Universal Retention Indices and Their Prediction in Reversed-Phase Liquid Chromatography Based On Principal Component Analysis and Target Testing

Charles E. Reese*, Lingyan Huang, Su-Hsiu Hsu, Sadhana Tripathi, and **C.H. Lochmüller**
Department of Chemistry, Duke University, Durham, North Carolina

Abstract

A universal solute and solvent retention index system for reversed-phase liquid chromatography (LC) has been developed and tested with a library of compounds and mobile phases as the base set. Examination of reversed-phase LC retention data by principal component analysis and target transformation factor analysis reveals that the data obtained from three different reversed-phase columns share a common factor space and that three factors are sufficient to describe these retention data. The resulting eigenvector matrix associated with analyte compounds from singular value decomposition is found to be characteristic of the retention behavior of the compounds and independent of the mobile phases and reversed-phase columns used for the measurement. The same is true for the mobile phase eigenvector matrix. Based on these observations, a reference retention index system is developed for both chromatographic solutes and solvents across different reversed-phase columns. Mean errors of retention prediction using this index system are within 4%.

Introduction

Numerous attempts have been made to construct a retention index system for reversed-phase liquid chromatography (LC). The earliest efforts were directed towards the use of retention calculated relative to the retention values for a homologous series (1–4) by analogy to the widely accepted Kováts Index of gas–liquid chromatography (GLC). This approach was confounded by the fact that, unlike GLC, the effect of the mobile phase is a major factor in LC retention; in other words, changing the mobile phase in LC is analogous to changing the stationary phase in GLC. This means that not only every stationary phase material but also every mobile phase requires a new index base. This problem can be alleviated to some degree by introducing a correction factor (5), but a more complex index system of limited accuracy results, even when the measurements are confined to a single column or column type.

* Author to whom correspondence should be addressed.

In recent years, researchers have tried to develop retention prediction systems by constructing solvent parameters (6–8) or by correlating a compound's molecular structure and properties to its retention behavior (9,10); however, in many cases, the retention characteristics of solvents and solutes are studied separately to simplify the problem. Often the success is limited to small groups of compounds, and the required parameters are difficult to determine; the results obtained using those indices can only provide a rough estimation of retention. There are a number of good reviews on recent developments in this area that focus on compound structure–retention relationships (11–13) and solvent composition–retention relationships (14,15).

One way to minimize the errors associated with imperfect models and parameters is to eliminate the model entirely and rely solely on relationships defined by the data themselves. Because differences in the retention of different compounds in different mobile phases are determined by different combinations of fundamental molecular forces and their interactions, it may be possible to find intrinsic retention characteristics for both compounds and mobile phases in one data set and relate these characteristics to other data sets. If this is indeed the case, a retention index system that is free of any preset physical models can be set up. Principal component analysis (PCA) is known to have the ability to reveal the rank of a data set and produce abstract factors that are intrinsic to that data set, and target transformation factor analysis (TTFA) can relate one data space to another by vector rotation (16). The work presented here applies PCA and TTFA to a library of retention data and shows that the indices obtained successfully characterize the retention properties of solute compounds and separation systems.

Theoretical

Upper case letters in bold are used throughout this paper to denote data matrices and vectors; lower case letters are used for scalar quantities.

Retention data bilinearity

A multivariate data matrix is factor analyzable only if the data set is bilinear, that is, each data point is a linear sum of factors, and each factor is weighted by an independent variable. So a data point in a bilinear data set can be expressed as

$$d = f_1 g_1 + f_2 g_2 + \ldots + f_n g_n$$

where f_i are the factors, g_i are the coefficients for f_i, and the vector $[f_1, f_2, \ldots, f_n]$ is orthogonal to the vector $[g_1, g_2, \ldots, g_n]$. Alternatively, g_i can be viewed as the factors, and f_i can be viewed as the weighting coefficients.

In reversed-phase LC, retention (the logarithm of the capacity factor, k', or $\ln k'$) has long been postulated to result from the uncorrelated effects of mobile phase solvents and analyte compounds. In numerous studies on retention models and indices, $\ln k'$ is typically expressed in a three-term equation as

$$\ln k' = f_1 g_1 + f_2 g_2 + f_3 g_3$$

where f_i denotes solvent parameters, and g_i denotes the compound-related parameters (17). It is reasonable to hypothesize that the retention data of reversed-phase LC are bilinear. With this bilinearity, the retention data matrix **D**, which is composed of $\ln k'$ and is arranged with compounds as the row designees and mobile phases as the column designees, can be decomposed by PCA into a row eigenvector matrix (associated with the compounds) and a column eigenvector matrix (associated with the mobile phases). The retention characteristics of the analyte compounds and the mobile phase solvents will be contained in the column eigenvector matrix and the row eigenvector matrix, respectively.

Principal component analysis

The singular value decomposition (SVD) method is one of the most often used PCA methods. In SVD, the row eigenvectors and column eigenvectors are obtained along with a diagonal matrix of singular values, so the data matrix is decomposed into three matrices:

$$\mathbf{D} = \mathbf{USV}^T \qquad \text{Eq 1}$$

The orthonormal matrices **U** and **V** contain the column eigenvectors and row eigenvectors, respectively, and **S** is the diagonal matrix of singular values. The matrices that contain the principal eigenvectors of **U** and $\mathbf{V}^T$ are designated $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^T$, and the respective singular values from **S** form matrix $\bar{\mathbf{S}}$. For a three-factor data set, $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^T$ contain the first three column vectors of **U** and $\mathbf{V}^T$. As the column eigenvector matrix and row eigenvector matrix, $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^T$ carry separately the retention characteristics of compounds and mobile phases. Geometrically, for a three-factor data set, the principal eigenvectors (which will be referred to as eigenvectors for short throughout this paper if not specified) of $\bar{\mathbf{U}}$ define a three-dimensional compound retention space, and the rows corresponding to each compound in matrix $\bar{\mathbf{U}}$ can be considered as the coordinates of the compound retention characteristics (CRC) in this space. Likewise, the rows of $\bar{\mathbf{V}}^T$ can be taken as the coordinates of corresponding mobile phase retention characteristics (MRC) in the space defined by the eigenvectors in matrix $\bar{\mathbf{V}}^T$. With the retention properties of the analyte compounds and mobile phases independent of each other, the relative geometric position of every CRC or MRC point in the three-dimensional space of $\bar{\mathbf{U}}$ or $\bar{\mathbf{V}}^T$ should be independent of individual mobile phases or compounds used to make the retention measurements as long as the compound retention space or the mobile phase retention space is truly spanned.

