Journal of Chromatography B, 877 (2009) 433–440

Contents lists available at ScienceDirect

Journal of Chromatography B

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

Standardization of retention time data for AMT tag proteomics database generation

I.A. Tarasova[a], V. Guryča[b,c], M.L. Pridatchenko[a], A.V. Gorshkov[d], S. Kieffer-Jaquinod[b,c], V.V. Evreinov[d], C.D. Masselon[b,c,*], M.V. Gorshkov[a]

[a] Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia
[b] CEA, iRTSV, Laboratoire d'Etude de la Dynamique des Protéomes, Grenoble F-38054, France
[c] INSERM, U880, Grenoble, F-38054, France
[d] N.N. Semenov's Institute of Chemical Physics, Russian Academy of Sciences, Moscow, Russia

ARTICLE INFO

Article history:
Received 4 September 2008
Accepted 17 December 2008
Available online 25 December 2008

Keywords:
Reversed phase liquid chromatography
Retention time correlation
Proteomics
AMT tag
Mass spectrometry

ABSTRACT

The combination of liquid chromatography (LC) with mass spectrometry (MS) has become a mainstream proteome analysis strategy. In LC–MS, measured masses possess their "universal" scale derived from atomic mass tables. In contrast, the observed LC retention times (RT) are not tied to a conventional time scale, and depend on experimental conditions. However, RT data, being explicitly orthogonal to MS, offer relevant information for proteome characterization. We present here a strategy for peptides RT data standardization, based on the generation of a standard scale using retention prediction models, which enables sharing of identification databases in the context of multi-laboratory research.

1. Introduction

Modern proteomics strategies often rely on the so-called "Shotgun" approach, based on a combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS) for identifying peptides in complex mixtures of digested proteins. Most often, the chromatographic separation is merely used as a mean of reducing complexity of the mixture delivered to the mass spectrometer. Nonetheless, the possibility to use both MS/MS and LC data for peptide identification and sequencing has attracted considerable interest [1–7], given that chromatography provides information about the primary structure which is complementary to the MS data.

Because it can provide high-quality separation for a great variety of chemical species, reverse-phase high performance liquid chromatography (RP-HPLC) is a preferred method for the separation of complex mixtures according to the analyte hydrophobicity and size. In proteomics, RP-HPLC using linear solvent gradients with aqueous/organic mobile phases is by far the most frequent, and provides superior results for proteins and peptides separation prior to mass spectrometry (MS). In these applications, the need to balance LC separation efficiency and MS detection requirements restricts the use of mobile phases. For the same reasons, columns for proteomics application have to conform to strict quality criteria since the weakly buffered mobile phases used can contribute to poor peak shape if free silanols or residual metals are present. It is therefore not surprising to see similar chromatographic methods in most reports on proteomics, with linear gradients from water to acetonitrile, using formic or acetic acid as ion pairing reagent, and separations at room temperature. Nevertheless, differences in RP-HPLC protocols published in the proteomics literature include changes in gradient steepness, flow rate, and column parameters such as length, diameter, particles diameter and pores size. This results in different observed retention times (RT) measured for the same species in different research laboratories.

A noted trend in proteome analyses is the increase in processing of data content during LC–MS/MS experiments, which are compiled into continuously updated databases. In particular, high throughput methodologies relying on Accurate Mass and retention Time (AMT) measurements are increasingly gaining momentum [8–11]. It is worth mentioning that identification database compilation is a labor-, sample-, and time-consuming task, which has to be repeated in each laboratory working on a given proteome due to the specificity of the measured RT. It would undoubtedly be very beneficial to translate these databases across laboratories working on the same biological material. In addition, there is growing awareness, in the proteomics community, of the need to provide means to fairly

* Corresponding author.
E-mail address: christophe.masselon@cea.fr (C.D. Masselon).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.12.047

compare data obtained across laboratories working on different instrumental platforms and using slightly different analytical protocols.

In "Shotgun" proteomics, the collected mass spectrometric data possess their absolute "universal" value, derived from atomic mass tables; but LC data, in the form of RT, are not tied to a conventional time scale, and may vary depending on the separation protocols used (gradient profile, flow rate, mobile phase composition), types of LC columns (column size, pore and particle sizes, adsorbent type, manufacturer), as well as the HPLC instrument. This can make the translation of identification databases problematic. Such a translation requires that a simple relationship between RT in different conditions be sought. In other words, there is a need for standardization of the RT obtained under particular experimental conditions, i.e. the introduction of a relative RT scale independent of the LC protocols, systems, or conditions used.

Previous efforts to implement standardization procedures for LC data mostly focused on ways to improve reproducibility of RT measurement on a particular instrumental setup using an established protocol [3,12,16,27]. The main idea behind these approaches was that fixing the LC protocol does not prevent retention time scattering between different HPLC runs for identical samples, because of column aging or variations in mobile phase preparation, etc.... Therefore, standardization becomes essential for data comparison especially for complex samples, as encountered in proteomics. Most standardization methods to date were based on the usage of an internal or external reference, or standard. One of the assumptions is that LC data scale linearly in day-to-day runs on a given instrument and for a given LC protocol (gradient profile, flow rate, etc.). Within this assumption, a standard sample can be used to obtain "relative retention time" using the following equation: $t_i = \mathrm{RT}_{i,\mathrm{exp}}/\mathrm{RT}_{st}$, where $\mathrm{RT}_{i,\mathrm{exp}}$ is the retention time of the sample compound, and RT_{st} the retention time of the internal standard. Note that this simple approach is limited to data obtained using the same LC protocol; and a change in gradient slope, for instance, may result in different relative retention times for the same compound. A more sophisticated standardization procedure using external standards was suggested by Sapirstein et al. [12] who proposed to use as standards several selected peaks from a specific protein sample which was analyzed before and after the sample of interest. The RTs of these peaks were then used as anchor points in a piecewise calibration algorithm to normalize the chromatograms of samples run in the interval between two of the standards. The proposed normalization algorithm demonstrated a fivefold improvement in the precision of chromatographic data over a period of several months of data collection. In another work, Petritis et al. [3], have proposed to use a Genetic Algorithm for normalization, which was set to optimize two variables of a linear equation, $y = ax + b$. The variable a normalized the gradient slope, and the variable b normalized the LC run start time (dead volumes, delay time, etc.). The optimization of these variables was performed for each separation and the normalization of RT into a 0–1 range and was based on 6 peptides chosen as calibration standard which were specific for the proteomes under study. Over the course of many experiments, the RT normalized using this procedure deviated from the mean by about 1% for the identified peptides. It is assumed that LC conditions in these experiments were the same or at least similar. In summary, previous efforts dealing with peptide RT standardization ranged from very simple to highly sophisticated. It is of particular significance that all these attempts were limited in scope to the effect of an unwanted change in LC separation on RT and offered time scale tied up to specific calibration standards separated under fixed LC conditions. In addition, most authors referred, at least implicitly, to a linear relationship between the measured RT [3,8,12,16,27].

