Linear Correlation Equation for Retention Factor of Nucleic Acid Using QSPR

Jinzhu Zheng, Soon Koo Han, and Kyung Ho Row*

*Department of Chemical Engineering, Inha University, Incheon 402-751, Korea. *E-mail: rowkho@inha.ac.kr*

Received January 6, 2005

In the reversed-phase chromatography, the retention time of sample was investigated based on the molecular structure of compound. Several descriptors that were related to retention factors were selected, and then the values of descriptors were calculated with several softwares. The effect of retention factor was measured with calculated values, and the results were obtained that each descriptors of molecular structure of compound have different effect on the retention factor. Therefore, the empirical equation for seven types of descriptors considered was obtained, and it has high values of correlation coefficient. Furthermore, the experimental data and calculated values have good agreement.

Key Words : Nucleic acid, Retention factor, QSPR, Linear equation, Correlation

Introduction

In the separation process using chromatography, the retention time of sample in column is an important parameter. The retention time of sample in column was determined by various parameters.[1] To predict the accurate retention time, complex equation should be calculated. Several researchers have proposed empirical equation of simple type for predicting retention time, previously.[2] In the liquid chromatography, the empirical equations related to the contents of organic modifier in mobile phases were mainly proposed. It has advantage of prediction of retention time with content of organic modifier, but it has disadvantage as to apply various empirical equations with sample. For this reason, the study that predicting retention time based on different structures of samples is performed.[3]

While quantitative structure activity relationships (QSAR) connect a variety of biological and medicinal properties to molecular structure, quantitative structure property relationships (QSPR) relate physicochemical properties such as boiling point, solubility, and partition coefficients to molecular structure.[4,5] And it was quantified the connection between the structure and retention property of molecules and allow the prediction retention parameters of solute using molecular structure parameters. The retention of a solute in chromatography is determined by different kinds of interaction among the stationary phase, the mobile phase and the solute molecules, such as dipole/dipole interaction, dipole/induced interaction, dispersion interaction, etc. The kind of interaction depends on the structure and properties of the stationary phase, the mobile phase and the solute molecules.[6]

There were several studies on the prediction of physicochemical properties for organic compounds by the linear logarithm relationship between the retention factors and physicochemical properties of the targeted compounds.[6]

In this study, the retention factors of nucleic acids were assumed to be governed by their several descriptors such as water solubility, polarizability, solvation free energy, Wiener index, molar refractivity etc. The research of the correlation relationships between descriptors and the chromatographic retention factors of nucleic acids were the purpose of this work. The correlations were established by the empirical equations. The retention factors of 2’-deoxyribonucleosides and deoxynucleotides were demonstrated to validate the feasibility of the correlations.

Experimental Section

Reagents. The standard chemicals of 2’-deoxynucleoribosides (dCyd, dUrd, dGuo, dThd, dAdo) and deoxynucleotides (5’-UMP, 5’-IMP, 5’-GMP, 5’-TMP) as shown in Figure 1 were purchased from Sigma (St. Louis, MO, USA) and Fluka (St. Louis, MO, USA). The HPLC grade methanol used was purchased from Duksan Co. (Seoul, Korea). The chromatographic column (3.9 × 300 mm) used was packed with C18 (15 μm, Merck, Germany). The sodium phosphate monobasic for buffer solution was purchased from Yakuri Pure-Chem. Co. (Osaka, Japan). Water was filtered by a Milipore ultra pure water system (Milipore, Bedford, MA, USA).

Sample Preparation. 2’-deoxynucleoribosides and deoxynucleotides each were dissolved in water and the concentration was 1 mg/mL each. A constant injection volume (20 μL) was used throughout the experiment.

