Journal of Structural Chemistry, Vol. 35, No. 6, 1994

FORMING TAXONOMIC GROUPS OF ORGANIC COMPOUNDS FOR CALCULATING GAS-CHROMATOGRAPHIC RETENTION INDICES FROM PHYSICOCHEMICAL CONSTANTS

I. G. Zenkevich

UDC 54.021:54.061:543.544.45

The traditional approach to calculations of gas-chromatographic retention indices from physicochemical constants of organic compounds in homologous series does not encompass all possible objects. Forming different taxonomic groups of compounds by variations in the formal unsaturation, the number of atoms of different elements within molecules, or any complex fragments of structure/composition makes it possible to calculate unknown retention indices by several independent methods from different arrays of initial data. The application of simple three-parameter equations provides an accuracy comparable to the modern interlaboratory error of determination of retention indices for any compounds.

INTRODUCTION

In gas-chromatographic identification of organic compounds, the main parameter is the retention index of a compound (RI, I_x in formulas) [1]. RI is in effect a way of representing retention parameters in a moving system of coordinates determined by reference components (usually n-alkanes):

$$I_x = I_1 + (I_2 - I_1)\frac{f(t_x)-f(t_1)}{f(t_2)-f(t_1)}, \quad (1)$$

where I_1 and I_2 are the postulated quantities of RI for reference components satisfying the interpolation condition $t_1 < t_x < t_2$ (for n-alkanes, $I_i = 100n_i$, where n_i is the number of carbon atoms in a molecule); the form of the function $f(t)$ depends on the temperature conditions of the analysis: in the general case of programmable temperature, $f(t) = t+q \lg (t-t_0)$, where t_0 is the retention time of a nonsorbate gas [2].

RI depend to the greatest extent on the polarity of the stationary phases or sorbents used for separating substances [1]. In deciding on standard phases (nonpolar polydimethylsiloxanes or polar polyethylene glycols), it is possible to consider polarity to be a property analogous to other scalar characteristics (boiling point, refraction index, relative density, etc.).

However, efficient use of RI for identification is absolutely impossible without maximally complete databases. In contrast to mass-spectrometry, where modern databases include more than 200,000 compounds, several of the most complete databases on RI cover a total of 10,000 to 12,000 compounds at most. Any significant extension of these databases by adding even experimental data is a long-term, laborious, and expensive job that cannot be currently done by one or even several laboratories. The problem may only be solved by widely employing calculation methods. Among the numerous known RI calculation techniques [3, 4], only those methods whose accuracy is comparable to the modern interlaboratory error of experimental RI determination (on the average, 5-10 RI units for standard nonpolar phases and 15-25 for polar ones) are suitable for extending databases. For any organic compounds, this accuracy is provided only by calculation techniques based on physicochemical constants. In homologous series, it seems most reasonable to use the three-parameter correlation equations suggested recently [5, 6]

St. Petersburg State University. Translated from *Zhurnal Strukturnoi Khimii*, Vol. 35, No. 6, pp. 176-182, November-December, 1994. Original article submitted November 19, 1993.

 0022-4766/94/3506-0906$12.50

$$\log I_x = a\log T_b + bA + c, \quad (2)$$

where T_b is the boiling point (K); A are the values of any additive characteristics of homologs (see discussion); a, b, and c are the coefficients calculated by the least-square fits to the data for compounds with known I_x and T_b.

All known RI calculation techniques [3-6] have some logical stereotype in their construction. Equations of the form of (2) as well as less convenient four-parameter relations [7, 8] are designed a priori for use in homologous series only. Of course, homologous series are the most "natural" taxonomic groups, and are most frequently separated from the whole variety of organic compounds. But when used alone, this approach severely limits the practical value of RI calculation techniques. This is due to the fact that homologs of many substances are not characterized by any of the two necessary parameters (I_x, T_b) or are nonexistent. For example, according to the known definitions of homology [9], tetrachloromethane and hexachlorobenzene are the sole members of the corresponding series. Consequently, to significantly raise the efficiency of RI calculation techniques, it is necessary to dismiss the viewpoint that homologous series are the sole taxonomic groups of organic compounds formed when dealing with the problem. This paper establishes other groups of compounds for which Eqs. (2) provide the accuracy that is sufficient for inclusion of the calculated RI values into databases along with experimental values.

