Journal of Chromatography A, 1218 (2011) 6742–6749

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles

Paul G. Boswell[a,*], Jonathan R. Schellenberg[a], Peter W. Carr[b], Jerry D. Cohen[a], Adrian D. Hegeman[a]

[a] Department of Horticultural Science and the Microbial and Plant Genomics Institute, University of Minnesota, 1970 Folwell Avenue, St. Paul, MN 55108, USA
[b] Department of Chemistry, University of Minnesota, 207 Pleasant Street SE, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 13 May 2011
Received in revised form 9 July 2011
Accepted 21 July 2011
Available online 30 July 2011

Keywords:
Retention prediction
Chemical identification
Gradient profile
Retention projection
Cross-instrument

ABSTRACT

Isocratic retention data should make a suitable foundation for an accurate, cross-instrument LC retention prediction system. Our previous work suggested that in order to accurately calculate (or "project") gradient retention times on a wide range of HPLC systems using a single set of isocratic retention data, the precise shape of both the gradient and flow rate profiles produced by each instrument must be properly taken into account. However, accurate measurement of these system properties is difficult and time-consuming. In this work, we describe an approach that uses the measured gradient retention times of a set of standard solutes spiked into the sample along with their known isocratic retention vs. eluent composition relationships to determine the effective gradient and flow rate profiles *by back-calculation*. Retention "projections" of 20 other solutes using these back-calculated profiles, under various chromatographic conditions typical of metabolomics experiments, were *remarkably* accurate (as good as 0.23% of the gradient time, R^2 up to 0.99996), being very near the level of retention reproducibility. Our calculations suggest that this level of accuracy will allow a quadrupole MS to identify 38-fold more compounds out of a simulated mixture of 7307; it would allow an FTICR-MS to improve its identification rate nearly two-fold with the same mixture. Moreover, *very* little effort is required of the user. This approach provides a simple way to correct for all instrument-related factors affecting retention, allowing dramatically streamlined and improved retention projection across gradients, flow rates, and HPLC instruments.

1. Introduction

The development of an accurate, robust method to predict (i.e. to align expected and experimental) gradient elution LC retention, one which could be used with different gradient programs and flow rates and on many different makes and models of HPLC instruments, would be an extremely useful tool for the identification of compounds in complex samples such as those common to metabolomic studies. Not only would it *dramatically* improve the ability of LC–MS to identify compounds, but it would require virtually no additional capital investment and very little user time to take advantage of its potential. Indeed, calculations presented previously [1] suggested that use of LC retention in combination with exact mass measurements would enable relatively inexpensive quadrupole mass spectrometers to identify more compounds than even an LC-Fourier Transform Ion Cyclotron Resonance (FTICR) MS. Such a method for retention prediction would make it possible for the majority of laboratories without access to the more expensive instrumentation to perform state-of-the-art experiments in areas such as metabolomics, a field in which further progress heavily depends on improvements in the ability of instrumentation to *identify* compounds in complex mixtures [2,3]. Finally, accurate LC retention prediction would improve the ability of all types of LC–MS instrumentation to distinguish between isobaric compounds, which are very common in many important metabolites classes. In fact, retention prediction is already becoming very important for peptide identification in proteomics. There, retention times of peptides are usually calculated from amino acid sequence [4–6], but the correlation between experimental and predicted retention times are no better than $R^2 = 0.99$ (about ±3% of the gradient time, t_G). Similar retention predictions made for metabolites based on their chemical structure are even worse because of their vast chemical diversity relative to peptides [2].

Many general LC retention prediction schemes have been proposed [7–9], but for a variety of reasons, none have gained wide

* Corresponding author at: 328 Alderman Hall, 1970 Folwell Avenue, St. Paul, MN 55108, USA. Tel.: +1 612 250 5188.
E-mail addresses: bosw0011@umn.edu (P.G. Boswell), jtschellenberg@gmail.com (J.R. Schellenberg), petecarr@umn.edu (P.W. Carr), cohen047@umn.edu (J.D. Cohen), hegem007@umn.edu (A.D. Hegeman).

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.07.070

use. Some schemes require use of a rigid set of experimental conditions which must be strictly followed to reproduce retention times [10–12]. However, these schemes require HPLC equipment identical to that used to create the library be used. Not only does this make the systems inaccessible to most potential users, but intrinsically the library must eventually become obsolete.

Others have advocated for the use of retention indices [7–9]. Unfortunately, the accuracy of LC retention indexing is *fundamentally limited* because retention indices depend strongly on solvent composition [13,14]. This causes gradient retention indices in LC to be different under different gradient and/or flow rate profiles just as programmed-temperature retention indices in gas chromatography are sensitive to the temperature program and carrier gas flow rate [15–17]. In order for gradient LC retention indices to be accurate and reliable, one must rigorously reproduce the original experimental conditions in which they were measured. This greatly limits the usefulness of retention indices in metabolomics, where gradient elution is invariably used due to the sample complexity and no standard experimental conditions have been widely adopted. Other peak alignment schemes, such as "time warping" algorithms [18,19], also make serious assumptions about the relative retention behavior of compounds. Their accuracy is also fundamentally rather limited.

In our previous work [1], we explored the use of isocratic retention factor (k) vs. volume fraction of organic modifier (ϕ) relationships as the basis of a retention prediction system for gradient elution. Precise isocratic log k vs. ϕ relationships were measured on one LC–MS system and gradient retention was "projected" from them for a different LC–MS system. The system is *theoretically sound* and therefore does not suffer from the limitations of retention indexing approaches. In order for this approach to work, one must carefully measure and painstakingly correct for the effect of both gradient non-idealities (including dwell time, gradient dispersion, and solvent mis-proportioning) and flow rate non-idealities on different instruments. This is particularly true under the low flow rate conditions typical of many metabolomics experiments, where gradient profiles produced even by new HPLC instruments can be severely distorted [1,20]. However, measuring these non-idealities is difficult, time-consuming, requires a knowledgeable operator, and may require special instrumentation.

