analytical chemistry

Article

pubs.acs.org/ac

Sequence-Specific Model for Peptide Retention Time Prediction in Strong Cation Exchange Chromatography

Daniel Gussakovsky,[†] Haley Neustaeter,[†] Victor Spicer,[‡] and Oleg V. Krokhin*[,‡,§]

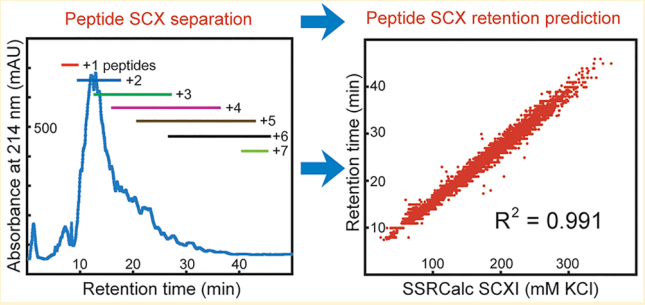
[†]Department of Chemistry, University of Manitoba, 360 Parker Building, Winnipeg, Manitoba R3T 2N2, Canada

[‡]Manitoba Centre for Proteomics and Systems Biology, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada

[§]Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, Manitoba R3E 3P4, Canada

Supporting Information

ABSTRACT: The development of a peptide retention prediction model for strong cation exchange (SCX) separation on a Polysulfethyl A column is reported. Off-line 2D LC-MS/MS analysis (SCX-RPLC) of *S. cerevisiae* whole cell lysate was used to generate a retention dataset of ∼30 000 peptides, sufficient for identifying the major sequence-specific features of peptide retention mechanisms in SCX. In contrast to RPLC/hydrophilic interaction liquid chromatography (HILIC) separation modes, where retention is driven by hydrophobic/hydrophilic contributions of all individual residues, SCX interactions depend mainly on peptide charge (number of basic residues at acidic pH) and size. An additive model (incorporating the contributions of all 20 residues into the peptide retention) combined with a peptide length correction produces a 0.976 R^2 value prediction accuracy, significantly higher than the additive models for either HILIC or RPLC. Position-dependent effects on peptide retention for different residues were driven by the spatial orientation of tryptic peptides upon interaction with the negatively charged surface functional groups. The positively charged N-termini serve as a primary point of interaction. For example, basic residues (Arg, His, Lys) increase peptide retention when located closer to the N-terminus. We also found that hydrophobic interactions, which could lead to a mixed-mode separation mechanism, are largely suppressed at 20−30% of acetonitrile in the eluent. The accuracy of the final Sequence-Specific Retention Calculator (SSRCalc) SCX model (∼0.99 R^2 value) exceeds all previously reported predictors for peptide LC separations. This also provides a solid platform for method development in 2D LC-MS protocols in proteomics and peptide retention prediction filtering of false positive identifications.



Strong cation exchange (SCX) separation of peptides is the second most popular mode of peptide separation in proteomics behind reversed-phase (RP) LC.[1] Due to the simplicity of its coupling to ESI-MS, the supremacy of RPLC is unlikely to be challenged. This leaves the remaining separation modes (hydrophilic interaction liquid chromatography (HILIC), SCX, anion-exchange, and high pH RP) to compete for the supporting role of the first separation dimension in 2D LC-MS methods or for use in peptide enrichment protocols. SCX possesses sufficient separation orthogonality with RPLC,[2] which prompted its wide use in bottom-up proteomics. Moreover, due to compatibility of the eluents, it can be used in both off-line[3] and on-line[4,5] 2D LC of complex peptide mixtures. Extensive literature in the field of proteomics gives clear indication of the dominant role of 2D (SCX-RP) LC-MS methodology in the past two decades.

Rapid developments in the field of proteomics have rejuvenated the interest of separation scientists in peptide retention modeling.[6] Peptide retention prediction has found applications in developing quantitative LC-MS protocols,[7] filtering false positive MS/MS identifications,[8−10] and guiding method development in multidimensional LC-MS.[11] The major efforts were understandably directed toward developing prediction models for RPLC.[6,12−14] However, recent reports indicate further advancements in modeling peptide retention for high pH RP,[10] capillary zone electrophoresis (CZE),[15] and HILIC[16−18] using proteomics derived data. Our lab has been active in the peptide retention prediction field since 2004, developing predictive models for RPLC,[10,12,19] methods for standardization of peptide separations,[20] and retention data collection using 2D[10] and 3D LC-MS approaches.[11] In 2017, we expanded the application of our Sequence-Specific Retention Calculator (SSRCalc) model into peptide CZE[15] and HILIC.[18] Attempting to develop a SSRCalc SCX model would be a natural continuation of our efforts in this direction.