Target transformation

Target transformation can project vectors from one data space to another. For an orthonormal column vector matrix $\bar{\mathbf{U}}$, the product of $\bar{\mathbf{U}}$ and $\bar{\mathbf{U}}^T$ generates a projection matrix **P** that spans the space defined by the column vectors of $\bar{\mathbf{U}}$:

$$\mathbf{P} = \bar{\mathbf{U}}\bar{\mathbf{U}}^T \qquad \text{Eq 2}$$

To test if a target column vector matrix **H** shares common space with $\bar{\mathbf{U}}$, the projection matrix **P** of space $\bar{\mathbf{U}}$ is obtained first by Equation 2; then matrix **P** is multiplied by the target vector matrix **H**, as shown in Equation 3, producing a new matrix $\hat{\mathbf{H}}$:

$$\hat{\mathbf{H}} = \mathbf{PH} \qquad \text{Eq 3}$$

$\hat{\mathbf{H}}$ is therefore the projection matrix of **H** in the vector space of **U**. If $\bar{\mathbf{H}}$ truly shares the same space with $\bar{\mathbf{U}}$, then each element of $\hat{\mathbf{H}}$ will equal the corresponding element of the target **H**.

Universal indices

In order to define universal indices, two new matrices, **C** and **M**, are introduced by the following equations:

$$\mathbf{C} = \bar{\mathbf{U}}\sqrt{\bar{\mathbf{S}}} \qquad \text{Eq 4}$$

$$\mathbf{M} = \sqrt{\bar{\mathbf{S}}}\,\bar{\mathbf{V}}^T \qquad \text{Eq 5}$$

Since $\bar{\mathbf{S}}$ is a diagonal matrix, Equations 4 and 5 are in fact scalar multiplication of the eigenvectors in $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^T$, and the column vectors in **C** and **M** thus contain weights of relative importance for each eigenvector. **C** and **M** generated from the base data set are defined as the retention index matrices: the rows of **C** are defined as the retention indices for the corresponding compounds, and the columns of **M** are defined as the retention indices of the corresponding mobile phase solvents.

Indices for compounds not in the library can be computed by the "free floating" method. The term *free floating* refers to a target transformation process in which one or more rows in the target column vector matrix are left blank, and the values of the missing rows are obtained from the projection of the target vectors in another data space that are shared by the target vectors. To achieve free floating, the available retention data of the new compounds are appended to the rows of the base data set first to form a new data matrix; then SVD is applied to the new data matrix using Equation 1, and new ma-

trices $\bar{\mathbf{U}}$ and $\bar{\mathbf{S}}$ are generated, which further forms a matrix $\hat{\mathbf{U}}$ according to Equation 6:

$$\hat{\mathbf{U}} = \bar{\mathbf{U}}\bar{\mathbf{S}} \qquad \text{Eq 6}$$

The matrix $\ddot{\mathbf{U}}$ is formed from $\hat{\mathbf{U}}$ by removing the row that corresponds to the new compound. With the original compound index matrix as **C**, the retention index for the new compound, which is a row in the new matrix $\hat{\mathbf{C}}$, can be calculated by Equation 7:

$$\hat{\mathbf{C}} = \hat{\mathbf{U}}(\ddot{\mathbf{U}}^{\mathrm{T}}\ddot{\mathbf{U}})^{-1}\ddot{\mathbf{U}}^{\mathrm{T}}\mathbf{C} \qquad \text{Eq 7}$$

Indices for mobile phase solvents that are not included in the library can be computed in a similar way.

Retention values for compounds in all mobile phases can be predicted by computation from the retention index matrices **C** and **M** according to Equation 8:

$$\bar{\mathbf{D}} = \mathbf{CM} \qquad \text{Eq 8}$$

Experimental

Ammonium nitrate was used as the dead volume marker for all data sets. Data set 1 was part of a data set previously reported (18); it was obtained on a Perkin-Elmer MCH-5 column (ODS packing) (Perkin-Elmer; Norwalk, CT). This column will be referred to as column 1. The elution time of ammonium nitrate was measured periodically, and its retention volumes were calculated from the measured retention times and flow rate. The 12 compounds and mobile phases used for this study were chosen based on a previous publication (19) by a combination target testing procedure which indicated that these compounds and mobile phases best spanned the factor space. The number of mobile phases and compounds was limited to 12 each so as to make the task of triplicate measurement for each value reasonable. As the number of factors needed to span the retention space did not vary from matrices of 6×6 to 35×38, a 12×12 matrix was a reasonable compromise between experimental effort and the desire to have a large matrix.

Data set 2 was collected using a Perkin-Elmer Series 4 pump, an LC-235 diode-array detector, and a Perkin-Elmer ISS-100 autosampler. The retention times were recorded using either a Perkin-Elmer LCI-100 or a Beckman 3390A integrator (Beckman; Fullerton, CA). The column used (column 2) was a YMC Type A column (YMC USA; Wilmington, NC) with brush-type C_{18} packing material. Each injection sample had a solvent composition close to that of the mobile phase used. This was achieved by diluting 10 µL of concentrated methanol solutions of the compound standards with 1 mL of the mobile phase. The retention times were recorded to within 0.001 min. The elution time of ammonium nitrate was measured for each mobile phase, and the average was used in all calculations. The retention volumes of the analytes were adjusted by using ketones as the internal standards, and the retention times of a homologous series of ketones were measured along with the flow rate for each mobile phase. The flow rate was measured within 0.001 mL/min using a burette attached to the detector to obtain accurate measurements of the retention volumes of the ketones. One or more of the ketones was included with each injection sample, and the retention volumes of the analytes were calculated from the known retention volumes of the ketones. The mean standard deviation ($n = 3$) for all sets was 0.28% for retention times and 0.2% for retention volumes. The mobile phases were premixed by weight according to the density of each solvent at 25.0°C. The column was maintained at 25.0°C with a circulating water bath. The mobile phases were pre-equilibrated with a precolumn (C_{18} packing material) that was inserted between the pump and autosampler. No drift in chromatographic retention was observed for the column during the period of these measurements.

Data set 3 was collected in a manner identical to that of data set 2 except that a Perkin-Elmer 3×3 (3-cm length and 3-mm particle size) C_{18} cartridge column (column 3) was used.

All calculations were performed using the MATLAB (The Mathworks Inc.; Natick, MA) matrix algebra computation package. The MATLAB singular value decomposition function was used for data matrix decomposition.

Results and Discussion

Table I contains the list of compounds and mobile phases used in the study, along with assigned numbers that identify them in the figures. The mobile phases are divided into two groups labeled v1 and v2 for the purpose of testing the prediction of retention in alternate mobile phase sets. In general, the mobile phases are in order of decreasing solvent strength insofar as this is indicated by the sum of the retention of the 12 compounds.