When expanding the scope of standardization methods to deliberate changes in separation conditions, the first problem one is faced with is the question of the reference: could one measured retention time constitute a reliable reference for all further measurements, calibrations and alignments? When comparing two or more runs obtained under identical or similar LC conditions, the choice is not so crucial. However, when multiple datasets acquired under variable conditions are to be brought to the same scale, it becomes important to carefully choose what to align with. It is clear that simply choosing an experimental dataset as a reference is not only arbitrary, but risky, since this particular dataset can be prone to errors in RT estimations. One way to deal with this issue has been independently proposed by McIntosh and co-workers [8] and by us [13]: it consists in the conversion of experimental RT values into a scale corresponding to an intrinsic property of the peptide sequence.

McIntosh et al. suggestion was based on linking peptide LC data with their predicted hydrophobicity values. Using peptides identified with high confidence, they estimated the parameters of a linear equation relating hydrophobicities with RT for a particular experiment. The RT normalization was performed using the Sequence Specific Retention Calculator (SSRCalc) [4], an RT prediction algorithm. In the underlying model of SSRCalc, peptides relative hydrophobicities are assumed to be proportional to RT. These authors claimed the independence of normalized RT on the separation conditions (e.g. the gradient slopes) to combine data from multiple different LC configurations into a single AMT database.

In the present work, we assess the feasibility of LC data standardization using a normalized RT scale tied up with aminoacid interaction energies, using a model introduced by Gorshkov et al [14]. This model of peptide separation is based on the Liquid Chromatography at Critical Conditions applied to biomolecules (BioLCCC) [6–7,15]. It takes into account exclusion effects during peptide separation and the corresponding normalized RT scale is considered sequence specific and generally independent of the LC protocols. Due to the fact that only a few phenomenological parameters are used in the model (determined from the number of aminoacid residues and C- and N- terminal groups) it can be easily adapted for a large variety of solid and mobile phases.

The key issue of RT standardization using predicted properties of peptide sequences is the assumption of linear correlation between experimental retention times acquired under different separation conditions. In the present work, following previous evidence by Casal et al [16], we tested this assumption for a range of experimental parameters such as columns parameters, mobile phase compositions and gradient slope typically used in proteomics experiments [17,18].

Finally, we demonstrate an approach for standardization of peptide RT by conversion of measured values to a standard scale, independent of the instrument or method used. While any of the sequence-dependent RT prediction algorithms [3–4,6–7,19–21] can be used for the purpose of this work, we have selected the additive model pioneered by Meek [22] and recently refined by Krokhin et al. [4], and the BioLCCC model proposed by Gorshkov et al [6]. Both models performed equally well.

2. Experimental

Cytochrome c digest and 6 protein digest were purchased from Dionex/LCPacking (Dionex, Amsterdam, Netherlands) and used as recommended. After a careful analysis of MS/MS data, we found that the molecular structures of two peptides differed from the sequences specified in the Dionex data sheet: IFVQKCAQCHTVEK should be designated correctly as CAQCHTVERL + heme, and KGEREDLIAYLK as GEREDLIAYLKK. The *Cytochrome c* peptides used as retention time calibrants are recapitulated in Table 3. The 6 protein digest standard includes *Cytochrome c*, lysozyme,

alcohol-dehydrogenase, bovine serum albumin, apo-transferrin, and beta-galactosidase. The samples were injected at concentrations of 200–500 fM/μl.

Peptides mixtures were separated on an "Ultimate 3000" nano-HPLC system (Dionex, Amsterdam, Netherlands) coupled to a hybrid linear ion trap - Fourier transform ion cyclotron resonance mass spectrometer LTQ-FT (ThermoFisher Scientific, Bremen, Germany). The mobile phases for gradient HPLC experiments were (A) ACN/water/formic acid (2:98:0.1, v/v) and (B) ACN/water/formic acid (80:20:0.08, v/v). All solvents were of HPLC purity and were purchased from Merck, Darmstadt, Germany (acetonitrile) or VWR International, London, England (formic acid).

The columns used throughout this work are listed in Table 1. The column equilibration time in nano-HPLC experiments was set to 25–35 min, shorter re-equilibration times resulting in lower reproducibility for the retention of hydrophilic peptides. With such a prolonged equilibration, the precision of retention for hydrophilic peptides in all experiments was within a few seconds.

Identification of peptides from the 6 protein digest was performed against the SwissProt database using Mascot software (Matrix Science Ltd.). Only identifications pointing to proteins known to be in the mixture and having Mascot scores above the identity threshold ($p < 0.05$) were considered for evaluation of the calibration procedure.

The in-house developed software package "Theoretical Chromatograph: BioLCCC/MS-MS" [23] was used to calculate retention times for the sequences found and scored by Mascot. The detailed background and basic equations behind the BioLCCC model are given in earlier publications [6–7,15]. The "Theoretical Chromatograph: BioLCCC/MS-MS" allows prediction of retention times based on peptide primary structures and the RP-HPLC conditions used. This software is available online at http://biolccc.mhost.ru/. The publicly available 3.0 version of Sequence Specific Retention Calculator (SSRCalc) developed at Manitoba Centre for Proteomics (http://hs2.proteome.ca/SSRCalc/SSRCalc.html) was used for comparison. In the SSRCalc algorithm [4,24], the peptide retention times were calculated using a linear equation RT = $a + b \times$ Hydrophobicity, where a and b are constants related to gradient delay time and slope, respectively.

