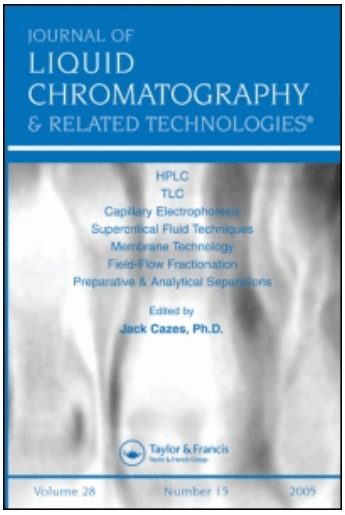

Journal of Liquid Chromatography & Related Technologies
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597273

Prediction of HPLC Retention Time Using Multiple Linear Regression: Using One and Two Descriptors

Taesung Moon[a]; Myung Whan Chi[a]; Song Ja Park[a]; Chang No Yoon[a]
[a] Bioanalysis and Biotransformation Research Center, Korea Institute of Science and Technology, Cheongryang, Seoul, Korea

Online publication date: 10 September 2003

To cite this Article Moon, Taesung , Chi, Myung Whan , Park, Song Ja and Yoon, Chang No(2003) 'Prediction of HPLC Retention Time Using Multiple Linear Regression: Using One and Two Descriptors', Journal of Liquid Chromatography & Related Technologies, 26: 18, 2987 — 3002
To link to this Article: DOI: 10.1081/JLC-120025413
URL: http://dx.doi.org/10.1081/JLC-120025413

JOURNAL OF LIQUID CHROMATOGRAPHY & RELATED TECHNOLOGIES®
Vol. 26, No. 18, pp. 2987–3002, 2003

Prediction of HPLC Retention Time Using Multiple Linear Regression: Using One and Two Descriptors

Taesung Moon, Myung Whan Chi, Song Ja Park, and Chang No Yoon*

Bioanalysis and Biotransformation Research Center,
Korea Institute of Science and Technology,
Cheongryang, Seoul, Korea

ABSTRACT

The quantitative structure-retention relationships (QSRR) studies on polycyclic aromatic hydrocarbons (PAHs) show that the multiple linear regression (MLR) models achieved in training sets have to be validated. The MLR models derived by using one or two descriptors showed high linear correlations between input descriptor(s) and high-performance liquid chromatography (HPLC) retention times, but showed very poor predictivity in test sets for one-descriptor models, whereas the models derived by using two descriptors (molecular connectivity and dipole moment) showed good predictivity. These results suggest that one-descriptor models are not sufficient to explain the retention time in spite of high r^2 values in training sets.

*Correspondence: Chang No Yoon, Bioanalysis and Biotransformation Research Center, Korea Institute of Science and Technology, P.O. Box 131, Cheongryang, Seoul 130-650, Korea; E-mail: cody@kist.re.kr.

DOI: 10.1081/JLC-120025413
1082-6076 (Print); 1520-572X (Online)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

In addition, the predictivity was affected by the solvent. High predictive r^2 values were obtained in conditions in which methanol was used as a solvent.

Key Words: HPLC; Retention time; QSRR; MLR; PAHs.

INTRODUCTION

The quantitative structure-activity (or structure-property) relationships (QSAR or QSPR) studies express physical, chemical, or biological property of molecules as a function of chemical structure. This concept is based on the assumption that the difference in the structural properties of molecules is able to account for the difference in their chemical or biological properties. The search for such relationships has been one of the most important applications of modeling techniques.[1–5] Polycyclic aromatic hydrocarbons (PAHs) are well known for their ubiquitous occurrences in the environment.[6] A large variety of combustion processes, for example, the burning of waste, coal, wood, and biomass, as well as traffic exhaust, spray PAHs in the environment.[7–10] The PAHs are included in carcinogenecity, unscheduled DNA synthesis, and free radical DNA damage. The propensity of PAHs to be carcinogenic depends on the ability to form the cation that arylates the nucleic acid base.[11,12] Nitrated polyaromatic hydrocarbons have been also recognized as a potent class of mutagen.[13] Nitropyrens have been reported to be the most potent mutagens.[14] It is known that nitrated PAHs can be formed by reaction of PAHs and nitrogen oxides, both of which are present in combustion effluents. Recent interest has focused on development of analytical methods for the determination of PAHs in environmentally significant samples. A lot of analytical methods have been used to determine PAHs. These include high-performance liquid chromatography (HPLC) with electrochemical, mass spectroscopy (MS), thin-layer chromatography (TLC) with fluorescence detection, gas chromatography (GC), and GC/MS. In chromatographic techniques, chromatographic retention is based on the interaction between solute and stationary phase. In this work, the quantitative structure-retention relationships (QSRR) studies were carried out to obtain further insight into the relationships between chemical structure and retention time of PAHs in HPLC. The multiple linear regressions (MLR) were performed to find the relationships.