Retention and selectivity

Retention and selectivity from different mobile phases and different columns were compared in order to give an overview

Table I. Compounds and Mobile Phases Used in the Study

Compounds	Assigned number	Mobile phase group	W/M/A* ratio	Assigned number
Anisole	1	Group v1	40:00:60	-
Benzene	2		40:15:45	3
p-Chlorotoluene	3		50:00:50	5
Dimethyl phthalate	4		50:37.5:12.5	7
m-Dinitrobenzene	5		60:00:40	9
2,4-Dinitrotoluene	6		60:40:00	11
2,6-Dinitrotoluene	7	Group v2	40:30:30	2
m-Fluoronitrobenzene	8		40:60:00	4
o-Fluoronitrobenzene	9		50:12.5:37.5	6
p-Fluoronitrobenzene	10		50:50:00	8
p-Methoxybenzaldehyde	11		60:30:10	10
Nitrobenzene	12		70:00:30	12

* W/M/A = water–methanol–acetonitrile.

of the difficulties in trying to classify reversed-phase solvents and columns by their retention or selectivity for different compounds. Figure 1 shows a plot of the k' values of Compound 1 obtained on the three columns versus the 12 mobile phases. As can be seen from Figure 1, the order of solvent strength is different for different columns, and the order of retention between columns also changes with changes in the mobile phase ratio.

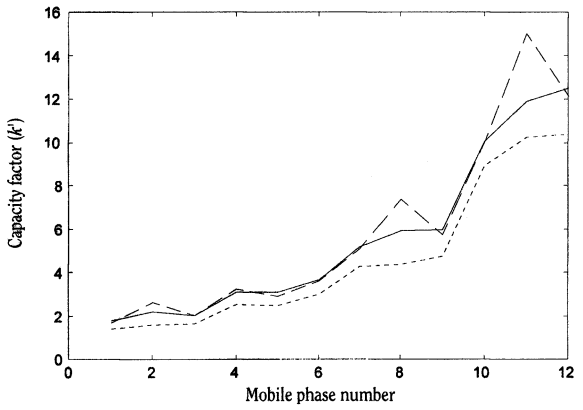

Figure 1. Capacity factor of compound 1 in 12 mobile phases on three columns: —, column 1; ···, column 2; - - -, column 3.

Table II. Results of Error Analysis for Deducing the Significant Number of Factors

Factors	Variance (%)	Imbedded error (×100)	Probability test
Data set 1			
1	99.30	4.23	0.000
2	0.51	3.26	0.015
3	0.18	0.93	0.000
4	0.01	0.58	0.029
5	0.00	0.36	0.042
6	0.00	0.21	0.051
Data set 2			
1	98.05	6.26	0.000
2	1.79	2.60	0.000
3	0.13	1.28	0.003
4	0.01	1.13	0.168
5	0.01	0.86	0.115
6	0.00	0.72	0.225
Data set 3			
1	98.94	5.23	0.000
2	0.89	3.06	0.002
3	0.14	1.58	0.004
4	0.02	1.23	0.090
5	0.01	1.03	0.180
6	0.00	0.69	0.094
Combined data set			
1	98.68	3.1	0.000
2	1.14	1.7	0.000
3	0.15	0.8	0.002
4	0.02	0.6	0.057
5	0.01	0.6	0.179
6	0.00	0.4	0.135

Figure 2 is the plot of selectivity (α) between Compounds 9 and 5 ($\alpha = k'_9 / k'_5$) versus the mobile phase number for the three columns. The order of selectivity among the three columns changed significantly when the mobile phase changed. Column 2 even showed a reversal of elution order ($\alpha < 1$) at low water concentrations.

The difficulty of extracting any pattern for retention and selectivity is apparent when the retention data from different mobile phases on different reversed-phase columns are studied. The characterization of the separation system is completely dependent on the reference compounds chosen. However, as will be shown, abstract factor analysis provides a means of classifying separation systems independent of any particular compound.

Number of significant factors

Three factors were found to be significant for reversed-phase LC retention data. Table II shows the results of significant factor tests (i.e., rank tests) for the three sets. The first three factors in all three data sets can explain 99.97% or more variance. Data sets 2 and 3, which were collected using internal standards for flow rate variation adjustment, show a clear cut for three significant factors at 5% confidence level of Malinowski's F-test (20). However, the F-test shows four significant factors at 3% confidence level for data set 1. Apparently, data set 1 has a larger noise level, which can be attributed to the fact that no flow rate adjustment was made and that no thermostat was used to keep the column temperature constant.

When the right number of significant factors was deduced, the data reproduced by the equation $\bar{\mathbf{D}} = \bar{\mathbf{U}}\bar{\mathbf{S}}\bar{\mathbf{V}}^{\mathbf{T}}$ gave better representation of the data because $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}^{\mathbf{T}}$, which were formed by only the principal eigenvectors, do not contain the secondary factors that are mostly associated with errors. The dif-

Table III. Error of Data Reproduction

	Data set 1	Data set 2	Data set 3
Mean ($\bar{\mathbf{D}}$–$\mathbf{D}$)	0.0128	0.0166	0.0211
STD ($\bar{\mathbf{D}}$–$\mathbf{D}$)	0.0162	0.0222	0.0274
Max ($\bar{\mathbf{D}}$–$\mathbf{D}$)	0.0534	0.0937	0.0895

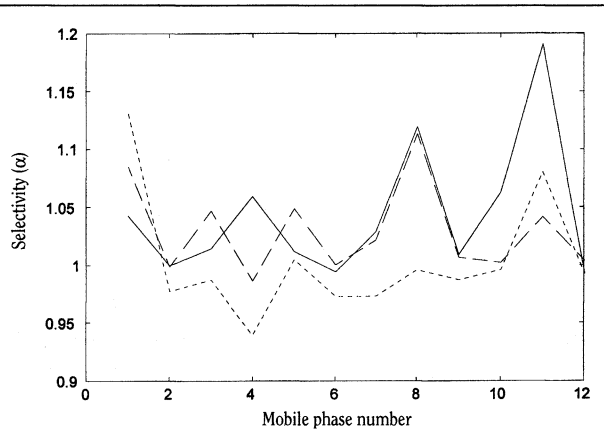

Figure 2. Selectivity between compounds 9 and 5 in 12 mobile phases on three columns: —, column 1; ···, column 2; - - -, column 3.

ference between $\bar{\mathbf{D}}$ and **D** reflects the error that exists in the space defined by the secondary factors and is often called extracted error (XE) or reproduction error. The reproduction errors for the three data sets are listed in Table III. It can be seen that the errors are reasonably small, which supports the finding that all three retention data sets have a rank of three. It should be pointed out that the use of experimental error in deducing factor ranking in reversed-phase LC retention can sometimes be misleading. There is a Pythagorean relationship between the real error (RE), XE, and the error imbedded into primary factor space (imbedded error or IE): $RE^2 = XE^2 + IE^2$ (21). Hence the extracted error is expected to be smaller than the real experimental error. However, if the standard deviations of retention measurements, which for data set 2 are under 0.3% for all data points, are taken as RE, then the extracted errors shown in Table III are well over the limit of RE. The explanation for this is that there are experimental errors that have not been accounted for. The main part of the unaccounted experimental error is likely due to the difficulty of measuring the real dead volume for all compounds and mobile phases. The real dead volume measurements are often complicated by the fact that the size exclusion of solutes from portions of the stationary phase surface may make the actual void volume of the column appear slightly different for different solutes; on the other hand, the variation of elution time of ammonium nitrate across different mobile phase compositions causes more problems in accurately measuring the dead volume. We have tried many other dead volume markers, and ammonium nitrate appears to be the best for us. However, the variation of its elution time was still as large as 12% from mobile phases with high water content to mobile phases with high organic content.