In this work, we present a new way to perform gradient retention prediction by use of an algorithm that enables significantly more accurate and precise retention projection while *drastically* reducing the amount of time and training required of a user. While considerable effort is required to develop a library, the end user only needs to measure the gradient retention times of a set of "instrument calibration solutes" which are spiked into the sample. The gradient retention times, in combination with their known log k vs. ϕ relationships (stored in the library, or "retention database") allow the back-calculation of *what the effective gradient and flow rate profiles must have been to produce the observed gradient retention times of the standards*. By applying the back-calculated profiles to *project* the gradient retention times of other solutes, flow rate non-idealities, dwell volume, solvent mis-proportioning, and gradient dispersion can be accounted for. Finally, since the instrument calibration solutes will be spiked into the actual sample, it can account for differences in any of these factors from run-to-run. As a test of the accuracy of this system, we use the algorithm with 35 chemically diverse solutes, designating 15 of them as instrument calibration solutes to allow accurate retention projection of the other 20.

In short, this system predicts the gradient retention of compounds in an isocratic retention database (a) among different HPLC/UHPLC instruments, (b) with chemically diverse solutes, (c) with different gradient profiles, and (d) with different flow rates, while other experimental conditions are fixed (standard make and model HPLC column, standard column temperature, standard eluents A and B). To further clarify how the system works, we also offer an online Java application with a step-by-step tutorial that walks you through the entire retention prediction process.

2. Experimental

2.1. Materials, methods, and equipments

Detailed experimental information is provided in the previous manuscript [1]. In short, a Waters Acquity BEH C_{18} (2.1 mm × 100 mm, 1.7 μm particle size) column was used for all experiments. The mobile phases were (A) 30 mM ammonium formate buffer (pH 2.80) in water (its preparation is described in the previous manuscript) and (B) 100% acetonitrile (ACN). The column temperature was held at 35 °C.

The primary LC–MS was comprised of a Thermo Fisher Scientific Inc. (Waltham, MA) Accela UHPLC pump and TSQ Quantum Access triple quadrupole MS. The secondary instrument was a Waters Corporation (Milford, MA) Acquity UPLC pump and Acquity SQD single quadrupole MS.

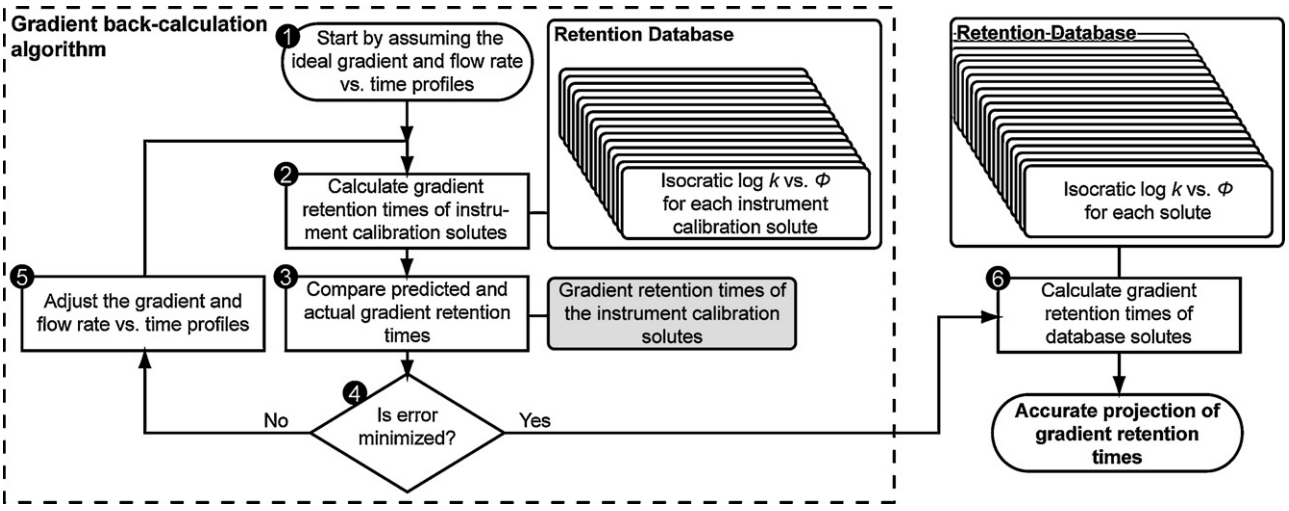
The 15 instrument calibration solutes used in this work were, in order of their elution: (1) adenosine, (2) *N*,*N*-dimethylacetamide, (3) *p*-toluenesulfonic acid, (4) *N*,*N*-diethylacetamide, (5) indole-3-acetic acid, (6) dimethyl phthalate, (7) indole, (8) diethyl phthalate, (9) diallyl phthalate, (10) di-*n*-propyl phthalate, (11) di-*n*-butyl phthalate, (12) di-*n*-pentyl phthalate, (13) di-*n*-hexyl phthalate, (14) di-*n*-heptyl phthalate, and (15) di-*n*-octyl phthalate. The 20 test solutes we used were: acetophenone, propiophenone, butyrophenone, valerophenone, hexanophenone, heptanophenone, octanophenone, nonanophenone, decanophenone, undecanophenone, dodecanophenone, *p*-coumaric acid, cortisone, indole-3-propionic acid, indole-3-butyric acid, 4-*n*-hexylaniline, nortriptyline, amitriptyline, 1-pentanesulfonic acid, and chlorogenic acid. The gradient retention times used in this work were not averaged from multiple experiments so as to allow run-to-run variation to be accounted for. The reproducibility of isocratic retention factors was 1.2%.

2.2. Software and calculations

The gradient back-calculation algorithm was written in Mathematica 7.0.1.0 (Wolfram Research, Inc., Champaign, IL). A Mathematica notebook containing this algorithm is included in the Supporting Information. The Retention Projector demonstration applet at http://www.retentionprediction.org was compiled for compliance with the Java 1.6 (Oracle, Redwood Shores, CA) runtime environment. It includes the Java OpenGL (JOGL) binding library version 1.1.1 (JogAmp, http://jogamp.org), JavaHelp version 2.0.05 (Oracle, Redwood Shores, CA), and SwingX library version 1.6.2 (Oracle, Redwood Shores, CA).

3. Theory

Scheme 1 shows a general flow chart for the gradient back-calculation algorithm. Each numbered step in the flow chart is described in the corresponding section below. In short, the algorithm begins by assuming that gradient and flow rate profiles are ideal. Then, through successive iterations (Fig. 1) of a sectioning search algorithm, effective gradient and flow rate profiles are adjusted to minimize variance in computed vs. measured retention times of the instrument calibration solutes. These optimized gradient and flow rate profiles can then be used to project retention for any solute whose log k vs. ϕ relationship has been characterized



Scheme 1. Flow chart for retention projection by gradient and flow rate profile back-calculation. The gray box, "Gradient retention times of the instrument calibration solutes", contains the only experimental measurements that must be supplied by the user.