EXPERIMENTAL

High-Performance Liquid Chromatography

A Hewlett–Packard Model 1090 HPLC was used for determination of the retention times of PAHs.[15] Retention times were measured at seven different

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

solvent conditions. Chromatographic conditions are shown in Table 1. The PAHs used in this study and their structural features are shown in Fig. 1.

Molecular Modeling

All the PAH compounds in this work were constructed by using InsightII package (Accelrys Inc.) and minimized by using va09a minimizers until maximum energy derivatives are less than 0.00001 kcal/Å. The minimized structures were then fully geometrically optimized using AM1 Hamiltonian in MOPAC software package. The semiemperically derived low energy structures were used to compute the values of QSRR descriptors. The following set of descriptors was used in multiple linear regression analysis: (1) molecular weight (MW); (2) molecular connectivity (χ); (3) length-to-breadth ratio (L/B); (4) volume; (5) Connolly surface area; (6) the largest principal moment of inertia (I_α); (7) dipole moment; (8) highest occupied molecular orbital energy (HOMO); and (9) lowest unoccupied molecular orbital energy (LUMO).

The molecular connectivity (χ) expresses a molecular structure based on a count of skeletal atom groupings, weighted by the degree of skeletal branch-

Table 1. HPLC conditions.[15]

HPLC chromatography	Hewlett-Packard Model 1090
Column	ODS Hypersil (100 mm × 4.6 mm I.D., 5 μm particles)
Injection volume	5 mL
Oven temperature	40°
Flow rate	1 mL/min
Solvent concentration	
Condition 1 ACN 50%	$\xrightarrow{0.5\%/\text{min}}$ 70%
Condition 2 ACN 50%	$\xrightarrow{1.0\%/\text{min}}$ 70%
Condition 3 MeOH 50%	$\xrightarrow{1.0\%/\text{min}}$ 70%
Condition 4 ACN 60%	$\xrightarrow{0.5\%/\text{min}}$ 80%
Condition 5 ACN 60%	$\xrightarrow{1.0\%/\text{min}}$ 80%
Condition 6 MeOH 60%	$\xrightarrow{0.5\%/\text{min}}$ 80%
Condition 7 MeOH 60%	$\xrightarrow{1.0\%/\text{min}}$ 80%

Note: ACN, acetonitrile; MeOH, methanol.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