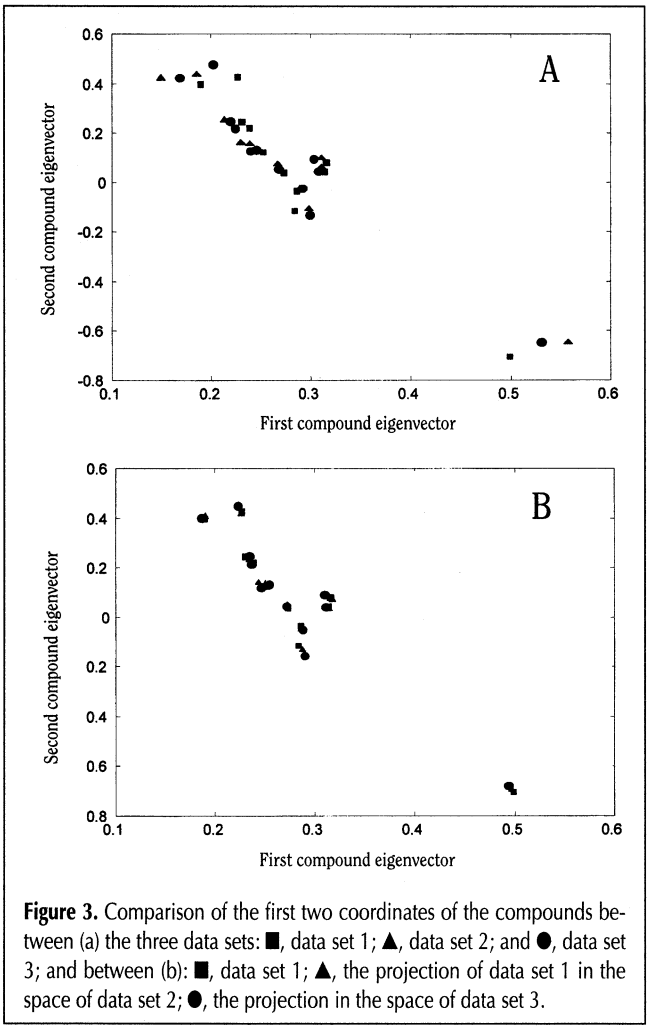

Figure 3. Comparison of the first two coordinates of the compounds between (a) the three data sets: ■, data set 1; ▲, data set 2; and ●, data set 3; and between (b): ■, data set 1; ▲, the projection of data set 1 in the space of data set 2; ●, the projection in the space of data set 3.

Common data space

The three data sets were combined into one large matrix (12×36) to check if the three data sets actually belong to the same data space. Here *data space* is a general term referring to either the compound space or the mobile phase space, which

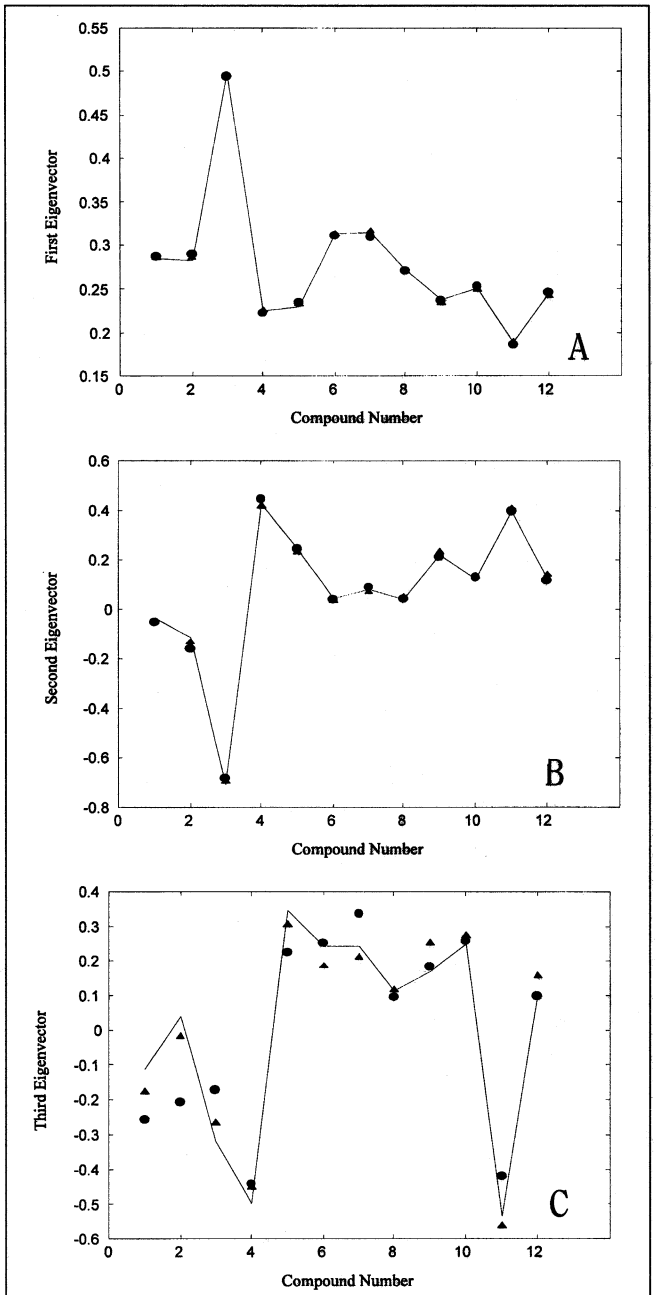

Figure 4. Eigenvector comparison for common compound space test using the first (A), second (B), and third (C) compound eigenvectors. Key: —, data set 1; ▲, the projection of data set 1 in the compound space of data set 2; ●, the projection in the space of data set 3.

will be elaborated on later. If the three individual data sets, which all have a rank of three, do not share a common data space, then one or more of the factors would be unique and would show up as extra factors in the combined data set. The results in Table II demonstrate that at a 5% confidence level, the combined data set still has three significant factors, which suggests that all the retention data sets, obtained from three different reversed-phase columns, not only have the same rank, but more importantly, belong to the same data space.