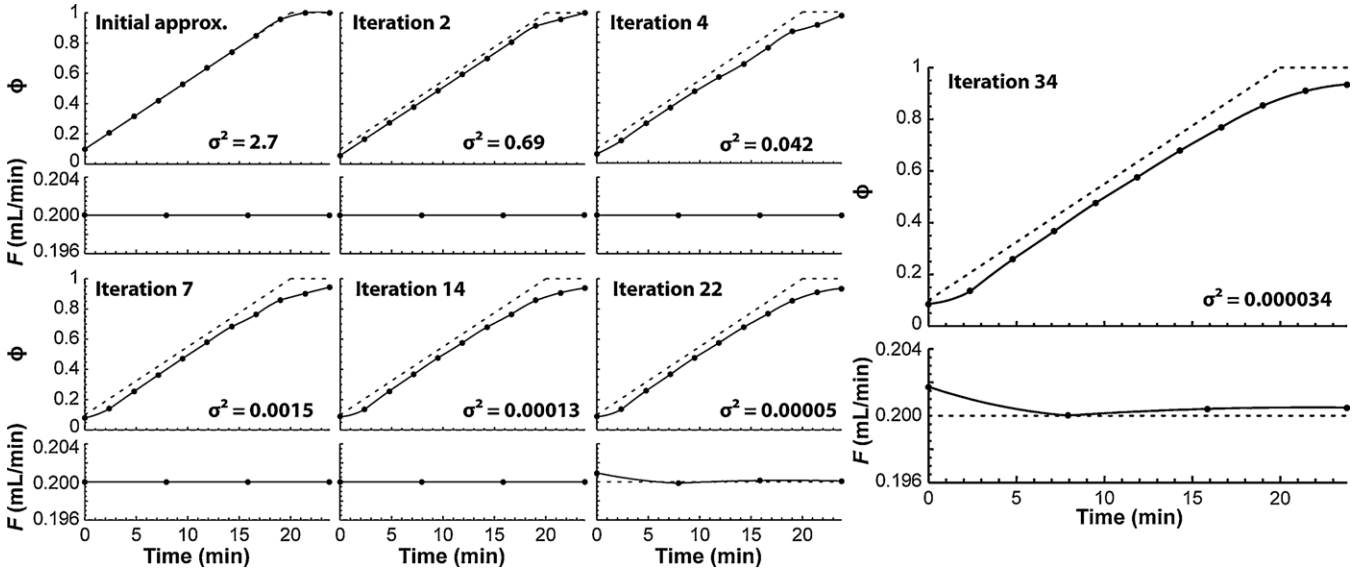


Fig. 1. The back-calculated gradient and flow rate profiles are determined iteratively. Here, selected iterations are shown (solid lines), overlaid on the programmed gradient and flow rate profiles (dotted lines). Variance was calculated from error in retention projections of the 15 instrument calibration solutes.

on the standard stationary phase. Each step of the back-calculation algorithm is explained in more detail below.

A hands-on demonstration of the retention projection system is available at http://www.retentionprediction.org. In addition to providing a step-by-step walkthrough of the retention prediction process, the website shows one possible implementation of the retention prediction system described here.

3.1. Step 1: collect information

The algorithm requires three pieces of information from the user: (1) the retention times of the instrument calibration solutes that were spiked into the sample, (2) the programmed gradient (the initial ϕ, ϕ_i, final ϕ, ϕ_f, and the gradient time, t_G) and (3) the programmed flow rate, F_{prog}. Using this information, initial working gradient and flow rate profiles are generated. We defined each profile using a set of data points spaced out evenly over time. Those points are interpolated (using second order Hermite interpolation) to yield a continuous function that represents the gradient profile or the flow rate profile (top left profiles in Fig. 1).

Because only 15 retention times are used to calculate both the gradient and the flow rate profiles, the number of interpolated data points that can be used to describe the two profiles together is limited to 15. Here, we allocate 11 and 4 data points to the gradient and flow rate profiles respectively. More data points are used to model the gradient profile because retention is more sensitive to errors in ϕ than in the flow rate, F, and because there are more features (changing slopes) to capture in the gradient profile than in a flow rate profile (the flow rate was programmed to be constant). In both profiles, the first interpolated data point is placed at time 0 and the last point is placed at the retention time of the latest eluting instrument calibration solute. The other points are distributed evenly between them; this works because the instrument calibration solutes are selected to elute over an approximately evenly spaced, wide range of retention times, so that there is information about the shape of the entire gradient.

3.2. Step 2: project retention times of the instrument calibration solutes

Using the new working gradient and flow rate profiles, the retention times of the instrument calibration solutes are calculated by numerical integration of the fundamental equation of gradient elution.

$$\int_0^{t_R - t_0} \frac{1}{t_0}\frac{dt}{k_\phi} = 1 \quad (1)$$

where t_0 is the dead time, k_ϕ is the isocratic retention factor of a solute with mobile phase composition ϕ, t_R is the gradient retention time of the solute, and dt is the time slice of the gradient program experienced by the solute. Eq. (1) calculates gradient retention times by considering gradient elution as a series of very small isocratic steps that together closely approximate the true shape of the gradient. It can also be written in an equivalent form that does not assume a constant dead time [21]:

$$\int_0^{t_R} \frac{dt_c}{t_{0,\phi}(1+k_\phi)} = 1 \quad (2)$$

where $t_{0,\phi}$ is the dead time at a certain ϕ and dt_c is the time that a solute is under the influence of a particular time slice of the gradient as it moves through the column:

$$dt_c = dt + \frac{dt}{k_\phi} \quad (3)$$

To account for variation in flow rate with time, we replace $t_{0,\phi}$ with $V_{m,\phi}/F_{t_c}$, where $V_{m,\phi}$ is the kinetic void volume [22,23] (see previous manuscript [1] for description and experimental details) at a certain ϕ and F_{t_c} is the flow rate at a certain t_c. This gives

$$\int_0^{t_R} \frac{F_{t_c}\,dt_c}{V_{m,\phi}(1+k_\phi)} = 1 \quad (4)$$

Using the following summation equation, analogous to Eq. (3), for numerical integration:

$$\sum_{i=1}^{n} \frac{F_{t_c}\,\delta t_c}{V_{m,\phi}(1+k_\phi)} \geq 1 \quad (5)$$

where n is the smallest integer that makes the inequality true, t_R can be calculated from:

$$t_R = \sum_{i=1}^{n} \delta t_c = \delta t \sum_{i=1}^{n} 1 + \frac{1}{k_\phi} \quad (6)$$

We used an implementation of Eqs. (5) and (6) to calculate the gradient retention times of each solute with the following step sizes, δt: for t_G = 5 min, δt = 0.025 min; for t_G = 20 min, δt = 0.05 min; for t_G = 80 min, δt = 0.1 min.

