A procedure for virtual fragmentation of molecules into functional groups

Laszlo Tarko

Center of Organic Chemistry "C.D.Nenitzescu" – Romanian Academy,
Spl. Independentei 202B, Sect. 6, Bucharest, Romania, PO Box 15-258, Fax 3121601,
E-mail: ltarko@cco.ro

(received 13 May 04; accepted 30 Aug 04; published on the web 02 Sept 04)

Abstract
This article presents a computer assisted procedure for virtual fragmentation of molecules. The proposed algorithm defines fragments which usually coincide with the classical functional groups. The fragmentation criteria are the bond order's value of the chemical bonds and the type of the connected atoms (hydrogen, carbon, heteroatom).

Keywords: Molecular fragment, functional group, bond order

Introduction

The study of the characteristic chemical properties of many molecules with very diverse chemical structures has led to the introduction of the functional group concept – a group of atoms causing the molecule to have certain "functions" (i.e. the capacity of participating in specific chemical reactions). During a specific reaction the structure of the functional group changes while the rest of the molecule remains unchanged. From the structural point of view the functional group is a group of atoms connected by certain types of chemical bonds.

In the last 150 years a large list of functional groups was created, i.e. the relationship between the structure of molecules and their properties was discovered.

In order to identify what are the functional groups of a molecule, one has to go through the molecular graph and to compare the groups of atoms found in the molecule with the list of functional groups mentioned above.[1] Other lists of fragments, which do not coincide with the classical functional groups, are often used. Sometimes the structure fragmentation is defined manually, either by dummy variables or fingerprints.[2] The procedures of fragmentation, as well as the lists of fragments, differ from author to author. Such lists are the starting point for the computation of various molecular descriptors, similarity functions, and for retro-synthesis procedures.[3-13]

In this article we propose an algorithm for the virtual fragmentation of molecules, an algorithm which does not need a previously established list of functional groups. The found fragments often coincide with the classical functional groups.

Methods and formulae

We use the following definitions:

- heavy atom = any atom other than hydrogen
- heteroatom = any heavy atom other than carbon
- B = the computed bond order value of a chemical bond
- k = the "border" value of B
- M = chemical bond type for which $B \geq k$
- S = chemical bond type for which $B < k$
- AS = heavy atom type connected to other heavy atoms only by S type bonds
- AM = heavy atom type connected to at least one heteroatom by M type bonds

When a certain molecule is analyzed the identification of the minimum potential energy conformer is important because the values of the bond orders *B* are characteristic to each conformer and, sometimes, they are very close to the border value *k*.

The geometry of the analyzed molecule was optimized by molecular mechanics using the GMMX procedure,[14] included in PCMODEL software.[15] Then, the geometry was optimized more exactly and the bond orders were computed with the PM3 method[16] included in the MOPAC package.[17] The following keywords string was used: "pm3 pulay gnorm=0.01 shift=50 geo-ok camp-king bonds mmok".

The MOPAC files data were then processed by a new version of the DESCRIPT software.[18] In this new version the previous fragmentation procedure is replaced by the fragmentation algorithm we propose here.

The fragmentation procedure has the following steps:

1) acquiring the molecular graph, the graph which contains only the heavy atoms
2) identification of the M and S bonds on the graph
3) identification of the AM and AS atoms on the graph
4) the definition of "internal" chemical bonds, i.e. the chemical bonds between atoms inside a single fragment:
   - *a)* the bonds involving hydrogen atoms
   - *b)* all M bonds
   - *c)* S bonds between AS carbon atoms
   - *d)* S bonds between AS heteroatoms
   - *e)* S bonds between an AS heteroatom and any AM atom

5) the definition of "external" chemical bonds, i.e. the chemical bonds between two fragments (any bond which is not "internal")

6) the removal of the "external" bonds from the graph

Removing the "external" bonds from the molecular graph we obtain a set of sub-graphs – the set of virtual fragments.

All computations were made on a Pentium 4 / 2400 MHz / 512 RAM.

