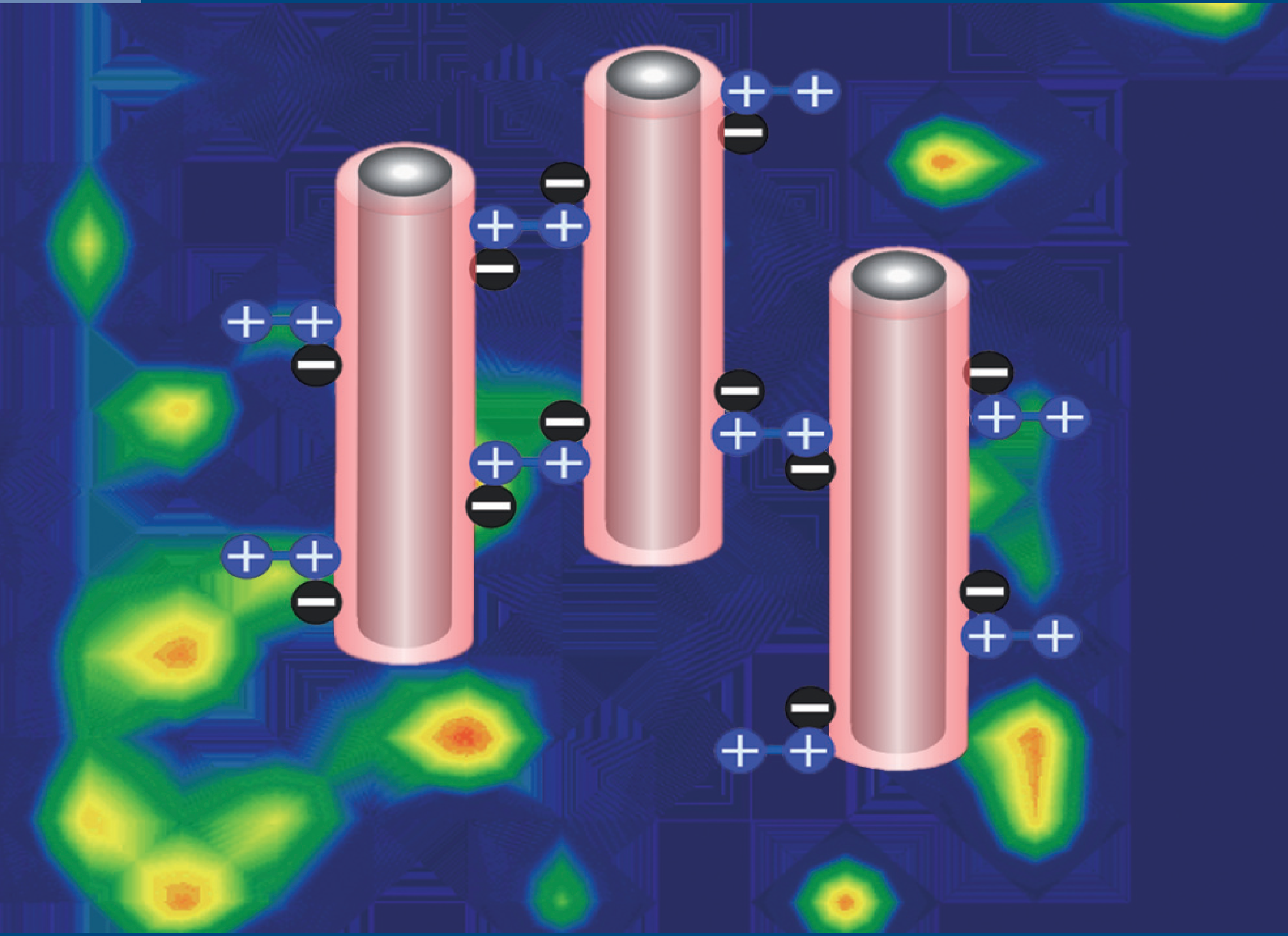
ISSN 1615-9306 · JSSCCJ 41 (18) 3503–3654 (2018) · Vol. 41 · No. 18 · September 2018 · D 10609

JSS

JOURNAL OF SEPARATION SCIENCE

18|18



Methods
Chromatography · Electroseparation

Applications
Biomedicine · Foods · Environment

www.jss-journal.com

WILEY-VCH

Received: 5 March 2018 | Revised: 17 July 2018 | Accepted: 18 July 2018

DOI: 10.1002/jssc.201800488

RESEARCH ARTICLE

JOURNAL OF SEPARATION SCIENCE

Visualization and application of amino acid retention coefficients obtained from modeling of peptide retention

Yassene Mohammed[1,2] | Magnus Palmblad[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands

[2] University of Victoria-Genome British Columbia Proteomics Centre, University of Victoria, Victoria, Canada

Correspondence
Dr. Yassene Mohammed, Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands.
Email: y.mohammed@lumc.nl

We introduce a method for data inspection in liquid separations of peptides using amino acid retention coefficients and their relative change across experiments. Our method allows for the direct comparison between actual experimental conditions, regardless of sample content and without the use of internal standards. The modeling uses linear regression of peptide retention time as a function of amino acid composition. We demonstrate the pH dependency of the model in a control experiment where the pH of the mobile phase was changed in controlled way. We introduce a score to identify the false discovery rate on peptide spectrum match level that corresponds to the set of most robust models, i.e. to maximize the shared agreement between experiments. We demonstrate the method utility in reversed-phase liquid chromatography using 24 datasets with minimal peptide overlap. We apply our method on datasets obtained from a public repository representing various separation designs, including one-dimensional reversed-phase liquid chromatography followed by tandem mass spectrometry, and two-dimensional online strong cation exchange coupled to reversed-phase liquid chromatography followed by tandem mass spectrometry, and highlight new insights. Our method provides a simple yet powerful way to inspect data quality, in particular for multidimensional separations, improving comparability of data at no additional experimental cost.

KEYWORDS

mass spectrometry, proteomics, retention time modeling, one-dimensional separation, two-dimensional separation

1 | INTRODUCTION

In a typical MS-based proteomics analysis, analytical separations like reversed-phase chromatography are used to reduce the complexity of the sample injected into the mass spectrometer [1,2]. Reversed-phase liquid chromatography (RPLC) is used to separate peptides based on their hydrophobicity before ESI and injection into the mass spectrometer [1,3,4]. This peptide separation accomplishes several tasks simultaneously, but most significantly it reduces the complexity of the mixture presented to the mass spectrometer at any given time. This makes it easier for the mass spectrometer to detect, select, and fragment more peptides for either identification and/or quantification depending on the experiment.