Received: August 23, 2017
Accepted: October 3, 2017
Published: October 3, 2017

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

SCX separation of peptides and proteins has a rich history going back to the 1980s, when it was recognized as one of the most potent methods of separation for these compounds.[21−23] Most of the peptide ion-exchange separations are performed using a salt gradient, often with the addition of an organic solvent to reduce hydrophobic interaction between the peptide and the stationary phase.[24] Efforts to understand separation mechanisms[25] and develop peptide retention prediction models[26] were based on the electrostatic interactions in ion-exchange chromatography of peptides and proteins driven by Coulomb's law:

$$F = Q_1Q_2/dr^2$$

where Q_1 and Q_2 are interacting charges of opposite sign, r is the distance between them, and d is the dielectric constant of the medium.[25] Therefore, peptide retention increases with peptide charge, a fact which was clearly recognized in seminal studies of peptide SCX. Studying the dependence of SCX retention of peptides with the same charge, but varying size, Hodges et al.[26] concluded that peptide retention time is proportional to $Q/\ln(N)$, where N is the number of residues. This model, however, was tested on a very limited number of peptides.

The introduction of mass spectrometry and proteomics has helped to increase the size of retention datasets available for modeling to thousands of peptides. Resing and co-workers[9,27] derived semiquantitative rules that describe peptide elution from a SCX column based on the number of charged residues. They found that complete separation of peptides based solely on charge is very hard to achieve. Nevertheless, simple correlation between the number of basic residues (BRs) and SCX retention allowed them to use SCX retention information for additional peptide retention filtering and improving confidence of MS/MS identification. Trinidad et al.[28] explored fractionation of nonmodified and phosphorylated peptides by SCX. They found a similar correlation between peptide charge and retention and concluded that phosphorylation decreases peptide retention due to the acidic character of the modifying group. This effect is widely utilized in phospho-peptide enrichment protocols.[29]

The first attempts to develop a sequence-dependent model for peptide SCX was undertaken by Petritis et al.[30] The authors used an Artificial Neural Network (ANN) approach with a combined retention dataset of ∼190 000 peptides acquired in 2250 LC-MS experiments. The ANN structure was based on 5 hidden layers and 1055 input nodes, which included parameters such as position of individual residues, peptide charge, pI, charge of peptide in gas phase, length, and hydrophobic moment. A resulting correlation of ∼0.9 R^2 value was demonstrated.

The most prominent sequence-specific feature, which has been included in all advanced peptide retention prediction models,[12−15,18] comes from the unique role of terminal residues in peptide interaction with the stationary phase. In RPLC, it manifests itself in reduced hydrophobic interactions of terminal amino acids due to the association of charged N-terminal amino group with hydrophilic counterions.[19] The use of separate retention coefficients for individual amino acids became a standard solution of this problem in RPLC,[12−14,19] HILIC,[18] and CZE[15] but requires a significantly larger dataset to avoid overfitting. Alpert et al.[31] explored these effects using the extended SCX retention dataset of Petritis et al.[30] and found a significant influence of peptide orientation in cation exchange and ERLIC separation modes. In the case of SCX, it originates from preferential interaction of positively charged N-termini with the stationary phase, thus increasing/decreasing the interaction of N-terminal basic (Lys, Arg, His)/acidic (Asp, Glu) residues. Mant et al.[32] demonstrated another sequence-dependent effect in peptide SCX: synthetic amphipathic peptides with four positively charged residues in the hydrophilic face showed increased retention compared to the non-amphipathic analog of identical composition.

The review of the literature shows that, despite a long history, a good understanding of the basic principles, and an abundance of SCX applications in proteomics, the modeling of peptide retention in SCX trails behind other peptide separation techniques. The goal of our study was to collect retention data using 2D LC-MS (SCX-RP, ∼40−50 fractions in the first dimension) of a complex tryptic digest and to develop a sequence-specific retention model to quantitatively describe peptide retention in cation-exchange mode. This closely follows the established methodology from our recent efforts to model HILIC separation.[18] Retention modeling using tens of thousands of data points should provide sufficient information for defining major features of cation-exchange separation and an in-depth look at sequence-specific retention features of peptides' SCX.

■ EXPERIMENTAL SECTION

Materials and Digest Preparation. Unless otherwise noted, all chemicals were sourced from Sigma Chemicals (St. Louis, MO). Eluents were prepared using HPLC-grade acetonitrile, deionized water, formic acid, and potassium chloride (Thermo Fisher Scientific (Toronto, ON)). Sequencing grade modified trypsin (Promega, Madison, WI) and 15 mL Amicon centrifugal filter units (Merck Millipore, Ireland) were used for the digestion. Chromatographic fractions were collected in siliconized 1.5 mL tubes (BioPlas, San Rafael, CA). The custom designed standard peptides P1−P6[20] as well as the synthetic peptides with different charges at acidic pH (LASAADFG (+1), LASAADFR (+2), LASAAHFR (+3), and LAHAAHFR (+4)) were synthesized by Bio-Synthesis Inc. (Lewisville, TX).

The *S. cerevisiae* tryptic digest was prepared with the FASP protocol scaled up for 15 mL centrifugal filter units.[33] The digest was acidified with trifluoroacetic acid (TFA), purified by reversed-phase SPE, aliquoted into vials with ∼200 μg of peptides in each vial (according to NanoDrop 2000 (Thermo-Fisher)), and lyophilized. ∼200 μg of digest was used for the first-dimension separation.