Results and Discussion

We take the bond order of the chemical bond as the fundamental criterion of the molecular fragmentation procedure. In our opinion this is imposed by the fact that in unsaturated molecules the conjugation of neighbouring functional groups leads to the formation of new functional groups: ester (carbonyl + ether), amide (carbonyl + amino), phenyl (ene + ene + ene), furan (ene + ether + ene) etc.

The computed bond order of the chemical bond between two heavy atoms is *B*. The DESCRIPT (SDFP)[18] standard virtual fragmentation procedure uses the following axiom:

two heavy atoms are part of the same fragment if they are connected by an "internal" bond; a chemical bond is "internal" if its B value is greater then the limit value, k, between the "single" bond and the "aromatic" bond

In the standard version of DESCRIPT the SDFP skips the above steps *3*, *4c*, *4d* and *4e*, as it defines as "internal" bonds only the bonds involving hydrogen atoms and the M bonds.

The SDFP axiom is insufficient for obtaining a coincidence between the virtual fragments and the classical functional groups.

Indeed the molecular fragments identified by SDFP (Figure 1, C + D zone) coincide with the classical functional groups (Figure 1, A + B + C zone) only if all the bonds between the heavy atoms of the classical functional groups have a bond order greater than *k* value (Figure 1, C zone).

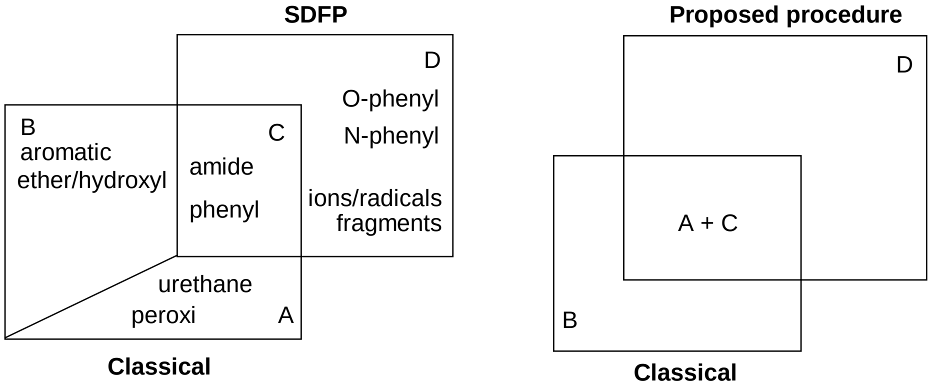


Figure 1. Classical functional groups and DESCRIPT virtual fragments.

SDFP identifies some fragments which are not considered classical functional groups: C_6H_5O (phenyl + ether), C_6H_5N (phenyl + substituted amino) etc. We note that none of the SDFP fragments Y – Z (Y=heteroatom, Z=Ar / ene, Y and Z are bonded by M bond) coincide with a classical functional group (Figure 1, D zone). In these fragments, denoted as YMZ, the conjugation modifies Y's and Z's chemical properties. Taking these YMS fragments as "functional groups" seems as justified as taking the amide or ester fragments as functional groups. However, these fragments are not classically considered single functional groups, but ensembles of functional groups (Figure 1, B zone).

There are certain classical functional groups which break the SDFP axiom: acid halide (carbonyl + halogen), lactone ester (carbonyl + ether), carbamate (amide + ether), peroxi (ether + ether), nitrate (ether + nitro) etc. According to SDFP these groups are ensembles of distinct fragments (Figure 1, A zone).

When, instead of the SDFP, we use the fragmentation procedure proposed here, the set of virtual fragments identified by DESPRIPT includes a greater number of the classical functional groups (figure 1, right).

The bond order can be computed by various methods (AM1, PM3, DFT, *ab initio*). When the bond orders are computed by PM3, as we have done, the $k = 1.017$ is used. This "border" value of B, which is used in aromaticity calculations by TPA (*Topological Path Aromaticity*) algorithm,[19] has been empirically established after we have analyzed the experimental aromaticity data for a great number of molecules with very diverse (aromatic and non-aromatic) structure. Also this k value is used by the last version of PRECLAV[20,21] program in QSPR/QSAR computations. SDFP use $k = 1.014$ value.