3.3. Step 3: compare projected and experimental retention times

In this step, the accuracy of the working gradient and flow rate profiles is determined by comparing the projected retention times with the experimental retention times of the instrument calibration solutes. The error is recorded as the variance in their retention projections (where the variance equals the average of the squares of the error among the instrument calibration solutes).

3.4. Step 4: test termination condition (Is error minimized?)

We found that if both the gradient and flow rate profiles are simultaneously optimized from the beginning of optimization, they often fall into a local minimum in which the flow rate profile is distorted. This is because the initial approximation of the gradient profile is significantly more inaccurate than that of the flow rate profile. During the first iterations, the flow rate profile becomes distorted to compensate for the error in the gradient profile and the distortion never works itself back out in later iterations.

Instead, a two-phase approach proved necessary. In phase I, only the gradient profile is optimized in each iteration while in phase II, the gradient profiles and the flow rate profiles are co-optimized in each iteration. Optimization begins in phase I and progresses to phase II when the sum squared error in retention projection improves by ≤10% from one iteration to the next. Once in phase II, the optimization terminates when the variance in the retention projections improves by ≤1% from one iteration to the next.

3.5. Step 5: adjust the gradient and flow rate profiles

During phase I optimization, the gradient profile is adjusted one point at a time. The value of ϕ for each point is optimized using a sectioning search algorithm. The algorithm starts with the earliest point in time and ends with the latest point in time. Because points are optimized one at a time, they often overcorrect for errors in other points, particularly at the beginning of optimization. To avoid extreme overcorrection, we defined the maximum allowable change in ϕ during any one iteration to be ±0.02. In all, three termination conditions were defined for optimization of each point in the profile: (1) the sectioning search algorithm optimized ϕ to within 0.00001 of the optimal value, (2) ϕ changed by more than ±0.02, or (3) further optimization would require ϕ to be greater than 1 or less than 0.

During phase II optimization, each iteration begins with optimization of each point in the gradient profile followed by each point in the flow rate profile. The flow rate profile is optimized in a similar way as the gradient profile, except that the termination conditions for optimization of each point in the profile are defined as: (1) the sectioning search algorithm optimized ϕ to within Fprog/100,000 of the optimal value, (2) F changed by more than $\pm F_{prog}/20$, or (3) further optimization would require F to be less than 0.

During the optimization process, as adjustments are made to the gradient and flow rate profiles to reduce error in gradient retention projections, one might imagine that this error could be equally compensated by changes in either the gradient profile or the flow rate profile. However, the equations governing gradient retention force the error to be properly compensated by the correct source. This is discussed in more detail in the Supporting Information.

3.6. Step 6: project gradient retention times of other solutes

Once the final termination condition for profile optimizations has been reached, the working gradient and flow rate profiles are fully optimized and can be used to accurately project retention of other solutes. The retention projections are performed in the same way as described in step 2.

4. Results and discussion

4.1. Back-calculated gradient and flow rate profiles

While gradient retention times were measured on both LC–MS instruments, isocratic retention data were only measured on one instrument, which we refer to as the "primary" LC–MS instrument. The other we call the "secondary" LC–MS.