3. Results and discussion

3.1. Correlation of RP-HPLC data for peptides separated under different experimental conditions

3.1.1. RP-HPLC data linearity concept

The main assumption in RT standardization is that LC data are linearly correlated within a wide range of experimental parameters (i.e. different separation conditions, columns, mobile and/or solid phases). Needless to say, this linear correlation is established for data generated on the same LC system under the same conditions in different experiments. The LC data linearity concept (which is familiar in the case of separations of low molecular weight compounds) can be considered a direct consequence of the mechanism behind biomacromolecules separation in a gradient RP-HPLC [6,7].

In case linearity is preserved, the measured RTs can be transferred to a conventional time scale where a given peptide will have a definite RT. The relationship between the RT_X and RT'_X measured for peptide X under different conditions can be expressed as:

$$RT_X = \alpha RT'_X + \beta \quad (1)$$

in which the coefficients α and β are defined by the experimental parameters. A similar concept been described by Petritis, et al. [3] who used a genetic algorithm to optimize a linear equation of RT normalization to generate accurate mass and time tags (AMT) peptide databases. We extended their assumption beyond the scope of the same LC system, separation conditions, and column parameters; and investigated the changes in peptides RT for a variety of LC conditions commonly employed in proteomics studies.

In the present work we initially performed a series of systematic experiments using commercially available mixture of standard protein digests to test the linearity assumption. The influence of column properties (C_{18} phase, length, and internal diameter) and other LC parameters such as gradient slope, flow rate has been evaluated. The linearity was tested by plotting RT measured in one experiment *versus* RT for the same mixture measured in other experiments. The resulted plots were fitted by a linear equation and the Pearson coefficient was used as a measure of data correlation. We arbitrarily considered data with $R^2 > 0.95$ as highly correlated and $R^2 < 0.8$ as showing poor correlation; and deliberately limited the study to linear gradients of water/acetonitrile with formic acid as ion pairing reagent at room temperature using C_{18} as a stationary phase. This range of parameters covers a very broad cross-section of proteomics applications.

3.1.2. Effect of column parameters

Reversed-phase columns used in standard proteomics experiments can differ in terms of supplier (commercial/home-made), column dimension (i.d., length), type of absorbent (monolith/particles), structure of stationary phase (particle size, pore size). We selected an assortment of ten columns with characteristics encompassing all these variables (see Table 1). For all selected columns, a linear 30 min gradient of 0–50%B was used (Fig. 1). In addition, some columns were tested with longer gradient, specifically, 0–35%B in 120 min, in order to test the linearity for different gradient durations (vide infra) (Fig. 2).

Results are presented in the form of a multi-correlation matrix to visualize RT covariance across all columns as shown in Fig. 1A. This matrix arrangement allows the display of all conditions without the bias due to selecting a particular column as a reference. The matrix of Pearson correlation coefficients is depicted in a cell plot Fig. 1B. Pearson coefficients were better than 0.982 for all column pairs; in other words, RT on a particular column were predicted at least at 96% by RT on the other column. High values of R^2 confirm

Table 1
Specifications of columns used in this study.

	Name	Supplier	ID μm	Column length cm	Particle size μm	Pore size A	Phase
I.	*PepMap*	LC Packings	75	15	3	100	C18
II.	*PepMap*	LC Packings	75	25	3	100	C18
III.	*AtlantisC18*	Waters	75	15	3.5	110	C18
IV.	*PepMap*	LC Packings	75	15	5	100	C18
V.	*PLRP-S*	LC Packings	75	15	5	300	C18
VI.	*Chromolith*	Merck	100	15	–	110	C18 (Monolith)
VII.	*Gemini*	Phenomenex	100	15	3	100	C18
VIII.	*Gemini*	Phenomenex	75	15	3	100	C18
IX.	*Proteo*	Phenomenex	75	15	4	100	C12
X.	*Jupiter*	Phenomenex	75	15	3	300	C18