The time a peptide is introduced to the mass spectrometer contains information about the peptides. However retention time of peptides in such experiments is still underused. Using retention time (RT) modeling, different candidates to the identity of an unknown peptide can be discriminated based on their arrival times at the mass spectrometer. Although less discriminating than high-quality MS/MS, the information is already available, without additional

Article Related Abbreviations: ACN, acetonitrile; FDR, false discovery rate; PSM, peptide spectrum match; RPLC, reversed-phase liquid chromatography; RT, retention time; SCX, strong cation exchange standards, analyses or hardware, and, unlike tandem mass spectra, is practically of uniform quality for all peptides [5]. In addition to using RT in peptide identification, it is also used in combination with synthetic peptide standards (of known chromatographic behavior) to control the quality of the separation or enhancing and mapping predicted retention time to different LC conditions, which is proven very helpful in targeted proteomics approaches [6–9].

We here introduce a novel method for using RT to obtain information about the conditions and quality of the experiment and peptide separation. We show how the simple linear regression modeling [10–14] can be used to visualize the relationship between mobile phase hydrophobicity and amino acid composition of peptides in an easily interpreted manner. We consider the amino acid coefficients obtained from modeling of RT and compare these coefficients between runs. Within a fixed experimental setting, these coefficients are stable and summarize the effect the incorporation of an amino acid in a peptide regarding that peptide RT, independent of the sample used. Unintended modifications, like a sudden or uncontrolled change of the mobile phase pH, will result in an immediate change in the coefficient values and their orders. To demonstrate our method, we generated and used two RPLC–MS/MS datasets. We also show the utility of our method with datasets obtained from PRIDE (a public repository for proteomics experimental data) [15].

2 | MATERIALS AND METHODS

To test our method we used datasets from two experiments, as well as several available from PRIDE with no previously known issues of data quality.

For our experiments we used whole-cell protein extract from *Escherichia coli*. In the first RPLC–MS/MS dataset we varied the mobile phase pH. In the second experiment we used SDS-PAGE to fraction the proteome and analyzed each fraction with RPLC–MS/MS following in-gel protein digestion.

In an attempt to challenge our method we used datasets acquired with 1D RPLC–MS/MS and online 2D strong cation exchange (SCX) coupled to RPLC–MS/MS obtained from PRIDE. In the following sections we describe the sample preparation and acquisition, the used data processing pipeline for the identification, RT modeling, determining the amino acid coefficients and their ranks.

2.1 | SDS-PAGE followed by RPLC–MS/MS experiments

2.1.1 | Preparation of the *E. coli* samples

E. coli cells were grown on LB medium (Life Technology^{TM}) washed with 1 × 0.3 M Sucrose, Hepes pH 7.0 and centrifuged into a pellet. Protein extraction was performed using 50 µl of 1% SDS (containing protease inhibitor and 1 µL benzonase of 25 U/µL), placed at 4°C for 30 min. Afterwards the samples were centrifuged at 16 000 × *g* at 4°C for 15 min and subsequently the supernatant was taken. The protein concentration was measured by a bicinchoninic acid protein assay kit (Thermo Fischer Scientific).

In the subsequent steps, the SDS used to lyse the cells was diluted in the SDS-PAGE running buffer (with 0.1% SDS) and removed during the washing of the (combined) gel slices as described in [16,17]. The peptides were further cleaned up on a trap column, where the wash was directed to waste, so very SDS or buffer from the cell lysis should reach the analytical column in this setup.

2.1.2 | In-solution digestion

Fifty microgram of the protein was used for a standardized tryptic digestion without pre-fractionation. The proteins were first reduced using 2 µL 60 mM dithiothreitol for 40 min at 60°C and alkylated by 4 µL 100 mM iodoacetamide for 1 h in the dark at room temperature. Afterwards proteins were digested overnight at 37°C using trypsin (sequencing grade, Promega, Madison, WI, USA). The digestion was quenched by addition of 2 µl 10% TFA.

2.1.3 | In-gel digestion

Forty five microgram of the protein was loaded on a 1 mm thick 10-well 4–12% NuPAGE® Bis-Tris gel (Invitrogen, Carlsbad, CA). Proteins were separated in the gel for 1 h at 180 V. The gel was stained in NuPAGE® Colloidal Blue (Invitrogen) overnight at room temperature and de-stained with milli-Q water until the background was transparent. The gel lanes were cut into 48 identical slices using a custom-made OneTouch Mount and Lane Picker (The Gel Company, San Francisco, CA). Each slice was placed in a well in a 96-well polypropylene PCR plate (Greiner Bio-One, Frickenhausen Germany). In-gel digestion and peptide extraction were performed as described previously [16,17] but using acetic acid with a factor 10 higher concentration than TFA. Consecutive sample wells were combined to obtain 24 samples.

2.1.4 | LC

RPLC was performed using a splitless NanoLC–Ultra 2D plus system (Eksigent, Dublin, CA), controlled by HyStar 3.4. Four different mobile phase pH buffers were used. For all acquisitions the same 45 min linear gradient was used with increasing the organic solution in the mobile phase from 4 to 35%. At pH 3 the buffering solution was generated with the aqueous solvent being 0.05% formic acid and the organic solvent being 95% acetonitrile (ACN) and 0.05% formic acid. At pH 5.0, the aqueous solvent was a 10 mM ammonium acetate buffer and the organic solvent 75% ACN and 40 mM ammonium acetate buffer, and ammonia was used to adjust the pH to 5.0. For the pH 8.5 experiment, the a mobile phase used was of 10 mM ammonium acetate buffer as aqueous solvent and 75%

ACN with 40 mM ammonium acetate buffer as the organic solvent and the pH was adjusted with ammonia to reach a value of 8.5. For the pH 10 experiment, the aqueous solvent of 10 mM ammonium bicarbonate and the organic solvent consisted of 40 mM ammonium bicarbonate in 75% ACN, and the pH was adjusted with ammonia to 10.0.

For each analysis, 10 μL of sample was loaded and desalted on a C18 PepMap 300 μm, 5 mm i.d., 300 Å precolumn (Thermo Scientific) and separated by RPLC using a 150 mm 0.3 mm i.d. ChromXP C18CL, 120 Å column. A volume of 5 μL of ultra-pure water was added to each sample fraction.