Target transformation factor analysis can be used to further test the existence of shared data space. As discussed in the theoretical section, SVD produces two orthonormal eigenvector matrices $\bar{U}$ and $\bar{V}^T$; the rows of $\bar{U}$ can be considered as the coordinates of the corresponding compound in the three-dimensional space defined by eigenvectors of $\bar{U}$; when a common space exists, the coordinates of each compound from different data sets are expected to fall into the same position. However, because there are different levels of variance in different data sets, the three-dimensional axes defined by the eigenvectors of $\bar{U}$ may appear different for different data sets, and the coordinates may appear to be rotated away from each other. If the eigenvectors of $\bar{U}$ from different data sets truly define the same space, then the axes of each individual data set can be rotated toward a common orientation through target transformation (Equations 2 and 3), and the resulting coordinates of the same compound should coincide (see Figure 3). In Figure 3A, the first two coordinates of compounds from different data sets do not overlap; instead, they are shifted away from each other. In Figure 3B, the first and second coordinates of compounds for data set 1 and its projections in the space defined by data sets 2 and 3 overlap as expected when a common space exists, which suggests that data set 1 can be represented by the space defined by itself and the space defined by data set 2 or data set 3 and that the three data sets actually define one common compound space. It should be emphasized that the target testing procedure using Equations 2 and 3 only relocates vectors in a new data space expressed by the projection matrix. In the case of Figure 3, where the eigenvectors of data set 1 are projected ($\mathbf{H} = \bar{U}_1$, with 1 denoting data set 1) into the spaces defined by data set 2 ($\mathbf{P} = \bar{U}_2\bar{U}^T_2$) or data set 3 ($\mathbf{P} = \bar{U}_2\bar{U}^T_2$), the resulting eigenvector matrix $\hat{\mathbf{H}}$ from Equation 3 defines the compound space of data sets 2 or 3, not that of data set 1. It is only because the data spaces are the same that $\hat{\mathbf{H}}$ is identical to $\mathbf{H}$. Therefore, $\hat{\mathbf{H}}$ is as equally valid as $\mathbf{H}$ in representing the retention characteristics of the compounds. Using $\hat{\mathbf{H}}$ as the new retention coordinates of the compounds does not modify the retention characteristics; it is simply a different viewpoint. In fact, any set of three vectors that spans the data space could be used to represent the compound retention characteristics, including the actual retention data vectors. It is this feature of eigenvector space transformation that lays the foundation for our universal index system.

Figure 4 further demonstrates that the compound retention space is independent of the reversed-phase column. In Figure 4, the three-compound eigenvectors from data set 1 and its projections in the compound space of data sets 2 and 3 are plotted for all the compounds. Only very small errors appear in the first eigenvector (Figure 4A), and the agreement between the second eigenvector and its projections is also good (Figure 4B). Although the fitting for the third compound eigenvector shows more scattering, it can be well explained by the fact that in SVD method, eigenvectors are extracted in decreasing order in the amount of variance for which they account. Thus, each successively extracted eigenvector is closer to the noise level of the data than its predecessor. Therefore more scatter is expected in the higher numbered eigenvectors. Singular values are a measure of the average amount of retention that is accounted for by the successive eigenvectors. The singular values for all three sets with data are listed in Table IV. It is shown that, on average, less than 4% of reversed-phase retention is due to the third factor. As the imbedded error is around 1% or more (IE values in Table II), it is not surprising that there is a

Table IV. Percentage of Average Retention Accounted for by Eigenvectors

	Data set 1	Data set 2	Data set 3
Eigenvector 1	89.75	85.33	88.30
Eigenvector 2	6.43	11.55	8.39
Eigenvector 3	3.82	3.12	3.30

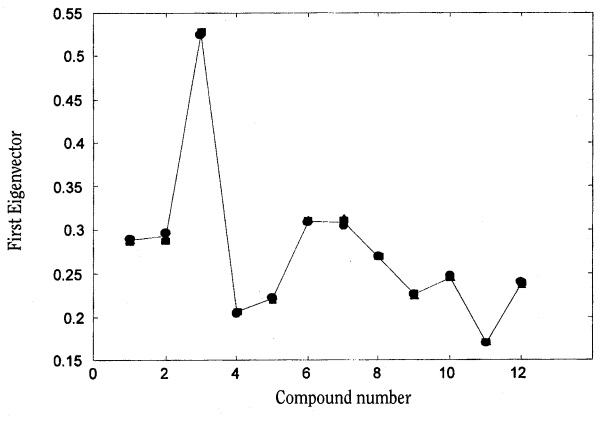

Figure 5. Eigenvector comparison for common compound space test. The first compound eigenvector for the combined data set is shown as a line. The projections of this eigenvector into the compound space of data sets 1, 2, and 3 are shown as ■, ▲, and ●, respectively.

Table V. Universal Compound Indices

Compound	Index 1	Index 2	Index 3
1	1.668	–0.061	–0.211
2	1.694	–0.242	–0.074
3	3.046	–1.258	–0.325
4	1.191	0.847	–0.542
5	1.281	0.453	0.358
6	1.794	0.091	0.278
7	1.788	0.164	0.328
8	1.557	0.110	0.132
9	1.307	0.450	0.238
10	1.419	0.266	0.310
11	0.986	0.794	–0.583
12	1.380	0.274	0.136

Table VI. Universal Separation System Indices

Mobile Phase	Index 1			Index 2			Index 3		
	Col. 1	Col. 2	Col. 3	Col. 1	Col. 2	Col. 3	Col. 1	Col. 2	Col. 3
1	0.357	0.208	0.305	-0.263	-0.550	-0.468	0.099	0.040	0.043
2	0.495	0.280	0.590	-0.230	-0.578	-0.355	0.054	-0.017	-0.007
3	0.437	0.305	0.412	-0.243	-0.544	-0.446	0.093	0.052	0.045
4	0.656	0.507	0.630	-0.242	-0.620	-0.464	-0.153	-0.081	-0.127
5	0.691	0.552	0.646	-0.038	-0.335	-0.238	0.201	0.155	0.138
6	0.801	0.672	0.780	-0.002	-0.291	-0.174	0.174	0.138	0.128
7	1.005	0.869	0.959	0.080	-0.245	-0.140	-0.027	-0.045	-0.024
8	1.043	0.834	1.185	0.001	-0.377	-0.059	-0.255	-0.122	-0.205
9	1.101	0.958	1.058	0.229	-0.066	0.027	0.305	0.242	0.252
10	1.416	1.322	1.397	0.383	0.075	0.170	-0.088	-0.079	-0.074
11	1.471	1.365	1.506	0.316	-0.006	0.208	-0.424	-0.309	-0.426
12	1.555	1.450	1.551	0.496	0.226	0.285	0.297	0.300	0.266

good deal of scatter for the third eigenvectors in Figure 4C.