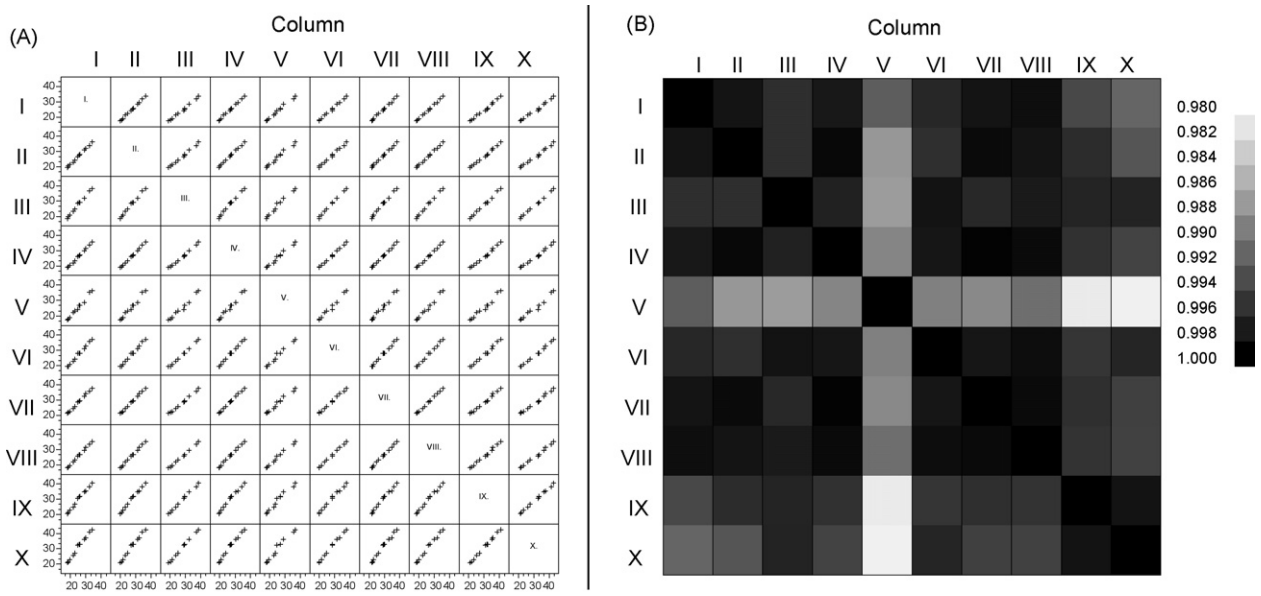

Fig. 1. (A) Scatterplot matrix illustrating the linearity between RT for 12 peptides of *Cytochrome c* obtained on various columns. (B) Cell plot representing the Pearson correlations between LC data represented in the scatterplot matrix. The worse Pearson correlation with the other columns was found for column V. However, these coefficients were still better than 0.98. These results were obtained for the 10 different columns described in Table 1 and the following LC gradient profile: 0–50%B in 30 min, 300 nl/min. Sample: 500 fmol *Cytochrome c* digest.

that a linear regression can be applied to translate RT across columns provided that the retention mechanism is conserved. This is in agreement with the linear solvent strength model proposed earlier by Snyder, et al. [25]. It is worth pointing out that slight selectivity changes during the separation do not exaggeratedly affect the overall correlation because they are limited to closely eluting peaks when working within the range of parameters that are typical for proteomics applications. For instance, in comparison of particulate columns with micro porous monolithic silica column, peptides TGPNLHGLFGR and MIFAGIK exhibited a switch in retention (i.e. change in selectivity), while the Pearson coefficient still held above 0.998. We also found that the linearity is robust to a change in the type of stationary phases, where selectivity is expected to change ($C_{18} \times C_{12}$, $R \approx 0.992$). Hence, our data convincingly confirm that for columns typically used in proteomics experiments, the linear correlation hypothesis is valid.

3.1.3. *Effect of mobile phase*

For the evaluation of the effect of mobile phase on experimental RT, we focused mainly on the mobile phase gradient slope and the flow rate. In doing so, we restricted the choice of mobile phase composition to water/acetonitrile with formic acid (pH~3) as ion pairing reagent, which covers a broad range of proteomics applications. Like in experiments described above, we recorded RT of *Cytochrome c* peptides obtained under variable conditions and plotted the linear correlation between the data obtained (see Fig. 2A).

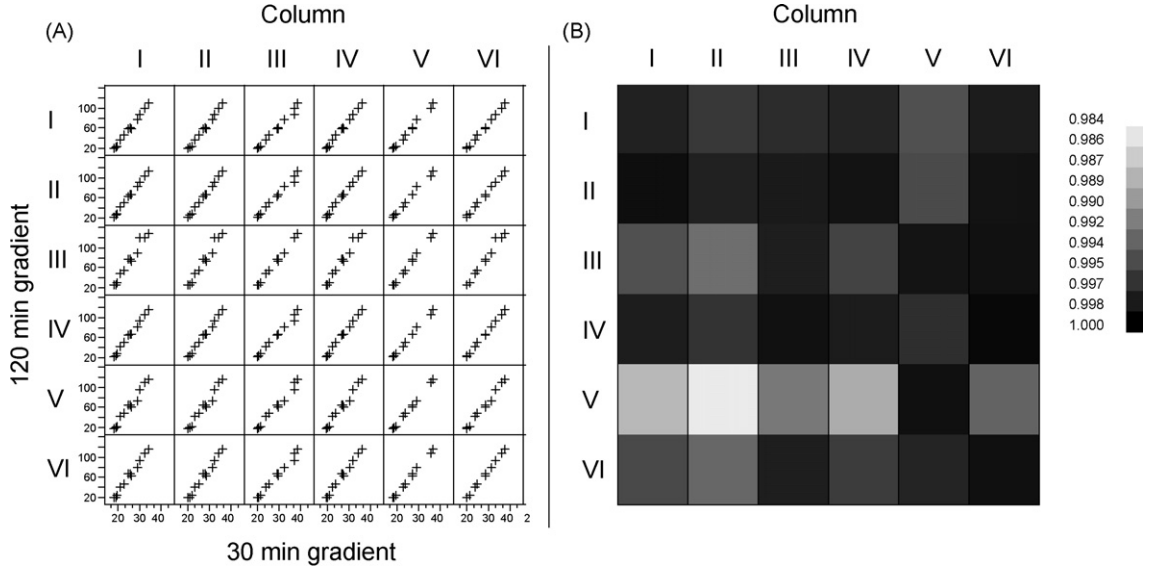

Fig. 2. (A) Scatterplot matrix representing the linearity between RT for 12 peptides of *Cytochrome c* obtained on 6 selected columns using gradients of 30 min and 120 min, corresponding to gradient slopes of 1.7%B/min and 0.3%B/min, respectively. (B) Cell plot representing the Pearson correlations between LC data represented in the scatterplot matrix. RT obtained on both gradients correlated with each other, even across columns, with Pearson coefficients greater than, 0.984. Sample: 500 fmol *Cytochrome c* digest.

Table 2
R^2 values in the correlation between experimental data measured for various gradient slopes and flow rates. Reference: PepMap (#1 in Table 1), under conditions of 30 min linear gradient (1.7%B/min), with 12 h of column equilibration. LC conditions are indicated in Fig. 2. Sample: 500 fmol *Cytochrome c* digest.

Gradient time [min]	Gradient slope [%B/min]	R^2	Flow-rate [nl/min]	R^2
15	3.3	0.995	300	0.993
30	1.7	0.995	500	0.992
60	0.8	0.995	700	0.989
120	0.3	0.995	900	0.988
360	0.1	0.987	3000	0.978

Fig. 2 shows the Pearson correlations between data obtained on 6 selected columns using two different gradients in the form of a multi-correlation matrix. The gradients in these tests were 30 min and 120 min, corresponding to gradient slopes of 1.7%B/min and 0.3%B/min, respectively. Excellent correlations were obtained with R^2 values ranging from 0.994 to 0.998 for all columns under both gradient slopes (Fig. 2B).

