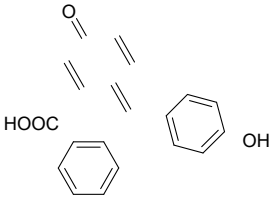


Figure 6 The eight fragments in Phenolphthalein (molecule **73**).

In molecules **1** – **8** the identified fragments and the classical chemical groups are similar.

In molecule **9** the C-C fragment is present because of the high calculated value of the bond order BO ~ 1.068. In molecule **8** the bond order of C-C link is much smaller, BO ~ 0.887.

In molecule **10** the bond order of C-F bonds is low, BO ~ 0.938. However, the CF_3 fragment is present because of the high value of the net charge of C atom in CF_3 group, s ~ 0.509 and thus $c = 0$ in formula (1). In molecules C_6H_5-CY_3 (Y = Cl, Br and I), not included in Table 3, the net charge of C atom in CY_3 group is much smaller and, consequently, there are 5 fragments C_6H_5, C and Y.

We note the low conjugation between, for instance, C_6H_5 and O fragments in diphenyl ether (BO ~ 0.98), between C_6H_3 and O fragments in molecule **47** (BO ~ 1.047 and 1.02), between C_6H_4 and OH fragment in molecule **59** (BO ~ 1.03) and between C_6H_5 and OH fragments in phenol (BO ~ 1.04); consequently, these groups are included in different fragments. On the contrary, in molecules **17**, **18**, **19**, **40, 43, 51, 52, 54, 55, 71** and **72**, the conjugation of OH/NH_2 and C_6H_x (x = 0, 1, 2, 3, 4 or 5) groups is higher (BO > k_1) and these groups are included in the same fragment.

There is a strong enough conjugation in molecule **24** and only one fragment is present. From the point of view of bond orders, in Dicyanoacetylene there are three "triple" bonds (BO ~ 2.90) and two "very weak aromatic" bonds (BO ~ 1.07). The chemical properties in pairs keto-enols / acids, keto-ethers / esters, keto-amines / amides, or aniline / benzylamine are quite different. It can be inferred that the chemical properties of dicyanoacetylene and 1,4-dicyano-2-butyne are quite different. The same situation is encountered in molecules **25** and **26**. There is only one fragment in any PAH (Polycyclic Aromatic Hydrocarbon).

In ion **66** the HC-N-BF_2-O fragment is present because of the high value of the bond order in HC – N link, BO ~ 1.193.

There is a wide conjugation area in anion **76**. Only groups CH_3 and $NHSO_2$ are outside this area.

In molecule **79** the value of the bond order for Ca-C link is low, BO ~ 0.907. On the contrary, in molecule **84** the value of the bond order for Ca-C link is much higher, BO ~ 1.164. Actually, calcium carbide **84** is calculated as a "molecule" including only covalent bonds, not a "salt" including ionic Ca-C bond and covalent C-C bond.

In many molecules in Table 3 the identified fragments are the effect of the presence of X-X "two heteroatoms bonds", regardless of the BO's value, for instance in molecules **13, 14, 63, 66**, **78, 81, 82, 83** and **85**.

The identified fragments are very different in pairs **68/69, 70/71, 72/73** and **74/75**.

The fragments identified according to formula (1) can be considered *chemical groups*. In addition, *two molecules which include the same fragments* (regardless of the number of fragments) *should be considered molecules in the same class from the viewpoint of the chemical structure*.

Figure 7 presents the fragmentation of the ammonium cation **67** by the SLASH algorithm [36], SDFP algorithm [24] and the proposed algorithm. One can observe the proposed algorithm identifies, as SLASH, the ester group, but does not aggregate the groups

CH, CH_2 and CH_3 into greater alkyl groups. SDFP does not identify the COO group as a "chemical group".

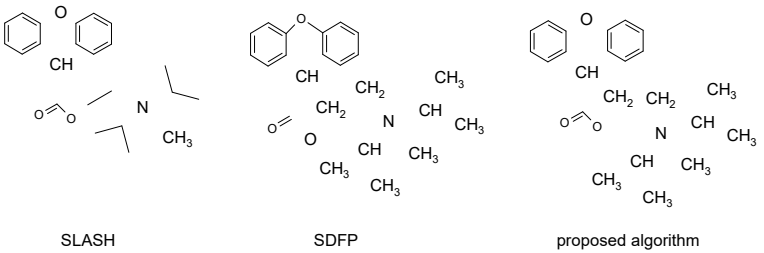


Figure 7 Comparison between SLASH, SDFP and the proposed algorithm.

According to the "Fragment-based Shape Signatures" method [22] (here called FbSS), the heavy atoms (and the attached hydrogen atoms) are included in four categories of fragments: (a) ring systems, regardless of the number of cycles, (b) fragments neighboring two or more ring systems, (c) fragments including more than five heavy atoms, neighboring a single ring system and (d) fragments including maximum five heavy atoms *and* neighboring a single ring system. Therefore, the fragments identified by the FbSS method differ from the classical chemical groups. Figure 8 presents the fragmentation of Novobiocin **62** by FbSS algorithm (only 5 fragments) and the proposed algorithm (22 fragments).

