TOOL FOR GROUP CONTRIBUTION METHODS – COMPUTATIONAL FRAGMENTATION

Zdeňka KOLSKÁ[a,*] and Pavel PETRUS[b,c]

[a] *Department of Chemistry, J. E. Purkinje University, České mládeže 8, 400 96 Ústí nad Labem, Czech Republic; e-mail: zdenka.kolska@ujep.cz*

[b] *Department of Physics, J. E. Purkinje University, České mládeže 8, 400 96 Ústí nad Labem, Czech Republic*

[c] *Institute of Chemical Process Fundamentals, Academy of Sciences of the Czech Republic, v.v.i., Rozvojová 2, 165 02 Prague 6, Czech Republic; e-mail: ppetrus@physics.ujep.cz*

Received October 27, 2009
Accepted January 12, 2010
Published online April 15, 2010

Dedicated to Professor Ivo Nezbeda on the occasion of his 65th birthday.

Group contribution methods are presently one of the universal and the most frequently used approach to estimate many physico-chemical properties of compounds. One of the important steps in development of group contribution method is a correct division of chemical structures of compounds into defined structural fragments. Computational program dividing automatically chemical structures of compounds (hydrocarbons and halogenated hydrocarbons) into structural fragments are now presented. For description of chemical structures of compounds and structural fragments we used SMILES format. New database of fragments and new record of fragments were created.

Keywords: Estimation methods; Group contribution method; SMILES format; Fragmentation.

To obtain values of physico-chemical properties of compounds we can apply experimental or non-experimental techniques. The former ones require appropriate equipment, necessary amount of measured compounds of sufficient purity and time to carry out experiment. The further are applied when experiment is not possible. Non-experimental approaches can be divided into calculation methods or estimation ones[1–3].

All estimation methods can be divided into two groups depending on the required input data[1,3]. QPPR methods (Quantity–Property–Property–Relationship) are input data-intensive. They require values of some known physico-chemical properties in order to calculate values of estimated prop-

doi:10.1135/cccc2009533

erty. The second groups involve the QSPR [1–3] methods (Quantity–Structure–Property–Relationship). These methods need only knowledge of the chemical structure of a compound to predict the estimated property. QSPR methods use some structural characteristics, such as number of fragments (atoms, bonds or group of atoms in a molecule), topological indices or other structural information, molecular descriptors, to express the relation between the property and molecular structure of compound[3]. Group contribution methods belong to these approaches. Due to these methods require only knowledge of chemical structure, they are one of the universal and one of the most frequently used.

Group contribution methods are based on the so called additive principle. The compound is divided into fragments, usually atoms, bonds or group of atoms. A fragment has a partial value called a contribution. A property of a compound is obtained by summing up the values of the contributions in the molecule. The simplest methods stem from the relationship between estimated property and the number of carbon atoms n_C or the number of methylene groups n_{CH2} in the molecule. More sophisticated methods are based on more complex fragments. Zero-order methods are based on the additivity of atomic fragments independent on mutual bonds. For cyclic and aromatic compounds, the ring fragment is considered to be an indivisible unit. First-order methods use bonds between neighbouring atoms as structural fragments. Again, cyclic fragments are considered to be unique units. Second-order methods are based on the additivity of groups. A group is defined as a polyvalent central atom surrounded by all its ligands[3]. Most group contribution methods permit an estimation at a single temperature (mostly at 293 or 298 K), some of them at several discrete temperatures or as a function of temperature. They can be developed for limited number of compound families, e.g. only for hydrocarbons, etc., or they can be universal, applicable to a wide variety of, mostly organic, compounds. They are mostly developed for pure compounds but they can be also used for mixtures. Group contribution methods are mostly proposed for estimation of only individual property, some of them are designed to estimate more properties[3].