The above *a) – e)* rules were obtained empirically, by analyzing a large number of molecules with very diverse structures. Our aim was to achieve a correspondence as good as possible between the list of virtual fragments and the list of classical functional groups.

The proposed algorithm is exemplified in figure 2 – the computation of the bond orders and then the virtual fragmentation of strychnine.

The bond orders in the benzene cycle (the average of $B = 1.387$) were computed in the interval [1.284, 1.491]. For the Ar – N bond $B = 1.030$, for the N – CO bond $B = 1.058$, and for the C=O bond $B = 1.802$. These values show the existence of an M bond and of the (Ar + N + CO) fragment. In a different area of the molecule there exists another M bond with $B = 1.911$. The bond orders of the S bonds turned out much lower.

Strychnine's "weakest" S bond ($B = 0.928$) as well as its "strongest" S bond ($B = 0.985$) are both shown in figure 2.

In figure 2a the M bonds are shown in red and the S bond are shown in black. Figure 2b shows the AS and AM atoms. In figure 2c the "internal" bonds are shown in magenta and the "external" bonds in black. The resulting virtual fragments are shown in figure 2d.

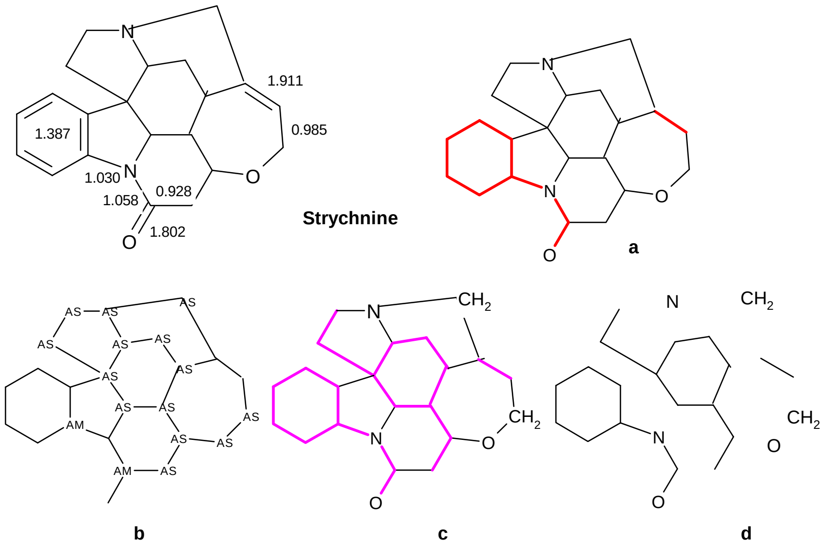


Figure 2. Fragmentation procedure applied on strychnine.

Additional examples are presented in Table 1. The molecules were chosen for their diversity of classical functional groups, M/S bonds, and AS/AM atoms. Thus, this set of molecules is, in our opinion, a broad illustration of the proposed algorithm.

Table 1. Identified fragments of some analyzed molecules

No.	Molecule	F*	Fragment(s)
1	*iso*-Butane	**1**	C_4H_{10}
2	Cyclohexane	**1**	C_6H_{12}
3	Butadiene	**1**	C_4H_6
4	Cyclohexene	**2**	HC=CH and $(CH_2)_4$
5	Ethylbenzene	**2**	C_2H_5 and C_6H_5
6	Azulene	**1**	$C_{10}H_8$
7	Fulvene	**1**	C_6H_6
8	Perchloroethylene	**5**	C=C and four Cl
9	Methyl phenyl ether	**2**	CH_3 and O-C_6H_5
10	Methyl phenyl thioether	**3**	CH_3, S and C_6H_5
11	Dimethyl ether	**3**	two CH_3 and O
12	Hexachlorobenzene	**1**	C_6Cl_6
13	Thiophene	**1**	C_4H_4S
14	Acetaldehyde	**2**	CH_3 and CH=O
15	Acrolein	**2**	CH_2 =CH and CH=O
16	Methyl ethyl ketone	**3**	CH_3CH_2, C=O and CH_3
17	Benzaldehyde	**2**	C_6H_5 and CH=O
18	Formyl iodide	**1**	HC(O)I
19	Acetyl chloride	**2**	CH_3 and C(O)Cl