In subsequent experiment, we selected the PepMap column #I (Table 1) to systematically test a broader range of gradient slopes. For this column we obtained correlations between data in a range of $R^2 \approx 0.9880$–0.9997 when gradient slope varied from 0.1 to 3.3%B/min as shown in Table 2. Even for the longest run used (duration of 6 hrs corresponding to 0.1%B/min gradient slope); the correlation was still as high as $R \approx 0.987$. The impact of changing the gradient slope, together with the effect of flow rate, is further illustrated in the results presented in Fig. 3. Because a linear correlation between RT obtained on any column and the standard PepMap column was observed previously (see above), we conclude that the present results are relevant to other columns as well. As can be seen in Fig. 3, a change in the gradient slope led to a change in the linear data correlation slope, whereas a change in the flow rate changed the intercept of the linear fit. In all cases, a high correlation coefficient was consistently obtained.

In conclusion, it appeared from these experiments that, within a quite broad range of column properties and gradient parameters, the assumption of linear correlation of RT is valid and a simple linear regression is sufficient to translate RT between conditions.

3.2. Toward a standard retention time scale

The above mentioned results demonstrate that there is a linear correlation of RT data across a range of separation conditions, column types, and LC protocols typical for proteomic studies. The high linear correlation with R^2 in a range between 0.98 and 0.99 provides the basis to introduce a standardized RT scale invariant across LC platforms, protocols, and conditions within the framework of the previously stated parameters.

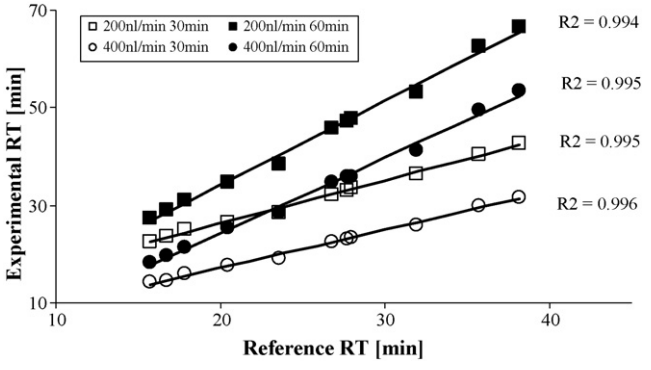

Fig. 3. Effects of the gradient slope and flow rate on the linear fit between corresponding RT data. The gradients were 30 min and 60 min, corresponding to gradient slopes of 1.7%B/min and 0.8%B/min, respectively. Sample: 500 fmol *Cytochrome c* digest.

The second issue to address is the choice of a suitable reference. While in practice, any data obtained under specified LC parameters and setup could be used, we would like to point out that the quality of the normalization is linked to the choice of the reference: this means that, any error made in the determination of the reference RT will propagate to all the data standardized using this reference. Moreover, the reference dataset could never be exactly reproduced even on the same system using the same separation parameters because of inherent measurement errors.

To circumvent the above mentioned issues, one attractive option is to tie up the measured RT to an intrinsic property of the peptides being measured [8,13]. One possible approach for the RT normalization, shown schematically in Fig. 4, would be the following: once the linear correlation between experimental LC data is established, one can choose a simple peptide mixture (standard) and predict their RT under specified LC conditions as "standard LC protocol" using RT prediction software. These predicted RT for the standard will be the reference data, $RT^{pred}_{ref,i}$. It is worth mentioning that the prediction software will generate the exact same RT for the reference peptides anywhere and anytime. These $RT^{pred}_{ref,i}$ can be further normalized by dividing all reference RT by e.g. the time for the Nth-peptide from the mixture with the largest predicted RT, $RT^{pred}_{ref,N}$:

$$RT^{norm}_{ref,i} = \frac{RT^{pred}_{ref,i}}{RT^{pred}_{ref,N}} \quad (2)$$

For normalization of any dataset, one can add the standard peptides into the sample and analyze the mixture under any chromatographic conditions within the linearity range. Alternatively, the standard can be analyzed before (and/or after) the sample under identical conditions. The latter is sometimes preferable as the sample mixture may be complex and not all standard peptides can be separated from peptides in the sample. In the subsequent step, the experimental values obtained for the standards ($RT^{exp}_{ref,i}$) are plotted *versus* the normalized reference ($RT^{norm}_{ref,i}$) and the resulting curve is fitted using a linear equation:

$$RT^{norm}_{ref,i} = aRT^{exp}_{ref,i} + b \quad (3)$$

The a and b coefficients obtained from the fit are then used to convert the experimental RT of the sample peptides into RT in a common time scale independent of the particular LC conditions and/or instrument. As for the standard mixture, we suggest, following van Midwoud et al. [26], a *Cytochrome c* digest, which is commercially available and quite inexpensive. The peptides in this digest cover a wide range of retention times. The only requirement here is that RT databases have to be generated using LC conditions within the linearity range. For example, using different ion-paring agents to generate databases for RT of peptides having different end-groups may not be acceptable.