To develop the reliable and accurate group contribution method it is necessary to realize three important steps: (i) to prepare correct and reliable input database, rather of critically assessed experimental data, from which parameters – group contributions, are calculated; (ii) to propose suitable structural fragments to describe all of compounds of input database; (iii) to provide the correct division of all chemical structures into defined structural fragments.

Some group contribution methods for estimation of several physical or physico-chemical properties of pure compounds we presented previously[4–6]. One of them was developed for a small number of compounds with a limited number of simple structural fragments[6]. The others were proposed for databases of several hundreds of compounds[4,5] and for group contribution values calculation we used databases of several hundreds[4] or thousands[5] of critically assessed experimental input data. As a base for these approaches were taken the group contribution methods by Marrero and Gani[7], firstly presented by Constantinou and Gani[8]. We modified this approach to estimate enthalpy of vaporization and entropy of vaporization at 298.15 K and at the normal boiling temperature T_b (ref.[4]), and further to estimate liquid heat capacity as a function of temperature[5]. This method involves a three-level calculation procedure, covering structural fragments of the first, second and third levels. In the first or primary level, contributions from simple groups are employed. This allows estimation of a wide variety of organic compounds. These fragments are insufficient to capture the proximity effect and the differences between isomers and are able only to estimate, correctly, values for simple and mono-functional compounds. Due to this Marrero and Gani[7], and before them, Constantinou and Gani[8] included the second[7,8] and third level[7] more complex groups, involving poly-functional and structural groups that provide more information about the molecular structure of the more complex compounds.

When database of chemical structures and structural fragments inclusive several members, it can be used a manual division of structures into fragments. When database contain hundreds of compounds and structural fragments are more complex (e.g. in our methods[4,5]), it is necessary an automatic fragmentation by computer program. We used ProPred program[9] for this division up to now. But this program does not include any families of structural fragments necessary for our purposes. Due to this it is necessary to develop new program for this division – fragmentation. To input chemical structures of compounds and structural fragments into program we used the so called SMILES format[10–13]. Many papers describing usage of SMILES formats and fragmentation for a wide variety of applications were presented previously (e.g., refs[4,5,7,8,14–26]). Authors applied this approach as a tool for databases of chemical compounds, libraries or information systems or as a tool for group contribution methods[14–20], for prediction of physico-chemical properties of compounds[4,5,7,8,14,19–24] or biophysical and toxicological properties[23,25,26]. They were developed either for one property estimation[21,22,26] or for more ones[4,5,7,8,24]. They can be used for a limited number of compounds[21,22] or for a wide spectrum of substances[4,5,7,8,14,23,24].

Almost all of them developed program for their own purposes without any availability to other users. Especially when presented fragments are complex, the usage by subsequent users can be difficult. We would like to develop a useful tool for users of group contribution methods for estimation of physico-chemical properties of pure compounds. Now we present a partial contribution to this – computational fragmentation of hydrocarbons and halogenated hydrocarbons into groups defined by Benson[27] and Marrero and Gani[4,5,7,8]. Only first-order groups are involved up to now.

COMPUTATIONAL PROTOCOL

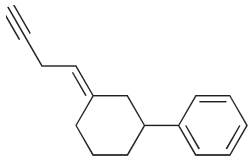
The main goal of our program is to provide a powerful tool for authors using group contribution methods for automatic fragmentation of chemical structures. The program is going to be independent on operation system or other programs. Also a possibility to affect databases of fragments of the program and to extend a family of fragments is great advantage of our approach.