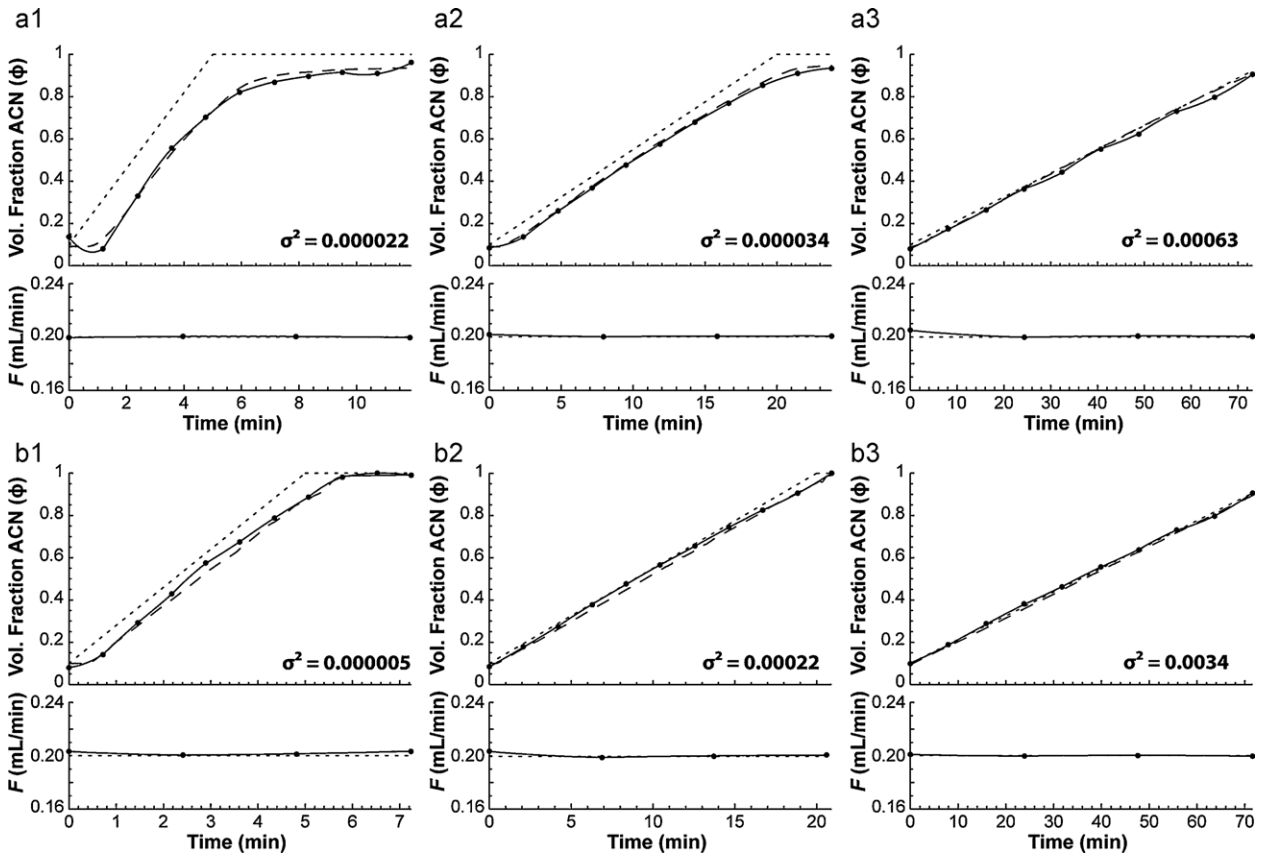


Fig. 2. Six of the back-calculated gradient and flow rate profiles (solid lines) overlaid on the ideal ones (dotted line) and those measured by UV absorbance (dashed lines). a1, a2, and a3 are 5 min/200 μL/min, 20 min/200 μL/min, and 80 min/200 μL/min gradients, all measured on the *primary* LC–MS. b1, b2, and b3 are the same respective gradients, measured on the *secondary* LC–MS. The variances (σ^2) of retention projections of the instrument calibration solutes are also shown.

The 35 solutes used in this study were selected to give a strong signal with electrospray-ionization MS and cover a wide range in retention. Most importantly, they were chosen to be very chemically diverse. They represent a range of all 5 types of interactions identified in the hydrophobic-subtraction model of reversed-phase LC which are known to contribute to retention on type-B alkylsilica stationary phases [24,25] and therefore encompass the retention characteristics of virtually any small molecule/metabolite (see previous manuscript [1]). Of the 35 solutes, we selected 15 to be used as instrument calibration solutes. The instrument calibration solutes were selected simply because they eluted evenly over a wide range of retention times (see Section 2). Otherwise, there is nothing special about the 15 compounds. The optimal number of instrument calibrations solutes was not determined, but will be the subject of future investigation; 15 proved sufficient to demonstrate the utility of this method of retention prediction. Their retention times, under each set of experimental conditions on the primary and secondary instruments, can be found in the previous manuscript [1].

Fig. 2 shows some of the back-calculated gradient and flow rate profiles (the rest are in Figs. S-2 and S-3) determined from these retention times. Each of the back-calculated gradient profiles are very similar to the same gradients measured by UV absorption, but as will be shown, the subtle differences between the profiles determined each way are *key* to providing improved accuracy in subsequent retention projections.

4.2. Retention projections from back-calculated gradient and flow rate profiles

Fig. 3 shows the error in gradient retention projection of the 20 test solutes (tabulated data are in Tables S-1–S-5). The projected gradient retention times using the back-calculated gradient and flow rate profiles are considerably more accurate than those we obtained by using measured gradient and ideal flow rate profiles as in our previous work [1]. Table 1 shows the standard deviation in retention projections for each LC–MS under each gradient condition (right column). For reference, the middle column shows the "baseline error", that is, the minimum possible error we could expect in our retention projections due to the random error in our isocratic retention measurements (see previous manuscript [1] for details

Table 1
Projected gradient retention time accuracy using the back-calculated gradient and flow rate profiles on the primary and secondary LC–MS.

	Baseline error[b]	Retention projection error[c]
Primary LC–MS gradient[a]		
5 min, 200 μL/min	0.29% (±0.86 s)	0.87% (±2.4 s, R^2 = 0.99990)
20 min, 200 μL/min	0.29% (±3.4 s)	0.28% (±3.3 s, R^2 = 0.99994)
80 min, 200 μL/min	0.29% (±13.8 s)	0.34% (±16 s, R^2 = 0.9998)
20 min, 100 μL/min	0.29% (±3.4 s)	0.35% (±4.2 s, R^2 = 0.99996)
20 min, 400 μL/min	0.29% (±3.4 s)	0.35% (±4.2 s, R^2 = 0.99990)
Secondary LC–MS gradient[a]		
5 min, 200 μL/min	0.29% (±0.86 s)	0.62% (±1.9 s, R^2 = 0.9998)
20 min, 200 μL/min	0.29% (±3.4 s)	0.25% (±3.0 s, R^2 = 0.99995)
80 min, 200 μL/min	0.29% (±13.8 s)	0.34% (±16 s, R^2 = 0.99990)
20 min, 100 μL/min	0.29% (±3.4 s)	0.23% (±2.8 s, R^2 = 0.99995)
20 min, 400 μL/min	0.29% (±3.4 s)	0.59% (±7.0 s, R^2 = 0.9997)

[a] t_G, flow rate. Gradients are linear from ϕ = 0.1 to 1.
[b] Baseline error is calculated from isocratic retention reproducibility, as discussed previously [1]. It is expressed as the standard deviation as a percentage of t_G; in parenthesis: one standard deviation.
[c] Error is expressed as the standard deviation among all test solutes as a percentage of t_G; in parenthesis: one standard deviation, R^2 correlation between measured and projected retention times.

of its calculation; gradient reproducibility was not considered here because gradient and flow rate profiles were back-calculated from retention measurements made in the *same* run as the retention projections were made).

While in our previous work [1], the retention projection accuracy averaged 4–7 times the baseline error, the retention projection accuracy obtained with the approach described here is nearly equal to the baseline error in 4 of the 5 gradient conditions (all but the 5 min gradient) on both LC–MS systems. The average of the retention projection error under those 4 conditions was 0.33% of t_G on the primary LC–MS and 0.35% of t_G on the secondary LC–MS, both extremely close to the 0.29% baseline error. *This indicates that under those conditions, virtually all instrument-dependent factors affecting retention were accounted for by the gradient and flow rate profiles back-calculated from the retention times of the standards*. Even in the case of the 5 min gradient, the retention projection errors were very small (±1.9 s and ±2.4 s), despite the fact that they were 2.2- and 2.8-fold greater than the baseline error (±0.86 s). It may be that the bigger error is caused by "solvent demixing", which describes a non-equilibrium process in gradient elution caused by selective sorption of the mobile phase (in which ϕ is steadily increasing) by the stationary phase [26,27]. This effect is more pronounced with steeper gradients and could not be fully accommodated by the back-calculated profiles.