In the present study we utilized a RT prediction program based on the BioLCCC model. Fig. 5 shows an example of correlation between the BioLCCC-predicted RT and experimental times for selected peptides from a *Cytochrome c* digest normalized to $RT^{BioLCCC}_{CYC12}$, the 12th *Cytochrome C* peptide GITWGEETLMEYLENPK. The correlation coefficient R^2 of 0.991 in this prediction is comparable with the correlation between experimental data obtained under different LC protocols and conditions in this work. Typical separation conditions were selected as "standard LC protocol": column 75 μm I.D. × 15 cm; particle size of 3 μm; pore size of 100 A; mobile phase (A) water/ACN (98:2), 0.1% Formic acid, (B) water/ACN

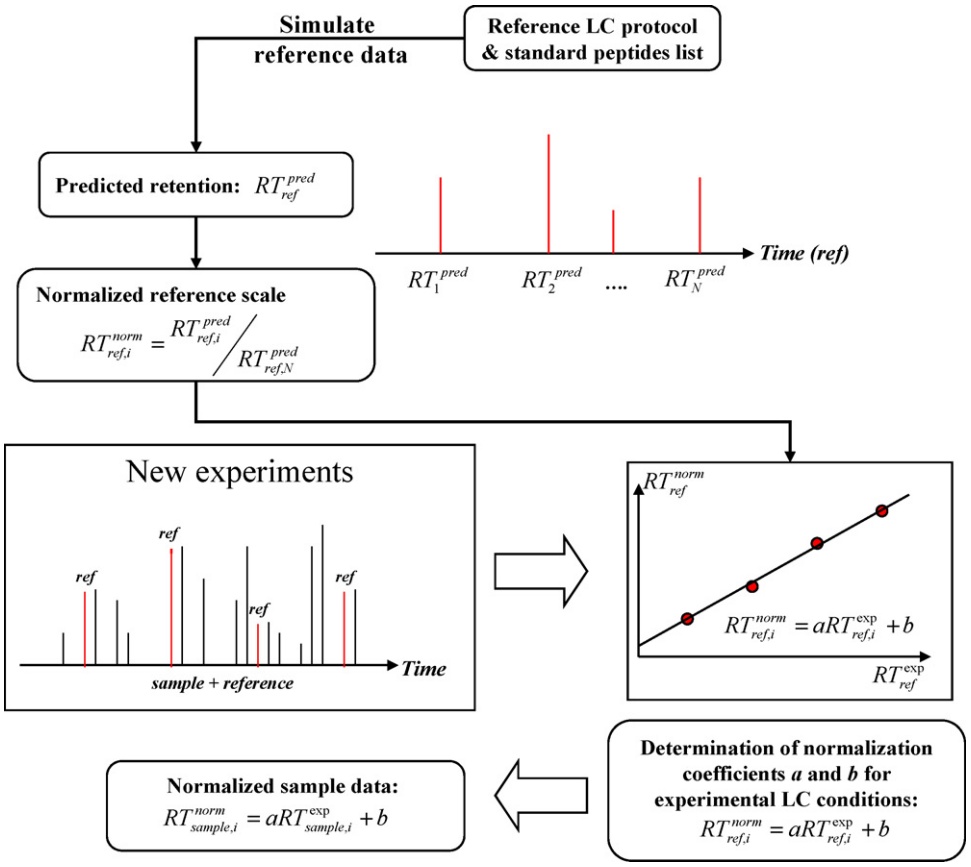

Fig. 4. Scheme of the procedure for normalization of LC data into the standard time scale based on the predicted RT. Initially, the reference RT scale is generated using an appropriate sequence-dependent RT prediction model. This time scale is based on the predicted RT of a standard mixture contained N known sequences separated under reference LC conditions. The time scale is then normalized to the RT of the Nth standard peptide for generality. The coefficients for translation between experimental RT of the sample under study and this normalized time scale are determined using the linear fit between the experimental RT of the standard peptides separated under LC conditions used for the sample analysis and the corresponding normalized RTs of the standard in the normalized RT scale. Using these coefficients all experimental RTs for the sample under study can then be converted into the normalized scale.

(20:80), 0.08% Formic acid; gradient 0–50%B in 60 min; sample concentration of 1 pmol/μl; and injection volume of 1 μl. Peptides 3, 5, 9, and 12 from *Cytochrome c* digest standard were used as a reference peptide mixture to determine the calibration coefficients *a* and *b*, according to the following equation:

$$\frac{\mathrm{RT}_{ref,i}^{BioLCCC}}{\mathrm{RT}_{CYC12}^{BioLCCC}} = a\mathrm{RT}_{ref,i}^{\exp} + b \quad (4)$$

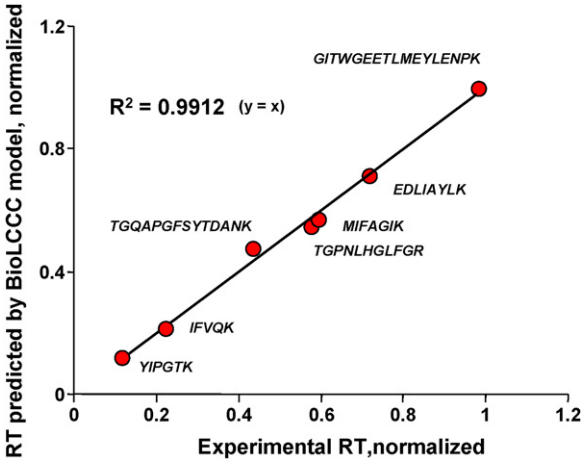

Fig. 5. Correlation between normalized retention times for peptides from *Cytochrome c* digest predicted by BioLCCC model and obtained experimentally. The LC protocol used in the model and to obtained experimental data was: Dionex's PepMap C18 column (75 μm ID, 150 mm long, 3 μm particles and 100 A pores), 4–50%B linear gradient over a 60 min period. This LC protocol was selected in this work as a reference LC protocol to generate normalized retention time scale.

Finally, by using *a* and *b* coefficients, experimental RT for the rest of the standard peptides were normalized according to Eq. (3). Note that in case some peptide from experimental sample elutes at a later time than the 12th *Cytochrome C* peptide, its normalized RT could be more than 1. However, this does not affect the generality of the method. Table 3 shows the results of normalization of experimental data for *Cytochrome c* digest obtained under different experimental conditions (specifically, different gradient profiles). We found that for these data the standard deviation between normalized RT does not exceed 1.6%.

The suggested normalization procedure can also be realized using other RT prediction tools. One of the widely used RT calculator is the publicly available 3.0 version of Sequence Specific Retention Calculator (SSRCalc) (http://hs2.proteome.ca/SSRCalc/SSRCalc.html). In Table 4, we compared the results of normalization for two prototype LC databases generated for a 6 protein standard digest using both BioLCCC model and SSRCalc algorithm. We can see from this Table that both RT prediction programs exhibited very close normalization results. In the time scales generated using RT prediction models, the standard deviation of the normalized RT for the same peptides separated under different HPLC conditions was in the range of 0.9–1.2%.

