QSRR Prediction of the Chromatographic Retention Behavior of Painkiller Drugs

Jahanbakhsh Ghasemi[1,*] and **Saadi Saaidpour**[2]

[1]Chemistry Department, Faculty of Sciences, K.N. Toosi University, of Technology, Tehran, Iran and [2]Chemistry Department, Faculty of Sciences, Razi University, Kermanshah, Iran

Abstract

Quantitative structure–retention relationship (QSRR) analysis is a useful technique capable of relating chromatographic retention time to the chemical structure of a solute. A QSRR study has been carried out on the reversed-phase high-performance liquid chromatography retention times (log t_R) of 62 diverse drugs (painkillers) by using molecular descriptors. Multiple linear regression (MLR) is utilized to construct the linear QSRR model. The applied MLR is based on a variety of theoretical molecular descriptors selected by the stepwise variable subset selection procedure. Stepwise regression was employed to develop a regression equation based on 50 training compounds, and predictive ability was tested on 12 compounds reserved for that purpose. The geometry of all drugs was optimized by the semi-empirical method AM1 and used to calculate different molecular descriptors. The regression equation included three parameters: *n*-octanol–water partition coefficient (log *P*), molecular surface area, and hydrophilic–lipophilic balance of the drug molecules, all of which could be related to retention time property. Modeling of retention times of these compounds as a function of the theoretically derived descriptors was established by MLR. The results indicate that a strong correlation exists between the log t_R and the previously mentioned descriptors for drug compounds. The prediction results are in good agreement with the experimental values.

Introduction

In the present study, we have selected some diverse drugs with different activity classifications such as antipyretic, antipsychotic, hypnotic, anticonvulsant, tranquilizer, antidepressant, antiparkinsonian, and other. These drugs were selected according to application and consumption for patients as painkillers. The fundamental processes of pharmacokinetics and pharmacodynamics (absorption, distribution, excretion, and receptor activation) are similar to the processes governing chromatographic separations. The same basic intermolecular interactions determine the behavior of chemical compounds in both the biological and chromatographic environments (1).

The retention is a measure of the speed at which a substance moves in a chromatographic system. In continuous development systems like high-performance liquid chromatography (HPLC) or gas chromatography (GC), where the compounds are eluted with the eluent, the retention is usually measured as the retention time t_R, the time between injection and detection. Reversed-phase high-performance liquid chromatography (RP-HPLC) consists of a non-polar stationary phase and a moderately polar mobile phase. One common stationary phase is silica which has been treated with RMe_2SiCl, where R is a straight chain alkyl group such as $C_{18}H_{37}$ or C_8H_{17}. The t_R is therefore longer for molecules which are more non-polar in nature, allowing polar molecules to elute more readily. The characteristics of the analyte molecule play an important role in its retention characteristics. In general, an analyte with a longer alkyl chain length results in a longer t_R because it increases the molecule's hydrophobicity (2–6).

Chemometric processing of chromatographic data can reveal systematic information both about the analytes (retention, physicochemical properties, and relative biological activity) and about the stationary phases studied (the molecular mechanism of separation operating in a particular chromatographic system, quantitative comparison of the retention properties of different stationary phases, the structural descriptors most suitable for predicting retention) (7).

Usually, molecular descriptors are divided into several classes, depending on their origin of calculation or on the structural item in the chemical structure (molecule, atom, or chemical bond). We classified, according to the origin of calculation, the most useful and famous descriptors and divided these into six conditional categories. Therefore, the present classification contains the following parameters: constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic. The descriptors that are created from the structure are believed to encode all the interactions that are responsible for the distribution of solutes on an HPLC or GC column. The solute–solute interactions, solute–stationary phase interactions, and solute–mobile phase interactions must be numerically encoded in order for a quantitative structure–retention relationship (QSRR) to be effective. If the mobile phase and stationary

* Author to whom correspondence should be addressed: email Jahan.ghasemi@gmail.com.

phase are the same for every solute, then only the differences in the structures of the solute molecules need to be encoded. Thus, all the numerical descriptors were derived from the chemical structures of the solute molecules (8–16).

A QSRR approach, as one of the all-important areas in modern chemical science, gives knowledge that is practical and necessary for drug design, combinatorial, and medicinal chemistries. HPLC is one of the most frequently used separation techniques in analytical chemistry. Of the liquid chromatographic techniques, RP-HPLC is the most popular. Retention behavior of solutes for this type of chromatography depends mainly on the type of nonpolar stationary phase and on the composition of the polar mobile phase. Retention mechanisms are often described by the difference in various solute hydrophobic and electronic interactions with both the stationary and mobile phases. RP-HPLC has been widely recognized as a valuable method for the extraction and quantitation of information about the structure and physicochemical properties of organic compounds. The concept of QSRR was reviewed by Kaliszan in 1987 (1). Numerous QSRR studies aimed at comparison of the retention mechanisms on alkyl silica reversed-phase materials for HPLC have employed several physically interpretable descriptors such as various parameters of hydrophobicity, polarity, hydrogen bonding ability, etc. (17–25). Buydens and Massart used the complete overlap differential method to correlate the retention index and topological, physicochemical, and quantum-chemically calculated electronic parameters, using multiple regression and factor analysis (26). Rohrbaugh and Jurs developed a four-descriptor QSRR model with multiple correlation coefficient of 0.997 for 86 alkenes (27). Bermejo and Guillen studied the relationships between retention indices and parameters related to electronic polarizability, such as molar refraction, refractive index, Van der Waals volume, and molar volume of alkenes (28). Voelkel correlated the retention indices of 85 alkenes with connectivity index, dipolar moment, and polarizability parameters, and also considered using multilinear regression (29). Hu and Zhang also developed a QSRR model for alkenes using solubility parameters, molar volumes, and number of carbon atoms (dummy descriptors) (30). An important step favored in this field is the work by Heinzen, Soares, and Yunes, who proposed a semi-empirical topological method for the prediction of the chromatographic retention of *cis*- and *trans*-alkene isomers and alkanes (31).