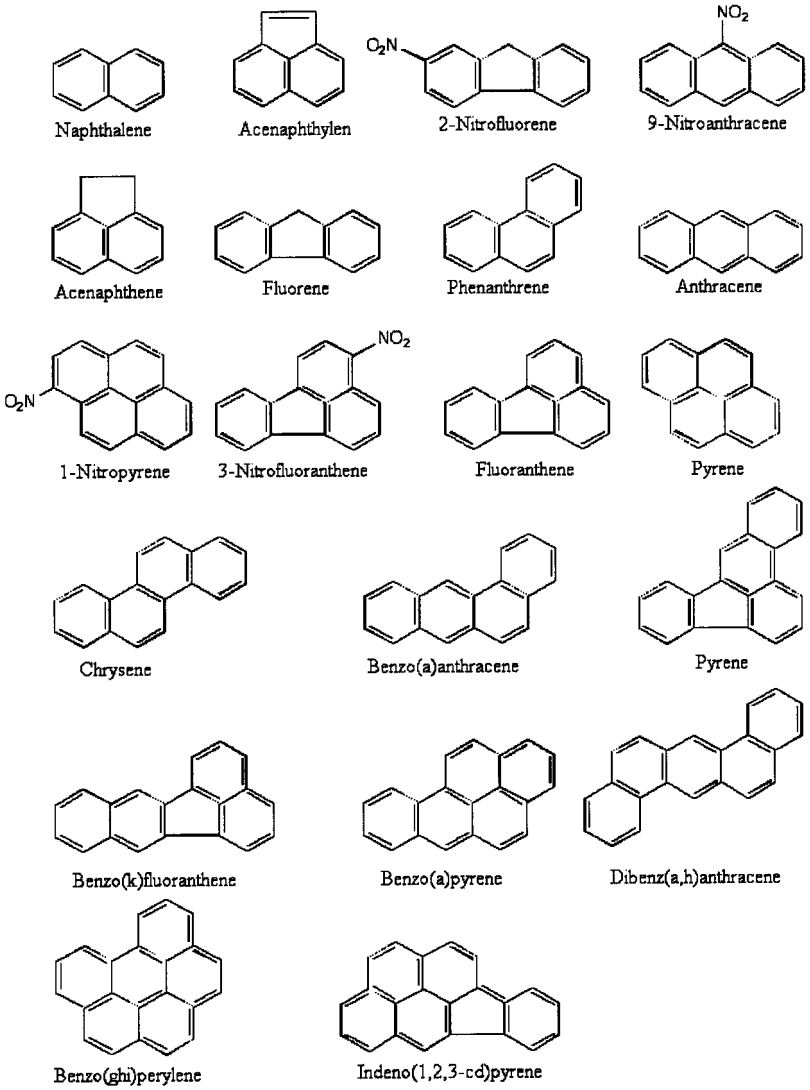

Figure 1. Structures of PAHs.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ing. It is known that molecular connectivity is correlated with solubility parameters, boiling points, densities, partition coefficients, and retention data.[3] In this study, the first-order molecular connectivity was used. This is calculated by:[3,16]

$$^1\chi = \sum(\delta_i\delta_j)^{-0.5} \tag{1}$$

where δ is equal to the difference between the number of valence electrons and the number of hydrogen atoms attached to that atom which is express as:

$$\delta = \frac{Z^v - h}{Z - Z^v - 1} \tag{2}$$

where Z^v is the number of valence electrons in the atom, Z is its atomic number, and h is the number of hydrogen atoms bound to the same atom. The L/B ratio is calculated from the rectangle enclosing the PAH molecule which maximizes L/B.[17]

Quantitative Structure-Retention Relationship Analysis

In order to ensure reliability of our model, the data set was divided into training and test sets. Several criteria have been used for selection of the best subset. One of the most popular criteria is the maximization of $|X'X|$, which is used in D-optimal design. In this work, the stepwise addition method was applied for the D-optimal design of Mitchell[18] for selection of the training set. The subsets having high $|X'X|$ were listed in the order of decreasing $|X'X|$. Then, one sample was added to the subsets and the new subsets were listed again. This procedure was repeated to reach a given number of members in the training set. Prior to the D-optimal design, the crossvalidation was performed in order to find the optimum number of descriptors. The MLR model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{3}$$

where the observed values of the y_is are the dependent variables, the x_{i1}s, x_{i2}s, ..., x_{ik}s are the sets of the k independent variables (or descriptors), β_0, $\beta_1, \ldots, \beta_k$ are the regression coefficients, and the ε_is are independently distributed normal errors. When MLR was performed by using two descriptors, all possible combinations of these descriptors were considered to find the best regression model. The multicollinearity among descriptors was identified

using the variance inflation factor (VIF).[19] The VIF for the ith regression coefficient is expressed as

$$\mathrm{VIF} = \frac{1}{1 - r_i^2} \tag{4}$$

where r_i is the coefficient produced by regressing the descriptor x_i against the other descriptors, x_j ($j \neq i$). The models of which VIF is greater than 10, were not considered. The predictive activity of the model was quantitated in terms of r^2 which is defined as

$$r^2 = 1.0 - \frac{\sum (y_{\mathrm{pred}} - y_{\mathrm{actual}})^2}{\sum (y_{\mathrm{actual}} - y_{\mathrm{mean}})^2} \tag{5}$$

where y_{pred}, y_{actual}, and y_{mean} are predicted, actual, and mean values of the target property, respectively.