Table 1. Continued

20	Phenol	**1**	C_6H_6O
21	Aniline	**1**	C_6H_7N
22	Nitromethane	**2**	CH_3 and NO_2
23	Glyceryl trinitrate	**4**	C_3H_5 and three NO_3
24	Formic acid	**1**	CH_2O_2
25	Oxalic acid	**2**	two COOH
26	Benzoic acid	**2**	C_6H_5 and COOH
27	Salicylic acid	**2**	C_6H_4OH and COOH
28	Ethyl acetate	**3**	C_2H_5, COO, and CH_3
29	Methoxyacetone	**5**	two CH_3, CO, CH_2 and O
30	Methylglycyl acetate	**5**	two CH_3, COO, C_2H_4 and O
31	Trifluoromethyl acetate	**6**	CH_3, COO, C and three F
32	δ-valerolactone	**2**	COO and $(CH_2)_4$
33	Phtalide	**3**	C_6H_4, CH_2 and COO
34	Dimethylformamide	**3**	two CH_3 and N-CH=O
35	Ethyl N-methylcarbamate	**3**	CH_3, NHCOO and C_2H_5
36	Phenyl N-methylcarbamate	**3**	CH_3, NHCOO and C_6H_5
37	γ-butyrolactam	**2**	$(CH_2)_3$ and NHCO
38	Succinic anhydride	**2**	$(CH_2)_2$ and O=C-O-C=O
39	Phtalic anhydride	**2**	C_6H_4 and O=C-O-C=O
40	Succinimide	**2**	$(CH_2)_2$ and O=C-NH-C=O
41	N-bromosuccinimide	**2**	$(CH_2)_2$ and O=C-N(Br)-C=O
42	Benzonitrile	**2**	C_6H_5 and CN
43	N-Hydroxypropionamide	**3**	C_2H_5, O=C-NH and OH
44	Di-*tert*-butyl peroxide	**3**	two C_4H_9 and O-O
45	Cyclohexane-1,2-dione	**3**	two CO and $(CH_2)_4$
46	2-Hydroxycyclohex-2-enone	**3**	$(CH_3)_3$, CO and CH=CH-OH
47	Dimethylsulfate	**3**	two CH_3 and SO_4
48	Dimethylsulfoxide	**3**	two CH_3 and SO
49	N,N-Dichlorobenzenesulfonamide	**2**	C_6H_5 and SO_2NCl_2
50	[1,4,3]-Oxathiazin-2-one	**2**	NHCO and S-CH=CH-O
51	THEIC (*tris*-hydroxyethyl-*iso*-cyanurate)	**7**	three OH, three $(CH_2)_2$ and $(NCO)_3$
52	Strychnine	**7**	$C_6H_4NC(O)$, $C_{10}H_{13}$, O, CH=CH, N and two CH_2
53	Saccharin	**3**	C_6H_4, SO_2 and NHCO
54	2-Amino ethanol	**3**	HO, $(CH_2)_2$ and NH_2

* The number of identified fragments in the analyzed molecule

In the molecules where all heavy atoms are AS carbon atoms the algorithm identifies a single fragment (e.g. **1** and **2**). The same happens when all the heavy atoms are connected by M bonds (e.g. **3**, **6**, **7, 13**, benzene, any PAH, pyridine, indole, urea, guanidine, thiourea etc.).

The CX_3 group (X = halogen) is identified as a collection of four fragments.