EXPERIMENTAL

RI databases for standard phases include information from all of the available original references (more than 300 publications dating from 1980 to 1993) and experimental values with subsequent statistical treatment of all data for every compound (for randomization principles, see [5, 6]). RI were experimentally determined to monitor the calculated values on a Biokhrom-1 chromatograph with flame-ionization detector and a glass capillary column $52\cdot10^3\times26\cdot10^{-2}$ mm with OV-101 in the linear mode of temperature programming from 50 to 200°C at a rate of 3 deg/min. n-Alkanes C_5-C_{16} were used in different combinations as reference components. The calculations of linear-logarithmic retention indices [formula (1) at $f(t) = t+q\lg(t-t_0)$] [2] and the least-squares calculations by Eqs. (2) were conducted using a CASIO PB 100 scientific calculator.

To reduce the amount of tabulated data, all examples of treatment of the dependences $I_x(T_b)$ are given only for polydimethylsiloxane standard nonpolar stationary phases, which are the most widely used phases in gas-chromatographic practice. If necessary, all tendencies and criteria found for formation of taxonomic groups may be extended to phases of any polarity.

DISCUSSION OF RESULTS

From the formally logical standpoint, the description of the dependences $I_x(T_b)$ by relations (2) is not associated with artificial separations of homologous series as the only possible taxonomic groups of organic compounds. These equations are abstract mathematical objects that merely reflect the one-to-one correspondence between the physicochemical constants and the chromatographic retention parameters within the group of objects chosen a priori from the whole variety of organic compounds and differing by the given composition of a fixed fragment. In homologous series, this fragment is the homologous difference CH_2. As was shown earlier [5, 6], the following equivalent characteristics of homologs may be chosen as the variable parameter A in Eqs. (2): molecular mass (M), the total number of carbon atoms in the molecule (n_C), indices of homologs in the series ($N = (M-M_1)/14 = n_C - n_C^1$, where M_1 is molecular mass of the simplest homolog and n_C^1 is the corresponding number of carbon atoms in the molecule), molar refractions ($MR_D \approx MR_D^1 + 4.62N$, where MR_D^1 is the molar refraction of the simplest homolog and 4.62 is the refraction of CH_2 [10]), or other quantities which depend linearly on each other and on the above parameters. Accordingly, some other characteristics of organic compounds in homologous series remain unchanged. Such invariants are as follows: the number of heteroatoms in molecules, the molecular formula $C_nH_{2n-m}X_xY_y\ldots$, and formal unsaturation

$$FU = (2n_{IV} + n_{III} - n_I + 2)/2, \quad (3)$$

TABLE 1. Main Types of Taxonomic Groups of Organic Compounds for Calculating Retention Indices

Taxonomic groups	Parameters A varied within the group	Group invariants	Examples of groups
I. Homologous series	M, n_C, N, MR_D	General molecular formula of the series FU, n_X	Alkanols $C_nH_{2n+1}OH$
II. Compounds with variable formal unsaturation	FU, n_H, M	n_C, n_X	$C_6H_{13}Cl \rightarrow \ldots \rightarrow C_6H_5Cl$
III. Compounds with a variable number of univalent heteroatoms	n_X, n_H, M	n_C, FU	$C_2H_6 \rightarrow \ldots \rightarrow C_2Cl_6$
IV. Compounds with a variable number of polyvalent heteroatoms	a) n_y b) n_y, n_H, FU c) n_y, n_H, n_C d) n_C	n_C, FU n_C FU n_H	$C_5H_{12} \rightarrow \ldots \rightarrow C_5H_{12}O_3$ $C_5H_{12} \rightarrow \ldots \rightarrow C_5H_6O_3$ $C_5H_{12} \rightarrow \ldots \rightarrow C_3H_8O_2$ $C_3H_8O \rightarrow \ldots \rightarrow C_{13}H_8O$
V. Compounds with a variable number of polyatomic functional groups	n_Z, M	No invariants	$C_2H_6 \rightarrow \ldots \rightarrow C_2H_2(OCH_3)_4$

where n_I, n_{III}, and n_{IV} represent the number of uni- tri-, and tetravalent atoms in the molecule, respectively.