Program requires four files for input and produces one file as an output of the program. First of the input file contains chemical structure of molecule written in SMILES format – in SMILES text string. Terms for writing of SMILES text strings are described below. Other three files contain databases of structural fragments (groups), from which the molecule can be created. First of database files is based on the method by Marrero and Gani[4,5,7,8] (further MG, 1), second one is based on the method by Benson[27] (2). The last input file is a new database of fragments created by ourselves (KP, 3) to complete missing groups. The output file contains (Figs 1 and 2): input SMILES format as a text string on the first line, chemical structure of molecule inclusive hydrogen atoms on the second line, the summary molecular formula and number of atoms on the next lines; information on applied database of fragments (1 or 2) follow and the list of fragments are presented on the next lines. Program user can choose between databases (1 or 2). In the case, that SMILES format of some molecule contains fragment, which is not included in chosen database (1 or 2), the third database (3) is applied. The output file then contains lines with fragments from our database. In Figs 1 and 2 are presented samples of two chemical structures with possible fragments of hydrocarbon (Fig. 1) and halogenated hydrocarbon (Fig. 2) with its fragmentation and output files. If any mistake occurs in SMILES input an error message is displayed on the screen.

We decided to create new own database because there are some first-order fragments, that are not described by both of methods and because the data-

bases of MG and Benson are different in distinction of fragments. While MG's database can describe fragments of the aliphatic, cyclic and aromatic chain, the Benson's database distinguishes only the aromatic and other fragments. Other difference is in differentiation of fragments for non-aliphatic groups: while the method by Benson considers the ring fragment (cyclic or aromatic) to be an indivisible unit, method by MG can divide all cyclic and aromatic rings into individually defined fragments.

We also introduce a new format to entry and to read structural fragments to unify two different formats for group contributions of MG and Benson and to cooperate with our program. To explain this format of fragments, we used a simple example. Group contribution CH_3- is input in program in the way of: CH3 F F F F F. This record consists of two parts; CH3 is the first part and F F F F F is the second part of the record. The first part contains information about form of fragment – a kind of a central atom (carbon atom C in this case) and all its surrounding – other atoms bonded to this central atom (hydrogen atoms H in our case) and their number in this fragment (3 in this case). After this number record can be followed by information about double or triple bond(s). There is one carbon atom C of normal

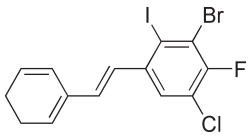


```
c1c(C2CC(=CCC#C)CCC2)cccc1
CH1CHOCH1CH2CHOCH1CH2CHOCH1CH2CH2CH2CH1CH1CH1CH1
Sumarni vzorec:C16H18
C           16
H           18
Zvolena databaze:data_1.txt
           1 ||CH2   F F F F F
           5 ||CH1=  F T F F T
           1 ||CHO=  F T F T F
           4 ||CH2   T F F F F
           1 ||CH1   T F F F F
Chybejici prispevky doplnene z KP databaze
           1 ||CH1=  F F F F F
           1 ||CH1#  F F F F F
           1 ||CHO#  F F F F F
           1 ||CHO=  T F F F F
```

FIG. 1

Example of the output file for one molecule of hydrocarbon and its division by our program. This sample molecule contains all forms of the first-order fragments for hydrocarbons – aliphatic chain, cyclic and aromatic rings. The molecular structure is followed by the output file of program for this sample molecule. The format of output file is written in the text above

valence with three hydrogen atoms H bonded to this carbon atom in our example. Due to this we have one free bond, which is single bond to another atom. We selected only the first-order groups for hydrocarbons in this first attempt. The second part of record gives five answers to five simple questions in this order: 1. Is the central atom part of a cyclic ring? 2. Is the central atom part of an aromatic ring? 3. Is a free bond of the central atom connected to any atom of aliphatic chain? 4. Is a free bond of the central atom connected to any atom of cyclic ring? 5. Is a free bond of the central atom connected to any atom of aromatic ring? Last 3 questions are related to only groups of aromatic rings of hydrocarbons. Possible answers to these questions are only yes (T) or no (F). When answers to the first two questions are "no" (F), that means the central atom of the fragment must be in aliphatic chain. When answers on the last three questions are "no" (F), that means there is no distinction among bonding atoms as it was in our example. The last three letters are related only to fragments of aromatic rings, when we apply the database by MG, because in any other case there is not so detailed differentiation for kind of bonding atom.




```
C1c1cc(C=CC2=CCCC=C2)c(I)c(Br)c1F
C1HOCHOCH1CHOCH1CH1CHOCH1CH2CH2CH1CH1CHOIHOCHOBrHOCHOFHO
Sumarni vzorec: C14H10Br1Cl1F1I1
C            14
F             1
Cl            1
Br            1
I             1
H            10
Zvolena databaze:data_1.txt
            1 ||CH1=   F T F F T
            5 ||CHO=   F T T F F
            2 ||CH2    T F F F F
            1 ||BrHO   F F F F F
            1 ||FHO    F F F F F
            1 ||ClHO   F F F F F
            1 ||IHO    F F F F F
Chybejici prispevky doplnene z KP databaze
            2 ||CH1=   F F F F F
            3 ||CH1=   T F F F F
            1 ||CHO=   T F F F F
```