Apparatus and Methods. The HPLC system was equipped with Waters 600S solvent delivery system including 616 solvent delivery pump and 600S controller and the 2487 UV dual channel detector, an injector (0.02 mL sample loop) of Rheodyne injector (Waters, Milford, MA, USA). The data acquisition system was Millenium32 (Waters) installed in a HP Vectra 500 PC. The mobile phases were composed of 8.9 mM of sodium phosphate buffer solution and methanol, and it was performed isocratic elution as 90/10(vol. %). The mobile phases were degassed with helium. The flow rate of mobile phase was 1 mL/min. The UV wavelength was fixed at 254 nm. And all experiment was performed at room temperature. The hold -up time was measured as retention time of 0.02 mL of acetonitrile.

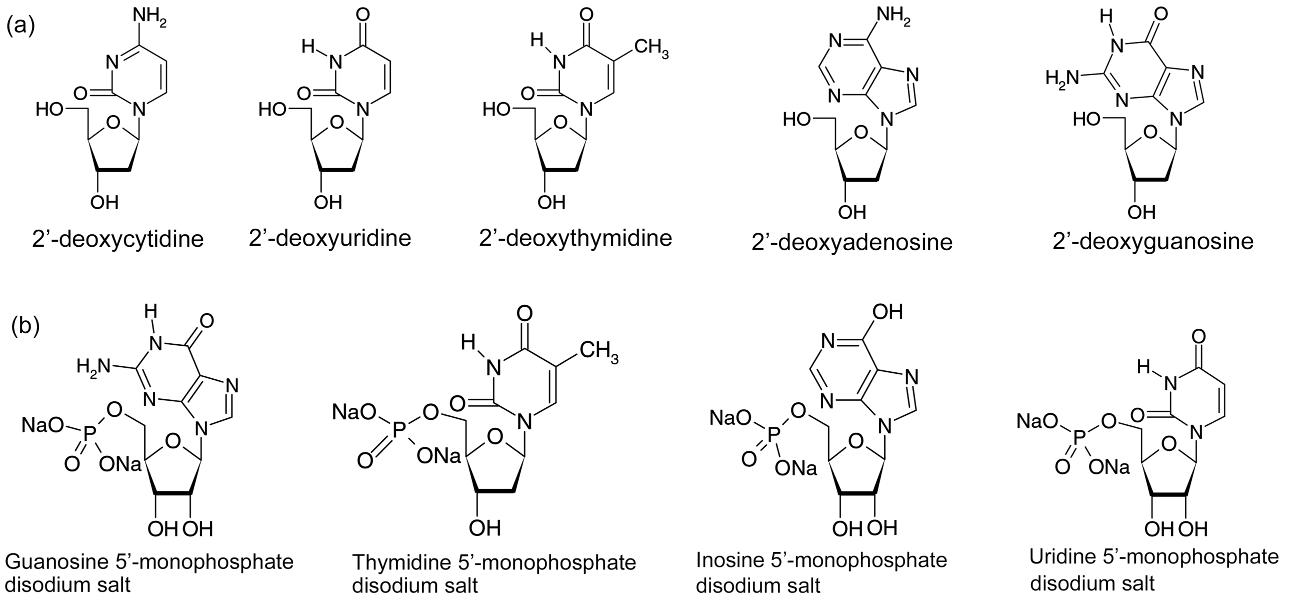
(a)
2’-deoxycytidine
2’-deoxyuridine
2’-deoxythymidine
2’-deoxyadenosine
2’-deoxyguanosine
(b)
Guanosine 5’-monophosphate disodium salt
Thymidine 5’-monophosphate disodium salt
Inosine 5’-monophosphate disodium salt
Uridine 5’-monophosphate disodium salt

Figure 1. Chemical structures of 2’-deoxyribonucleosides (a) and deoxynucleotides (b).

Theoretical Background

A. Retention Factor. The retention factor can be calculated according to Eq. (1).

$$k = \frac{V_R - V_M}{V_M} = \frac{t_R - t_M}{t_M} \qquad (1)$$

where, t_M is the hold-up time, t_R is the retention time, k is retention factor.

B. Physicochemical Descriptor. The physicochemical descriptor was calculated with HyperChem[7,8] and PreADME. The value of polarizability(α) of standard chemicals was calculated according to a semi-empirical method, in HyperChem, and the calculation of water solubility and solvation free energy were performed with PreADME.