As a result, there is increasing interest within the chromatography community in the development of QSRR models based on linear or nonlinear modeling techniques, including principal component regression (32), multiple linear regression (MLR) (33), partial least-squares (34), support vector machine (35), and artificial neural networks (36,37).

In our previous papers, we reported on the application of quantitative structure property/activity relationship (QSPR/QSAR) techniques in the development of a new, simplified approach to the prediction of compounds' properties using different models (38–42). In this study, the MLR technique was used for modeling the RP-HPLC t_R data of 62 drugs. The predictive power of the resulting model is demonstrated by testing them on unseen data that were not used during model generation. A physicochemical explanation of the selected descriptors is also given.

Materials and Methods

The QSRR model for the estimation of the t_R of various drug compounds is established in the following six steps: molecular structure input and generation of the files containing the chemical structures stored in a computer-readable format; quantum mechanics geometry optimization with a semi-empirical (AM1) method; structural descriptor computation; structural descriptor selection; structure-retention model generation with the MLR method; and statistical analysis.

Data set

Apparatus and analysis conditions for RP-HPLC t_R data are shown in Table I. T_R of 62 drug compounds were taken from the Toyohashi University of Technology website (43), and are presented in Table II. These values were converted from t_R (min) to logarithm of retention time (log t_R). The data set was split into a training set and a prediction set. The training set of 50 compounds was used to adjust the parameters of the models, and the test set of 12 compounds was used to evaluate its prediction ability.

Computer hardware and software

All calculations were run on an HP Pavilion dv6000 laptop computer with AMD Turion64 X2 Mobile Technology CPU running Windows XP operating system. The ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft Corporation; Cambridge, MA) software was used for drawing the molecular structures (44). The optimizations of molecular structures were done by the MOPAC 7.0 (AM1 method) (45), and descriptors were calculated by Molecular Modeling Pro Plus (MMPP) Version 6.0 (ChemSW, Inc.; Fairfield, CA) software (46). A stepwise procedure was used for selection of descriptors using the SPSS/PC software package (SPSS Inc.; Chicago, IL) (47). MLR was performed by using a routine from the Unscrambler version 7.6 package (CAMO Process; Trondheim, Norway) (48), and other calculations were performed in the MATLAB (version 7.0, MathWorks, Inc.; Natick, MA) environment.

Molecular modeling and theoretical molecular descriptors

The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds. In order to

Table I. Apparatus and Analysis Conditions for RP-HPLC Retention Time Data

1	Mobile phase: (10mM $HClO_4$ + 10mM $NaClO_4$ 70%) + (CH_3CN 30%)
2	Flow-rate: 1.0 mL/min.
3	Column: FineSIL C18T (25 cm × 4.0 mm i.d.) [monomeric octadecyl silica (ODS), particle size 5 × 10–6 m] (Jasco; Hachioji, Japan)
4	Wavelength: 210–350 nm
5	Column temperature: 50°C
6	880 PU LC pump (Jasco)
7	System controller 801-SC
8	Gradient device 880-02
9	Detector MULTI-320
10	Data processing system DP-L320/98 (Jasco) (Time accumulation 0.8 s)

calculate the theoretical descriptors, molecular structures were constructed with the aid of ChemDraw Ultra version 9.0, and optimized using AM1 algorithm (49). The computational chemistry software Chem3D Ultra version 9.0 with MOPAC was used to build the molecules and perform the necessary geometry optimizations. A gradient cutoff of 0.01 was used for all geometry optimizations. We have chosen descriptors associated with the neutral molecules of drugs in our calculations. As a result, a total of 54 theoretical descriptors were calculated for each compound in the data sets (62 compounds) by MMPP Version 6.0 (ChemSW, Inc.) software.

Stepwise regression for descriptor selection

The selection of relevant descriptors which relate the t_R to the molecular structure is an important step in constructing a predictive model. In this work, stepwise MLR was used as the feature selection method to select the best-calculated descriptors among 54 theoretical descriptors using MMPP software. All descriptors with zero values or constant and near constant values for all the molecules in the data set were eliminated. The correlation matrix was calculated between the descriptors; one of the two descriptors with a pairwise correlation coefficient above 0.6 ($r >$ 0.6) and a large correlation coefficient with the other descriptors was eliminated.

In order to select the subset of descriptors that best explain drug t_R, we used stepwise regression (50–52). This method combines both forward and backward procedures. Stepwise model-building techniques for regression designs with a single dependent variable involve identifying an initial model, repeatedly altering the model from the previous step by adding (forward stepwise) or removing (back stepwise) a predictor variable, and terminating the search when stepping does not further improve the model. The forward stepwise method employs a combination of the forward entry of independent variables and backward removal of insignificant variables. The best single predictor, which is the most significant variable, was used for the initial linear regression step. Next, descriptors were added one at a time, always adding the one that most improved the fit, until the fit was not significantly improved. Once all the significant variables were determined, the regression equation was constructed. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables from the model.