RESULTS AND DISCUSSION

Single-Descriptor Multiple Linear Regression Models

The number of training set members was set to 11, because the largest determinant is zero when the number of training set members is greater than 11 in D-optimal design by using one descriptor. Prior to D-optimal design, the input descriptors were selected by using crossvalidation. Since the cross-validated r^2 values of molecular weight, molecular connectivity, volume, surface area, and the largest moment of inertia were greater than 0.5 in all conditions, D-optimal designs were performed with each of these descriptors. The values of retention times and descriptors used in MLR models are listed in Table 2, which shows that retention times have a tendency to increase as the values of molecular weight, molecular connectivity, volume, and surface area increase. The selected training set members for each descriptor and the statistical summary of one-descriptor MLR models are shown in Tables 3 and 4. The MLR models achieved by using each of molecular connectivity, volume, and surface area show very high r^2 values (greater than 0.9) in training sets. The molecular connectivity had the highest linear correlations with retention times in all conditions, which seems to be caused by the fact that the molecular connectivity has the information about molecular size as well as molecular shape. In spite of high linear correlations in training sets, the predictivity in test sets was very low (Table 4). Most predictions gave negative r^2 values, which mean the lack of predictivity. This low predictivity is caused by the fact that

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 2. Set of data used in MLR.

No.	Compounds	Retention time						
		Cond-1	Cond-2	Cond-3	Cond-4	Cond-5	Cond-6	Cond-7
1	Naphthalene	5.614	5.328	10.305	3.417	3.340	6.249	5.856
2	Acenaphthylen	6.716	6.259	12.477	3.922	3.805	7.700	7.025
3	2-Nitrofluorene	7.694	7.019	14.322	4.217	4.063	8.983	7.976
4	9-Nitroanthracene	8.958	8.089	16.111	5.075	4.847	10.410	8.863
5	Acenaphthene	8.958	8.089	16.605	5.075	4.847	11.389	9.831
6	Fluorene	9.370	8.323	17.199	5.075	4.847	11.831	10.120
7	Phenanthrene	10.147	8.970	18.352	5.545	5.247	12.918	10.828
8	Anthracene	11.003	9.609	19.433	5.962	5.606	14.098	11.604
9	1-Nitropyrene	12.021	10.328	20.374	6.401	5.987	15.125	12.228
10	3-Nitrofluoranthene	12.500	10.643	21.242	6.401	5.987	16.202	12.901
11	Fluoranthene	13.190	11.183	22.012	7.065	6.534	17.231	13.603
12	Pyrene	13.952	11.763	22.788	7.573	6.968	18.265	14.277
13	Chrysene	18.036	14.434	26.751	9.591	8.546	24.104	17.531
14	Benzo(a)anthracene	18.670	14.819	27.190	9.902	8.777	24.824	17.905
15	Benzo(b)fluoranthene	23.178	17.673	30.476	12.616	10.816	30.445	20.914
16	Benzo(k)fluoranthene	23.947	18.124	30.876	13.053	11.123	31.223	21.274
17	Benzo(a)pyrene	23.947	18.452	31.218	13.496	11.491	31.670	21.639
18	Dibenz(a,h)anthracene	29.018	21.030	34.318	16.159	13.237	37.381	24.628
19	Benzo(ghi)perylene	29.935	21.824	35.105	17.211	14.116	37.906	25.469
20	Indeno(1,2,3-cd)pyrene	29.935	21.824	35.105	17.211	14.116	37.906	25.469

(*continued*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 2. Continued.

No.	Compounds	Molecular weight	χ[a]	L/B[b]	Volume	Area[c]	PI[d]	Dipole moment
1	Naphthalene	128	3.405	1.24	108.15	162.231	9.491	0.001
2	Acenaphthylen	152	4.149	1.08	123.36	179.368	12.484	0.344
3	2-Nitrofluorene	211	5.111	1.79	155.31	216.612	35.457	6.056
4	9-Nitroanthracene	223	5.320	1.41	163.88	221.375	29.52	5.203
5	Acenaphthene	154	4.445	1.11	128.60	181.959	12.87	0.551
6	Fluorene	166	4.612	1.57	137.57	197.282	18.090	0.368
7	Phenanthrene	178	4.815	1.46	146.01	204.450	20.345	0.009
8	Anthracene	178	4.809	1.57	146.06	206.446	22.429	0.008
9	1-Nitropyrene	247	6.065	1.08	177.94	234.615	38.384	6.232
10	3-Nitrofluoranthene	247	6.071	1.43	178.71	240.031	41.441	5.783
11	Fluoranthene	202	5.565	1.22	161.01	222.484	25.179	0.249
12	Pyrene	202	5.559	1.27	160.43	215.770	23.374	0.014
13	Chrysene	228	6.226	1.72	183.61	246.670	38.338	0.009
14	Benzo(a)anthracene	228	6.220	1.58	183.81	248.504	39.682	0.036
15	Benzo(b)fluoranthene	252	6.976	1.40	198.78	263.429	43.898	0.301
16	Benzo(k)fluoranthene	252	6.970	1.48	198.87	265.719	47.769	0.132
17	Benzo(a)pyrene	252	6.970	1.50	198.15	258.456	42.459	0.030
18	Dibenz(a,h)anthracene	278	7.631	1.79	221.58	267.413	65.724	0.009
19	Benzo(ghi)perylene	276	7.720	1.12	212.63	269.701	43.888	0.064
20	Indeno(1,2,3-cd)pyrene	276	7.720	1.40	213.23	274.093	50.286	0.470