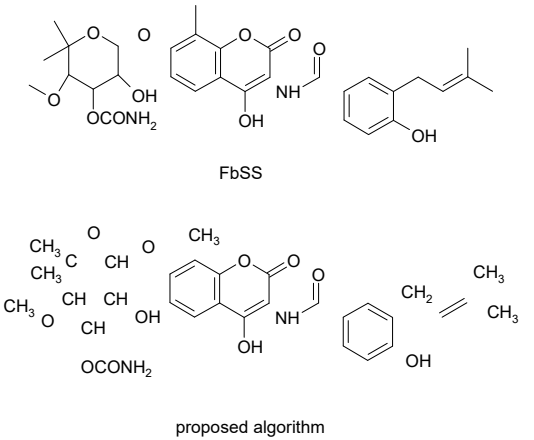


Figure 8 Comparison between FbSS and the proposed algorithm.

Should the carbon atoms in alkans be included in the same fragment?

Should the carbon and halogen X atoms be included in the same CX_3 group?

How many fragments are there in Dicyanoacetylene molecule?

Should the sulfur atoms in Cyclooctasulphur be included in the same S_8 group?

Should the three atoms in Calcium carbide be included in the same fragment?

Should the five heteroatoms in Thiotepa be included in the same fragment?

Is there only one fragment or two fragments in Phenol and Aniline?

The proposed formula is a unique criterion which can answer to above questions, regardless of the analyzed chemical structure and *without any previously established list of fragments.*

Any reader can verify the utility and effectiveness of the proposed formula including the algorithm in his own software, to verify the effect of this type of fragmentation on various calculated properties of the analyzed molecules.

4. Conclusions

The proposed algorithm uses a mathematical formula for the identification of molecular fragments as "chemical groups", allowing automatic virtual fragmentation of the analyzed molecules, organometallics, inorganics and ions.

The proposed algorithm does not need for identification of the fragments any previously established list of fragments.

The non-conjugated fragments coincide with the classical functional groups. The neighboring classical groups are included in the same fragment if their conjugation is strong enough or if they are bonded by heteroatoms. The aggregate of the classical chemical groups can be considered a new chemical group.

Two molecules which include the same fragments, regardless of the number of fragments, should be considered molecules in the same class, from the point of view of the chemical structure.

References

[1] K. Satoh, S. Azuma, H. Satoh, K. Funatsu, Development of a program for construction of a starting material library for AIPHOS, *J. Chem. Softw.* **4** (1997) 101–107.

[2] P. Japertas, R. Didziapetris, A. Petrauskas, Fragmental methods in the design of new compounds. applications of the advanced algorithm builder, *Quant. Struct. Act. Relat.* **21** (2002) 23–37.

[3] C. E. Berkoff, R. D. Cramer, G. Redl, Substructural analysis. Novel approach to the problem of drug design, *J. Med. Chem.* **17** (1974) 533–535.

[4] G. W. Adamson, J. A. Bush, Method for relating the structure and properties of chemical compounds, *Nature (London)* **248** (1974) 406–408.

[5] G. W. Adamson, J. A. Bush, Evaluation of an empirical structure–activity relationship for property prediction in a structurally diverse group of local anaesthetics, *J. Chem. Soc., Perkin I* (1976) 168–172.

[6] P. Willett, A review of chemical structure retrieval systems, *J. Chemom.* **1** (1987) 139–155.

[7] R. I. Geran, G. F. Hazard, L. Hodes, S. A. Richman, A statistical–heuristic method for automated selection of drugs for screening, *J. Med. Chem.* **20** (1977) 469–475.

[8] L. Hodes, Selection of molecular fragment features for structure–activity studies in antitumor screening, *J. Chem. Inf. Comput. Sci*. **21** (1981) 132–136.

[9] R. E. Carhart, D. H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure–activity studies: definition and applications, *J. Chem. Inf. Comput. Sci.* **25** (1985) 64–73.

[10] G. Klopman, O. T. Macina, Computed–automated structure evaluation of antileukemic 9-anilinoacridines, *Mol. Pharmacol.* **31** (1987) 457–476.

[11] A. Ormerod, P. Willett, D. Bawden, Comparison of fragment weighting schemes for substructural analysis, *Quant. Struct. Act. Relat.* **8** (1989) 115–129.

[12] G. Klopman, H. S. Rosenkranz, Toxicity estimation by chemical substructure analysis: the Tox II program, *Toxicol. Lett.* **79** (1995) 145–155.

[13] N. M. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, Stigmata: an algorithm to determine structural commonalities in diverse datasets, *J. Chem. Inf. Comput. Sci.* **36** (1996) 862–871.

[14] J. L. Gázquez, A. Cedillo, B. Gómez, A. Vela, Molecular fragments in density functional theory, *J. Phys. Chem. A* **110** (2006) 4535–4537.

[15] P. de Silva, M. Giebułtowski, J. Korchowiec, Fast orbital localization scheme in molecular fragments resolution, *Phys. Chem. Chem. Phys.* **14** (2012) 546–552.