In line with the theoretical statement above, the variable fragments of molecular composition or structure in the taxonomic groups within which Eqs. (2) provide the unique relationship $I_x \leftrightarrow T_b$, may be chosen using other than homologous differences. The sets of invariant parameters in the groups are varied in accordance with this. Table 1 lists the parameters that are the most essential in dealing with the problem of RI calculation. Thus, the variable parameter FU at constant n_C and n_X may be chosen as the criterion for formation of similar groups. Variation of FU by unity changes the number of hydrogen atoms and hence the molecular mass by 2. In this way one can form sets of compounds with the following molecular formulas:

$$C_{10}H_{22}(FU = 0) \rightarrow C_{10}H_{20} \rightarrow \ldots \rightarrow C_{10}H_{10} \rightarrow C_{10}H_8(FU = 7),$$

$$C_6H_{14}O(FU = 0) \rightarrow C_6H_{12}O \rightarrow \ldots \rightarrow C_6H_6O \rightarrow C_6H_4O(FU = 5).$$

For the first sequence, the coefficients of (2) were found from data for 33 compounds with different values of FU and are $a = 1.7926$; $b = 6.1240 \cdot 10^{-4}$; $c = -1.7506$. The average error of RI calculation on standard nonpolar phases $\Delta = I_{calc} - I_{exp}$ is $\overline{|\Delta|} = 10$ RI units. In the second sequence, the compounds need to be additionally divided into two subgroups depending on the presence of so-called active hydrogen atoms of hydroxyl groups –OH. In the general case, compounds with carboxyl functional groups –COOH and those with hydrogen atoms in the α-position relative to the nitro and cyano groups, i.e., $\rangle CH\text{–}NO_2$ and $\rangle CH\text{–}CN$, should form separate taxonomic groups. Other exceptions, which are difficult to predict from theoretical statements alone, will be revealed after the database of experimental RI is extended.

For 10 compounds from the second series having no active hydrogen atoms, $a = 1.3451$, $b = 5.0149 \cdot 10^{-3}$, $c = -0.6156$ at $\overline{|\Delta|} = 9$. Other examples of varying FU in the groups of halogen-containing organic compounds are given in Table 2.

As in homologous series, we can use proportional M and n_H values instead of FU when calculating RI by Eqs. (2). As a result of this substitution, the coefficients a in Eqs. (2) remain unchanged.

The third type of taxonomic group is based on varying the number of univalent heteroatoms at constant n_C and FU values. These groups include all halogen derivatives of organic compounds whose RI maybe in one-to-one correspondence with T_b of (2). This is certainly an important conclusion. When several halogen atoms are introduced in sequence into a molecule, the increments of indices differ greatly for phases of any polarity, so that none of the

TABLE 2. Examples of the Dependences $I_x(T_b)$ in Taxonomic Groups I-V for Retention Indices on Standard Nonpolar Polydimethylsiloxane Stationary Phases

Taxonomic group	No. of compounds	A	Coefficients of (2)			$\lvert\bar{\Delta}\rvert$
			a	$b\cdot10^3$	c	
I. I. Homologous series:						
Thiophenes	18	n_C	1.7437	6.3992	−1.6642	4
Chloroalkanes	11	n_C	1.9948	−4.9866	−2.2511	7
II. $C_5H_{11}Cl \rightarrow C_5H_7Cl$	7	FU	1.7988	−0.4784	−1.7762	5
$C_6H_{12}Cl_2 \rightarrow C_6H_4Cl_2$	7	FU	1.4631	1.0592	−0.8797	7
III. $C_2H_6 \rightarrow C_2Cl_6$	10	n_{Cl}	1.7147	8.3957	−1.5839	6
$C_6H_6 \rightarrow C_6H_2Cl_4$	9	n_{Cl}	1.7158	4.1130	−1.5541	6
IV. $C_6H_{14} \rightarrow C_6H_{14}O_3$	9	n_O	1.6759	8.1332	−1.4700	6
a. $C_4H_8 \rightarrow C_4H_8S_2$	5	n_S	2.1881	−59.809	−2.7213	4
b. $C_6H_{14} \rightarrow C_6H_{10}O_2$	10	n_O, FU	1.6134	7.0524	−1.3135	9
c. $C_6H_{10} \rightarrow C_3H_7N$	6	n_N, n_C	1.6348	−2.9851	−1.3579	8
d. $C_3H_7Cl \rightarrow C_7H_7Cl$	7	n_C	1.8310	−0.4890	−1.8581	9
V. $C_2H_6 \rightarrow C_8H_{18}O_13$	7	$n_{C_2H_4O}$	1.6896	20.583	−1.5303	6
$CH_4 \rightarrow C_4H_7Cl_3$	5	n_{CHCl}	1.4811	33.130	−1.0331	3

available additive schemes can effectively depict the tendencies of their variation. In the approximation of $I_x(T_b)$ by Eqs. (2), the accuracy of RI estimates compares well with the modern interlaboratory error of their experimental determination. For example, in the case of methane chloro-derivatives, the increments ΔI of sequential replacement of hydrogen by chlorine change by a factor of 3-5 for nonpolar phases and sorbents and even reverse sign for polar phases. At the same time, the average errors of RI approximation by (2) are only 8-12 RI units. This is quite tolerable for this group of low-boiling substances.