[a]χ, molecular connectivity.
[b]L/B, length-to-breadth ratio.
[c]Connolly surface area.
[d]PI, the largest principal moment of inertia ($\times 10^{-38}\,g\,cm^2$).

Table 3. Training set members selected by D-optimal design.

Descriptor	Training set members										
MW[a]	1	4	5	6	7	8	14	16	18	19	20
χ[b]	1	2	5	6	8	15	16	17	18	19	20
Vol[c]	1	2	5	6	7	15	16	17	18	19	20
Area[d]	1	2	5	6	7	15	16	17	18	19	20
PI[e]	1	3	5	6	7	8	15	16	18	19	20
χ Dip[f]	1	3	4	5	9	10	16	17	18	19	20

[a]MW, molecular weight.
[b]χ, molecular connectivity.
[c]Vol, volume.
[d]Area, Connolly surface area.
[e]PI, the largest principal moment of inertia ($\times 10^{-38}$ g cm^2).
[f]Dip, dipole moment.

these models cannot predict the retention times of nitrated PAHs (2-nitrofluorene, 9-nitrofluoranthene, 1-nitropyrene, and 3-nitrofluoranthene). In the case of the models derived by using each of molecular weight and the largest principal moment of inertia, r^2 values in training sets were relatively low but the predictivity in test sets was somewhat higher than others (Table 4). It is also shown that these models are not able to predict retention times of nitrated PAHs in test sets. The moment of inertia ($I_a = \sum_n m_i r_i^2$) can be expressed as a function of mass (m_i), so it generally increases with increasing retention time. It was reported that molecular connectivity is linearly related to gas chromatographic retention indices of nitrated PAHs and can be used to predict gas chromatographic retention characteristics of nitrated PAHs.[3,4] But in this study, molecular connectivity may not be sufficient to predict retention times of HPLC because both nitrated and non-nitrated PAHs are contained in the data set. The L/B had little relationships with retention time (Table 2), because the molecular weights of PAHs in a training set are very different. The L/B is capable of differentiating retention for a set of isomeric PAHs but not for combinations of PAHs with different molecular weights.[20]

Two-Descriptors Multiple Linear Regression Models

The number of training set members was set to 11, the same number as one-descriptor models. Prior to D-optimal design, the input descriptors were selected by using crossvalidation. Since molecular connectivity and dipole moment showed the highest crossvalidated r^2 values in all conditions,

Table 4. Statistical summary of the MLR models.

Conditions	1	2	3	4	5	6	7
			Molecular weight				
r^2	0.871	0.878	0.880	0.860	0.868	0.864	0.862
$q^{2\,a}$	0.825	0.840	0.848	0.804	0.819	0.821	0.822
F^{b}	60.485	64.913	65.675	55.384	59.147	57.227	56.066
s^{c}	3.309	2.118	2.984	1.952	1.457	4.412	2.644
pred $r^{2\,d}$	−0.017	0.118	0.316	−0.193	−0.068	0.112	0.185
sdep[e]	5.783	3.741	5.020	3.476	2.625	7.513	4.337
			Molecular connectivity				
r^2	0.987	0.992	0.994	0.977	0.983	0.992	0.994
q^2	0.976	0.985	0.990	0.959	0.970	0.984	0.989
F	681.040	1109.423	1421.583	373.827	517.654	1071.950	1449.998
s	1.083	0.569	0.729	0.820	0.542	1.135	0.584
pred r^2	−0.573	−0.235	0.315	−2.235	−0.854	−0.260	−0.005
sdep	4.423	2.764	3.398	3.261	2.035	5.787	3.243
			Volume				
r^2	0.978	0.983	0.989	0.962	0.969	0.985	0.985
q^2	0.965	0.973	0.983	0.942	0.952	0.976	0.977
F	390.370	506.048	822.504	229.802	279.994	582.543	608.444
s	1.433	0.845	0.963	1.044	0.736	1.544	0.903
pred r^2	−1.257	−0.806	−0.050	−2.217	−1.617	−0.793	−0.462
sdep	5.205	3.282	4.139	3.199	2.377	6.799	3.856