2.1.5 | MS

MS was performed on an amaZon speed high-capacity 3D ion trap (Bruker Daltonics, Bremen, Germany), with CID as the fragmentation method, and precursor ion selection window of 5 *m/z* units. After each MS scan, up to ten abundant multiply charged species in the *m/z* 300–1300 range were automatically selected for MS/MS (ignoring singly charged species). After an ion is selected twice consecutively, it was excluded for 1 min. The MS was controlled by amaZon ion trap by trapControl 7.0 (Bruker).

2.2 | 1D RPLC–MS/MS and 2D strong cation exchange–RPLC–MS/MS datasets

The dataset was obtained from PRIDE (accession number PXD000705). In their work [15], Marino et al. acquired data from HEK293 cells digest on 1D RPLC–MS/MS as well as online 2D SCX coupled to RPLC–MS/MS. The detailed method and data acquisition parameters are in the original paper, and we briefly mention here few aspects about Marino et al. experiments and why they were considered to evaluate our method.

In their work Marino et al. built on the original work of Washburn et al. on online coupling of SDX to RPLC–MS/MS (MudPIT) [18]. The main objective was assessing the total analysis time, proteome coverage, and sample usage in two competing workflow; an online 2D SCX-RP-UHPLC–MS/MS versus a 1D long gradient RP-UHPLC–MS/MS analysis. Importantly, the two workflows used the same setup without any changes except bypassing the SCX column when measuring in the 1D long gradient RP-UHPLC–MS/MS setup. For both experiments, the authors used an Orbitrap Q-Exactive mass spectrometer (Thermo Scientific) recording data-dependent acquisitions with a top 10 method (top 10 most abundant precursors were chosen for MS/MS in every MS scan). In the 1D experiments Marino et al. acquired multiple datasets of the same HEK293 sample using increasing gradients of 45, 60, 90, 180, 240, 360, 480, and 600 min (including the washing steps, column equilibration, and loading, which were reported to sum up to 20–25 min). In the 2D experiments, they compared a short and a long second dimension gradient of 37 and 157 min, respectively, both following six salt plugs containing ammonium acetate at concentrations of 5, 10, 20, 50, 100, and 500 mM (with 5% ACN and 0.1% FA).

The main reason to use Marino et al. dataset to evaluate our method is to exploit the various aspect of what our RT modeling is designed for, namely compare multiple experiments ran on the same system under different conditions. Having 1D with various gradients is interesting to test how RT coefficient ranks behave. The 2D experiments contain different set of peptides, similar to our SDS-PAGE followed by LC–MS/MS approach, but in an online setup. Additionally, the authors reported variable peptide and protein coverage in the various experiments which is an interesting aspect to challenge RT modeling methods.

2.3 | Mass spectra data analysis preparation

The raw data were converted to mzXML [19] using compassXport 3.0 (Bruker) in the case of the *E. coli* datasets and ProteoWizard [20] in case of the HEK293 datasets obtained from PRIDE. All datasets were searched with X! Tandem [21, 22] as delivered in Trans-Proteomic Pipeline [21] with the k-score plugin (2013.06.15.1 – LabKey, Insilicos, ISB). X! Tandem output with peptide identifications and scores were then converted to pepXML [21], and processed using PeptideProphet to obtain the probability of each peptide-spectral match [23]. In the case of *E. coli* dataset, the X! Tandem search was performed against the UniProtKB *E. coli* reference set (2010-01-21) allowing a precursor monoisotopic mass error tolerance of 0.5–2.5 Da and fragment monoisotopic mass error of 0.4 Da. For the HEK293 datasets the search was performed against UniProtKB human reference database (2017-03-29) allowing a precursor monoisotopic mass error tolerance of 50 ppm and fragment monoisotopic mass error of 0.05 Da. Cysteine carbamidomethylation was fixed and methionine oxidation was set as a variable modification. The peptide spectrum match probability assigned by PeptideProphet by mixture modeling and the error rate, estimated at each probability threshold, were used in further analysis to filter the peptide spectrum matchs (PSMs) at specific false discovery rate (FDR).

2.4 | Linear regression modeling coefficients

Peptide retention modeling is performed using a linear regression model of the amino acid composition of the PSMs passing a given FDR threshold. The modeling coefficients are determined by solving the linear regression

$$t_{\mathrm{j}} = a_{0\mathrm{j}} + \sum_{\mathrm{i}=1}^{\mathrm{i}=20} n_{\mathrm{ij}}\, a_{\mathrm{ij}} \tag{1}$$

by minimizing the square error cost function

$$Cost\,(\boldsymbol{a}) = \sum_{\mathrm{j=1}}^{\mathrm{j=all\ peptides}} \left| t_{\mathrm{j}} - \left(a_{0\mathrm{j}} + \sum_{i=1}^{i=20} n_{\mathrm{ij}}\ a_{\mathrm{ij}} \right) \right|^2 \quad (2)$$

to find the optimal coefficients

$$\hat{\boldsymbol{a}} = argmin\,(Cost\,(\boldsymbol{a})) \quad (3)$$

where t_j is the RT of the peptide j, n_i is the number of AA_i occurrences in the peptide sequence, a_i is the coefficient of AA_i, and a_0 is the offset (***a*** in bold font refers to the coefficient matrix) [10]. The training of the linear model is done using PSMs with unmodified peptides only [11]. As methionine may oxidize during the measurement [24–26], we excluded all methionine-containing peptides in the analyses. Extending the model to include peptide with modified amino acids and terminal modifications is possible by adding additional terms in the equations for each modifications. This possibility will be discussed in Section 3.6.

In this work, we care mainly about obtaining and visualizing the coefficients a_i, and not the RTs. Whenever comparing multiple experiments, we consider the retention coefficient *rank*, which is the position of the amino acid RT coefficient in the sorted list of all other coefficients. For visualizing the amino acid coefficients, we use the Lesk color scheme [27] to represent the basic physicochemical properties of each amino acid (polar, small nonpolar, hydrophobic, acidic, basic).