The atoms N and Br in molecule **41**, the atoms N and Cl in molecule **49** and the O atoms in molecule **44** are AS atoms. So these atoms are in the same fragment according to rule *d).*

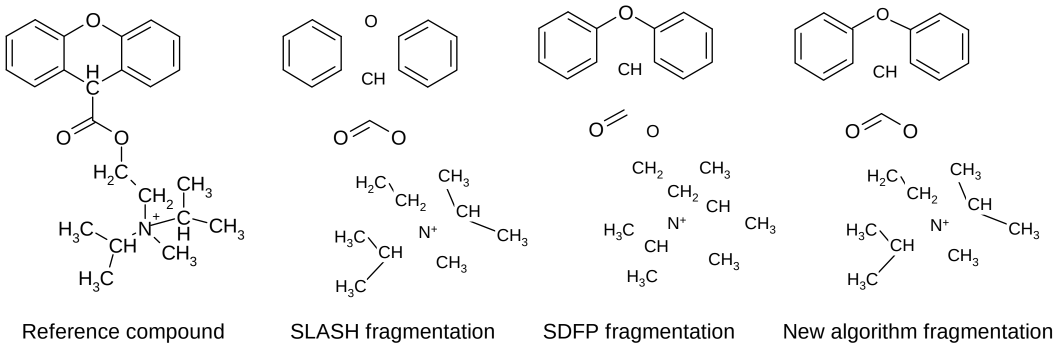
The COX group is identified as a single fragment (acid halide) according to rule *e).* The same rule also works in the case of phosgene, which is identified as a single fragment. The thiophosgene molecule is also identified as a single fragment but this time due to rule *b).* From the point of view of the proposed fragmentation procedure phosgene belongs to the same category as acetyl chloride and thiophosgene to the same category as thiourea.

In the molecules **9**, **12**, **20**, **21, 27** and **52**, the heteroatoms are connected to the aromatic cycle with an M bond. According to the proposed procedure, the conjugation determines the increase of the chemical bond order value and leads to the inclusion of the heteroatom in the same virtual fragment as the aromatic cycle to which it is connected. This does not happen in case of thioether **10**, urethane **36**, sulfonamide **49**, or saccharin **53**.

In the molecules **46** and **50**, the heteroatoms are connected to the non-saturated structures with an M bond. In the case of the exotic cycle **50**, not yet synthesized, the functional groups are difficult to identify by intuition.

The **45** and **46** molecules are in keto-enolic equilibrium.

In figure 3 we present the virtual fragmentation of an ammonium salt obtained with SLASH[22] algorithm, with SDFP, and with the procedure proposed here.



Reference compound SLASH fragmentation SDFP fragmentation New algorithm fragmentation

Figure 3. DESCRIPT / SLASH fragmentation.

In contrast to the SLASH algorithm, the DESCRIPT fragmentation identifies the Ph-O-Ph fragment (this is an YMS fragment).

There exists no obstacle of principle in applying this new algorithm to ions, radicals or ion-radicals. Due to the lack of experimental data – needed for the parameterization of the method – we used in the analysis of these species the same value for *k*. In Table 2 we present the virtual fragmentation of some ions, radicals and ion-radicals. For ions the charge $C \neq 0$, for radicals the multiplicity $M > 1$.

The Table 2 species can be formed by losing / capturing electrons, hydrogen ions or hydrogen atoms, in chemical reactions, inside mass spectrometers etc. Their electronic structure is very different from that of their provenance molecules. This is why the structure of virtual fragments is also very different. These fragments cannot be considered classical functional groups and in figure 1 they are placed in area D.