			Surface area				
r^2	0.955	0.965	0.980	0.936	0.947	0.965	0.970
q^2	0.931	0.948	0.971	0.903	0.920	0.948	0.955
F	188.564	250.701	449.167	130.596	160.768	251.378	289.899
s	2.037	1.190	1.297	1.366	0.961	2.327	1.299
pred r^2	−1.629	−1.120	−0.207	−2.766	−2.049	−1.070	−0.688
sdep	5.618	3.556	4.439	3.462	2.566	7.305	4.142
			Largest moment of inertia				
r^2	0.788	0.783	0.774	0.766	0.765	0.789	0.773
q^2	0.707	0.696	0.679	0.676	0.671	0.710	0.684
F	33.449	32.404	30.892	29.507	29.210	33.644	30.708
s	4.362	2.931	4.241	2.598	2.013	5.673	3.491
pred r^2	0.195	0.318	0.484	0.012	0.137	0.296	0.365
sdep	4.456	2.869	3.885	2.711	2.026	5.900	3.420
			Molecular connectivity, dipole moment				
r^2	0.993	0.997	0.998	0.986	0.991	0.997	0.997
q^2	0.983	0.993	0.996	0.970	0.980	0.993	0.995
F	592.213	1442.790	1765.132	283.940	427.673	1499.746	1520.333
s	0.767	0.327	0.423	0.630	0.397	0.635	0.376
pred r^2	0.770	0.867	0.978	0.552	0.693	0.858	0.934
sdep	2.378	1.247	0.795	1.748	1.165	2.592	1.066

[a]q^2, crossvalidated r^2 in training set.
[b]F, F-value in training set.
[c]s, standard error in training set.
[d]pred r^2, predictive r^2 in test set.
[e]sdep, standard deviation of error of prediction in test set.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

D-optimal designs were performed with using these descriptors. The cross-validated r^2 values were 0.966, 0.981, 0.991, 0.940, 0.957, 0.980, and 0.988 for conditions 1, 2, 3, 4, 5, 6, and 7, respectively, and condition 3 had the greatest value. The results of MLR derived by using molecular connectivity and dipole moment suggest that these descriptors have very high linear correlations with retention times. The r^2 values in training sets were 0.993, 0.997, 0.998, 0.986, 0.991, 0.997, and 0.997 for conditions 1, 2, 3, 4, 5, 6, and 7, respectively, and condition 3 had the greatest r^2 value of 0.998 (Table 4). The best model (condition 3) is

$$\begin{aligned} \mathrm{RT} &= 5.748[\chi] - 0.917[\mathrm{Dip}] - 9.049 \\ &\quad (\pm 0.113) \quad (\pm 0.057) \qquad (\pm 0.754) \\ N &= 11;\ r^2 = 0.998;\ F = 1765.132 \end{aligned}$$

where χ is molecular connectivity and Dip is dipole moment. HPLC retention time generally increased with increasing molecular connectivity. For example, the short retention time (10.305 in condition 3) of naphthalene correlates with small molecular connectivity (3.405), while the longer retention time (30.105 in condition 3) of indeno(1,2,3-cd)pyrene correlates with larger molecular connectivity (7.720). The correlations between experimental and calculated retention times of training and test sets in condition 3 are plotted in Fig. 2, which shows that the two-descriptors models had high predictivity. The predictive r^2 values in test sets were 0.770, 0.867, 0.978, 0.552, 0.693, 0.858, and 0.934 for conditions 1, 2, 3, 4, 5, 6, and 7, respectively. These results show that the predictivity in test sets had large differences with chromatographic conditions, in spite of a little difference in r^2 values with the chromatographic conditions in training sets. The differences in the chromatographic conditions are the kinds of solvent, the solvent concentration, and gradient elution (Table 1). Conditions 1, 2, 4, and 5 used acetonitrile (ACN) as a solvent and conditions 3, 6, and 7 used methanol (MeOH), of which polarity is weaker than that of ACN. It could be observed that the predictivity is generally higher in MeOH than in ACN (Tables 1 and 4). Moreover, very high predictive r^2 values were obtained in conditions 3 and 7, in which solvent is MeOH and gradient elution is 1%/min. This shows good agreement with the experiment, in which using MeOH as a solvent gives better separation than using ACN.[15]