The final test of the existence of the common compound space is to make projections of the eigenvectors of the combined data set into the compound space of the individual data sets and to see if all these compound spaces coexist. Figure 5 shows that the projections of the first compound eigenvector for the combined data set in the compound space of individual data sets fits well with the original compound eigenvector. Once again, this provides strong evidence that the three data sets share a common compound space.

Although all of these discussions are focused on compound space, the conclusions can be applied in the same fashion to the mobile phase space as well. Because the eigenvector matrices **U** and **V** are equally placed, their properties are expected to be interchangeable and the corresponding compound space and mobile phase space of **U** and **V** should follow the same rules. Therefore it can be concluded that the retention data from the three different reversed-phase columns share a common data space and the compound eigenvector matrix and the mobile phase eigenvector matrix are representative of the retention characteristics of the compounds and the mobile phases, respectively.

Universal indices

Because the eigenvector matrix **U** or **V** resulting from SVD each contains the retention characteristics of compounds or mobile phases, it should be able to characterize and systematize the retention properties of analyte compounds or separation systems. Unfortunately, as previously demonstrated, different data sets that share the same data space may produce different eigenvectors. However, if a base set can be established with a library of retention data and projections of all data spaces can be made into the data space of this base set, then a universal reference will be established to describe the common retention space shared by retention data from different reversed-phase columns. The coordinates of this reference data space that map the retention properties of the compounds and mobile phases can be used as universal retention indices.

In this paper, all the compounds and mobile phases used are considered as components of the library collection, and the combined data set is taken as the base set. The resulting matrices **C** and **M** from Equations 4 and 5 are defined as the index matrices: the three eigenvectors in matrix **C** are defined as the three indices of analyte compounds, and the three eigenvectors in matrix **M** are defined as the three indices of the separation system. The universal compound indices for the base set are shown in Table V, and the separation systems indices are listed in Table VI. Index 1, Index 2, and Index 3 in the tables refer to the retention characteristics that correspond to the first, second, and third principal factors, respectively.

The retention indices for a compound that is not included in the library can be calculated from Equations 6 and 7, provided its retention data from three or more mobile phases are available and these mobile phases belong to the data library and adequately span the mobile phase space. Retention indices for new mobile phases can be obtained in a similar way with the retention measurements of three or more library compounds in the desired mobile phases.

The separation indices in Table VI are independent of the specific compounds used to make the retention measurements, just as the compound indices are independent of the specific mobile phases used to determine them. To test this, the combined data are grouped into two data sets, each containing compounds with odd and even assigned numbers, as listed in Table I. SVD are then performed on the two data sets, and the number of significant factors is obtained. The results of the error tests for these two data sets are shown in Table VII. The significant factors for both sets with mobile phases are found to be three, the same as the number of significant factors found for the combined data set. Because the two data sets do not contain the same mobile phase solvents when they have the same number of significant factors as the combined data set, the factors in the individual set must be common to both.

Table VII. Results of Error Analysis

Factors	Variance (%)	Imbedded error (×100)	Probability test
Odd mobile phase data set			
1	98.48	3.7	0.000
2	1.36	1.7	0.017
3	0.15	0.7	0.042
4	0.01	0.5	0.352
5	0.01	0.2	0.326
6	0.00	0.0	1.000
Even mobile phase data set			
1	99.26	2.1	0.000
2	0.63	1.2	0.029
3	0.09	0.6	0.066
4	0.01	0.3	0.211
5	0.00	0.2	0.545
6	0.00	0.0	1.000

To further test the commonalty of the space defined by the data sets of odd- and even-numbered mobile phases, the separation system indices of the combined data set were target tested in the space defined by the separation system eigenvectors of the odd- and even-numbered compound sets. The results are shown in Figure 6. Between the predicted and the target vector, Index 1 has a very good fit; Index 2 and Index 3 show some degree of scatter. The large scatter can be attributed to the imbedded error from eigenvector extraction and to the limited range of the capacity factor for compounds in any single mobile phase. In this study, compounds with fairly similar retention values were selected so that a wide range of solvent strengths could be studied while the maximum retention times were kept reasonable. It is expected that with a larger and more diverse data library, the accuracy of Index 2 and Index 3 can be considerably improved.

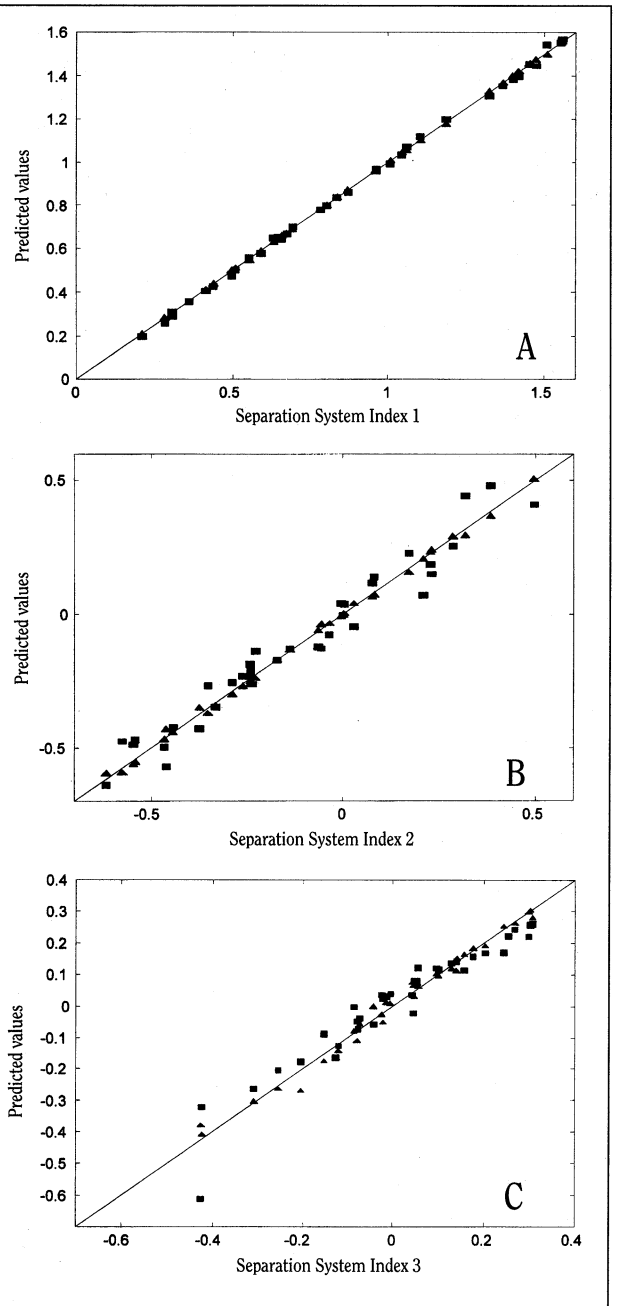

Figure 6. Prediction of separation system indices from all odd-numbered compounds (▲) and all even numbered compounds (■) for Index 1 (A), Index 2 (B), and Index 3 (C).