By using these criteria, 51 out of 54 original descriptors were

Table II. Experimental Retention Times and Molecular Descriptor Values for 62 Drug Compounds

No.	Drug	log t_R	Log *P*	SM	HLB	No.	Drug	log t_R	Log *P*	SM	HLB
1	Acetaminophen	0.38	0.49	10.72	14.95	32	Flurazepaml	1.13	3.68	23.30	4.54
2	Acetylpheneturide*	0.89	0.97	20.20	11.24	33	Glutethimide*	0.97	2.05	16.70	8.54
3	Acetylsalicylic acid	0.62	1.39	13.00	13.18	34	Haloperidol	1.19	3.49	24.20	5.11
4	Alprazolam*	1.18	2.47	19.75	8.00	35	Haloxazolam	0.65	2.00	17.20	10.40
5	Amitriptyline	1.43	4.14	20.96	0.00	36	Hydroxyzine	1.35	4.31	25.70	5.9
6	Amobarbital	0.89	2.06	17.67	7.74	37	Imipramine	1.35	3.85	20.88	0.00
7	Barbital	0.48	0.39	13.63	11.05	38	Levomepromazine	1.45	3.35	23.04	0.00
8	Biperiden*	1.29	4.21	23.24	1.44	39	Maprotyline	1.35	5.12	19.77	3.82
9	Bromazepam	0.66	1.93	16.21	10.16	40	Medazepam	1.11	4.31	18.70	2.67
10	Bromocriptine	1.50	3.53	27.90	4.00	41	Mephobarbital	0.92	1.54	16.96	9.98
11	Bromperidol*	1.25	2.46	25.10	4.97	42	Metharbital	0.66	0.61	13.63	9.80
12	Bromvalerylurea	0.68	–0.20	13.60	8.55	43	Mianserin*	1.08	4.33	17.95	1.14
13	Caffeine	0.44	–1.64	13.62	14.28	44	Nimetazepam	1.07	1.96	18.36	7.53
14	Carbamazepine	0.95	2.20	15.80	8.28	45	Nitrazepam	0.85	1.68	16.90	11.37
15	Carpipramine*	1.16	4.49	23.80	6.01	46	Nortriptyline	1.35	4.48	20.70	4.48
16	Chlordiazepoxid	0.77	2.40	18.72	7.50	47	Oxazepam	0.98	2.69	16.57	8.47
17	Chlormezanone	0.81	2.32	16.20	10.76	48	Pentobarbital	0.87	1.84	17.67	7.23
18	Chlorpromazine	1.54	4.43	22.50	0.07	49	Perphenazine	1.25	4.20	24.70	7.24
19	Clocapramine	1.42	5.55	27.10	4.80	50	Phenacetin	0.74	1.61	13.54	7.51
20	Clofedanol	0.98	3.11	20.59	0.18	51	Phenobarbital	0.66	1.38	15.52	11.82
21	Clomipramine*	1.62	4.57	24.20	0.00	52	Phenytoin	0.88	1.98	15.83	10.54
22	Clonazepam*	1.07	2.39	18.02	9.02	53	Primidone	0.51	–0.62	15.35	11.71
23	Clotiazepam	1.11	2.30	20.28	6.15	54	Promethazine	1.23	3.40	19.52	0.00
24	Cloxazolam*	0.68	2.32	17.90	10.24	55	Propericyazine	1.18	2.05	23.60	7.07
25	Desipramine	1.27	3.21	19.45	6.08	56	Secobarbital	0.96	1.83	18.51	8.50
26	Diazepam	1.09	2.93	17.46	4.76	57	Sulpiride*	0.44	–2.69	15.00	12.34
27	Estazolam	1.06	2.77	18.00	8.40	58	Timiperone	0.99	3.40	21.10	7.6
28	Ethenzamide*	0.71	0.91	14.10	11.97	59	Triazolam	1.29	4.06	19.52	7.08
29	Etizolam	1.31	4.21	20.44	5.15	60	Trihexyphenidyl	1.38	4.33	23.87	1.3
30	Fludiazepam	1.32	3.08	19.40	4.59	61	Trimethadione	0.57	–1.97	12.30	10.16
31	Flunitrazepam	1.17	2.11	18.83	7.30	62	Trimipramine	1.47	4.25	22.24	0.0

* Data set for prediction.

eliminated and the remaining descriptors were used to generate the models using the SPSS/PC software package. The result shows that the three calculated descriptors are the most feasible ones. The selected descriptors are *n*-octanol–water partition coefficient (log *P*), molecular surface area (SM), and hydrophilic-lipophilic balance (HLB) (53–58).

MLR

The general purpose of multiple regressions is to quantitate the relationship between several independent or predictor variables and a dependent variable. A set of coefficients defines the single linear combination of independent variables (molecular descriptors) that best describes drug t_R. The t_R value for each drug would then be calculated as a composite of each molecular descriptor weighted by the respective coefficients. A multilinear model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \qquad \text{Eq. 1}$$

where k is the number of independent variables, $\beta, \ldots, \beta$ are the regression coefficients and y is the dependent variable. Regression coefficients represent the independent contributions of each calculated molecular descriptor. The algebraic MLR model is defined in equation 1 and in matrix notation:

$$y = Xb + e \qquad \text{Eq. 2}$$

When X is of full rank, the least squares solution is: $\hat{b} = (X^T X)^{-1} X^T y$, where $\hat{b}$ is the estimator for the regression coefficients in $\hat{b}$.

A single MLR model was developed for drug compounds using the Unscrambler version 7.6 software. MLR model was constructed with remaining descriptors based on stepwise feature selection. The MLR model was built using a training set and validated using an external prediction set. MLR techniques based on least-squares procedures are often used for estimating the coefficients involved in the model equation (59).