Compared with one-descriptor models, the r^2 values in training sets were also very high. For example, the one-descriptor model derived by using molecular connectivity showed a high r^2 value of 0.994 in condition 3 and the two-descriptors model derived by using molecular connectivity and dipole moment showed a value of 0.998 (Table 4). The difference between the models is only 0.004 in a training set. However, the predictivity in test sets was highly improved when two descriptors were used. The MLR models derived by using molecular connectivity and dipole moment showed good predictivity in test

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

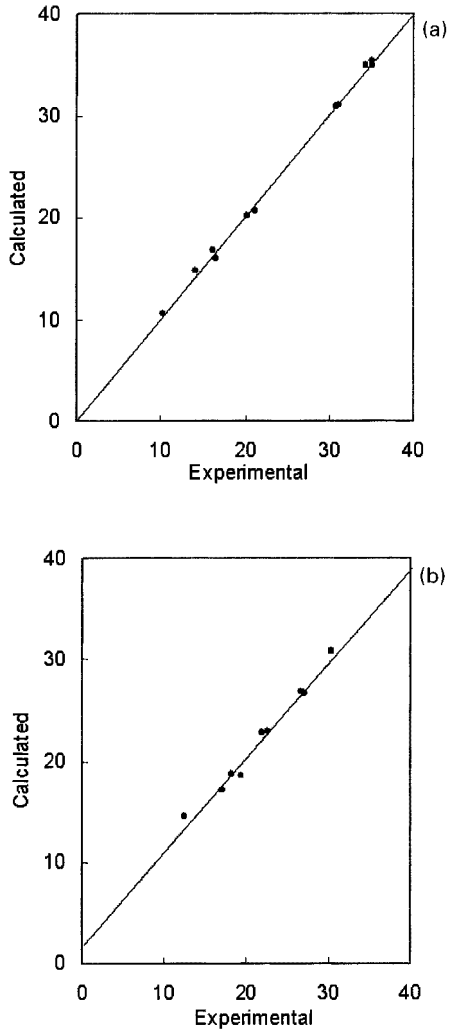

Figure 2. Plots of experimental vs. calculated retention time in condition 3 (two descriptors of molecular connectivity and dipole moment); (a) training set, (b) test set.

sets (r^2 values from 0.552 to 0.978), whereas the predictive r^2 values were almost negative in one-descriptor models (Table 4). This lack of predictivity in one-descriptor models is caused from the failure in prediction of retention times of nitrated PAHs (2-nitrofluorene, 9-nitrofluoranthene, 1-nitropyrene, and 3-nitrofluoranthene). It is noticeable that the dipole moment has a great effect

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

on predictivity cooperatively with molecular connectivity although r^2 values of dipole moment alone are very low (r^2 values of almost zero).

CONCLUSION

In this study, MLR was carried out with two kinds of models using one and two descriptors to obtain the factors explaining HPLC retention times of PAHs. The D-optimal design was tried for selection of training sets. It could be found that two descriptors of molecular connectivity and dipole moment had highly linear correlations with retention times in training sets and showed good predictivity in test sets. The predictivity was mainly affected by the solvent. High predictive r^2 values were obtained in conditions in which MeOH was used as a solvent. In the case of one-descriptor models, the r^2 values in training sets were very high but the predictivity in test sets was very poor. These results imply that MLR models achieved in training sets have to be validated or tested.

ACKNOWLEDGMENTS

This work was supported by the National Research Laboratory program of the Ministry of Science & Technology, Korea.

REFERENCES

1. Hansch, C.; Fujita, T. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. **1964**, *86*, 1616–1626.
2. Sneath, P.H.A. Relationships between chemical structure and biological activity in peptides. J. Theor. Biol. **1966**, *12*, 157–195.
3. Doherty, P.J.; Hose, R.M.; Robbat, Jr. A. Relationship between gas chromatographic retention indices and molecular connectivities of nitrated polycyclic aromatic hydrocarbons. Anal. Chem. **1984**, *56*, 2697–2701.
4. Robbat, A., Jr.; Corso, N.P.; Doherty, P.J.; Marshall, D. Multivariate relationships between gas chromatographic retention index and molecular connectivity of mononitrated polycyclic aromatic hydrocarbons. Anal. Chem. **1986**, *58*, 2072–2077.
5. Coleska, Z.N.; Suturkova, L.; Dorevski, K.; Krbavcic, A.; Solmajer, T. Quantitative structure-activity relationship of flavonoid inhibitors of $p56^{lck}$ protein tyrosine kinase: a classical/quantum chemical approach. Quant. Struct.—Act. Relat. **1998**, *17*, 7–13.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