First Dimension Separation Conditions. An Agilent 1100 series HPLC system with UV detector (214 nm) and a 200 μL injection loop was used for SCX separations. A 2.1 mm × 100 mm Polysulfethyl A, 5 μm 200 Å column (PolyLC, Columbia, MD) was also used with a 300 μL/min flow rate. Eluent A consisted of 80:20 water/acetonitrile and 0.1% formic acid. Eluent B was identical to eluent A plus 500 mM of KCl. Separation conditions were optimized to fit a 50 min separation window: linear increase of eluent B from 0% to 100% in 60 min or gradient increase of 8.5 mM KCl per minute. The gradient was followed by a 5 min wash with 100% eluent B and a 40 min equilibration step with 100% eluent A. We collected 46 1 min fractions, which were then lyophilized.

Fractions were resuspended in 0.1% TFA in water and desalted using a C18 4.6 mm guard cartridge (Phenomenex,

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

Torrance, CA). Once desalted, the fractions were lyophilized once more, resuspended in buffer A (0.1% formic acid in water) for the second-dimension separation, and spiked with approximately 200 fmol of the standard P1−P6 peptides. The volume of dilution buffer was adjusted on the basis of NanoDrop 2000 measurements and the UV profile of the LC trace to ensure injections of ∼1 μg or less of peptides per injection.

Second Dimension LC-MS/MS. The 2D LC Ultra system (Eksigent, Dublin, CA) delivered buffers A and B through a 100 μm × 200 mm analytical column packed with a 3 μm Luna C18(2) (Phenomenex) at a 500 nL/min flow rate. Samples representing each individual fraction were loaded on a 300 μm × 5 mm PepMap 100-trap column (ThermoFisher). The gradient program was as follows: a linear increase from 0.5% to 37% buffer B (acetonitrile) in 78 min, 5 min at 90% buffer B, and then 7 min at 0.5% buffer B for column equilibration (90 min total analysis time). Both buffers A and B contained 0.1% formic acid.

A TripleTOF5600 mass spectrometer (Sciex, Concord, ON) in standard MS/MS mode was used for data-dependent acquisition; settings used were: 250 ms survey MS spectra (m/z 375−1250) followed by up to 20 MS/MS measurements on the most intense parent ions (400 counts/second threshold for charged states between +2 and +5, m/z 100−1600 mass range for MS/MS, and 100 ms each). Previously targeted parent ions were excluded for 12 s from repetitive MS/MS acquisition.

Data Analysis and Retention Time Assignment. Protein/peptide identification was performed by the X!Tandem algorithm. The search parameters included 20 and 50 ppm mass tolerance for parent and daughter ions, respectively, and constant modification of Cys with iodoacetamide. All potential modifications were excluded for peptide identification.

Within the first dimension, retention times were assigned as being equal to the fraction number in which the peptide was found. When the peptide signal was distributed between two or more fractions, an intensity weighted average fraction number was used. The retention times of peptides in the second dimension were converted into HI (% acetonitrile) units using the established retention values of the standard peptides.[20]

■ RESULTS AND DISCUSSION

Selection of Chromatographic Conditions in SCX Mode. Our literature review showed that the majority of SCX separations are performed using a salt gradient involving either potassium chloride or ammonium formate at acidic pH. Burke et al.[24] observed that the addition of acetonitrile to the eluents in SCX separations reduces hydrophobic interactions. Therefore, peptide separations in proteomics have been mostly performed using 20−30% organic solvent in the eluent. A notable exception from this rule can be found in on-line coupling of SCX and RP, where acetonitrile concentration is usually kept at 5% to ensure peptide retention on RP phase.[4,5,34] We used a 0−500 mM KCl gradient at acidic pH, alongside 20% acetonitrile in both eluents A and B. Additionally, the gradient slope had to provide sufficient peptide separation to fit the expected ∼50 min elution window. Figure 1A shows the separation of four peptides with different charges (+1 to + 4) using an optimized gradient slope of 8.5 mM KCl per minute. This provided a desired separation window of the yeast tryptic digest as shown in Figure 1B. Figure 1C also shows the distribution of identified peptides of various charges across the salt gradient, which coincides with the distribution of synthetic peptides in Figure 1A.

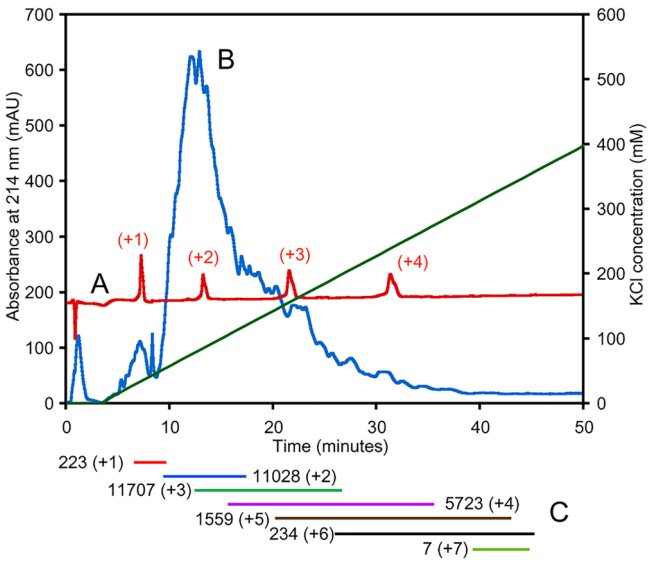


Figure 1. Selection of chromatographic conditions for the SCX separation of a complex digest. (A) Separation of four synthetic peptides with different charges (+1 to + 4). (B) Separation of *S. cerevisiae* tryptic digest (salt gradient profile at the exit of the column is shown in green). (C) Distribution of tryptic peptides with different charges across the chromatogram.