Polarizability: Polarizability(α) is the ease of distortion of the electron cloud of a molecular entity by an electric field (such as that due to the proximity of a charged reagent). It is the ratio of induced dipole moment (μ_{ind}) to the field (E),

$$\alpha = \mu_{ind}/E \qquad (2)$$

In ordinary usage the term refers to the “mean polarizability”, *i.e.*, the average over three rectilinear axes of the molecule. Polarizability is seen in certain modern theoretical approaches as a factor influencing chemical reactivity, etc.

Water solubility: Water solubility (log S), also known as aqueous solubility, is the maximum amount of a substance dissolved in water at equilibrium at a given temperature and pressure. Water solubility values are usually expressed as moles of solute per liter. Water solubility has been correlated to various chemical parameters used to determine the fate of chemicals in the environment.

Solvation free energy: To estimate the solubility of the metabolizable substance of component, water solvation free energy was usually used. It is the sum of each atom parameter divided from atom type in molecular and defined as,

$$\log S = \sum_i n_i S_i \qquad (3)$$

where, n_i is the number of atoms of type i, a_i is the atomic log S contribution.[9]

C. Geometrical Descriptor. Topological polar surface area (PSA) was calculated with PreADME. Relation to the molecular absorption, PSA is used to estimate molecular transport properties recently. PSA is defined the sum of the polar atoms in one molecular. Polar atom type and atomic parameter were invented by Peter Ertl *et al.*.[10]

D. Electronic Descriptor. The dipole length was calculated with Chem-Office.[11] It was determined by the dipole moment. The dipole moment is the first derivative of the energy with respect to an applied electric field. It measures the asymmetry in the molecular charge distribution and is reported as a vector in three dimensions. The dipole length is the summarized vector of three dimensions. If the branches have narrow dimension, calculated dipole length will be high, if wide enough to be a plane, the value is 0.

E. Thermodynamic Descriptor. The molar refractivity was calculated with HyperChem. The molar refractivity is the molar volume corrected by the refractive index. It was represented by size and polarizability of a fragment or molecule. The molecular refractivity can be rewritten as follows,

$$MR = \frac{(n^2 - 1)M}{(n^2 + 2)d} \qquad (4)$$

Where n is the refractive index, M is the molecular weight and d is the density.

F. Topological Descriptor. Topological descriptor was selected Wiener index. The Wiener index (w) is one of the oldest molecular-graph-based structure descriptors and its chemical applications are well documented.[12] If G is a molecular graph, u and v are two vertices of G, and $d(uv/G)$ is the distance between them value of edges in a shortest path connecting u and v, then the Wiener index is defined as,

$$W = w(G) = \Sigma\, d(uv/G) \tag{5}$$

It is summarized over all pairs of vertices of G. The fact that there are good correlations between varieties of physicochemical properties of organic compounds (boiling point, heat of evaporation, heat of formation, chromatographic retention times, surface tension, vapor pressure, partition coefficients, etc.) could be rationalized by the assumption that W is roughly proportional to the Van der Waals surface area of the respective molecule.

G. Regression. The regression equation is formed as follows,

$$k = f(D) = \sum_{i=1}^{N} a_i D_i \tag{6}$$

Where, D is descriptor, and a is each coefficient of descriptor, i is number of descriptor, respectively. The regression equation was obtained by correlating the experimental data of retention factors with the calculated all descriptors through linear regression analysis.

H. Standard Deviation. The standard deviation (σ) was defined as,

$$\sigma = \sqrt{\left\{ \sum_{k=1}^{n} (x_k - \bar{x})^2 \right\} / n} \tag{7}$$

Where x is the average of all the experimental values. x_k is the one of the experimental value. If the standard deviation is 0, the observations have same values. In the case that standard has higher values, the observation has the error with average. Therefore, the standard deviation was described degree of distributions of observations.