2.5 | Comparing models from a set of experiments

To compare models obtained from multiple experiments and provide a measure of robustness, we define the single amino acid RT coefficient rank change as

$$S_{\mathrm{i,\ \Delta rank}} = \sum_{\mathrm{j=1,k=2}}^{\mathrm{j=L-1,k=L}} \left| Rank\left(AA_{\mathrm{i,j}}\right) - Rank\left(AA_{\mathrm{i,k}}\right) \right| \quad (4)$$

where $Rank\ (AA_{i,j})$ is the rank of the amino acid AA_i in experiment j, and L is the total number of experiments. $S_{i,\ \Delta rank}$ is the sum of all changes in the rank of amino acid i between each two experiments. We also define the sum of all amino acid delta rank scores as

$$S_{\Delta\mathrm{rank}} = \sum_{\mathrm{i=1}}^{\mathrm{i=20}} S_{\mathrm{i,\Delta rank}} \quad (5)$$

This is a dimensionless value and reflects the overall changes in the ranks of all amino acids between experiments, i.e. the more one (or more) amino acid changes its rank, the higher delta rank score is. $S_{i,\ \Delta rank}$ for a set of L experiments is estimated at a specific FDR value at the PSM level (for the model training set of peptide). To compare multiple delta rank scores obtained at various FDR values, scaling can be performed.

3 | RESULTS AND DISCUSSION

3.1 | Using a simple regression model for retention time modeling

The simple linear regression model of RT allows associating each amino acid with a single value. In this work the objective is not to predict the retention time of peptides, but to reveal hidden properties of the data deriving from the liquid separation. For this, a simple approach is critical, as it allows comprehendible visualization and is robust for small and large sets of training peptides. For example, the average number of peptides used in the modeling with FDR 1% at the PSM level was 425 for RPLC–MS/MS in our short ion trap datasets. In the Marino et al. 1D data, the average was 12 000 peptides (varying from 3000 to 23 000 with the increasing gradient time), and for the six experiments of 2D the average number of peptides with FDR 1% was 6100 for the short gradients of 37 min and 8800 for the long gradient of 157 min.

For the actual purpose of predicting RTs, models using artificial neural networks [28–30] or including more variables than the amino acid composition of peptides [31–33] have been discussed. The choice of the simple model in our method is entirely on purpose, as assigning a single coefficient to each amino acid allows ranking the amino acid contributions in the model and in turn a direct and visual comparison between the models of multiple datasets. In other words, while artificial neural networks are possibly better in predicting RT (given sufficiently large training dataset), prediction is not the goal here. Furthermore, it is not very meaningful to compare the thousands of coefficients (or connectivities) of two or more artificial neural networks. This is also true for regression models where additional peptide properties are considered (like pI and helicity). To that end, using a simple linear regression model allows deterministic assignment of values to variables, which in our approach are the amino acids themselves.

3.2 | Amino acid regression coefficients

In addition to comparing predicted and actual RTs (Figure 1A), we can plot the RT regression coefficients using an amino acid color scheme, such as the one defined by Lesk (Figure 1B). The colored circles show the average effect of each amino acid residue to the retention time of any peptide containing it. At pH 3.0, the basic residues (blue) contribute negatively to the RT, making any peptide containing them elute earlier, while the acidic (red) residues do not significantly contribute to the RT and the hydrophobic (green) amino acids contribute positively to RT, making any

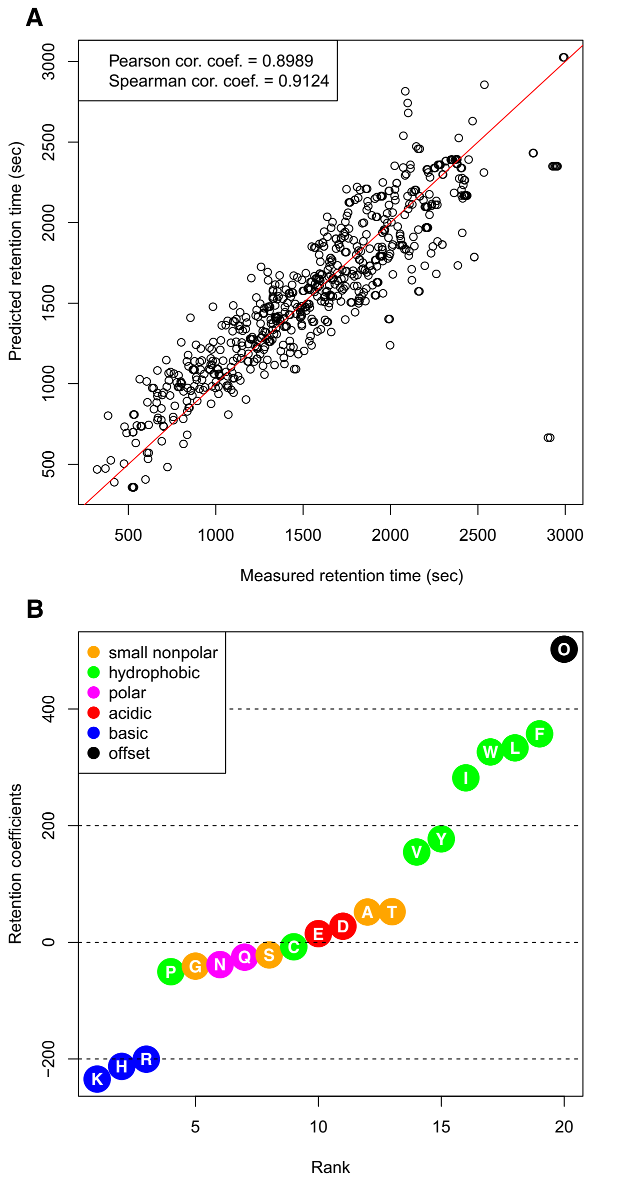


FIGURE 1 Retention time coefficients using a linear model on a peptide training set of an RPLC–MS/MS experiment at pH 3 and 1% FDR at the PSM level. Panel A shows the correlation between the predicted and measured retention time. Panel B shows the conversion between values and ranks of the amino acid retention coefficients. The amino acids are colored according to Lesk. The cysteines were carbamidomethylated, making these residues more hydrophilic. Proline constrains the conformation of the peptide and has an atypical influence on retention time. The offset typically takes on a positive value capturing the void time