LC-MS/MS Analysis in the Second Dimension: Identification Output. Each collected fraction was desalted, lyophilized, and submitted to the second dimension LC-MS analysis. ∼1 μg (or less) of peptides was injected for each fraction, which required an adjustment of the dilution volume depending on the UV profile shown in Figure 1B. The fractions that were analyzed corresponded to 69 h of instrument time. In total, the acquisition of 552 954 MS/MS spectra resulted in identification of 196 470 of them corresponding to 34 454 unique peptides (log (e) < −3) and 4185 proteins (log (e) < −3). Table 1 compares identification output to previously reported 2D LC-MS/MS analyses of peptide retention modeling using the same MS platform.[10,18]

Retention Time Prediction Filtering. The development of retention prediction models requires high quality retention data. The preferable option is to analyze synthetic peptides or digests of purified proteins with known sequences. However, this introduces time and cost constraints when larger datasets are required. Our experience shows that 2D LC-MS/MS analysis of complex digests with retention time prediction filtering in both dimensions acts as a compromise between the quality and the size of the retention dataset.[10,18] In this work, we used high confidence peptide identifications with log (e) score < −3. Next, analyzing the peptides with the highest prediction errors in both dimensions (intermediate version of SCX model was used), we removed suspected chromatographic outliers. Most of them represented peptides with unanticipated missed cleavage sites. At this step, we excluded ∼0.1% of identifications (43 peptides) from modeling. The remaining population of 30 482 peptides (Table S-1) was used for the model optimization where tryptic peptides in the dataset were 6−49 residues long (16 on average), carrying 1−8 positive charges at acidic pH.

Optimization and Major Features of Additive SCX Model. Ion exchange separation, as the literature suggests, is

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

Table 1. Identification Output of 2D (SCX-RP), 2D (HILIC-RP), and 2D (RP-RP)-LC-MS/MS for the Analysis of Whole Cell Yeast Tryptic Digest[a]

separation mode	number of fractions	total LC-MS time (h)	amount injected (μg)	# of MS/MS	# of identified peptides	# of nonredundant peptide IDs	# of protein IDs
SCX-RP	46	69	~35	552 954	196 470	34 454	4185
HILIC-RP	38	57	~30	389 917	171 844	34 832	4218
RP-RP[b]	20	30	~30	226 386	103 586	27 286	4093

[a]Confidence score log (e) < −3 or better was used for both peptides and proteins. [b]A standard 2D LC-MS/MS (high pH to low pH) with fraction concatenation applied in our lab.[10]

driven by the Coulombic interaction between the peptide and the stationary phase.[25] As a result, the larger the charge of the peptide, the stronger is the interaction, and therefore, longer is the retention time. Specifically, at acidic conditions, the peptide is positively charged, and the stationary phase is negatively charged. Thus, in general, the more basic residues the peptide contains (Arg, Lys, His), the greater is its retention time. For our first approximation of the SCX model, we counted the number of basic residues and then added one charge for the N-terminus to determine the peptide charge. Table 2 and Figure 2 show a step-by-step optimization process of the SSRCalc SCX model. Correlation between the charge of the peptide and the fraction number the peptide eluted in showed an R^2 value of 0.858. Additional examination of the distribution of peptides in each respective charge groups revealed that the larger peptides elute prior to the shorter ones, in complete agreement with Coulomb's law (Figure 2A).

Table 2. Optimization of SSRCalc SCX Model

optimization step	model information	number of variables	R^2 value	prediction error standard deviation (min)
1	Q	0	0.858	2.30
2	$Q/\ln(N)$	0	0.943	1.46
3	$Q \times (1 + 9.571/\ln(N))$	0	0.952	1.35
4	$Q \times (1 + C_Z/\ln(N))$	8	0.955	1.29
5	$Q \times (1 + C_Z/\ln(N))$ + composition	28	0.976	0.94
6	$Q \times (1 + C_Z/\ln(N))$ + composition + position-dependent R_C's	148	0.9862	0.72
7	reoptimized	148	0.9868	0.70
8	reoptimized + $i + 3$; $i + 4$ interactions	150	0.987	0.69
9	polynomial correction	174	0.991	0.64

To correct for the peptide length, a few approaches had been applied before reaching an optimal solution (Table 2). Hodges et al.[26] proposed correction for the peptide length was based on the division of Q, the net charge of the peptide, by the natural logarithm of N, the number of amino acids in the peptide sequence: $Q/\ln(N)$. Application of this correction resulted in a 0.943 R^2 value for our dataset. We improved the resulting correlation by introducing a slightly different length correction of $Q \times (1 + 9.571/\ln(N))$, where the coefficient 9.571 was optimized to fit all peptides (R^2 value 0.952). Because of slightly different behavior of groups of peptides with different charges, we introduced variables C_Z instead of the constant 9.571. These coefficients were slightly different for each individual charge from +1 to +7 (Table S-2). This improved the R^2 value to 0.955 when plotting the dependence of the fraction number versus $Q \times (1 + C_Z/\ln(N))$ as shown in Figure 2B.