Results and Discussion

In this paper, the quantitative relationships between the chromatographic retention factor of nucleic acids and several descriptors are studied. All nucleic acids were divided into four groups: pyrimidines with phosphate group (I), pyrimidines without phosphate group (II), purines with phosphate group (III), purines without phosphates group (IV). Retention factors of each component were calculated with Eq. (1). The retention factors and calculated physicochemical properties of nucleic acids were listed in Table 1.

In the liquid chromatography, the descriptors that affect retention factor were researched. The retention factor was related to several parameters such as particle size and pore size of stationary phase and the composition of mobile phase. Generally, the particle size and type of the stationary phase was determined at the beginning of the experiment. So it was important to determine composition of the mobile phase for the reason that the retention factor of sample was caused by interaction between sample and mobile phase in the surface of stationary phase. The difference of affinity of sample is caused by the polarity of mobile phase and molecular structure of sample, which influences the retention time of samples. Therefore, 12 descriptors that related properties of sample were selected. In the case the descriptors were calculated, they were considered the arrangements of molecule and structure of compound. Because the chemical bond was flexible, the structure of compound was varied. As result, the calculated values of the descriptors would have different values. Therefore, the

Table 1. The values of descriptors and retention factor of nucleic acid constituents

Descriptor		Physicochemical descriptor			Geometrical descriptor	Electronic descriptor	Thermodynamic descriptor	Topological descriptor	Retention factor
Group		Polarizability (D_1)	Water solubility (D_2)	Solvation free energy (D_3)	Topological PSA (D_4)	Dipole length (D_5)	Molar refractivity (D_6)	Wiener index (D_7)	k
Unit NO.*	Compound	A^3	g/mL	kcal/mol	A^2	eV	A^3	–	–
I	5’-UMP	22.800	289,176	–18.050	159.120	10.684	61.040	1,234	0.016
	5’-TMP	24.000	100,732	–13.560	138.890	36.940	63.890	1,261	0.455
II	dCyd	21.180	59,649	–21.230	110.600	9.634	53.270	434	1.360
	dUrd	20.330	145,805	–23.300	104.550	11.220	51.090	434	1.672
	dThd	22.170	54,642	–22.530	138.890	11.051	55.440	505	3.976
III	5’-IMP	25.780	5,714	–12.230	168.090	19.422	69.900	1,523	0.192
	5’-GMP	27.000	4,819	–14.940	193.850	15.672	72.540	1,686	0.202
IV	dGuo	24.530	2,573	–20.190	139.280	11.140	65.580	671	3.718
	dAdo	24.030	4,768	–16.580	119.310	7.247	61.890	582	18.044
Standard deviation		4.902	69.048	3.774	10.923	2.697	7.867	24.125	

*(I: Pyrimidines with phosphate group, II: Pyrimidines without phosphate group, III: Purines with phosphate group, IV: Purines without phosphates group)