FIG. 2
Example of the output file for one molecule of halogenated hydrocarbon and its division by our program. The molecular structure is followed by the output file of program for this sample molecule. The format of output file is written in the text above

Program consists of the following steps, the schematic presentation of individual steps is given in Fig. 3.

The first step is a reading of SMILES text string from the input file. We use the following basic rules for SMILES format record creation[10–13]. Atoms of elements are represented by their atomic symbols. Non-hydrogen atoms are specified independently by their atomic symbol enclosed in square brackets. Elements of the "organic subset" B, C, N, O, P, S, F, Cl, Br, and I are written without brackets if the number of attached hydrogen atoms conforms to the lowest normal valence consistent with explicit bonds. Atoms in aromatic rings are specified by lower case letters. Double and triple bonds are represented by the symbols = and #, respectively, single and aromatic bonds are not signed. Branches are specified by enclosing them in parentheses. Cyclic structures are represented by breaking one bond in each ring and these bonds are numbered in any order by a digit immediately following the atomic symbol at each ring. Numbers of "opened" cycles in one moment of writing or reading SMILES cannot be higher than 9. Program cannot recognize SMILES of aromatic rings written in Kekule's form. When SMILES is written by this, it is evaluated as a normal unsaturated cycle. Hydrogen atoms are not written when there is a normal valence of atoms and normal atom isotopes.

In the second step, the text string of SMILES is transformed into a vector of integer numbers (backbone vector). Atoms are assigned by positive num-

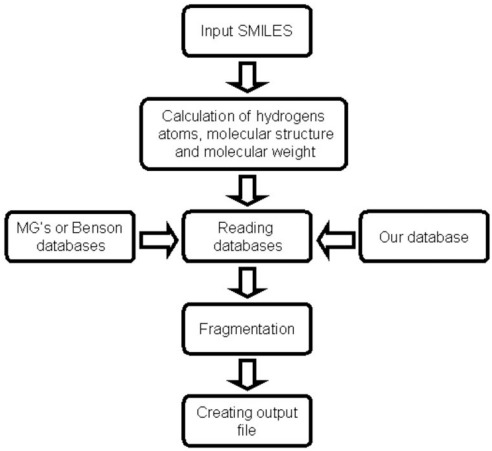


FIG. 3
Schematic presentation of operation of our program

ber, which corresponds to proton number of the atoms and other symbols of SMILES (e.g. =, #) are classified by a negative number. The length of this backbone vector is equal to number of symbols at SMILES. Exception cases are represented by atoms with double-letter symbols. The backbone vector then must be reduced about 1 for any double-letter symbols atoms. At this step all of atoms are assigned according to their binding capacity and atoms are identified as a part of: aliphatic string, cyclic or aromatic ring.