While it is difficult to compare the accuracy of these retention projections to those in the literature (they are all determined for only a small set of compounds and under experimental conditions more favorable for making accurate retention projections), the most accurate retention projections until now seem to have error of approximately ±1% of t_G [26,28,29]. Here, projection error was consistently below that, even on the secondary instrument. In fact, the error was low enough to be on the order of the peak width (which was between 3 s and 6 s, full width at half maximum).

The accuracy achieved on the secondary LC–MS is more than enough to *substantially* improve compound identification. To test how much, we did a similar simulation as in the previous manuscript [1]. Briefly, we used a set of 7307 compounds from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [30–32] and made very rough estimates of their retention times on a C_{18} column in a 60 min gradient. From that, we calculated the number of compounds that could be identified with ≥99% confidence by MS (using exact mass alone) and then by MS *in combination with* LC retention data. With the most accurate retention predictions *achieved here* (0.23% of t_G), an LC–quadrupole MS (resolution = 0.4 FWHM, mass accuracy = 0.5 amu), went from being able to identify only 88 compounds (1%) to 3355 compounds (46%). Similarly, an LC–FTICR MS (resolving power = 1,000,000, mass accuracy = 1 ppm), went from being able to identify 2682 compounds (37%) to 5116 compounds (70%). Even in the case of the worst error achieved here (0.62% of t_G), exact mass in combination with LC retention data could allow the quadrupole MS to identify 1853 compounds (25%) and the FTICR MS to identify 4512 compounds (62%). To further illustrate the potential significance of such accurate retention predictions, consider that peptide retention predictions are already being used to improve confidence in peptide identification despite maximum R^2 correlations between predicted and experimental retention times of 0.99. Here, we achieved R^2 correlations of up to 0.99996.

The only instrument parameter that had a big effect on either the baseline error or the projection error was the gradient time. At 5 min, retention projection accuracy began to deteriorate. Besides that, one might expect that viscous frictional heating [33] would cause significant error in retention projections among different flow rates, especially when using a stationary phase of 1.7 μm particles [34]. One might also expect that pressure-induced changes in the retention behavior [35,36] of different compounds could give rise to a significant loss of retention projection accuracy at different flow rates. However, in the limited flow rate range used here (100 μL/min to 400 μL/min; almost a 4-fold change in back-pressure), we observe no significant loss in retention projection accuracy.

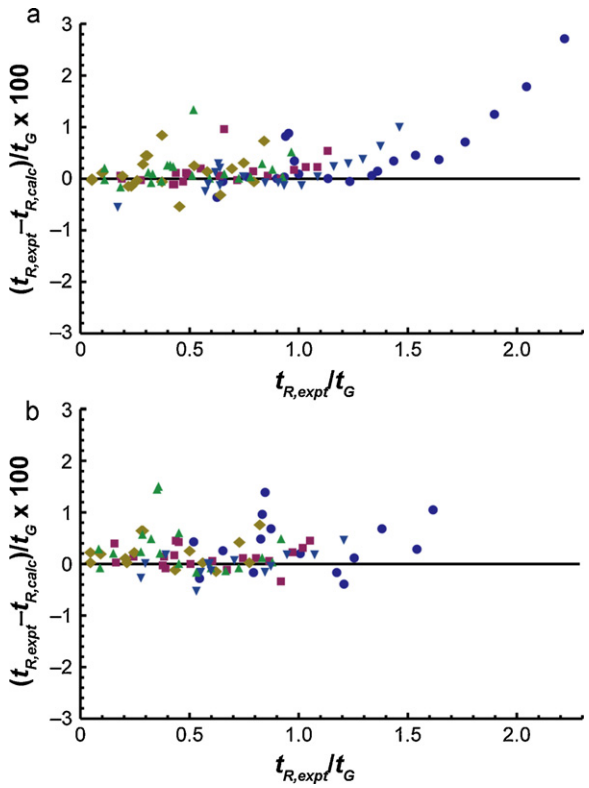


Fig. 3. Retention projection error of each test solute under the five gradient conditions (● = 5 min/200 μL/min, ■ = 20 min/200 μL/min, ◆ = 80 min/200 μL/min, ▼ = 20 min/100 μL/min, ▲ = 20 min/400 μL/min gradient) on (a) the primary LC–MS and (b) the secondary LC–MS.

Among the 20 test solutes whose retention times were projected, some of the largest errors in those projections were for the charged solutes. For example, if the projections for chlorogenic acid, amitriptyline, nortriptyline, and 4-*n*-hexylaniline are discarded in the 20 min/200 μL/min gradient set, the error in the projection of the remaining compounds decreases from ±3.3 s to ±2.2 s on the primary instrument and from ±3.0 s to ±2.4 s on the secondary instrument. We expected this based on our observations in the previous manuscript [1] that variances in the isocratic retention times of charged solutes were greater than those of neutral solutes. This indicated changes in the stationary phase selectivity, specifically the portion of stationary phase selectivity resulting from Coloumbic interactions, which are known to drift [37,38].

4.3. Comparison with retention indexing

The question remains: how does the described retention projection system compare with a retention indexing system? First, we give a brief description of retention indexing.

In isocratic LC retention indexing systems [7–9], a set of standard compounds from a homologous series are spiked into the sample and the retention behavior of a solute is characterized relative to the two standards that "bracket" it. The relative retention of the solute is called a retention index (RI), defined as:

$$\mathrm{RI} = 100C_n + 100(C_{n+i} - C_n)\frac{\log k_x - \log k_n}{\log k_{n+i} - \log k_n} \quad (7)$$

where k_x is the retention factor of a solute x, k_n and k_{n+i} are the retention factors for the standard solutes n and $n+i$ eluting before and after the solute x, and C_n and C_{n+i} are the number of carbon atoms in standard solutes n and $n+i$. Isocratic retention indices are a considerably more reproducible measure of retention because many experimental factors (e.g. flow rate, temperature) affecting the absolute retention of a solute also affect the absolute retention of the standards, cancelling out much of their effect. However, the indices are only accurate at the mobile phase composition in which they were measured because they do not account for differences in the log k vs. ϕ relationships among solutes. For example, a 20% change in the mobile phase volume percentage of methanol caused the RI of aspirin to drop from 302 to 8 [14].