All advanced peptide retention prediction models are based on accounting for interaction of individual amino acids with the stationary phase through the introduction of individual retention coefficients (R_C). We assumed that, in SCX, each residue will alter the effective charge of the peptide and optimized retention coefficients for each amino acid contributing to overall charge Q (internal position in Table 3). The starting point for this optimization step was the original assumption that Lys, His, and Arg had an effective charge of +1 ($R_C = 1$) and all other amino acids had an effective charge of 0. The optimized retention coefficients are shown in Figure 3A, confirming once again the dominant role of basic residues. The resulting R^2 value of the model was improved to 0.976 (Figure 2C). Among other trends, the positive contribution of Trp and Asn should be highlighted. Reoptimization of our model to fit Trinidad et al.[28] data (see Application of SCX Prediction Model) resulted in virtually the same accuracy of the final model (R^2 value ~0.984) and showed characteristic changes in R_C values for some residues (Figure 3B). Trinidad's data was collected using 30% acetonitrile in the eluent in contrast to 20% acetonitrile in our model. The reoptimized retention coefficients showed a consistent decrease in contribution of hydrophobic residues, especially aromatic ones (Trp, Phe, Tyr). Meanwhile, the retention coefficient of Asn and other neutral hydrophilic residues remained constant. This change, admittedly minor, shows that hydrophobic interactions between peptide and Polysulfethyl A stationary phase are largely extinct at 20% acetonitrile in the eluent but still visible, driving the difference between eluents with different contents of organic solvent. Note that the vast majority of tryptic peptides are not retained on the more hydrophobic C18 phase at 30% acetonitrile. Hydrophobic interactions are expected to be even less pronounced on the hydrophilic Polysulfethyl A phase.

Position-Dependent Retention Coefficients. As shown in our previous publications on RP,[10] CZE,[15] and HILIC,[18] the position of amino acids, relative to the ends of the peptides, is an important characteristic in peptide separation modeling. The terminal location allows the amino acid residues to interact more freely with the stationary phase in comparison to the internal amino acids. In RP, and to a lesser degree in HILIC, hydrophobicity/hydrophilicity of N-terminal residues is altered due to the interaction of positively charged N-termini with eluent counteranions (acetate, formate). Additionally, acidic residues (Asp, Glu) near N-termini significantly reduce effective peptide charge in CZE due to an induction effect, thus reducing electrophoretic mobility. SCX is a surface-based as well as a charge-based separation; therefore, one might expect different behavior of the basic and acidic residues in the terminal positions. Optimized position-dependent coefficients (Table 3) confirm these assertions. The R_C's of basic amino acids are