Table 3
Results of a self-consistency test using the normalized RT scale generated using *Cytochrome c* digest as a standard mixture and the reference LC protocol described in the text. The standard scale was generated using the sequence-dependent RT prediction algorithm based on the BioLCCC model. Experimental data for *Cytochrome c* digest were obtained using different LC gradient slopes. The average accuracy of normalized data was ~1.6%. The four shaded peptides were used as internal calibrants, and the others to evaluate the calibration error.

		RT ^{exp}, min		$RT_{norm}^{BioLCCC}$	RT ^{norm}	
	SEQUENCE	1.7%B/min	0.3%B/min		1.7%B/min	0.3%B/min
1	*KYIPGTK*	17.77	19.39		0.172	0.199
2	*YIPGTK*	18.6	21.47		0.215	0.217
3	*IFVQK*	19.35	24.36	0.247	0.254	0.243
4	*KTGQAPGFSYTDANK*	20.88	36.65		0.333	0.352
5	*TGQAPGFSYTDANK*	22.63	46.25	0.474	0.423	0.438
6	*KGEREDLIAYLK*	24.8	58.75		0.535	0.549
7	*TGPNLHGLFGR*	25.5	61.43		0.571	0.573
8	*MIFAGIK*	25.88	58.91		0.591	0.550
9	*EDLIAYLK*	28.61	77.52	0.718	0.732	0.716
10	*IFVQKCAQCHTVEK+heme*	29.62	86.6		0.784	0.797
11	*GITWGEETLMEYLENPKK*	31.84	99.32		0.898	0.910
12	*GITWGEETLMEYLENPK*	33.91	110.07	1	1.005	1.006

Table 4
An example of LC data from a 6 protein mixture AMT database prototype with experimental RT converted into normalized RT scale using the procedure described in this work and depicted in Fig. 4. Two different sequence-dependent RT prediction algorithms were used to generate the normalized scale, BioLCCC and SSRCalc.

SEQUENCE	PROTEIN	Mascot score	RTexp min	RT norm BioLCCC	RT norm SSRCalc
K.IGDYAGIK.W	ADH1_Yeast	64	23.95	0,526	0,526
K.EKDIVGAVLK.A	ADH1_Yeast	42	27.75	0.590	0.590
K.VVGLSTLPEIYEK.M	ADH1_Yeast	81	40.56	0.807	0.805
K.TCVADESHAGCEK.S + 2 Carboxymethyl (C)	ALBU_BOVIN	72	15.22	0.378	0.379
R.LCVLHEK.T + Carboxymethyl (C)	ALBU_BOVIN	42	20.92	0.475	0.475
K.CCTESLVNR.R + 2 Carboxymethyl (C)	ALBU_BOVIN	63	24.94	0.543	0.542
K.YLYEIAR.R	ALBU_BOVIN	45	30.47	0.636	0.635
K.DDPHACYSTVFDK.L + Carboxymethyl (C)	ALBU_BOVIN	43	31.36	0.651	0.650
K.LGEYGFQNALIVR.Y	ALBU_BOVIN	94	41.36	0.820	0.818
R.FNDDFSR.A	BGAL_ECOLI	48	24.56	0.536	0.536
R.VDEDQPFPAVPK.W	BGAL_ECOLI	80	31.80	0.659	0.658
R.IGLNCQLAQVAER.V + Carboxymethyl (C)	BGAL_ECOLI	90	37.36	0.753	0.751
R.KTGQAPGFSYTDANK.N	CYC_BOVIN	75	21.99	0.493	0.493
K.TGPNLHGLFGR.K	CYC_BOVIN	61	33.16	0.682	0.680
R.EDLIAYLK.K	CYC_BOVIN	44	40.06	0.798	0.796
R.HGLDNYR.G	Lysc_chick	40	14.89	0.373	0.373
R.NTDGSTDYGILQINSR.W	Lysc_chick	110	36.05	0.730	0.729
K.LCQLCAGK.G + 2 Carboxymethyl (C)	TRFE_BOVIN	56	23.50	0.518	0.518
K.ELPDPQESIQR.A	TRFE_BOVIN	53	28.00	0.594	0.594
K.DKPDNFQLFQSPHGK.D	TRFE_BOVIN	56	29.51	0.620	0.619
K.HSTVFDNLPNPEDR.K	TRFE_BOVIN	80	32.67	0.673	0.672
K.TYDSYLGDDYVR.A	TRFE_BOVIN	77	34.33	0.701	0.700
K.CACSNHEPYFGYSGAFK.C + 2 Carboxymethyl (C)	TRFE_BOVIN	66	36.56	0.739	0.738
K.SVTDCTSNFCLFQSNSK.D + 2 Carboxymethyl (C)	TRFE_BOVIN	109	40.75	0.810	0.808
K.CGLVPVLAENYK.T + Carboxymethyl (C)	TRFE_BOVIN	83	40.83	0.811	0.809

4. Conclusions

Our results demonstrate the feasibility to calibrate any HPLC system working within a range of experimental parameters in such a way that RT data can be standardized to a scale independent of the separation conditions and/or instruments. The standardization procedure is based on the assumption that in a broad range of experimental conditions there is a linear correlation between experimental LC data. This assumption is a direct consequence of the mechanism behind the biomacromolecules separation in gradient RP-HPLC and has been verified here for a variety of C18 and similar columns and LC protocols, common in proteomics experiments. Using sequence-dependent RT prediction tools, a standard RT scale can be generated, which is invariant to the experimental conditions, separation protocols, or instrument platforms. We propose to tie up this RT scale to a standard tryptic peptide mixture of *Cytochrome c* digest. The standard RT scale can be normalized for convenience by assigning the 1.0 value to the

predicted retention time for the 12th peptide from this digest. After the RT scale is generated, the experimental data for complex peptide mixtures can be transformed into this time scale using a linear function. The coefficients of this function can be obtained from the calibration of the instrument used and particular experimental conditions by RP-HPLC run for the standard peptide mixture.