Table 2. Empirical equation considered with five and six types of descriptors

Equation No.	Coefficient								
	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	r^2
1	41.6462		1.40×10^{-5}	2.3398	0.1189	−0.2533	0.2500	−0.0276	0.7917
2	16.9616	4.5862		1.6146	−0.0322	−0.3067	−1.1642	−0.0125	0.8072
3	−159.1299	19.0170	6.67×10^{-5}		−0.3474	−0.1980	−3.8195	−0.0029	0.9419
4	−66.3176	12.3251	5.41×10^{-5}	1.8130		−0.1710	−2.7050	−0.0243	0.9564
5	−125.3876	16.3719	6.82×10^{-5}	1.0400	−0.1098		−3.3819	−0.0205	0.9338
6	−3.5506	2.6127	2.77×10^{-5}	1.8135	0.0181	−0.2292		−0.0250	0.8342
7	−133.7868	18.3272	5.92×10^{-5}	0.2706	−0.3561	−0.2798	−3.8825		0.9438
8	−24.6370	3.3970	1.23×10^{-5}	0.3757	−0.2866	−0.3804			0.7083
9	−185.9597	21.4831	6.99×10^{-5}	−0.4136	−0.3866		−4.4521		0.8478
10	−22.7808	3.0623	3.49×10^{-5}	1.7603	0.0918			−0.0322	0.7792
11	23.5979	7.0293	-7.28×10^{-6}	1.2614		−0.3832	−2.5435		0.5279
12	−6.2640	2.7393	2.78×10^{-5}	1.7376		−0.2346		−0.0239	0.8339
13	−96.4034	14.0469	6.28×10^{-5}	1.5313			−2.9738	−0.0266	0.9237
14	−152.9608	19.0943	6.36×10^{-5}		−0.3738	−0.2383	−3.9185		0.9382
15	−72.3073	5.0310	2.88×10^{-5}		−0.2471	−0.2155		−0.0074	0.7223
16	−175.4432	19.5765	7.35×10^{-5}		−0.2786		−3.8829	−0.0095	0.9007
17	−135.1348	12.8290	5.04×10^{-5}			0.0641	−2.4392	−0.0193	0.6671
18	−11.2544	7.2611		0.8646	−0.1983	−0.3670	−1.7406		0.7797
19	36.1177	0.7696		1.8492	0.0299	−0.2887		−0.0189	0.7790
20	8.4319	3.7192		1.6032	0.0901		−0.9117	−0.0219	0.7028
21	22.1780	4.1633		1.7485		−0.2967	−1.0927	−0.0146	0.8063
22	−47.3879	8.1845			−0.2800	−0.3049	−1.6643	0.0051	0.7239
23	34.6972		-8.73×10^{-6}	0.8987	−0.1914	−0.4363	0.3083		0.6318
24	56.6866		9.78×10^{-6}	2.5732	0.1569	−0.2655		−0.0280	0.7836
25	25.7090		2.03×10^{-5}	2.3321	0.2124		0.3508	−0.0361	0.7238
26	33.0511		1.07×10^{-5}	1.8596		−0.3039	0.4208	−0.0199	0.7745
27	−19.4057		1.95×10^{-5}		−0.1945	−0.2506	0.9337	−0.0041	0.5704
28	50.9748			2.1565	0.0857	−0.2869	0.0612	−0.0220	0.7716

lowest energy conformation of each structure calculated using the HyperChem was used to analyze the effect of descriptors. To investigate the effect of descriptors on retention factor, the standard deviation was calculated. When standard deviation of descriptor has high values, it has influence on retention factor more than other. The seven descriptors were selected according to the standard deviation from large to small ranking. It was shown in Table 1.

More influence of the descriptors on retention factor had the correlation coefficients of empirical equation was higher value. Therefore, the effects of descriptor on retention factors were investigated by correlation coefficients. Two descriptors were selected and then linear equation was obtained. In order to study the correlation coefficient the effect of polarizability, water solubility and molar refractivity was investigated. It could be said that the correlation coefficient of linear empirical equations which was considered polarizability, molar refractivity have higher value in the group of I, II and III, IV. And the correlation coefficients of empirical equations have lower values in the phosphate (III, IV). But, empirical equation that the dipole length and topological PSA was considered has high values of correlation coefficient. According to the results, it was obtained that the descriptor of each group has different effect on retention factor. Therefore, the empirical equation for

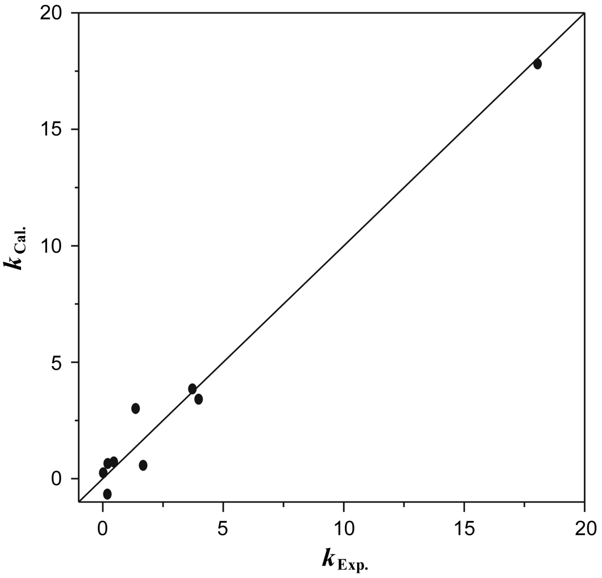

Figure 2. Comparison of experimental and theoretical retention factor.

prediction of retention factor of nucleic acid should be used above three descriptors at least.