In the third step of approach, a number of bonded hydrogen atoms for each atom of SMILES is calculated. The process is realized using the so call auxiliary vector which is created for any atom of SMILES. This vector consist of 11 symbols (inclusive a part of the backbone vector) and relevant atom is in a center of this and we test the left and the right sides of this vector. After calculation of bonded atoms on the relevant atom on the left and on the right sides we determine the number of free bonds up to binding capacity and these free bonds are filled with hydrogen atoms. This approach is carried out with respect to the valid rules for creation of SMILES format described above.

In the next step, databases of group contributions (structural fragments) are read line by line and important information about the fragments are obtained (if the fragment is from aliphatic, cyclic or aromatic strings, if it is bonded to the aliphatic, cyclic or aromatic atom, and number of hydrogen atoms for relevant fragment). The next step is the fragmentation of the backbone vector through the auxiliary vector creation again. The relevant atom is placed again in the centre of this vector and around him the atoms and other symbols from the backbone vector are placed. After this, the auxiliary vectors of fragments and of SMILES are compared. When the auxiliary vector of SMILES is equal to auxiliary vector of any fragment, this fragment is counted.

In the next step, all information are printed into the output file, which is described above. If any mistake occurs through this approach (a lot of controls verify these), an error message displays on the screen.

We rewrite some first-order group contributions of the methods by MG and Benson for hydrocarbons and halogenated hydrocarbons to our new format. They are presented in Table I (MG) and Table II (Benson). There is the kind of fragment in the first column, following by the 5 answers to 5 questions discussed above in the second column and the original definition of original group contribution in the third column. In Table III, we present our fragments database (KP) for hydrocarbons and halogenated hydrocarbons.

TABLE I
List of fragments of the first-order group contribution fragments for hydrocarbons and halogenated hydrocarbons from database by Marrero and Gani. New record of fragment is written in the first two columns

Form of fragment	5 Answers to 5 questions	Original record of group contribution
CH3	F F F F F	CH3
CH2	F F F F F	CH2
CH1	F F F F F	CH
CH0	F F F F F	C
CH1=	F T F F T	aCH
CH0=	F T F F T	aC fused with aromatic ring
CH0=	F T F T F	aC fused with nonaromatic subring
CH0=	F T T F F	aC except as above
CH2	T F F F F	CH2 (cyclic)
CH1	T F F F F	CH (cyclic)
CH0	T F F F F	C (cyclic)
BrH0	F F F F F	-Br
FH0	F F F F F	-F
ClH0	F F F F F	-Cl
IH0	F F F F F	-I

TABLE II
List of fragments of the first-order group contribution fragments for hydrocarbons and halogenated hydrocarbons from database by Benson. New record of fragment is written in the first two columns

Form of fragment	5 Answers to 5 questions	Original record of group contribution
CH2=	F F F F F	Cd-(H)2
CH1#	F F F F F	Ct-(H)
CH0==	F F F F F	Ca
CH1=	F T F F F	CB-(H)

TABLE III
List of fragments of new database by Kolska and Petrus writen in our new format for record of fragments

Form of fragment	5 Answers to 5 questions
CH3	F F F F F
CH2	F F F F F
CH1	F F F F F
CH0	F F F F F
CH2=	F F F F F
CH1=	F F F F F
CH0=	F F F F F
CH0==	F F F F F
CH1#	F F F F F
CH0#	F F F F F
CH2	T F F F F
CH1=	T F F F F
CH0#	T F F F F
CH0=	T F F F F
CH0	T F F F F
CH1	T F F F F
CH0==	T F F F F
CH1=	F T F F T
CH0=	F T F F T
CH0=	F T F T F
CH0=	F T T F T
BrH0	F F F F F
FH0	F F F F F
ClH0	F F F F F
IH0	F F F F F

RESULTS AND DISCUSSION

New computer program for automatic division of chemical structures of hydrocarbons and halogenated hydrocarbons written in SMILES formats into structural fragments was developed. New database of fragments and new record of fragments for hydrocarbons and halogenated hydrocarbons were created. For the present, only hydrocarbons, aliphatic, cyclic, saturated, unsaturated, aromatic ones and halogenated hydrocarbons are possible for this program. We also used only first-order group contributions. Further it will be extended for other families of organic compounds inclusive of new families of industrially important substances, as ionic liquids, organometallic compounds, etc. We will use all groups defined by Marrero and Gani in the first-, second- and third-level estimation and by Benson. This program will also be supplemented about databases of group contributions calculated previously for some physico-chemical properties. Later it will be extended to calculate group contributions of new properties. This program will help to other users for quick estimation of many properties of pure compounds.