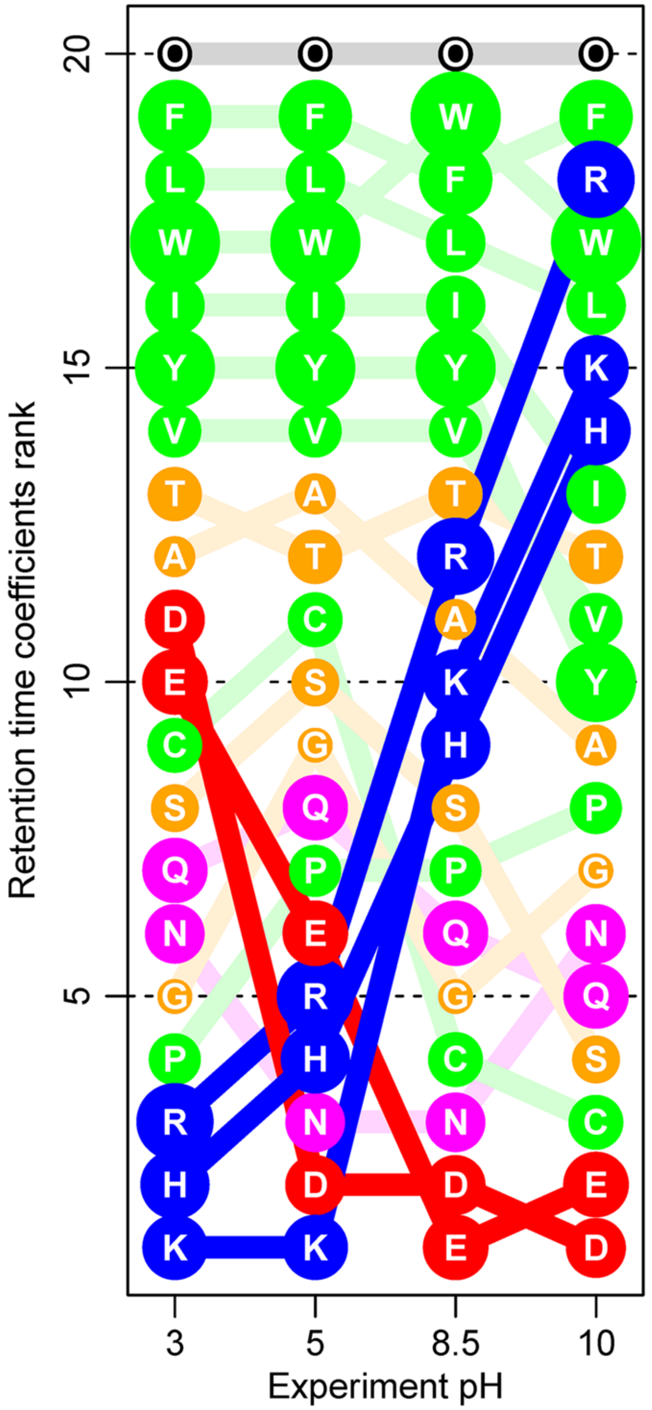


FIGURE 2 Retention time coefficient ranks as function of pH showing the changes in average hydrophobicity of the three basic (Arg, Lys, His) and two acidic (Asp, Glu) residues. These residues serve as intrinsic pH indicators in any RPLC separation of peptides

peptide containing them to elute later. This is expected as the basic residues are protonated and charged, and therefore hydrophilic, at low pH. Proline has an atypical influence on RT compared to the other hydrophobic residues, likely due to conformational effects on the peptide.

When comparing multiple experiments, the actual values of the amino acid coefficients are of minor importance and need to be transformed to allow the comparison. Here we introduced the coefficient ranks as in Figure 2 that shows the RT coefficient ranks of the amino acid residues with increasing mobile phase pH. The influence of pH on the amino acid RT coefficients can be immediately derived from the plot, with the acidic and basic residues exhibiting the largest change in

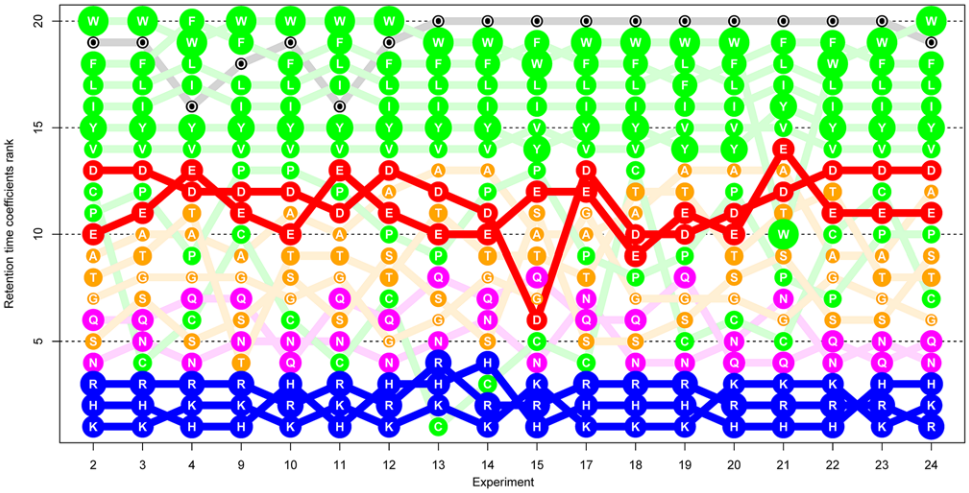


FIGURE 3 Retention time coefficient ranks across multiple experiments for which a model was produced at a training set FDR of 1.6%. At a 1.6% FDR we observed a local minimum (Figure 4), which is the FDR we used to select the peptide training list for all 24 runs that most produce the most robust model or minimizes the differences in retention time coefficient ranks across experiments

RT coefficient rank with changing pH of the mobile phase. This is because as the pH increases, the basic residues change from positive to neutral, and therefore become hydrophobic, and the acidic residues change from neutral to negative, and therefore become hydrophilic. The plot also shows that the basic and acidic residues have similar hydrophobicity around pH 5.

These effects of the individual amino acid residues are the average effects from measurement and identification of many peptides in a bottom-up proteomics experiment. Cysteine is categorized as hydrophobic by Lesk, but here we are measuring carbamidomethyl cysteine, not natural cysteine. The RT coefficients of the hydrophobic, polar and small nonpolar amino acids are barely influenced by the pH.

3.3 | Rank change and false discovery rate value optimization

To demonstrate the applicability of the method and show its independence of the dataset used in the training, we used two sets of 24 in-gel digested SDS-PAGE fractions. Analogous to Figure 2 from the first experiment, Figure 3 shows the amino acid coefficient ranks after peptide identification and FDR filtering from the second experiment. It is clear that even a small error in mobile phase preparation, for example being off by one pH unit, should be picked up by a quick inspection of the amino acid residue coefficients.