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

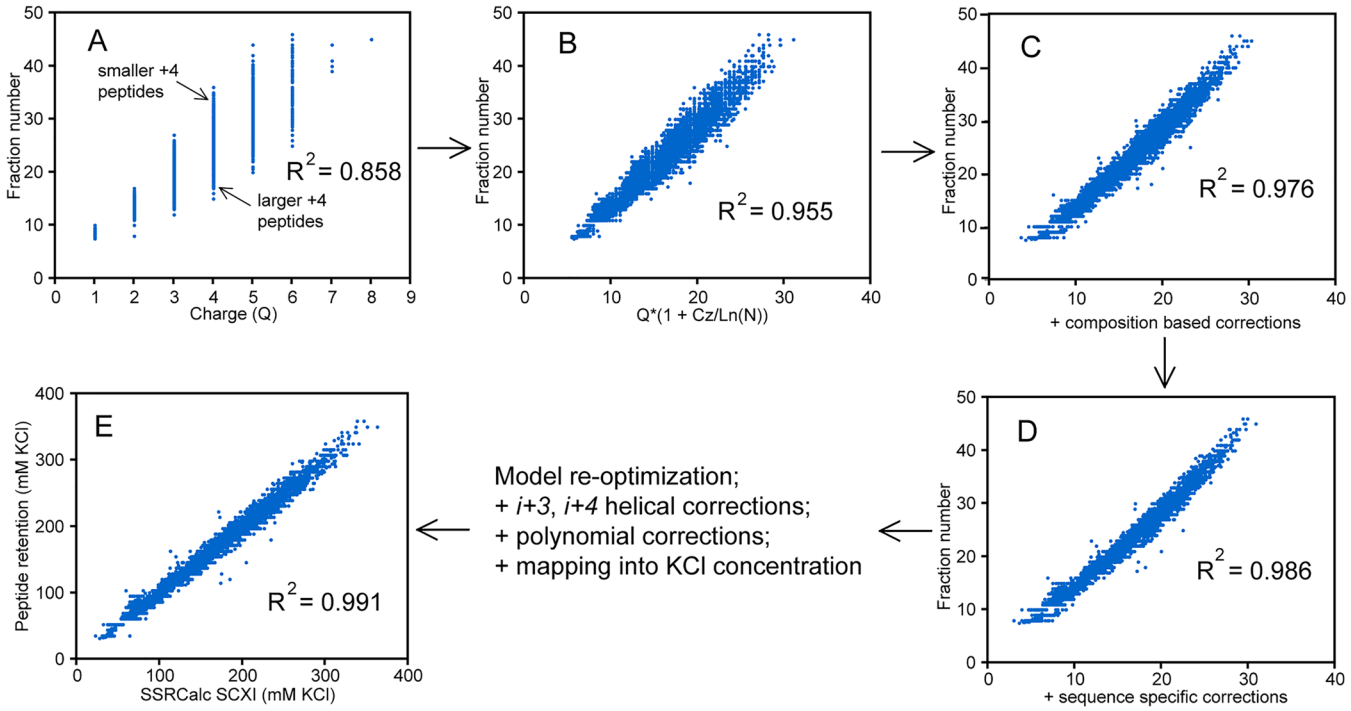


Figure 2. Step-by-step optimization of SSRCalc SCX model. (A) Correlation between peptide charge and retention time. (B) Retention time vs $Q \times (1 + C_Z/\ln(N))$. (C) Correlation for additive SSRCalc SCX model, taking into account peptide composition. (D) Correlation after incorporation of position-dependent retention coefficients. (E) Final SSRCalc SCX model.

Table 3. Position-Dependent Retention Coefficients for Individual Residues

residue	N-terminal	$N + 1$	$N + 2$	internal	C-2	C-1	C-terminal[a]
R[b]	1.271	1.267	1.217	1.085	1.090	1.095	1.069
H[b]	1.192	1.199	1.162	1.038	1.043	0.980	0.921
K[b]	1.096	1.103	1.043	0.972	0.969	0.953	0.974
W	0.075	0.112	0.125	0.105	0.092	0.101	0.016
N	−0.008	0.004	0.027	0.036	0.033	0.037	0.085
Y	−0.037	0.000	0.019	0.028	0.014	0.018	0.097
G	−0.051	−0.027	0.022	0.028	0.019	0.019	0.134
C	−0.016	0.009	0.015	0.024	0.025	0.009	0.054
F	−0.051	−0.010	0.006	0.020	0.007	0.005	0.004
D[b]	−0.150	−0.043	−0.003	0.012	0.009	0.018	0.031
S	−0.053	−0.031	0.000	0.011	0.007	0.000	0.089
E	−0.081	−0.054	−0.025	0.008	−0.003	0.001	0.041
Q	−0.066	−0.036	−0.018	0.002	−0.013	−0.009	0.078
M	−0.076	−0.055	−0.035	−0.007	−0.033	−0.023	−0.056
A	−0.106	−0.063	−0.032	−0.010	−0.024	−0.022	0.042
T	−0.089	−0.069	−0.037	−0.018	−0.024	−0.019	0.033
L	−0.136	−0.088	−0.058	−0.032	−0.053	−0.040	0.009
I	−0.121	−0.085	−0.068	−0.040	−0.054	−0.049	0.003
V	−0.136	−0.090	−0.060	−0.043	−0.055	−0.045	0.034
P	−0.124	−0.068	−0.062	−0.054	−0.057	−0.056	0.049

[a]C-terminal retention coefficients have been assigned with lower confidence due to a low number of peptides, which are not terminated by Lys or Arg. [b]Residues showing the largest effect of position relative to N-termini.

highest near the N-terminus. This can also be attributed to the orientation of the peptide as suggested by Alpert et al.[31] Since positively charged N-termini serve as the primary point of interaction with the stationary phase, N-terminal location of basic residues leads to decreased distance between its side chain and the sorbent, thus increasing the Coulombic interactions. While the basic residues increase retention near N-termini, the acidic amino acids decrease it. In the CZE model,[15] the effect of Asp on the N-terminus is the greatest relative to all other amino acids: its N-terminal position lowers peptide charge by ∼0.27 units. In SCX, the retention coefficient of Asp shows a decrease of ∼0.15 units in the N-terminal position as shown in Table 3. This can be attributed to the decrease of basicity of the N-terminus, decreasing its charge and therefore Coulombic interaction with the stationary phase. The positive contribution of Trp to peptide retention is independent of its position, except for a slightly lower R_C for the N-terminus. All hydrophobic residues exhibit lower retention coefficients