Results and Discussion

QSRRs describe the effects of an analyte's molecular structure on chromatographic t_R, and also explore the mechanisms of absorption in stationary phase and elution from the column. Chromatographic t_R in RP-HPLC is affected by a large number of system variables which may be divided into four groups. The first group consists of variables characterizing the stationary phase; for instance, the type of stationary phase material, column diameter, and column length. The second group consists of physical variables such as temperature, pressure, and mobile phase flow. A third group consists of variables defining the mobile phase composition. The fourth group consists of different molecular descriptors of solutes. In this study, the first three groups of variables are constant. Systematic modeling of retention is basically concerned with the search for the relation between t_R and this fourth group of variables.

All descriptors were calculated for the neutral species. The log t_R is assumed to be highly dependent upon the log *P*, SM, and HLB. Linear correlations are observed between log t_R and molecular descriptors of solutes that are not homologues. The correlation coefficients between experimental log t_R and the log *P*, SM, and HLB are 0.85, 0.84 and –0.84, respectively. We used these descriptors to generate linear QSRR for the RP-HPLC t_R of a set of drugs.

MLR analysis

The software package used for conducting MLR analysis was Unscrambler 7.6. MLR analysis has been carried out to derive the best QSRR model. The MLR technique was performed on the molecules of the training set shown in Table II. After regression analysis, a few suitable models were obtained among which the best model was selected and presented in equation 3. A small number of molecular descriptors (log *P*, SM, and HLB) proposed were used to establish a QSRR model. Additional validation was performed on an external data set consisting of a 12-drug compound. MLR analysis provided a useful equation that can be used to predict the log t_R of drugs based upon these parameters. The best equation obtained for the retention time of the drug compounds is:

$$\log t_R = 0.4904 + 0.0372 \log P + 0.0278 SM - 0.0261 HLB \qquad \text{Eq. 3}$$
$$n = 50,\ R^2 = 0.94,\ R^2_{adj} = 0.88,\ s^2 = 0.0123,\ F = 113.2$$

where n is the number of compounds used for regression, R^2 is the squared correlation coefficient, s^2 is the standard error of the regression, and F is the Fisher ratio for the regression. The squared correlation coefficient, R^2, is a measure of the fit of the regression model. Correspondingly, it represents the part of the variation in the experimental data that is explained by the model; the higher the value of correlation coefficient, the better the model. The correlation coefficient values closer to 1 represent the better fit of the model. The s^2 is the standard error measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line. Thus, the standard error measures the model error. If the model is correct, it is an estimate of the error of the data variance, s^2. The F-test reflects the ratio of the variance explained by the model and the variance due to the error in the model, and high values of the F-test indicate that the model is statistically significant.

Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of log t_R, whereas negative values indicate that the greater the value of the descriptor, the lower the value of log t_R. In other words, increasing the HLB will decrease log t_R and increasing the log *P* and SM increases the extent of log t_R of the drug's organic compounds. The order of significance of the descriptors is: log *P* > SM > HLB.

In the present study, the QSRR model was generated using a training set of 50 molecules. A test set of 12 molecules (Table II) with regularly distributed log t_R values was used to assess the predictive ability of the QSRR model produced in the regression. For evaluation of the predictive power of the generated MLR, the optimized model was applied for the prediction of log t_R values of 12 compounds in the prediction set which were not used in the optimization procedure. For the constructed model, the predic-

tive ability of the MLR model was evaluated by calculation of statistical parameters. The predicted values of log t_R, residuals, and the percent relative errors (%RE) of prediction obtained by the MLR method are presented in Table III. The plots of predicted log t_R versus experimental log t_R; the residuals (experimental log t_R–predicted log t_R) versus experimental log t_R value obtained by the MLR modeling; and the random distribution of residuals about zero mean are shown in Figure 1. The stability and validity of the model was tested by prediction of the response values for the prediction set. This model is applicable for prediction of log t_R from 0.38 (2.41 min) to 1.62 (41.65 min). The average %RE of prediction and R^2 are –3.28% and 0.90 for the MLR model, respectively.

The effect of the selected descriptors on the t_R

The QSRR developed indicated that log *P*, SM, and HLB significantly influence drug t_R. The effect of each selected descriptor on the t_R can be interpreted according to the entity and type of descriptors. In the following section, these effects are explained with regard to the values of the coefficients of each descriptor in the MLR model presented in equation 3.

Table III. Experimental log t_R, Predicted log t_R, Residuals, and Percent Relative Error Values for Train and External Test Sets