The delta rank score, or the sum of all distances between any two models (coefficients) in a set of models, has a minimum at the FDR value that produces best training set. The best training set implies the least discrepancies between the identifications used from the different experiments, which also means high agreement between experiments on the amino acid coefficient ranks. The experiments are performed on pre-fractionated sample, i.e. while the proteins and peptide training sets differ between the fractions, they contain the similar information on the underlying mechanism of the liquid separation.

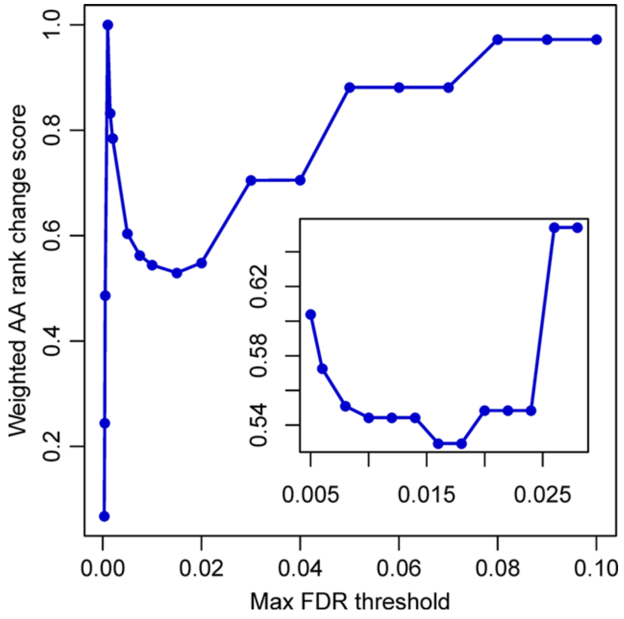


FIGURE 4 Model robustness measured by the sum of amino acid delta rank scores at different FDR with a local minimum is observed at 1.6% FDR

Figure 4 illustrates that datasets with too few peptides that in training a model. Increasing the FDR threshold will increase the number of peptides, but including too many false identifications in the training set will produce less robust models. For the RPLC data, the optimum FDR was 1.6% as shown in Figure 4. This FDR was used to filter the peptides to train each of the models shown in Figure 3.

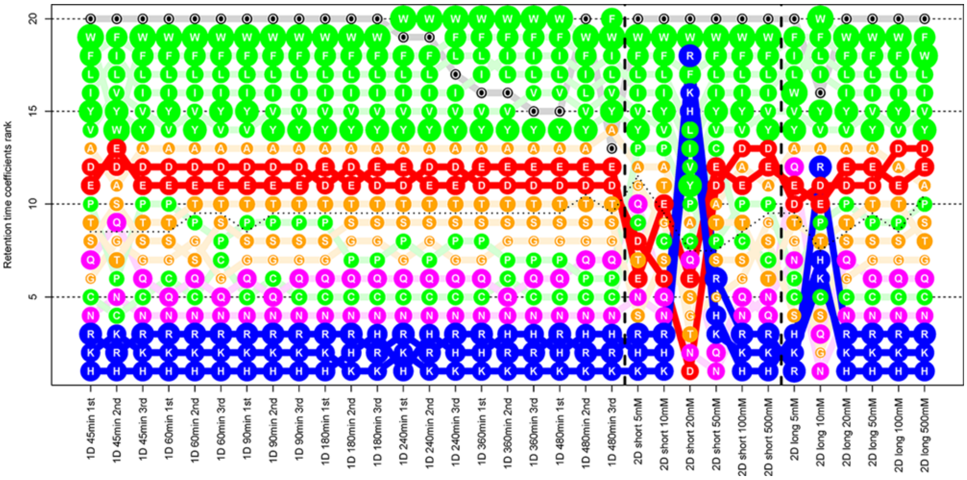


FIGURE 5 Retention time coefficient ranks across the 1D RPLC–MS/MS and 2D online SCX-RPLC–MS/MS experiments (data from Marino et al. [15]). The 1D RPLC–MS/MS were performed with triplicates and increasing retention time. The 2D online SCX-RPLC–MS/MS runs were for six ammonium acetate plugs ranging from 5 to 500 mM and using one short and one long gradients of 37 and 157 min respectively. All peptides used in the training sets for the retention time models were selected with 1% FDR threshold at the peptide spectrum match level

3.4 | Ranks reveal pH inconsistencies in 2D experiments

Using our method we produced 33 models for RT in the PRIDE dataset (Figure 5). These consist of 21 models for the 1D datasets, and 12 for the 2D datasets. The 21 models derived from the 1D datasets are very similar except for the offset moving to lower ranks in longer gradients. This behavior of the offset rank is expected as the void time is shorter relative to the longer gradients. The consistency between the 21 models of the 1D acquisitions is in agreement with the expectations that changing the gradient, and in turn the peptide RTs, should generate reproducible amino acid retention coefficient ranks and allow comparison between experiments. However, two of the models for 2D datasets are noticeably different (2D short 20 mM and 2D long 10 mM). Both of these anomalies are similar to the high pH (8.5–10) in Figure 2, suggesting there was a change in buffering and the actual pH during the separation in these two chromatographic separation. This was not noticed or mentioned by the authors of the original paper, suggesting the utility of the method presented here for simple QC of chromatographic separations in proteomics experiments. Under such conditions, a priori predicted RTs would not be accurate. For the original study these RT shifts were probably inconsequential as the goal was to compare the numbers of peptide and protein identifications between two methods. When looking into the data, the GRAVY scores of the peptides identified at 1% FDR (those used to build the models) were very different in the two acquisitions (2D short 20 mM and 2D long 10 mM) relative to all other acquisitions. The peptides in the 2D short gradient dataset with 20 mM salt plug were on average more hydrophobic than those in the other short runs, while those from the 2D long gradient with 10 mM salt plug were more hydrophilic than the other long runs. Interestingly, the cumulative distribution functions of GRAVY score of all 2D datasets (short and long) are almost encapsulated completely within the ones from the short run with 20 mM salt plug to the higher hydrophobicity side, and the long 10 mM to the lower side (Supporting Information Figure S1).

3.5 | Training set size

To investigate the conversion and stability of the model in regard to the size of the training set, we trained consecutive models with increasing number of peptides in the training set. We used the in-solution digestion *E. coli* dataset (used for Figure 1) at 1% FDR and with each new model we added ten random peptides. Supporting Information Figure S2 shows that the model starts to converge when using around 100 peptides, with some discrepancies in the ranks of the acidic amino acids, although at that small training set the basic amino acid retention coefficients already found their places. Having around 200 peptides in the training set the basic and acidic amino acids retention coefficient ranks already take their final rank positions, and the rest of the amino acids retention coefficient ranks converge when using around 400 peptides. It has been shown previously that the simple linear regression model for RT modeling needs around 50 occurrences of each amino acid in the training set for conversion [34].