We have found that using the suggested procedure the standardized RT for experimental LC data can be obtained with a relative accuracy of less than 1.2%. This standard RT scale can be useful in AMT tag database generation requiring extensive and prolonged collaborative efforts. Moreover, it could allow users to tap data repositories for peptide identifications and use them to construct AMT tag databases.

Acknowledgements

This work was supported in part by the Russian Foundation for Basic Research (RFBR), the U.S. Civilian Research and Development Foundation (CRDF) (grants RFBR 08-04-01339 and RFBR 08-04-91121-CRDF, respectively), an International Association INTAS grant (Genomics 05-10000004-7759), and the Russian Academy of Sciences (OHNM 4.2). An EU International Reintegration Grant to CDM is gratefully acknowledged (Marie Curie Actions contract MIRG-CT-2006-030810).

References

[1] M. Palmblad, M. Ramstrom, K.E. Markides, P. Hakansson, J. Bergquist, Anal. Chem. 74 (2002) 5826.
[2] E.F. Strittmatter, P.L. Ferguson, K. Tang, R.D. Smith, J. Am. Soc. Mass Spectrom. 14 (9) (2003) 980.
[3] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Yu. Shen, R. Zhao, R.D. Smith, Anal. Chem. 75 (2003) 1039.
[4] O.V. Krokhin, R.V. Craig, V. Spicer, W. Ens, K.G. Standing, R.C. Beavis, J.A. Wilkins, Mol. Cell. Prot. 3 (9) (2004) 908.
[5] Y. Shi, R. Xiang, C. Horvath, J.A. Wilkins, J. Chrom. A 1053 (2004) 27.
[6] A.V. Gorshkov, I.A. Tarasova, V.V. Evreinov, M.M. Savitski, M.L. Nielsen, R.A. Zubarev, M.V. Gorshkov, Anal. Chem. 78 (2006) 7770.
[7] A.V. Gorshkov, V.V. Evreinov, I.A. Tarasova, M.V. Gorshkov, Polym. Sci. B 49 (3–4) (2007) 93.
[8] D. May, M. Fitzgibbon, Y. Liu, T. Holzman, J. Eng, C.J. Kemp, J. Whiteaker, A. Paulovich, M. McIntosh, J. Proteome Res. 6 (7) (2007) 2685.
[9] K.C. Leptos, D.A. Sarracino, J.D. Jaffe, B. Krastins, G.M. Church, Proteomics 6 (2006) 1770.
[10] J.S.D. Zimmer, M.E. Monroe, W.-J. Qian, R.D. Smith, Mass Spectrom. Rev. 25 (2006) 450.
[11] L. Paša-Tolić, M.S. Lipton, C.D. Masselon, G.A. Anderson, Y. Shen, N. Tolić, R.D. Smith, J. Mass Spectrom. 37 (2002) 1185.
[12] H.D. Sapirstein, M.G. Scanlon, W. Bushuk, J. Chrom. A 469 (1989) 127.
[13] I.A. Tarasova, V. Guryca, M.L. Pridatchenko, S. Kieffer-Jaquinod, C.D. Masselon, J. Garin, A.V. Gorshkov, V.V. Evreinov, M.V. Gorshkov, Proceedings of the 55th ASMS Conference on Mass Spectrometry and Allied Topics, Indianapolis, IN, 2007.
[14] A.V. Gorshkov, V.V. Evreinov, M.V. Gorshkov, Proceedings of the 52nd Conference of American Society for Mass Spectrometry and Allied Topics, Nashville, Tennessee, 2004, MPX447.
[15] I.A. Tarasova, A.V. Gorshkov, V.V. Evreinov, C. Adams, R.A. Zubarev, M.V. Gorshkov, Polym. Sci. A 50 (3) (2008) 309.
[16] V. Casal, P.J. Martin-Alvarez, T. Herraiz, Anal. Chim. Acta 326 (1–3) (1996) 77.
[17] S. Kieffer-Jaquinod, A.V. Gorshkov, M.V. Gorshkov, M. Court, S. Brugiere, C. Bruley, M. Ferro, J. Garin, C.D. Masselon, Proceedings of the 23rd LC/MS Montreux Symposium, Montreux, Switzerland, 2006.
[18] V. Spicer, A. Yamchuk, J. Cortens, S. Sousa, W. Ens, K.G. Standing, J.A. Wilkins, O.V. Krokhin, Anal. Chem. 79 (2007) 8762.
[19] T. Baczek, P. Wiczling, M. Marszall, Y.V. Heyden, R. Kaliszan, J. Proteom. Res. 4 (2005) 555.
[20] A.A. Klammer, X. Yi, M.J. MacCoss, W.S. Noble, Anal. Chem. 79 (2007) 6111.
[21] N. Pfeifer, A. Leinenbach, C.G. Huber, O. Kohlbacher, BMC Bioinformatics 23 (13) (2007) 1273.
[22] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (3) (1980) 1632.
[23] I.A. Tarasova, A.V. Gorshkov, V.V. Evreinov, A.A. Goloborodko, S.S. Shitov, M.L. Nielsen, R.A. Zubarev, M.V. Gorshkov, Proceedings of the 17th International Conference on Mass Spectrometry, Prague, Czech Republic, 2006, TuP-252.
[24] O.V. Krokhin, Anal. Chem. 78 (22) (2006) 7785.
[25] L.R. Snyder, J.W. Dolan, Adv. Chromatogr. 38 (1998) 115.
[26] P.M. van Midwoud, L. Rieux, R. Bischoff, E. Verpoorte, H.A.G. Niederlander, J. Proteom. Res. 6 (2007) 781.
[27] N. Jaitly, M.E. Monroe, V.A. Petyuk, T.R.W. Clauss, J.N. Adkins, R.D. Smith, Anal. Chem. 78 (22) (2006) 7397.