	Train set								
No.	log t_R (exp.)	log t_R (pred.)	Residuals	RE%	No.	log t_R (exp.)	log t_R (pred.)	Residuals	RE%
1	0.38	0.43	–0.05	13.61	44	1.23	1.28	–0.05	4.06
3	0.44	0.38	0.06	–14.35	45	1.25	1.30	–0.05	3.76
5	0.48	0.61	–0.13	27.67	46	1.25	1.24	0.01	–1.04
6	0.51	0.57	–0.06	11.42	47	1.27	1.11	0.16	–12.69
7	0.62	0.61	0.01	–1.13	48	1.30	1.15	0.15	–11.58
9	0.65	0.84	–0.19	29.45	49	1.31	1.23	0.08	–5.88
10	0.66	0.82	–0.15	23.04	50	1.32	1.14	0.18	–13.82
12	0.66	0.66	0.01	–0.90	51	1.35	1.37	–0.02	1.41
13	0.66	0.71	–0.05	7.53	52	1.35	1.32	0.03	–2.37
14	0.68	0.63	0.05	–7.89	53	1.35	1.35	0.00	0.30
16	0.74	0.79	–0.05	6.19	54	1.35	1.28	0.08	–5.69
17	0.77	0.99	–0.22	28.87	55	1.38	1.44	–0.06	4.21
18	0.81	0.83	–0.02	2.47	56	1.42	1.53	–0.10	7.16
19	0.85	0.79	0.07	–7.85	58	1.43	1.38	0.05	–3.50
20	0.87	0.93	–0.06	7.04	59	1.45	1.38	0.07	–4.97
23	0.88	0.80	0.08	–9.19	60	1.47	1.42	0.05	–3.60
25	0.89	0.93	–0.04	4.96	61	1.50	1.42	0.08	–5.33
26	0.92	0.81	0.11	–11.43	62	1.54	1.44	0.10	–6.43
27	0.95	0.88	0.07	–7.70	No.	Test set			
29	0.97	0.92	0.05	–4.97	2	0.89	0.83	0.06	–6.43
30	0.98	1.29	–0.31	31.90	4	1.18	1.01	0.17	–14.38
31	0.98	0.93	0.05	–5.50	8	1.29	1.41	–0.12	9.40
32	0.99	1.13	–0.14	14.42	11	1.16	1.32	–0.17	14.24
34	1.06	0.98	0.09	–8.19	15	1.62	1.50	0.12	–7.53
35	1.07	0.95	0.13	–11.73	21	1.07	0.93	0.14	–12.66
36	1.09	1.07	0.02	–2.20	22	0.68	0.79	–0.11	15.64
37	1.11	1.26	–0.15	13.14	24	0.71	0.64	0.08	–10.92
38	1.12	1.06	0.05	–4.75	28	0.97	0.88	0.09	–9.07
39	1.13	1.29	–0.16	13.76	33	1.08	1.18	–0.09	8.68
40	1.17	0.98	0.20	–16.62	43	0.44	0.39	0.05	–11.62
41	1.18	1.11	0.07	–5.92	57	0.57	0.48	0.08	–14.66
42	1.19	1.29	–0.09	7.89					

Log *P*

Hydrophobicity (lipophilicity) was the initial physicochemical property to be defined, and remains the only one for which methods of prediction have been developed and widely accepted in the field of pharmaceutical research and medicinal chemistry. Hydrophobicity is understood to be a measure of the relative tendency of a molecule "to prefer" a nonaqueous over an aqueous environment. The log *P* is a reference system that provides the most commonly recognized hydrophobicity measure: the common logarithm of the partition coefficient (log *P*). Log *P* has become the standard scale for hydrophobicity. Traditionally, experimental log *P* measurement involves dissolving a compound within a biphasic system comprised of aqueous and organic layers and then determining the molar concentration of the compound in each layer. The organic solvent used is typically, but not exclusively, 1-octanol. Lipophilicity is the measure of the partitioning of a compound between a lipidic and an aqueous phase. The partition coefficient, *P*, is the ratio between the concentration of a drug or other chemical substance in two phases: one aqueous, the other an organic solvent:

$$P = \frac{[\text{drug}]_{org}}{[\text{drug}]_{aq}} \qquad \text{Eq. 4}$$

where $(\text{drug})_{org}$ and $(\text{drug})_{aq}$ are the concentrations of the solute in organic and aqueous phases, respectively. The partition coefficients are usually transformed into a logarithmic form as:

$$\log P = \log \frac{[C]_{org}}{[C]_{aq}} \qquad \text{Eq. 5}$$

The logarithmic value of the *n*-octanol–water partition coefficient is called log *P* (53,54). The log *P* is frequently used to estimate the membrane permeability and bioavailability of compounds, because an orally-administered drug must be lipophilic enough to cross the lipid bilayer of the membranes, and on the other hand, must be sufficiently water-soluble to be transported in the blood and lymph. The log *P* is frequently used in QSRR as a measure of the lipophilic character of the molecules. Log *P* is used in QSRR studies and rational drug design as a measure of molecular hydrophobicity. Hydropho-bicity affects drug absorption, bioavailability, hydrophobic drug-receptor interactions, and metabolism of molecules, as well as their toxicity. Lipophilicity is approximately correlated to passive transport across cell membranes and the ability of a compound to partition through a membrane, because membranes are composed largely of lipids. Log *P* is well-established as a key parameter for

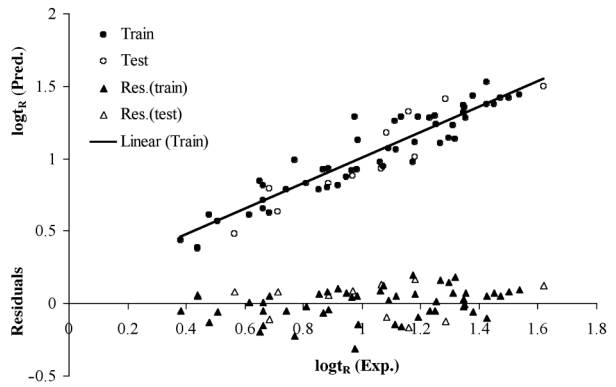

Figure 1. Plot of predicted log t_R and residuals estimated by MLR modeling versus experimental log t_R of drug compounds.

describing lipophilicity, uptake, and distribution in biological systems. With increased log *P*, hydrophobic interactions between the solutes and nonpolar stationary phase and t_R in RP-HPLC increase.

SM

The molecular volume and the SM (56) are used mostly as bulk/cavity terms. There is no unique way to define the van der Waals surface area or surface area, but most approaches try to define a surface contour similar to the van der Waals volume.

$$SM = \sum_i S_{VW}^{(i)} - S_{ov} \qquad \text{Eq. 6}$$

where $S^{(i)}_{VW}$ is van der Waals area of the *i*-th constituent atom of a molecule and S_{ov} is van der Waals area of atoms inside overlapping atomic envelopes.