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

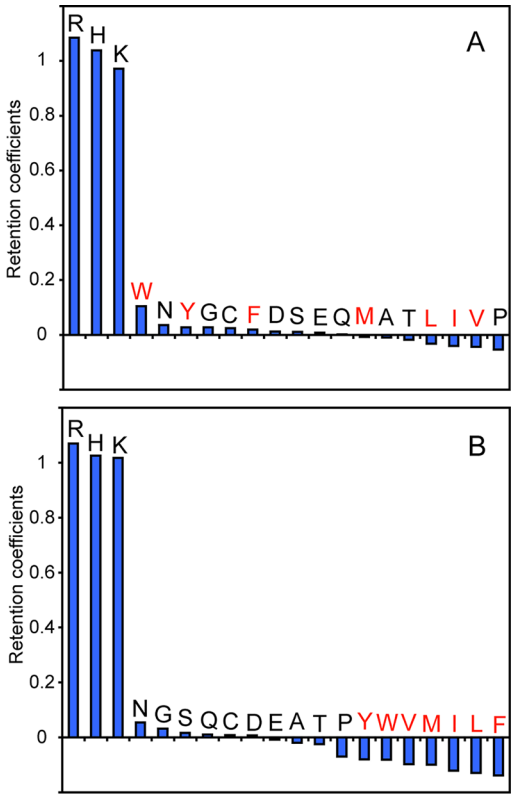


Figure 3. Retention coefficients in additive SCX models. (A) R_C values for our retention data (20% acetonitrile in eluents). (B) R_C values for reoptimized SSRCalc SCX using Trinidad et al.[28] data (30% acetonitrile).

when located at peptide N-termini, the primary point of contact with the more hydrophilic stationary phase.

The application of position-dependent corrections further improved accuracy of the model from 0.976 to 0.9862 R^2 value (Figure 2D). The addition of the six terminal position-dependent coefficients for 20 different amino acids adds 120 variables to our model, for 148 variables in total (Table 2). To avoid overfitting, the dataset is required to be significantly larger compared to the number of variables. In our case, this criterion is fulfilled: our modeling dataset contains over 30 000 peptides, which provides enough data to accurately model the retention coefficients. For example, our *least abundant* amino acids (Cys, Trp, Met) are present in the N-terminal position 281, 375, and 442 times, respectively.

Helical Interactions of Basic Residues. After multiple individual layer implementations, we reoptimized the model involving all variables and layers of the model simultaneously. This slightly increased the R^2 value to 0.9868. Next, we incorporated a correction related to possible $i + 3$ and $i + 4$ interactions between basic residues (Table S-2), which resulted in minor improvement of the R^2 value to 0.987. Mant et al.[32] showed that the presence of four positively charged Lys in the hydrophilic face of synthetic amphipathic helical peptides significantly increase retention time in SCX. Our attempted correction of this effect did not provide significant improvement. We explain this by the relatively low number of internal Arg and Lys residues found in tryptic peptides. Most of them occupy C-terminal positions, while internal ones are often followed by a Pro residue, known as a "helix-breaker". All of these factors contribute to the smaller effect of amphipathic helicity and helical corrections when incorporated into our prediction model.

Empirical Corrections for Nonlinearity. After implementation of all optimization steps, a slight nonlinearity in charge-specific subsets of peptides was still visible. For example, correlation plots for groups of peptides with different charges had slightly different slopes, and some of them showed convex character (Figure S-1). To correct this, we devised a simple Monte Carlo method to adjust the final predicted value for each peptide depending on charge Q as a polynomial: $A \times (\text{model output})^2 + B \times (\text{model output}) + C$. The combination of variables A, B, and C is specific for each charge group (Table S-2). These values were determined with the dual optimization goal of both improving the overall correlation while not significantly perturbing the overall model's slope and intercept. Our final model after all optimization steps showed an R^2 value of 0.991.

Expression of Peptide Fraction Elution in SCX Separation. All of the steps of the model optimization utilized a unitless expression of predicted SCX retention as shown in Figure 2B−D. For practical purposes, expressing SCX retention using an eluent parameter allows one to better visualize and describe SCX separation. RP-HPLC[10] and HILIC[18] models often use acetonitrile percentage for peptide hydrophobicity and water percentage for peptide hydrophilicity degree, respectively. Similarly, we propose to use the