6. Grimmer, G. Ed. *Environmental Carcinogens, Polycyclic Aromatic Hydrocarbons, Chemistry, Occurrence, Biochemistry, Carcinogenecity*; CRC Press Inc.: Boca Raton, FL, 1983.
7. Taylor, P.; Dellinger, B.; Lee, C.C. Development of a thermal stability-based ranking of hazardous organic compound incinerability. Environ. Sci. Technol. **1990**, *24*, 316–328.
8. Grimmer, G.; Naujac, K.W.; Dettbarn, G.; Brune, H.; Deutsch-Wenzel, R.; Misfeld, J. In *Analysis of Balance of Carcinogenic Impact from Emission Condensates of Automobile Exhaust, Coal Heating, and Used Engine Oil by Mouse Skin Painting as a Carcinogen Specific Detector*, Polynucleic Aromatic Hydrocarbons: Physics, Biology, and Chemistry 6th International Symposium; Cooke, M., Dennis, A.J., Fisher, G.L., Eds.; Bartelle Press: Columbus, OH, 1981; 335 pp.
9. Freeman, D.J.; Cattell, F.C.R. Woodburning as a source of atmospheric polycyclic aromatic hydrocarbons. Environ. Sci. Technol. **1990**, *24*, 1179–1186.
10. Fromme, H.; Oddoy, A.; Piloty, M.; Krause, M.; Lahrz, T. Polycyclic aromatic hydrocarbon (PAH) and diesel engine emission (elemental carbon) inside a car and a subway train. Sci. Total Environ. **1998**, *217*, 165–173.
11. Geacintov, N.E. In *Reaction pathways of benzo[a]pyrene diol epoxide and DNA*, Polynucleic Aromatic Hydrocarbons: Physics, Biology, and Chemistry 6th International Symposium; Cooke, M., Dennis, A.J., Fisher, G.L., Eds.; Bartelle Press: Columbus, OH., 1981; 311 pp.
12. Hodgson, R.M.; Cary, P.D.; Grover, P.L.; Sims, P. Metabolic activation of chrysene by hamster embryo cells; evidence for the formation of a "bay-region" diol-epoxide-N^2-guanine adduct in RNA. Carcinogenesis (London) **1983**, *4*, 1153.
13. Wang, Y.Y.; Rappaport, S.M.; Sawyer, R.F.; Talcott, R.E.; Wei, E.T. Direct-acting mutagens in automobile exhaust. Cancer Lett. **1978**, *5*, 39–47.
14. Mermelstein, R.; Kiriazides, D.K.; Butler, M.; McCoy, E.C.; Rosenkranz, H.S. The extraordinary mutagenicity of nitropyreres in bacteria. Mutant Res. **1981**, *89*, 187–196.
15. Hong, J.E. *Study on the Analysis of Polycyclic Aromatic Hydrocarbons by Gas Chromatography/Mass Spectroscopy and Liquid Chromatography*, Department of Chemistry, Kyunghee University, 1996; Master Dissertation.
16. Sabljić, A.; Horvatić, D. GRAPH III: a computer program for calculating molecular connectivity indices on microcomputers. J. Chem. Inf. Comput. Sci. **1993**, *33*, 292–295.
17. Wise, S.A.; Bonnett, W.J.; Guenther, F.R.; May, W.E. A relationship between reverse-phase C18 liquid chromatographic retention and the

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

shape of polycyclic aromatic hydrocarbons. J. Chromatogr. Sci. **1981**, *19*, 457–465.
18. Mitchell, T.J. An algorithm for the construction of "D-optimal" experimental designs. Technometrics **1974**, *16*, 203–210.
19. Myers, R.H. *Classical and Modern Regression with Applications*; PWS/KENT: Boston, 1990.
20. Collantes, E.R.; Tong, W.; Welsh, W.J.; Zielinski, W.L. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. Anal. Chem. **1996**, *68*, 2038–2043.

Received April 27, 2003
Accepted May 24, 2003
Manuscript 6162

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016