Molecular surface area determines transport characteristics of molecules, such as intestinal absorption or blood-brain barrier penetration. Molecular surface area is therefore often used in QSRR studies to model molecular properties and t_R. The steric effects characterize bulk properties of a molecule and can be described with molecular surface area. The molecular surface area is clearly an important descriptor for log t_R. In order for a solute to enter into aqueous solution, a cavity must be formed in the solvent for the solute molecule to occupy. Water as a solvent would much prefer to interact with itself or other hydrogen bonding or ionic species than with a non-polar solute, so there is an increasing penalty (and thus higher log t_R) for larger solutes. Increasing molecular surface area leads to increasing cavity formation energy in water; the larger the solute, the greater the energy demand to make a cavity, and the lower the solubility in a polar mobile phase. In RP-HPLC, the dispersive interactions of large molecules are stronger with the bulky hydrocarbon chains of the stationary phase than with the small molecules of solutes. Apparently, increasing the SM increases the extent of log t_R of drugs' organic compounds.

HLB

A measure of the proportion of a molecule's mass is hydrophilic. A parameter of utmost importance in the development of pharmaceutical emulsions is the evaluation of their critical HLB. For nonionic molecules, the minimum value is 0 and the maximum value is 20; a number on the scale of one to 20 according to the HLB system, introduced by Griffin (57,58). The HLB system is based on the concept that some molecules have hydrophilic groups, other molecules have lipophilic groups, and some have both. Hydrophilic compounds have a high HLB value (generally over 10), whereas lipophilic compounds have values ranging from 1 to 10. Compounds with self-balance between their lipophilic and hydrophilic portions are extremely efficient as emulsifying agents because they tend to concentrate at the oil/water interface. The HLB of compounds is a measure of the degree to which it is hydrophilic or lipophilic, determined by calculating values for the different regions of the molecule, as described by Griffin in 1949 and 1954. Griffin's method for nonionic compounds as described in 1954 works as follows:

$$HLB = 20 \times {}^{Mh}\!/_{M} \qquad \text{Eq. 7}$$

where *Mh* is the molecular mass of the hydrophilic portion of the molecule, and *M* is the molecular mass of the whole molecule, giving a result on an arbitrary scale of 0 to 20. An HLB value of 0 corresponds to a completely hydrophobic molecule, and a value of 20 corresponds to a molecule made up completely of hydrophilic components.

Thus, drugs with high values of HLB are highly water-soluble. A very hydrophilic drug resides in the polar mobile phase; a very lipophilic drug resides in the stationary phase. The lower the HLB number, the more oil-soluble the product; in turn, the higher the HLB number, the more water-soluble the product. The results indicate that the HLB increases (increase of polar interactions and hydrophilic interactions between the solutes and mobile phase) as log t_R decreases.

Statistical parameters

For the constructed model, four general statistical parameters were selected to evaluate the prediction ability of the model for t_R. For this case, the predicted log t_R of each sample in the prediction step were compared with the actual t_R. Root mean square error of prediction (RMSEP) is a measurement of the average difference between predicted and measured response values at the prediction stage. RMSEP can be interpreted as the average prediction error, expressed in the same units as the original response values. The RMSEP was obtained by the following formula:

$$RMSEP = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right]^{0.5} \qquad \text{Eq. 8}$$

The second statistical parameter was relative error of prediction (REP), which shows the predictive ability of each component, and is calculated as:

$$REP(\%) = \frac{100}{\bar{y}} \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right]^{0.5} \qquad \text{Eq. 9}$$

The predictive applicability of a regression model is described in various ways. The most general expression is the standard error of prediction (SEP), which is given in the following formula:

Table IV. Statistical Parameters Obtained by Applying the MLR Method to the Test Set

Parameter	RMSEP	REP (%)	SEP	R^2	R^2Adj	%RE
Value	0.11	11.54	0.12	0.90	0.88	–3.28

$$SEP = \left[\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1} \right]^{0.5} \quad \text{Eq. 10}$$

R^2, which indicated the quality of fit of all the data to a straight line, is calculated for the checking of test set, and is calculated as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad \text{Eq. 11}$$

where y_i is the experimental log t_R of the drug in the sample i, represented the predicted log t_R of the drug in the sample i, is the mean of true log t_R in the prediction set, and n is the total number of samples used in the prediction set. The statistical results (RMSEP, REP, SEP, R^2, %RE) are summarized in Table IV.

Conclusions

The investigation of the QSRR of drugs is an important issue in chromatographic science and medicinal chemistry, as well as in drugs discovery. The same fundamental intermolecular interactions determine the behavior of chemical compounds in both biological and chromatographic environments. A predictive QSRR model which is based on molecular descriptors is proposed in this study to correlate the t_R of drug compounds. Application of the developed model to a testing set of 12 compounds demonstrates that the new model is reliable, with good predictive accuracy and simple formulation. Because the QSRR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the t_R of drug compounds. MLR analysis provided useful equation that can be used to predict the log t_R of drugs based upon log P, SM, and HLB parameters. We have developed here a useful QSRR equation derived from theoretical descriptors associated with t_R property. A MLR is successfully presented for the prediction of t_R property (log t_R) of various drug compounds with diverse chemical structures. A model with high statistical quality and low prediction errors was obtained. The model could accurately predict the t_R property of the drug compounds. This procedure allowed us to achieve a precise and relatively fast method for the determination of log t_R of different series of drug compounds and to predict with sufficient accuracy the log t_R of new drug derivatives.