In gradient elution, retention factors have no practical meaning because they change with the gradient program. Instead, retention indices may be calculated according to the following equation [39]:

$$\mathrm{RI} = 100C_n + 100(C_{n+i} - C_n)\frac{t_{R(x)} - t_{R(n)}}{t_{R,(n+i)} - t_{R(n)}} \quad (8)$$

where $t_{R(x)}$ is the retention time of a solute x, and $t_{R(n)}$ and $t_{R(n+i)}$ are the retention times of the standard solutes n and $n+i$ eluting before and after the solute x. Just as isocratic retention indices change with the solvent composition, gradient retention indices are inherently sensitive to changes in the gradient profile.

To test the accuracy of retention indexing compared to the retention projection system proposed in this work, we used the same 15 instrument calibration solutes as retention index standards. Since these compounds are not all part of a homologous series, the standards were assigned retention indices based on the order of their elution, n. Therefore, C_n and C_{n+i} in Eq. (8) were replaced by n and $n+1$. The retention index of each test solute was calculated from its measured retention time under one gradient condition (gradient time and flow rate) to predict its retention times under different conditions on the same instrument. When the 20 min/200 μL/min gradient on the primary instrument was used to calculate retention indices, the average error in the predicted retention times (expressed as % of t_G) on the same instrument were as follows: in the 5 min/200 μL/min gradient, 4.6% error; in the 80 min/200 μL/min gradient, 2.4% error; in the 20 min/100 μL/min gradient, 1.7% error; in the 20 min/400 μL/min gradient, 1.5% error. All of these show more than 4-fold greater error than observed with the retention projection method described above. That is, a simple retention indexing system cannot tolerate changes in the gradient conditions. When the difference between the gradient profile used to calculate the retention indices and the gradient profile used to predict retention times is even larger, the error is even greater, e.g. using retention indices calculated from the 5 min/200 μL/min gradient, the average retention time prediction accuracy in the 80 min/200 μL/min gradient was 4.9%.

On the other hand, when retention indices calculated from data collected on the primary instrument were used to predict retention under the *same* gradient conditions on the secondary instrument, average retention prediction error was considerably less. For example, using retention indices calculated from the 20 min/200 μL/min gradient on the primary instrument, average retention time prediction accuracy for the same gradient on the secondary instrument was only 0.31%. However, in the case of the 5 min/200 μL/min gradients, where the profiles produced by the two instruments were quite different, the average error was 1.1%, which is 30% more error than was observed using the retention projection system under the same gradient.

So while retention indexing may be sufficient to accurately predict gradient retention times between different HPLC instruments *when the gradient programs are the same*, it is considerably less accurate than the retention projection system when gradient programs are different. This means that in order to reliably and accurately predict retention, a user of a retention indexing system would always need to use the same gradient time, initial and final compositions, column size, and flow rate as used to measure the retention indices, leaving virtually no room for method development. Even then, the retention index system is less accurate when gradient profiles differ significantly between HPLC instruments. In contrast, the retention projection system proposed here enables accurate retention prediction over a wide range of gradient programs and flow rates, and between different HPLC instruments.

To be fair, one could imagine using a more complicated version of retention indexing in which isocratic retention indices are precisely known as a function of solvent composition. However, to predict gradient retention times from the RI vs. ϕ relationships, the precise gradient and flow rate profiles would need to be known to properly use these relationships in the calculation of gradient RIs. This brings us back to the original difficulty of acquiring precise gradient and flow rate profiles. The back-calculation algorithm presented here could also be applied in that situation to collect the information, but then the system becomes nearly identical to the one we already discussed.

4.4. Feasibility of building an isocratic retention database of k vs. ϕ

Building an isocratic retention database containing k vs. ϕ relationships for each compound would require a lot of data to be collected, particularly if each compound is characterized on more than one standard stationary phase. For example, if isocratic retention factors were measured at 11 different eluent compositions and on 2 different standard stationary phases (possibly providing additional means for compound identification if the stationary phases have sufficiently orthogonal selectivity [12]), then at maximum, 22 measurements could be made for each compound. If measurement of each isocratic retention factor required 1 h, it would take approximately 1 day to measure all the retention data for each compound.

Fortunately, this process could be greatly streamlined by measuring data simultaneously in groups of ~30 compounds (LC–MS could be used to distinguish between them). It could be optimized further by taking advantage of the fact that there is only a narrow window of solvent compositions in which most compounds are usefully retained. By grouping compounds based on their retention times in a scouting gradient, isocratic retention could be measured at fewer solvent compositions. Once retention data is measured for a large number and wide variety of compounds, a suitable fitting function could be determined to further reduce the number of necessary measurements for each compound. Many such fitting functions exist [40], but it is unclear whether the published equations would adequately fit the log k vs. ϕ relationships of a very diverse set of compounds run under the chromatographic conditions in which the database is to be built. Furthermore, a fast flow rate could be used to speed up these isocratic retention measurements. A flow rate 3 times faster would reduce measurement time 3-fold. All together, these optimizations mean that the speed of the measurements would probably not be a rate-limiting step – almost 1260 compounds could theoretically be measured per HPLC instrument per week (assuming isocratic measurements at 6 solvent compositions per compound, each taking 20 min to complete at the fast flow rate).

5. Conclusions

The retention projection system described here clearly provides the means for very accurate, cross-instrument retention predic-

tion, even under challenging experimental conditions (i.e. low flow rate, small particle size, small column inner diameter) commonly used in metabolomics experiments. Evidently, this method accounts for nearly all significant instrument-dependent factors controlling retention as the retention projection accuracy was virtually equal to the random error expected based on the isocratic retention reproducibility (except in the case of the steepest gradient). To the best of our knowledge, the accuracy achieved here (from 0.23% to 0.87% of t_G) is significantly better than any previous reports of retention projection. In fact, it is more than sufficient to be useful as a supplementary tool for LC–MS compound identification. Our calculations indicate that supplementing exact mass information with retention information (accurate to that achieved here) could enable an LC–single quadrupole MS to identify significantly more compounds than an LC–FTICR MS not so assisted whereas an LC–FTICR MS with such information could *double* its rate of identification. Yet despite the emphasis on high accuracy, the biggest benefit of this system may be that it is extremely simple to use – the only experimental information the user must provide are the gradient retention times of a small group of instrument calibration solutes (15 in this work) spiked into the sample.