This work was supported by the Grant Agency of the Academy of Sciences of the Czech Republic (No. IAA 400720710).

REFERENCES

1. Baum E. J.: *Chemical Property Estimation: Theory and Practice*. Lewis Publisher, Boca Raton 1998.
2. Poling B. E., Prausnitz J. M., O'Connell J. P.: *The Properties of Gases and Liquids*, 5th ed. McGraw–Hill, New York 2001.
3. Zábranský M., Kolská Z., Růžička V., Malijevský A.: *Heat Capacities: Liquids, Solutions and Vapours*, Chap. 19. The Royal Society of Chemistry, London, in press.
4. Kolská Z., Růžička V., Gani R.: *Ind. Eng. Chem. Res.* **2005**, *44*, 8436.
5. Kolská Z., Kukal J., Zábranský M., Růžička V.: *Ind. Eng. Chem. Res.* **2008**, *47*, 2075.
6. Randová A., Bartovská L., Hovorka Š., Poloncarzová M., Kolská Z., Izák P.: *J. Appl. Polym. Sci.* **2009**, *111*, 1745.
7. Marrero J., Gani R.: *Fluid Phase Equilib.* **2001**, *183–184*, 183.
8. Constantinou L., Gani R.: *AIChE J.* **1994**, *40*, 1697.
9. *Program ProPred*, Version 3.5. Department of Chemical Engineering, DTU Denmark. Presented: May 2002.
10. Weininger D.: *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
11. Weininger D., Weininger A., Weininger J.: *J. Chem.. Inf. Comput. Sci.* **1989**, *29*, 97.
12. Weininger D.: *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237.
13. http://www.daylight.com
14. Liao C., Liu B., Shi L., Zhou J., Lu X.-P.: *Eur. J. Med. Chem.* **2005**, *40*, 632.

15. Bador P., Lardy J. P.: *New J. Chem.* **1990**, *14*, 805.
16. Raymond J. W., Rogers T. N.: *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 463.
17. Qu D. L., Su J. M., Muraki M., Hayakawa T.: *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 448.
18. Homer R. W., Swanson J., Jilek R. J., Hurst T., Clark R. D.: *J. Chem. Inf. Model.* **2008**, *48*, 2294.
19. Rowley J. R., Oscarson J. L., Rowley R. L., Wilding W. V.: *J. Chem. Eng. Data* **2001**, *46*, 1110.
20. Drefahl A., Reinhard M.: *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 886.
21. Green F. M., Dell E. J., Gilmore I. S., Seah M. P.: *Int. J. Mass. Spectrom.* **2008**, *272*, 38.
22. Green F. M., Gilmore I. S., Seah M. P.: *Appl. Surf. Sci.* **2008**, *255*, 852.
23. Vidal D., Thormann M., Pons M.: *J. Chem. Inf. Model.* **2005**, *45*, 386.
24. Rowley J. R., Wilding W. V., Oscarson J. L., Rowley R. L.: *Int. J. Thermophys.* **2007**, *28*, 824.
25. Karwath A., De Radet L.: *J. Chem. Inf. Model.* **2006**, *46*, 2432.
26. Pugh W. J., Hadgraft J.: *Int. J. Pharm.* **1994**, *103*, 163.
27. Benson S. W., Buss J. H.: *J. Chem. Phys.* **1958**, *29*, 546.