3.6 | Considering modified amino acids and terminal modifications

In addition to amino acid retention coefficients, considering RT coefficients for post transnationally modified amino acids and peptide terminal modifications is also possible. While

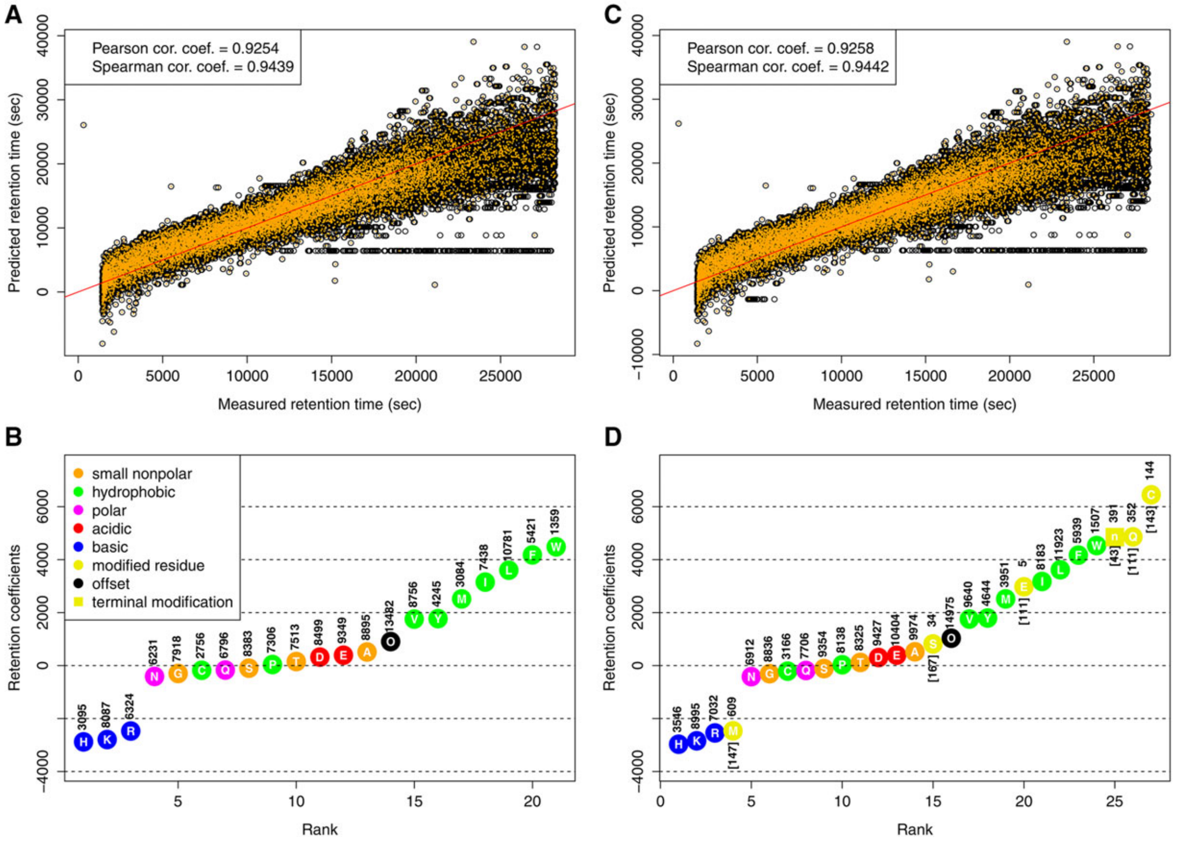


FIGURE 6 The effect of including post-translationally modified peptides in the retention time model. The 1D RPLC–MS/MS dataset form Marino et al. with the long gradient of 480 min was used to build two models, one without (panels A and B) and one with modified peptides (panels C and D). The numbers above the symbols indicate the number of unique peptides in the training set which included the (modified) amino acid. The numbers below the symbols denote the mass of the modified residue or terminal adduct

this improves the model slightly, it requires that the training set includes enough peptide entries with the modifications. In Figure 6 we used the 1D RPLC–MS/MS dataset form Marino et al. with the long 480 min gradient to build two models, one considers the peptides with no modification shown in panels A and B, and the second considers peptides with as well as without modifications in panels C and D. From the values of the correlation coefficients we can imply that the model with more retention coefficient entries, i.e. including the modifications, is slightly better. Comparisons and plots like in Figures 3 and 5 can be extended to consider peptides with modifications, however when comparing models from multiple experiments, it is necessary to have enough representative peptides in the training set for the modified amino acids (as well as the unmodified ones). This is probably the case with technical replicates and comparable measurements, but not necessary the case for experiments with variable gradients.

4 | CONCLUDING REMARKS

We have demonstrated the use of a simple and robust model of peptide behavior in RPLC separations for easy data quality inspection and detection of reproducibility issues across many datasets. Our method also uses a data-derived FDR value that maximizes the agreements of acquired data, increasing robustness. We demonstrated the utility of the method by applying it on datasets from five different experiments, including modifying mobile phase pH, and comparing RPLC–MS/MS with 2D SCX-RPLC–MS/MS. The method is not limited to RPLC and we were able to apply its data obtained using CE in place of RPLC (data not presented in this work). The visualization captured the effect of changing pH by revealing changes in the ranks of the amino acid coefficients. RPLC–MS/MS analysis of 24 in-gel digested SDS-PAGE fractions have demonstrated how our modeling allows to derive an optimal FDR threshold that maximize the agreement between multiple experiments on the same system. We believe this approach has the potential to be a method for data/experiment-derived FDR threshold for accepting spectrum peptide matches by maximizing the agreement between experiments measured on the same system. We note that the FDR threshold obtained in our approach, i.e. 1.6% is close to the 1% value commonly accepted in the proteomics community, suggesting that using higher values would introduce discrepancies in identified sets of peptides. We were also able to show changes in pH in online 2D SCX-RPLC–MS/MS experiments obtained from PRIDE when comparing the models of the SCX fractions between each other and with those of 1D RPLC–MS/MS on the same system. In contrast to extended modeling methods like support victor machines or artificial neural networks, using a simple modeling approach produces comprehendible amino acid retention coefficients and facilitates for visualization that allows direct comparison between experiments. Importantly, our method does not require any additional modification of the experimental setup or addition of RT standards. It is simply making use of already available information for additional QC, which can also be applied retrospectively. An implementation of the described method in R statistical language is available online under https://cpm.lumc.nl/yassene/rt_modeling/.