[16] R. F. Nalewajski, Entropy/information bond indices of molecular fragments, *J. Math. Chem.* **38** (2005) 43–66.

[17] A. J. Bridgeman, C. J. Empson, Bond orders between molecular fragments, *Chem. Eur. J.* **12** (2006) 2252–2262.

[18] I. Mitra, A. Saha, K. Roy, Quantification of contributions of different molecular fragments for antioxidant activity of coumarin derivatives based on QSAR analyses, *Can. J. Chem.* **91** (2013) 428–441.

[19] M. A. Collins, M. W. Cvitkovic, R. P. A. Bettens, The combined fragmentation and systematic molecular fragmentation methods, *Acc. Chem. Res.* **47** (2014) 2776–2785.

[20] M. A. Collins, Can Systematic Molecular fragmentation be applied to direct ab initio molecular dynamics?, *J. Phys. Chem. A* **120** (2016) 9281–9291.

[21] H. A. Le, H. J. Tan, J. F. Ouyang R. P. A. Bettens, Combined fragmentation method: A simple method for fragmentation of large molecules, *J. Chem. Theory Comput.* **8** (2012) 469–478.

[22] R. J. Zauhar, E. Gianti, W. J. Welsh, Fragment–based shape signatures: a new tool for virtual screening and drug discovery, *J. Comput. Aid. Mol. Des.* **27** (2013) 1009–1036.

[23] L. Tarko, A statistical method for calculation of intramolecular synergy, *MATCH Commun. Math. Comput. Chem.***75** (2016) 533–558.

[24] L. Tarko, Virtual fragmentation of molecules and similarity evaluation, *Rev. Chim.* **55** (2004) 539–546.

[25] L. Tarko, A new manner to use application of Shannon entropy in similarity computation, *J. Math. Chem.* **49** (2011) 2330–2344.

[26] L. Tarko, A procedure for virtual fragmentation of molecules into functional groups, *Arkivoc* **xiv** (2004) 74–82.

[27] PCModel program is available from J. J. Gajewski, K. E. Gilbert, Serena Software, Box 3076, Bloomington, IN, USA.

[28] MOPAC program is available from J. J. P. Stewart, 15210 Paddington Circle, Colorado Springs, CO 80921; MrMOPAC@OpenMOPAC.net http://www.openmopac.net/, accessed in February 2017.

[29] J. J. P. Stewart, Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements, *J. Mol. Model.* **13** (2007) 1173–1213.

[30] according to ISO 80000-2-2009, page 6, the category "natural numbers" includes the number "0".

[31] L. Tarko, Aromatic molecular zones and fragments, *Arkivoc* **xi** (2008) 24–45.

[32] in-house PRECLAV program for QSPR calculations, documentation included, is available from Center of Organic Chemistry – Bucharest – Romanian Academy; ltarko@ccocdn.ro; tarko_laszlo@yahoo.com.

[33] http://www.chem.sis.nlm.nih.gov/chemidplus/, accessed in February 2017.

[34] T. Zhang, Y. M. Jia, C. Y. Yu, Z. T. Huang, Synthesis of BF_2 complex of 3-methylthio enaminones, *Arkivoc* **xiv** (2009) 156–170.

[35] J. D. Walker, M. C. Newman, M. Enache, *Fundamental QSARs for Metal Ions*, CRC Press, Boca Raton, 2013, pp. 101–101.

[36] D. A. Cosgrove, P. Willett, SLASH: a program for analysing the functional groups in molecules, *J. Mol. Graph. Model.* **16** (1998) 19–32.

[37] P. de Silva, M. Giebułtowski, J. Korchowiec, Fast orbital localization scheme in molecular fragments resolution, *Phys. Chem. Chem. Phys.* **14** (2012) 546–552.

[38] S. Saaidpour, Computational model for chromatographic relative retention time of polychlorinated biphenyls using sub–structural molecular fragments, *Comput. Meth. Sci. Techn.* **22** (2016) 41–53.

[39] D. Barbara, B. Guzowska–Swider, Fuzzy Definition of molecular fragments in chemical structures, *J. Chem. Inf. Comput. Sci.* **40** (2000) 325–329.

[40] M. Naderi, C. Alvin, Y. Ding, S. Mukhopadhyay, M. Brylinski, A graph–based approach to construct target-focused libraries for virtual screening, *J. Cheminf.* **8** (2016) 1–14.

[41] L. Chen, C. Chu, K. Feng, Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization, *Comb. Chem. High Throughput Screen.* **19** (2016) 136–143.

[42] S. Sharma, K. Karri, I. Thapa, D. Bastola, D. Ghersi, Identifying enriched drug fragments as possible candidates for metabolic engineering, *BMC Med. Gen.* **9** (2016) 167–177.

[43] K. Apoorva, S. Rounak, Retrosynthetic analysis – a review, *Int. J. Pharm. Techn.* **3** (2011) 1463–1476.

[44] S. E. Robertson, K. Sparck–Jones, Relevance weighting of search terms, *J. Am. Soc. Inf. Sci.* **27** (1976) 129–146.