Retention prediction

Theoretically, three retention measurements would be sufficient to make a retention prediction because the retention space is three-dimensional. Usually twice that (i.e., six measurements) would be more practical. The first step in retention prediction is to calculate the retention indices for a desired compound. To test the accuracy of index calculation and prediction, all of the mobile phases in the study were recombined

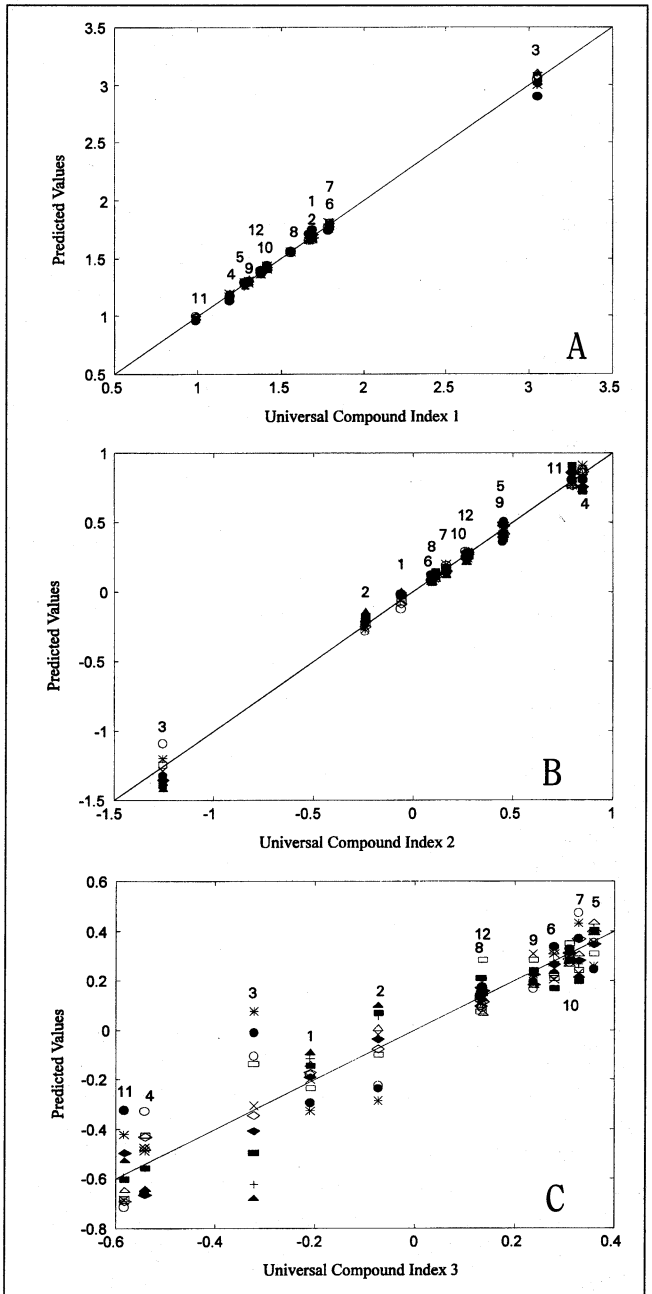

Figure 7. Prediction of universal compound index. Indices 1, 2, and 3 are shown in A, B, and C, respectively. Key: υ, all odd solvents; ◊, all even solvents; +, all data set 1; x, all data set 2; ∗, all data set 3; ▲, odd solvents (v1) of data set 1; △, even solvents (v2) of data set 1; ■, odd solvents (v1) of data set 2; □, even solvents (v2) of data set 2; ●, odd solvents (v1) of data set 3; ○, even solvents (v2) of data set 3.

into different groups, and within each group, the free floating method was applied to calculate the indices of a compound purposely removed. In total, there were 11 sets, including two sets with 18 mobile phases, three sets with 12 mobile phases, and six sets with 6 mobile phases. The sets with 18 mobile phases contained the odd- and even-numbered mobile phases from the overall 36 mobile phases; the sets with 12 mobile phases were the original three data sets; and the sets with 6 mobile phases were formed from the odd- and even-numbered mobile phases from the three individual data sets. The newly calculated or predicted indices from these 11 sets were compared with the original indices, and the results were plotted in Figure 7. It can be seen that even the sets with six mobile phases give satisfactory predictions for the first index (Figure 7A) and fairly good prediction for the second index (Figure 7B). For the third index, which is shown in Figure 7C, the predictions from the sets with 6 mobile phases and sets with 12 mobile phases are scattered, but the predictions from the sets with 18 mobile phases appear to be acceptable.

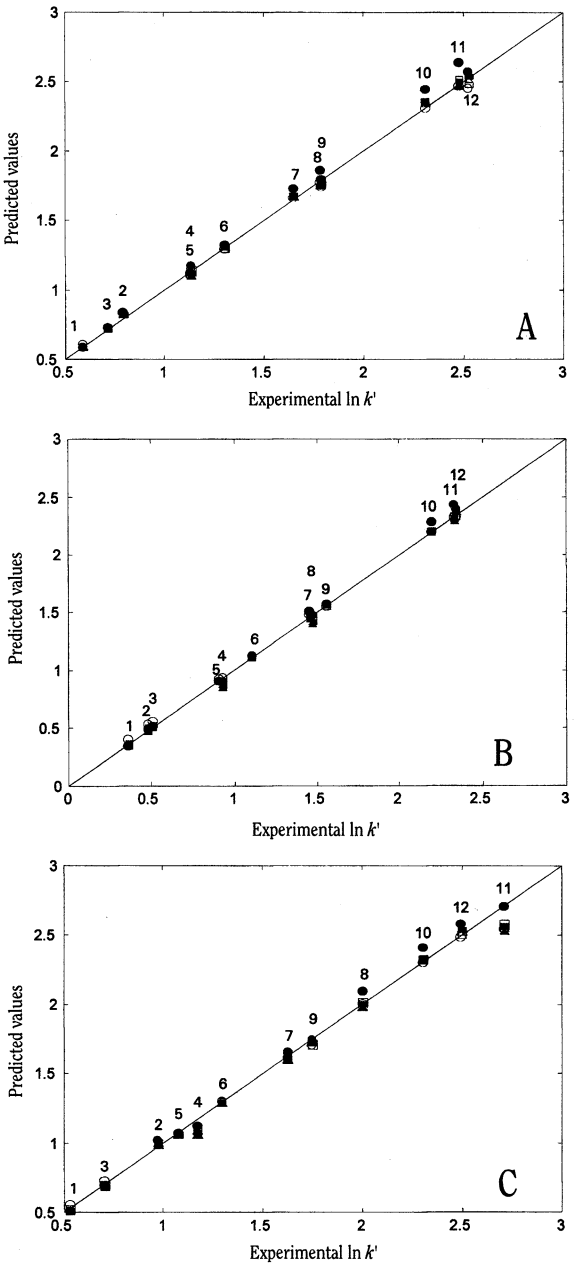

Figure 8. Retention prediction for compound 1 on column 1 (A), column 2 (B), and column 3 (C) predicted from the following: ▲, odd solvents (v1) of data set 1; △, even solvents (v2) of data set 1; ■, odd solvents (v1) of data set 2; □, even solvents (v2) of data set 2; ●, odd solvents (v1) of data set 3; ○, even solvents (v2) of data set 3.