Our comparison of retention projection with retention indexing showed that both systems have high cross-instrument accuracy *when the gradient and flow rates are the same*, but that retention projection is considerably more accurate when predicting retention under gradient and flow rate profiles different than those used to calculate the retention indices. The flexibility of being able to use a range of gradient profiles and flow rates should make the described retention projection system much more practical for end users.

To demonstrate how the retention prediction system works and show a possible software implementation of the algorithm, we offer an online Java application with a step-by-step tutorial. Also, from the Java application, one can clearly see another possible benefit of using the retention projection system: when the back-calculated gradient and flow rate profiles are made visible to the user, they can serve as diagnostics of instrument performance. This might allow users to quickly discover instrument problems based on changes in the profiles.

Though the retention projections in this work were remarkably accurate, they might be further improved. Since the accuracy seems to be limited by the isocratic retention reproducibility (particularly of charged solutes [1]), the greatest improvement in accuracy may come from selection of extremely robust, fast-equilibrating [41] stationary phases with which to build the retention database, by selecting eluent compositions that promote fast column equilibration (e.g. by also adding buffer to eluent B) [41], and by developing a means to account for stationary phase selectivity drift. In addition, accounting for solvent demixing may improve retention projection accuracy under steep gradient conditions. In future work, we plan to address these issues.

Acknowledgements

We thank the National Science Foundation [IOS-0923960 and MCB-0725149], the Minnesota Agricultural Experiment Station, and the Gordon and Margaret Bailey Endowment for Environmental Horticulture for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2011.07.070.

References

[1] P.G. Boswell, J.R. Schellenberg, P.W. Carr, J.D. Cohen, A.D. Hegeman, J. Chromatogr. A, this issue.
[2] J.L. Griffin, Philos. Trans. Roy. Soc. B 361 (2006) 147.
[3] O. Fiehn, Plant Mol. Biol. 48 (2002) 155.
[4] O.V. Krokhin, V. Spicer, Anal. Chem. 81 (2009) 9522.
[5] T. Bączek, R. Kaliszan, Proteomics 9 (2008) 835.
[6] K. Petritis, L.J. Kangas, B. Yan, M.E. Monroe, E.F. Strittmatter, W.-J. Qian, J.N. Adkins, R.J. Moore, Y. Xu, M.S. Lipton, D.G. Camp, R.D. Smith, Anal. Chem. 78 (2006) 5026.
[7] R.M. Smith, J. Chromatogr. A 236 (1982) 313.
[8] J.K. Baker, Anal. Chem. 51 (1979) 1693.
[9] M. Bogusz, R. Aderjan, J. Chromatogr. A 435 (1988) 43.
[10] T.R. Sana, S.M. Fischer, S. Jenkins, M.T. Matyska, J.J. Pesek, Two pH Optimized LC–MS Methods for Metabolomics Analysis of Hydrophilic Compounds on Silica Hydride Stationary Phases, Presented at the 58th Annual Conference on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, May 26, 2010.
[11] S. Moco, R.J. Bino, O. Vorst, H.A. Verhoeven, J. de Groot, T.A. van Beek, J. Vervoort, C.H.R. de Vos, Plant Physiol. 141 (2006) 1205.
[12] J. Zeng, X. Zhang, Z. Guo, J. Feng, X. Xue, X. Liang, J. Chromatogr. A 1218 (2011) 1749.
[13] R.M. Smith, T.G. Hurdley, R. Gill, A.C. Moffat, Chromatographia 19 (1984) 401.
[14] R.M. Smith, N. Finn, J. Chromatogr. A 537 (1991) 51.
[15] L. Weber, J. High Resol. Chromatogr. 9 (1986) 446.
[16] H. van Den Dool, P.Dec. Kratz, J. Chromatogr. A 11 (1963) 463.
[17] S. Yiliang, Z. Ruiyan, W. Qingqing, X. Bingjiu, J. Chromatogr. A 657 (1993) 1.
[18] E. Lange, R. Tautenhahn, S. Neumann, C. Gropl, BMC Bioinformatics 9 (2008) 375.
[19] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[20] M.A. Quarry, R.L. Grob, L.R. Snyder, J. Chromatogr. A 285 (1984) 1.
[21] P. Nikitas, A. Pappa-Louisi, Anal. Chem. 77 (2005) 5670.
[22] M. Wang, J. Mallette, J.F. Parcher, Anal. Chem. 80 (2008) 6708.
[23] J.H. Knox, R. Kaliszan, J. Chromatogr. A 349 (1985) 211.
[24] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, J. Chromatogr. A 961 (2002) 171.
[25] L.R. Snyder, J.W. Dolan, P.W. Carr, J. Chromatogr. A 1060 (2004) 77.
[26] M.A. Quarry, R.L. Grob, L.R. Snyder, J. Chromatogr. A 285 (1984) 19.
[27] P. Jandera, J. Chromatogr. A 965 (2002) 239.
[28] P. Jandera, M. Kucerová, J. Chromatogr. A 759 (1997) 13.
[29] L.R. Snyder, J.W. Dolan, High-Performance Gradient Elution, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
[30] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, Nucleic Acids Res. 34 (2006) D354.
[31] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, Nucleic Acids Res. 38 (2010) D355.
[32] M. Kanehisa, S. Goto, Nucleic Acids Res. 28 (2000) 27.
[33] H. -jye Lin, S. Horváth, Chem. Eng. Sci. 36 (1981) 47.
[34] F. Gritti, M. Martin, G. Guiochon, Anal. Chem. 81 (2009) 3365.
[35] M.M. Fallas, U.D. Neue, M.R. Hadley, D.V. McCalley, J. Chromatogr. A 1209 (2008) 195.
[36] X. Liu, D. Zhou, P. Szabelski, G. Guiochon, Anal. Chem. 75 (2003) 3999.
[37] D.H. Marchand, L.A. Williams, J.W. Dolan, L.R. Snyder, J. Chromatogr. A 1015 (2003) 53.
[38] D.H. Marchand, L.R. Snyder, J.W. Dolan, J. Chromatogr. A 1191 (2008) 2.
[39] P. Kuronen, in: R.M. Smith (Ed.), Retention and Selectivity in Liquid Chromatography—Prediction, Standardisation and Phase Comparisons, Elsevier, Burlington, MA, 1995, p. 209.
[40] U.D. Neue, H.-J. Kuss, J. Chromatogr. A 1217 (2010) 3794.
[41] A.P. Schellinger, D.R. Stoll, P.W. Carr, J. Chromatogr. A 1192 (2008) 54.