ACKNOWLEDGMENTS

The authors thank Dana Ohana, Hans Dalebout, Wendy Plugge for their help in conducting part of the experimental work. The authors also thank Dr. Oleg. Mayboroda for the suggestions on how to improve this work. This work was funded in large part by the NWO Vidi grant 917.11.398 (MP, YM).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Yassene Mohammed http://orcid.org/0000-0003-3265-3332
Magnus Palmblad http://orcid.org/0000-0002-5865-8994

REFERENCES

1. Yates, J. R., 3rd, Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* 2004, *33*, 297–316.
2. Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198–207.
3. Fonslow, B. R., Yates, J. R., 3rd, Capillary electrophoresis applied to proteomic analysis. *J. Sep. Sci.* 2009, *32*, 1175–1188.
4. Zhao, Y., Kong, R. P., Li, G., Lam, M. P., Law, C. H., Lee, S. M., Lam, H. C., Chu, I. K., Fully automatable two-dimensional hydrophilic interaction liquid chromatography-reversed phase liquid chromatography with online tandem mass spectrometry for shotgun proteomics. *J. Sep. Sci.* 2012, *35*, 1755–1763.
5. Moruz, L., Kall, L., Peptide retention time prediction. *Mass Spectrom. Rev.* 2017, *36*, 615–623.
6. Krokhin, O. V., Spicer, V., Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem.* 2009, *81*, 9522–9530.
7. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J., Rinner, O., Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 2012, *12*, 1111–1121.
8. Bodzioch, K., Dejaegher, B., Baczek, T., Kaliszan, R., Vander Heyden, Y., Evaluation of a generalized use of the log Sum(k+1)AA descriptor in a QSRR model to predict peptide retention on RPLC systems. *J. Sep. Sci.* 2009, *32*, 2075–2083.
9. Tyteca, E., De Vos, J., Vankova, N., Cesla, P., Desmet, G., Eeltink, S., Applicability of linear and nonlinear retention-time models for reversed-phase liquid chromatography separations of small molecules, peptides, and intact proteins. *J. Sep. Sci.* 2016, *39*, 1249–1257.
10. Meek, J. L., Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. U. S. A.* 1980, *77*, 1632–1636.
11. Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P., Bergquist, J., Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* 2002, *74*, 5826–5830.
12. Krokhin, O. V., Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300-and 100-angstrom pore size C18 sorbents. *Anal. Chem.* 2006, *78*, 7785–7795.
13. Gilar, M., Jaworski, A., Retention behavior of peptides in hydrophilic-interaction chromatography. *J. Chromatogr. A* 2011, *1218*, 8890–8896.
14. Gilar, M., Xie, H., Jaworski, A., Utility of retention prediction model for investigation of peptide separation selectivity in reversed-phase liquid chromatography: impact of concentration of trifluoroacetic acid, column temperature, gradient slope and type of stationary phase. *Anal. Chem.* 2010, *82*, 265–275.
15. Marino, F., Cristobal, A., Binai, N. A., Bache, N., Heck, A. J., Mohammed, S., Characterization and usage of the EASY-spray technology as part of an online 2D SCX-RP ultra-high pressure system. *Analyst* 2014, *139*, 6520–6528.
16. Mostovenko, E., Deelder, A. M., Palmblad, M., Protein expression dynamics during Escherichia coli glucose-lactose diauxie. *BMC Microbiol.* 2011, *11*, 126.
17. van der Plas-Duivesteijn, S. J., Mohammed, Y., Dalebout, H., Meijer, A., Botermans, A., Hoogendijk, J. L., Henneman, A. A., Deelder, A. M., Spaink, H. P., Palmblad, M., Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J. Proteome. Res.* 2014, *13*, 1537–1544.
18. Washburn, M. P., Wolters, D., Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.
19. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 2004, *22*, 1459–1466.
20. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T.,

Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 2012, *30*, 918–920.

21. Keller, A., Eng, J., Zhang, N., Li, X. J., Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* 2005, *1*, 1–8.
22. Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.
23. Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.
24. Morand, K., Talbo, G., Mann, M., Oxidation of peptides during electrospray ionization. *Rapid Commun. Mass Spectrom.* 1993, *7*, 738–743.
25. Boys, B. L., Kuprowski, M. C., Noel, J. J., Konermann, L., Protein oxidative modifications during electrospray ionization: solution phase electrochemistry or corona discharge-induced radical attack? *Anal. Chem.* 2009, *81*, 4027–4034.
26. Pei, J., Zhou, X., Wang, X., Huang, G., Alleviation of electrochemical oxidation for peptides and proteins in electrospray ionization: obtaining more accurate mass spectra with induced high voltage. *Anal. Chem.* 2015, *87*, 2727–2733.
27. Lesk, A. M., Introduction to bioinformatics, Oxford University Press, Oxford, United Kingdom 2014.
28. Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasa-Tolic, L., Lipton, M. S., Auberry, K. J., Strittmatter, E. F., Shen, Y., Zhao, R., Smith, R. D., Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* 2003, *75*, 1039–1048.
29. Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E., Strittmatter, E. F., Qian, W. J., Adkins, J. N., Moore, R. J., Xu, Y., Lipton, M. S., Camp, D. G., 2nd, Smith, R. D., Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* 2006, *78*, 5026–5039.
30. Strittmatter, E. F., Ferguson, P. L., Tang, K., Smith, R. D., Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* 2003, *14*, 980–991.
31. Klammer, A. A., Yi, X. H., MacCoss, M. J., Noble, W. S., Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* 2007, *79*, 6111–6118.
32. Moruz, L., Tomazela, D., Kall, L., Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome. Res.* 2010, *9*, 5209–5216.
33. Pfeifer, N., Leinenbach, A., Huber, C. G., Kohlbacher, O., Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *Bmc Bioinformatics* 2007, *8*.
34. Palmblad, M., Retention time prediction and protein identification. *Methods Mol. Biol.* 2007, *367*, 195–207.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Mohammed Y, Palmblad M. Visualization and application of amino acid retention coefficients obtained from modeling of peptide retention. *J Sep Sci* 2018;41:3644–3653. https://doi.org/10.1002/jssc.201800488