Recall that, on average, more than 85% of reversed-phase retention is accounted for by the first factor (Index 1), less than 15% is due to the second factor (Index 2), and less than 4% of the total retention is due to the third factor (Index 3), as indicated by the singular values in Table IV. For the purpose of method development, which requires only close estimation of retention, the accuracy of the third index is not critical. Therefore, the sets with six mobile phases can be expected to provide adequate accuracy for retention predictions across different columns. In other words, measurements of the retention of a compound with six mobile phases on a single column will give reasonably good estimation for the retention of the compound in other mobile phases and on other reversed-phase columns. For the purpose of compound identification or classification, where accurate evaluation of retention is required, it is necessary that the second compound index be obtained with better accuracy.

Retention value prediction is the ultimate test of the index system. The indices calculated from the sets with six mobile phases were used to calculate compound retentions across different columns by Equation 8. The ln k' values of Compound 1 on column 1 and the predicted ln k' values of Compound 1 on the other two columns are shown in Figure 8. The accuracy of the prediction is satisfactory; the mean errors for the predicted capacity factors of Compound 1 were 2.5% on column 1, 2.3% on column 2, and 3.5% on column 3.

Conclusion

It has been demonstrated that with PCA and TTFA, the retention properties of compounds can be described independently of any particular column or mobile phase and that the retention properties of separation systems can be described independently of any particular compound; the retention properties of compounds and mobile phases from different data sets can be represented in a common space defined by a reference base set. The prediction made from the indices based on the combined data set in this study gave satisfying results.

Although the small base set used in this study was sufficient to prove the principle of the PCA- and TTFA-based universal retention index system, a practical library, which is under construction in our laboratory, should include a wide variety of compounds, columns, and mobile phases. The collection of such a library would allow the identification of critical compounds and mobile phases so that indices for new compounds or separation systems can be calculated with a minimum number of additional measurements. A sufficiently large library would also permit method development in reversed-phase LC to be reduced to a simple matrix multiplication. Be-

sides making retention prediction and identification of compounds possible, the universal retention index system is also anticipated to be useful in stationary phase quality control, QSAR relationships study, fragmental functional group index development, mobile phase solvent strength and selectivity relationship study, and novel bonded phase material classification.

References

1. R.M. Smith. Alkylarylketones as a retention index scale in liquid chromatography. *J. Chromatogr.* **236:** 313–20 (1982).
2. M. Popl, V. Dolansky, and J. Coupek. Chromatography of aromatic hydrocarbons on macroporous polystyrene gel. *J. Chromatogr.* **130:** 195–204 (1977).
3. M.N. Hasan and P.C. Jurs. Computer-assisted prediction of liquid chromatographic retention indexes of polycyclic aromatic hydrocarbons. *Anal. Chem.* **55:** 263–69 (1983).
4. J.K. Baker and C. Ma. Retention index scale for liquid chromatography. *J. Chromatogr.* **169:** 107–15 (1979).
5. P. Jandera. Correlation of retention and selectivity of separation in reversed-phase high-performance liquid chromatography with interaction indices and with lipophilic and polar structural indices. *J. Chromatogr.* **656:** 437–67 (1993).
6. W.J. Cheong and P.W. Carr. Limitations of all empirical single-parameter solvent strength scales in reversed-phase liquid chromatography *Anal. Chem.* **61:** 1524–29 (1989).
7. H. Patel, D. King, and T. Jefferies. Application of solute and mobile phase partition coefficient to describe solute retention in reversed-phase high-performance liquid chromatography. *J. Chromatogr.* **555:** 21–31 (1991).
8. P. Jandera and J. Rozkošná. Method for characterization of selectivity in reversed-phase liquid chromatography. *J. Chromatogr.* **556:** 145–48 (1991).
9. R. Smith and C. Burr. Retention prediction of analytes in reversed-phase high-performance liquid chromatography based on molecular structure. *J. Chromatogr.* **550:** 335–56 (1991).
10. A. Tchapla, S. Héron, E. Lesellier, and H. Colin. General view of molecular interaction mechanisms in reversed-phase liquid chromatography. *J. Chromatogr.* **656:** 81–112 (1993).
11. T. Hanai. Structure–retention correlation in liquid chromatography. *J. Chromatogr.* **550:** 313–24 (1991).
12. R. Kaliszan. Quantitative structure–retention relationships applied to reversed-phase high-performance liquid chromatography. *J. Chromatogr. A* **656:** 417–35(1993).
13. R. Smith. Function group contributions to the retention of analytes in reversed-phase high-performance liquid chromatography. *J. Chromatogr.* **656:** 381–415 (1993).
14. C.H. Lochmüller, C.R. Reese, A.J. Aschman, and S.J. Breiner. Current strategies for prediction of retention in high-performance liquid chromatography. *J. Chromatogr.* **656:** 3–18 (1993).
15. F.D. Anitia and C. Horváth. Dependence of retention on the organic modifier concentration and multicomponent adsorption behavior in reversed-phase chromatography. *J. Chromatogr.* **550:** 411–24 (1991).
16. E.R. Malilnowski. *Factor Analysis in Chemistry.* Wiley, New York, NY, 1991, p 65.
17. K. Valkó, L.R. Snyder, and J.L. Glajch. Retention in reversed-phase liquid chromatography as function of mobile-phase composition. *J. Chromatogr.* **656:** 501–20 (1993).
18. C.H. Lochmüller, S.J. Breiner, C.E. Reese, and M.N. Keol. Characterization and prediction of retention behavior in reversed-phase chromatography using factor analytical modeling. *Anal. Chem.* **61:** 367–75 (1989).
19. C.H. Lochmüller, C.E. Reese, and S. Hsu. Cross-column retention prediction in reversed-phase liquid chromatography using factor analytical modeling. *Anal. Chem.* **66:** 3805–13 (1994).
20. E. Malinowski. Statistical *F*-tests for abstract factor analysis and target testing. *J. Chemom.* **3:** 49 (1988).
21. E. Malinowski. Theory of errors in factor analysis. *Anal. Chem.* **49:** 606–12 (1977).

Manuscript accepted August 24, 1995.