An empirical equation was obtained with three types of descriptors. The correlation coefficient of the empirical equations has lower values. It was deemed that descriptors which related to retention factor were not includes. For this reason, the empirical equation was obtained with four types of descriptor. It could be seen that the correlation coefficient was increasing. Because the suitable number of descriptors that were related retention factor was small. There were error between the experimental data and calculated values. For this reason, the descriptor, which has influence on retention factors, was not suitable to the variable of empirical equation. When the empirical equation was considered with five types of descriptors (Eq. No. 7-28 in Table 2), the a few value of correlation coefficient was above 0.9, as shown in Table 2. In this case, the experimental data was good agreement with theoretical values. When the numbers of descriptors were six (Eq. No. 1-6 in Table 2), the values of correlation coefficient of almost empirical equations were higher. When the number of descriptors was seven, the value of correlation coefficient of empirical equation was above 0.97. Moreover, the experimental data and theoretical values have good agreement, as shown in Figure 2. The empirical equation for prediction of retention factor of nucleic acid was obtained as Eq. (8).

$$k = -104.8153 + 15.5663D_1 + 6.07 \times 10^{-5}D_2 + 1.1099D_3 - 0.1710D_4 - 0.2088D_5 - 3.2813D_6 - 0.0143D_7 \quad (8)$$

Descriptors, D_1 to D_7, were shown as follows;

D_1 : Polarizability.

D_2 : Water solubility.

D_3 : Solvation free energy.

D_4 : Topological PSA.

D_5 : Dipole length.

D_6 : Molar refractivity.

D_7 : Wiener index.

Conclusion

Several descriptors were selected according to the standard deviation to forecast the retention factor. The effect of retention factor was estimated with calculated values, and the results showed that each descriptor with different molecular structure has different effect. The empirical equation was estimated from two to seven types of descriptors, respectively. When seven types of descriptor were used, a higher value of correlation coefficient of the empirical equation was obtained. Furthermore, the experimental data and the theoretical values have good agreement.

Acknowledgment. This work was supported by INHA UNIVERSITY Research Grant (INHA-2005).

References

1. Lee, S. K.; Polyakova, Y.; Row, K. H. *Bull. Korean Chem. Soc*. **2003**, *24*, 1757.
2. Han, S. K.; Zin, E.; Row, K. H. *Bull. Korean Chem. Soc*. **2004**, *25*, 1807.
3. Lee, S. K; Row, K. H. *Bull. Korean Chem. Soc.* **2003**, *24*, 1265.
4. Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; Am. Chem. Soc.: Washington, DC, 1995.
5. Katritzky, A. R.; Karelson, M.; Lobanov, V. S. *Pure Appl. Chem*. **1997**, *69*, 245.
6. Yao, X.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *Talanta* **2002**, *57*, 297.
7. Dewar, M. J. S.; Helay, E. J. *J. Comput. Chem*. **1983**, *4*, 158.
8. Dewar, M. J. S.; Helay, E. J.; Stewart, J. *J. Comput. Chem*. **1988**, *5*, 358.
9. Viswanadhan, V. N.; Ghose, A. K.; Singh, C.; Wendoloski, J. J. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 405.
10. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
11. *CS Chem Office Pro for Microsoft Windows*; Cambridge Scientific Computing Inc.: 875 Massachusetts Avenue, Suite 61, Cambridge MA, 02139, USA, 2003.
12. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.