The macroscopic (bulk) activities/properties of chemical compounds clearly depend on their microscopic (structural) characteristics. Development of QSPR/QSAR on theoretical descriptors is a powerful tool not only for the prediction of the chemical, physical, and biological properties/activities of compounds, but also to gain a deeper understanding of the detailed mechanisms of interactions in the complex systems that predetermine these properties/activities.

References

1. R. Kaliszan. *Quantitative Structure-Chromatographic Retention Relationships*. Wiley–Interscience, New York, NY, 1987.
2. T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, and Y. Everingham. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemom. Intell. Lab. Syst.* **76:** 185–196 (2005).
3. J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, M. Rosés, E. Bosch, A.M. Zissimos, and M.H. Abraham. Analysis of a solute polarity parameter in reversed-phase liquid chromatography on a linear solvation relationship basis. *Anal. Chim. Acta* **515:** 209–227 (2004).
4. K.A. Lippa, L.C. Sander, and S.A. Wise. Chemometric studies of polycyclic aromatic hydrocarbon shape selectivity in reversed-phase liquid chromatography. *Anal. Bioanal. Chem.* **378:** 365–377 (2004).
5. M.F. Silva, L.F. Chipre, J. Raba, and J.M. Luco. Amino acids characterization by reversed-phase liquid chromatography. Partial least-squares modeling of their transport properties. *Chromatographia* **53:** 392–400 (2001).
6. Q.S. Wang, L. Zhang, M. Zhang, X.D. Xing, and G.Z. Tang. A system for predicting the retentions of *O*-alkyl, *n*-(1-methylthioethylideneamino) phosphoramidates on RP-HPLC. *Chromatographia* **49:** 444–448 (1999).
7. R. Kaliszan. *Handbook of Analytical Separations*, Vol. 1. Elsevier, Amsderdam, The Netherlands, 2000, pp. 503–533.
8. T.F. Woloszyn and P.C. Jurs. Prediction of gas chromatographic retention data for hydrocarbons from naphthas. *Anal. Chem.* **65:** 582–587 (1993).
9. R. Gautzsch and P. Zinn. Use of incremental models to estimate the retention indexes of aromatic compounds. *Chromatographia* **43:** 163–176 (1996).
10. O. Ivanciuc, T. Ivanciuc, D.J. Klein, W.A. Seitz, and A.T. Balaban. Quantitative structure–retention relationships for gas chromatographic retention indices of alkylbenzenes with molecular graph descriptors. *SAR QSAR Environ. Res.* **11:** 419–452 (2001).
11. R.H. Rohrbaugh and P.C. Jurs. Prediction of gas chromatographic retention indexes for diverse drug compounds. *Anal. Chem.* **60:** 2249–2253 (1988).
12. L.S. Anker, P.C. Jurs, and P.A. Edwards. Quantitative structure-retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups. *Anal. Chem.* **62:** 2676–2684 (1990).
13. T.F. Woloszyn and P.C. Jurs. Quantitative structure–retention relationship studies of sulfur vesicants. *Anal. Chem.* **64:** 3059–3063 (1992).
14. A.R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, and M. Karelson. Prediction of gas chromatographic retention times and response factors using a general quantitative structure-property relationship treatment. *Anal. Chem.* **66:** 1799–1807 (1994).
15. T. Ivanciuc and O. Ivanciuc. Quantitative structure-retention relationship study of gas chromatographic retention indices for halogenated compounds. *Internet Electron. J. Mol. Des.* **1:** 94–107 (2002).
16. J. Olivero and K. Kannan. Quantitative structure–retention relationships of polychlorinated naphthalenes in gas chromatography.