Table 2. Identified fragments in some ions, radicals and ion-radicals

Ion, radical, ion-radical	C^{*}	M^{**}	Identified fragment(s)			
			SDFP		Proposed procedure	
			F^{***}	fragment(s)	F^{***}	fragment(s)
$C_6H_5 – CH_2 – CH_2$	+1	1	2	C_6H_5 and $CH_2 – CH_2$	2	C_6H_5 and $CH_2 – CH_2$
$(C_6H_5)_3 – C$	0	2	1	unic fragment	1	unic fragment
$CH_3 – O – CH – O – CH_3$	+1	1	3	2 fragments CH_3 and $O – CH – O$	3	2 fragments CH_3 and $O – CH – O$
$CH_3 – N – C(O) – CH_3$	-1	1	2	$CH_3 – N – C(O)$ and CH_3	2	$CH_3 – N – C(O)$ and CH_3
$CH_3 – NH – C(O) – CH_3$	-1	2	4	CH_3 , NH, C(O) and CH_3	3	CH_3 , NHC(O) and CH_3
$CH_3 – O – CH_2 – O – CH_3$	+1	2	3	2 fragments CH_3 and $O – CH_2 – O$	3	2 fragments CH_3 and $O – CH_2 – O$
$C_6H_5 – OH_2$	+1	1	2	C_6H_5 and OH_2	2	C_6H_5 and OH_2
$C_6H_5 – C(CH_3) = OH$	+1	1	1	unic fragment	2	$C_6H_5C(OH)$ and CH_3

* The charge of the analyzed species.

** The multiplicity of the analyzed species.

*** The number of identified fragments.

The quality of the fragmentation method proposed here can be verified with various computation procedures which, taking this fragment identification as their starting point, compute values for logP, solubility, molar refraction etc. One can then check whether these computed values are in good agreement with the experimental data.

Conclusions

The procedure we proposed here does not need a previously established list of functional groups or fragments, allows automatic virtual fragmentation; once MOPAC has characterized the molecules, the users' assistance is no longer needed and can fragment any molecules, ions, radicals, and ion-radicals.

If the analyzed species is a molecule the identified fragments usually coincide with the classical functional groups. The conjugated classical functional groups should be always considered a single fragment – a new functional group.

References

1. Satoh, K.; Azuma, S.; Satoh, H.; Funatsu, K. *J. Chem. Software* **1997**, *4,* 101.
2. Japertas, P.; Didziapetris, R.; Petrauskas, A. *Quant.Struct.-Act.Relat.* **2002**, *21*, 23.
3. Berkoff, C. E.; Cramer, R. D.; Redl, G. *J. Med. Chem.* **1974**, *17*, 533.
4. Geran, R. I.; Hazard, G. F.; Hodes, L.; Richman, S. A. *J. Med. Chem.* **1977**, *20*, 469.
5. Hodes, L. *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 132.
6. Adamson, G.W.; Bush, J. A. *Nature (London)* **1974**, *248*, 406.
7. Adamson, G.W.; Bush, J. A. *J. Chem. Soc., Perkin I* **1976**, 168.
8. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
9. Ormerod, A.; Willett, P.; Bawden, D. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115.
10. Robertson, S. E.; Sparck-Jones, K. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129.
11. Klopman, G.; Rosenkranz, H. S. *Toxicol. Lett.* **1995**, *79*, 145.
12. Klopman, G.; Macina, O.T. *Mol. Pharmacol.* **1986**, *136*, 67.
13. Blankley, C. J.; Humblet, C.; Shemetulskis, N. E.; Weininger, D.; Stigmata, Y. J. J. *J. Chem.Inf. Comput. Sci.* **1996**, *36*, 862.
14. Saunders, M.; Houk, K.N.; Wu, Y.-D.; Still, W.C.; Lipton, M.; Chang, G.; Guida, W. *J. Am. Chem. Soc*. **1990**, *112*, 1419.
15. Gilbert, K.; Gajewski, J. J. *Serena Software, Box 3076, Bloomington, IN, USA.*
16. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
17. Stewart, J. J. P. *QCMP175, software MOPAC 7.0*
18. Tarko, L. *Rev. Chim.(Bucharest)* **2004**, *55,* 539.
19. Tarko, L. *Rev. Roum. Chim.* **2003**, *48,* 745.
20. Tarko, L. *Rev. Roum. Chim.* **2000**, *45,* 809.
21. Tarko, L.; Ivanciuc, O. *MATCH* **2001**, *44,* 201.
22. Cosgrove, D. A.; Willett, P. *J. Mol. Graph. Model.* **1998**, *16,* 19
23. DESCRIPT software is available from Center of Organic Chemistry (CCO) –Bucharest; e-mail: pfilip@cco.ro