	CH_4		CH_3Cl		CH_2Cl_2		$CHCl_3$		CCl_4	$\lvert\bar{\Delta}\rvert$
T_b °C	−161.5		−24.2		39.8		61.2		76.7	
1. Standard nonpolar phases:										
I_{exp}	100		326±16		515±7		609±7		658±14	
ΔI		226		189		94		49		
I_{calc}	99		338		503		591		675	12
2. Porapak Q [11]:										
I_{exp}	100		301±18		467±10		559±4		628±12	
ΔI		201		166		92		69		
I_{calc}	99		309		458		548		639	8
3. Silipor 600 [12]:										
I_{exp}	100		513±2		674±7		699±3		650±7	
ΔI		413		161		25		−49		
I_{calc}	101		498		701		690		648	11

This is the method that was found most effective for calculating RI of unavailable methane and ethane halogen derivatives [13].

The fourth type of taxonomic group is based on varying the number of polyvalent atoms (O, S, N, etc.). This type of modification of molecular composition may be realized in different ways: by varying the number of polyvalent atoms at constant FU and n_C in the set of molecular formulas formed or by choosing only one of these parameters as an invariant, for example,

a. FU = const, n_C = const:

$$C_4H_{10} \rightarrow C_4H_{10}O_2 \rightarrow C_4H_{10}O_2 \rightarrow C_4H_{10}O_3(FU = 0);$$

b. n_C = const:

$$C_6H_{14}(FU = 0) \rightarrow C_6H_{12}S \rightarrow C_6H_{12}S_2(FU = 2);$$

c. FU = const:

$$C_7H_{12} \rightarrow C_6H_{11}N \rightarrow C_5H_{10}N_2 \rightarrow C_4H_9N_3(FU = 2).$$

Varying the number of tetravalent carbon atoms may belong to the same type of composition changes within a group. In this case, neither FU nor n_C is constant in the sets of molecular formulas formed, and n_H is the invariant, for example,

d. n_H = const:

$$C_3H_8O(FU = 0) \rightarrow C_4H_8O \rightarrow \ldots \rightarrow C_{12}H_8O \rightarrow C_{13}H_8O(FU = 10).$$

Table 2 also lists the coefficients of Eqs. (2) and the accuracy estimates, $\overline{|\Delta|}$, for some sets of organic compounds of these types. For all oxygen- and nitrogen-containing compounds, the grouping should obey the above-mentioned prohibition to unite compounds having no active hydrogen atoms and containing –OH, $-CO_2H$, $\rangle CH-NO_2$, and $\rangle CH-CN$ fragments. Data processing for polar phases may possibly expand the list of such "anomalous" functional groups, but this demands special investigation.

Finally, the logical consequence of the above treatment is the general case of forming taxonomic groups of organic compounds by varying any polyatomic fragments of molecular composition (CH_2O, C_2H_4O, CO, CO_2, CHN, CH_3N, CF_2, CHCl, CCl_2, etc.) or structure (OH, OR, CO, CO_2, NH, NR, SR, etc.). The method is "naturally" restricted only by the presence of data on boiling temperatures at atmospheric pressure for complex polyfunctional compounds. As is evident from the examples in Table 2, the accuracy of RI estimates by Eqs. (2) when the C_2H_4O and CHCl composition fragments are varied is in agreement with the estimates of the accuracy of calculations for any other taxonomic groups (3-9 RI units).

Using this concept of forming different taxonomic groups of compounds, we can substantiate and implement in practice the new principle of RI calculation for experimentally undefined, unknown, or unavailable compounds. Since any compound may be included in several nonoverlapping sets of objects, its retention parameters may thus be calculated from several independent sets of experimental data. Coincidence of the results of such calculations is the simplest and at the same time the strictest criterion of their correctness, and it has not been used before.