J. Chromatogr. A **849:** 621–627 (1999).
17. Y. Polyakova, L.M. Jin, and K.H. Row. Linear regression based QSPR models for the prediction of the retention mechanism of some nitrogen containing heterocycles. *J. Liq. Chromatogr. Relat. Technol.* **29:** 533–552 (2006).
18. T. Baczek and R. Kaliszan. Predictive approaches to gradient retention based on analyte structural descriptors from calculation chemistry. *J. Chromatogr. A* **987:** 29–37 (2003).
19. N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, and P.W. Carr. Column selectivity in reversed-phase liquid chromatography: I. A general quantitative relationship. *J. Chromatogr. A* **961:** 171–193 (2002).
20. K. Azzaoui and L. Morin-Allory. Comparison and quantification of chromatographic retention mechanisms on three stationary phases using structure-retention relationships. *Chromatographia* **42:** 389–395 (1996).
21. P.C. Sadek, P.W. Carr, R.M. Doherty, M.J. Kamlet, R.W. Taft, and M.H. Abraham. Study of retention processes in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method. *Anal. Chem.* **57:** 2971–2978 (1985).
22. P.W. Carr, R.M. Doherty, M.J. Kamlet, R.W. Taft, W. Melander, and C. Horvath. Study of temperature and mobile-phase effects in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method. *Anal. Chem.* **58:** 2674–2680 (1986).
23. T. Baczek, R. Kaliszan, K. Novotna, and P. Jandera. Comparative characteristics of HPLC columns based on quantitative structure–retention relationships (QSRR) and hydrophobic-subtraction model. *J. Chromatogr. A* **1075:** 109–115 (2005).
24. M.A. Alhaj, P. Haber, R. Kaliszan, B. Buszewski, M. Jezierska, and Z. Chilmonzyk. Mechanism of separation on cholesterol–silica stationary phase for high-performance liquid chromatography as revealed by analysis of quantitative structure–retention relationships. *J. Pharm. Biomed. Anal.* **18:** 721–728 (1998).
25. M.H. Abraham, M. Roses, C.F. Poole, and S.K. Poole. Hydrogen bonding. 42. Characterization of reversed-phase high-performance liquid chromatographic C18 stationary phases. *J. Phys. Org. Chem.* **10:** 358–368 (1997).
26. L. Buydens and D.L. Massart. Prediction of gas chromatographic retention indexes with topological, physicochemical, and quantum chemical parameters. *Anal. Chem.* **55:** 738–744 (1983).
27. R.H. Rohrbaugh and P.C. Jurs. Prediction of gas chromatographic retention indexes of selected olefins. *Anal. Chem.* **57:** 2770–2773 (1985).
28. J. Bermejo and M.D. Guillen. Biparameter equations for calculating Kovats retention indices of hydrocarbons. *Int. J. Environ. Anal. Chem.* **23:** 77–86 (1985).
29. A. Voelkel. Influence of structure of alkenes on their retention on different stationary phases. *Chromatographia* **25:** 655–658 (1988).
30. Z.D. Hu and H.W. Zhang. Prediction of gas chromatographic retention indices of alkenes from the total solubility parameters. *J. Chromatogr. A* **653:** 275–282 (1993).
31. V.E.F. Heinzen, M.F. Soares, and R.A. Yunes. Semi-empirical topological method for the prediction of the chromatographic retention of *cis*- and *trans*-alkene isomers and alkanes. *J. Chromatogr. A* **849:** 495–506 (1999).
32. A.R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, and E. Benfenati. Interpretation of quantitative structure-property and -activity relationships. *J. Chem. Inf. Comput. Sci.* **41:** 679–685 (2001).
33. Y. Ren, H. Liu, X. Yao, and M. Liu. An accurate QSRR model for the prediction of the GC×GC–TOFMS retention time of polychlorinated biphenyl (PCB). *Anal. Bioanal. Chem.* **388:** 165–172 (2007).
34. M.P. Montana, N.B. Pappano, N.B. Debattista, J. Raba, and J.M. Luco. High-performance liquid chromatography of chalcones: Quantitative structure-retention relationships using partial least-squares (PLS) modeling. *Chromatographia* **51:** 727–735 (2000).
35. M. Song, C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, and N. Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J. Chem. Inf. Comput. Sci.* **42:** 1347–1357 (2002).
36. J.M. Sutter, T.A. Peterson, and P.C. Jurs. Prediction of gas chromatographic retention indices of alkylbenzenes. *Anal. Chim. Acta* **342:** 113–122 (1997).
37. Y.L. Loukas. Artificial neural networks in liquid chromatography: efficient and improved quantitative structure–retention relationship models. *J. Chromatogr. A* **904:** 119–129 (2000).
38. J. Ghasemi, S. Saaidpour, and S.D. Brown. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct. (Theochem.)* **805:** 27–32 (2007).
39. J. Ghasemi and S. Saaidpour. QSPR prediction of aqueous solubility of drug-like organic compounds. *Chem. Pharm. Bull.* **55:** 669–674 (2007).
40. J. Ghasemi, S. Asadpour, and A. Abdolmaleki. Prediction of gas chromatography/electron capture detector retention times of chlorinated pesticides, herbicides, and organohalides by multivariate chemometrics methods. *Anal. Chim. Acta* **588:** 200–206 (2007).
41. J. Ghasemi and Sh. Ahmadi. Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures. *Ann. Chim.* **97:** 69–83 (2007).
42. J. Ghasemi, S. Shahmirani, and E.V. Farahani. Development of a model to predict partition coefficient of organic pollutants in cloud point extraction process. *Ann. Chim.* **96:** 327–337 (2006).
43. Jinno Laboratory, School of Materials Science, Toyohashi University of Technology, Japan. Web: http://chrom.tutms.tut.ac.jp/.
44. ChemOffice 2005, CambridgeSoft Corporation, Web: http://www.cambridgesoft.com.
45. Web: http://www.psu.ru/science/soft/winmopac/.
46. Web: http://www.chemsw.com/.
47. Web: http://www.spss.com/.
48. The Unscrambler version 7.6, 2000, Web: http://www.camo.com.
49. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, and J.J.P. Stewart. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107:** 3902–3909 (1985).
50. R.B. Darlington. *Regression and Linear Models*. McGraw-Hill Higher Education, New York, NY, 1990.
51. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics*, Part A. Elseiver, Amsterdam, The Netherlands, 1997.
52. L. Xu and W.J. Zhang. Comparison of different methods for variable selection. *Anal. Chim. Acta* **446:** 475–481 (2001).
53. C. Hansch and A. Leo. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley & Sons, New York, NY, 1979.
54. A. Leo, C. Hansch, and D. Elkins. Partition coefficients and their uses. *Chem. Rev.* **71:** 525–616 (1971).
55. R.O. Potts and R.H. Guy. The influence of molecular volume and hydrogen-bonding on peptide transport across epithelial membranes. *Pharm. Res.* **10:** 635–637 (1993).
56. M. Karelson. *Molecular Descriptors in QSAR/QSPR*. J. Wiley & Sons, New York, NY, 2000.
57. W.C. Griffin. Classification of surface-active agents by HLB. *J. Soc. Cosmetic Chem.* **1:** 311–326 (1949).
58. W.C. Griffin. Calculation of HLB values of non-ionic surfactants. *J. Soc. Cosmetic Chem.* **5:** 249–256 (1954).
59. H. Martens and T. Naes. Multivariate Calibration. Wiley, Chichester, U.K., 1989.

Manuscript received June 18, 2007;
Revision received September 28, 2007.