Example. Calculate the retention index of 3-chlorotetrahydrofuran C_4H_7OCl with T_b 129.2°C [14] on standard polydimethylsiloxane stationary phases.

3-Chlorotetrahydrofuran belongs to the ill-defined homologous series of monochloro-derivatives of cyclic ethers (no experimental data are available for this series). Hence its RI may not be calculated using the scheme applicable only to homologous series [5, 6] because of the lack of the necessary additional information. However, the molecular formula of this compound may be included in at least three other taxonomic groups, whose members were defined by experimental values of RI:

I. $C_4H_9OCl \rightarrow C_4H_7OCl \rightarrow C_4H_5OCl \rightarrow C_4H_5OCl$; FU is the variable parameter, $a = 2.0072$; $b = 7.5671 \cdot 10^{-3}$, $c = -2.3330$; at FU = 1, $I_x = 799$.

II. $C_4H_8O \rightarrow C_4H_7OCl \rightarrow C_4H_6OCl_2$; n_{Cl} is the variable parameter, $a = 1.7489$, $b = 9.0483 \cdot 10^{-3}$, $c = -1.6586$; at $n_{Cl} = 1$, $I_x = 805$.

III. $C_4H_7Cl \rightarrow C_4H_7OCl \rightarrow C_4H_7O_2Cl$; n_O is the variable parameter, $a = 1.8156$, $b = -17.678 \cdot 10^{-3}$, $c = -1.8167$; at $n_O = 1$, $I_x = 784$.

All three values of I_x were calculated from different sets of initial parameters and are basically equivalent. Consequently, the final result should include their statistical treatment with an estimation of the average value and its standard deviation. In the final analysis, the desired RI of 3-chlorotetrahydrofuran is 796 ± 11. No other method can give such accurate estimates for any compounds with known T_b.

The concept of forming different taxonomic groups of organic compounds for the purpose of calculating RI is a prerequisite for significantly extending the available chromatographic databases by adding values calculated from the more detailed and well-classified physicochemical databases [14]. This concept ensures multiple verification of all calculated values. The results of such verification confirm the accuracy of RI calculation, which is comparable to the modern interlaboratory error of RI determination.

This work was carried out as part of the Chemical Informatics Program with partial financial support from the Scientific-Technical Center on Chemical Informatics, Siberian Branch, Russian Academy of Sciences.

REFERENCES

1. M. S. Wigdergaus, L. V. Semenchenko, V. A. Ezrets, and Yu. N. Bogoslovskii, in *Qualitative Gas-Chromatographic Analysis* [in Russian], Nauka, Moscow (1977).
2. I. G. Zenkevich, *Zh. Anal. Khim.*, **39**, No. 7, 1297-1307 (1984).
3. R. Kaliszan, in *Quantitative Structure – Chromatographic Retention Relationships*, Wiley, New York (1987), p. 303.
4. V. M. Nabivach and V. P. Dmitrikov, *Usp. Khim.*, **62**, No. 1, 27-38 (1993).
5. I. G. Zenkevich and L. M. Kuznetsova, *Collect. Czech. Chem. Commun.*, **56**, 2042-2056 (1991).
6. I. G. Zenkevich and L. M. Kuznetsova, *Zh. Anal. Khim.*, **47**, No. 6, 982-993 (1992).
7. K. Heberger, *Chromatogr.*, **29**, No. 7/8, 375-384 (1990).
8. K. Heberger, *Anal. Chim. Acta*, **223**, 161-174 (1989).
9. Yu. A. Zhdanov, in *Homology in Organic Chemistry* [in Russian], Moscow State University, Moscow (1950).
10. B. V. Ioffe, *Refractometric Methods in Chemistry* [in Russian], Khimiya, Leningrad (1974).
11. I. G. Zenkevich and S. V. Konyukhova, *Vestn. Sankt-Peterb. Gos. Univ., Ser. Fiz. Khim.*, No. 1, 66-70 (1992).
12. I. G. Zenkevich, I. A. Tsibulskaya, and A. A. Rodin, *Zh. Alan. Khim.*, **46**, No. 1, 101-110 (1991).
13. I. G. Zenkevich, S. V. Konyukhova, and B. N. Maksimov, *Zh. Fiz. Khim.*, **67**, No. 7, 1474-1479 (1993).
14. *Beilstein Handbuch der Organischen Chemie*, Vol. I-XXXI, 4th edition, Springer-Verlag, Berlin (1918).

Translated by L. Smolina