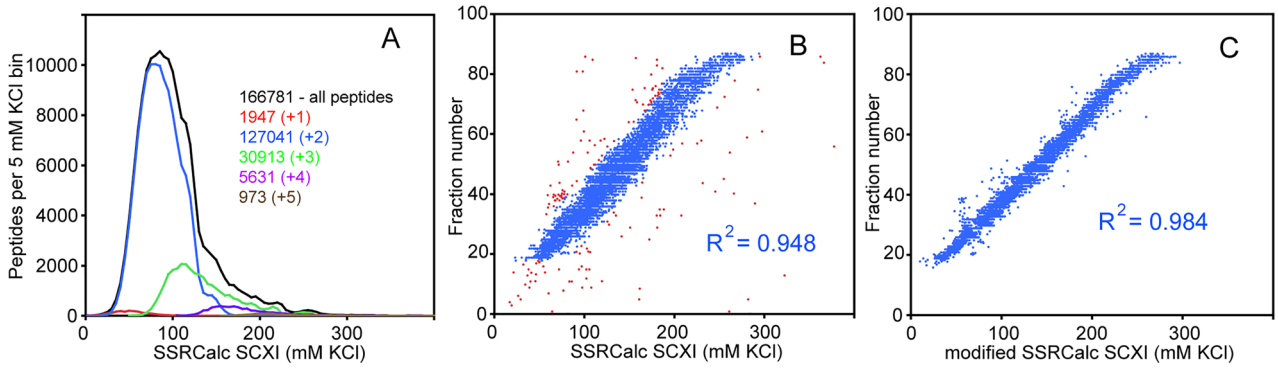


Figure 4. Applications of the SSRCalc SCX model. (A) Theoretical distribution of 166 781 peptides from in silico digest of *S. cerevisiae* (>4 residues, no missed cleavages) across the SCX separation scale. (B, C) Application of nonmodified and reoptimized SSRCalc SCX model to the data from Trinidad et al.,[28] respectively. Potential false positive IDs shown in red were excluded (B); 8135 peptides in blue were used for the model adjustment.

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

concentration of KCl (mM) required for elution of a particular peptide from the SCX column as a measure of Strong Cation Exchange Retention Index (SCXI). Retention times (or fraction numbers) were converted into concentrations of KCl using the known values of the gradient delay time of the LC system (3.3 min at 300 μL/min) and experimental gradient slope. The unitless output of the predictive model was then converted into SCXI units by introducing a mapping slope and intercept to provide a slope of 1 and intercept of 0 in final dependence: experimental retention (mM KCl) vs SSRCalc SCXI (mM KCl) is shown in Figure 2E.

Application of SCX Prediction Model. Having an accurate prediction model developed provides many applied options: the ability to predict the separation of individual peptides or groups of peptides, estimate orthogonality between various peptide separation techniques, and use peptide retention time as an additional filter in identification protocols. Figure 4A shows predicted distribution of peptides from in silico digested yeast proteome throughout SCX separation space. This graph closely resembles the experimental distribution of peptides within experimental chromatographic space, Figure 1B,C. Figure S-2 shows a comparison of experimental and in silico predicted orthogonality plots between SCX and RPLC separation dimensions for the whole collection of ~30 000 peptides. The high degree of similarity between these plots shows great potential of accurate SSRCalc models in estimating separation orthogonality between different peptide separation modes and guiding development of 2D LC-MS protocols.

Finally, we applied the SSRCalc SCX prediction to external datasets to estimate its applicability and gauge the degree of influence of other eluent parameters on separation selectivity. Webb et al.[34] have reported application of the MudPIT protocol to the yeast whole cell digest with a 39-step ammonium acetate salt gradient. Application of SSRCalc SCX to this data showed a very poor correlation (Figure S-3); this is most likely the consequence of using low (5%) acetonitrile in SCX buffer systems in MudPIT. We believe that the hydrophobic interactions with the SCX matrix as proposed by Burke et al.[24] were the greatest difference between the datasets in question. To verify this, we applied our prediction model to the data from Trinidad et al.[28] that was collected using 30% acetonitrile and observed a satisfactory R^2 value of ~0.95 after exclusion of some obvious SCX outliers (Figure 4B). Following removal of false positives, we reoptimized the model using 8135 peptides and obtained an R^2 value of 0.984 (Figure 4C). Adjustment of the model revealed another subset of possible SCX outliers and plateau at the end of the plot due to the rapid increase of KCl at the end of the gradient. Otherwise, prediction accuracy of these two models is comparable. The major difference between them consisted of a decreased contribution of hydrophobic aromatic residues (Figure 3B), in complete agreement with variation of organic solvent content used (20% vs 30%).

■ CONCLUSIONS

Our findings confirm the majority of the conclusions made in prior literature on the major factors driving peptide cation-exchange separation: the influence of peptide charge and length, and concentration of organic solvent. Compared to modeling studies in the 1980s and 1990s, we have access to a much larger collection of SCX retention data. We used it to explore the fine details of the separation mechanism: contribution of individual residues and sequence-specific features. The R^2 values for the additive model (0.976, Figure 2C) and final version of the SSRCalc SCX algorithm (0.991, Figure 2E) are significantly higher than the algorithms of similar complexity for either HILIC[18] or RP[10] but lower than for peptide CZE.[15] This indicates that the separation mechanism for SCX is simpler compared to other peptide LC separation techniques. Most of the residues contribute to the RP and HILIC retention mechanisms, whereas in SCX the basic ones dominate. This is true for SCX separations using eluents with high (20−30%) acetonitrile content, which suppresses hydrophobic interactions. Separations using a low concentration of organic solvent will likely exhibit features of mixed-mode (SCX/hydrophobic) interactions with subsequent complications in the modeling. At the same time, mixed-mode separations, as well as peptide SCX at different pHs, will produce significantly altered separation selectivity. Exploring these features in SCX should constitute a significant portion of future modeling studies and may result in the discovery of separation systems with unique selectivity and optimal orthogonality to RPLC.

We expect that our SSRCalc SCX model will be applicable to other similar separation systems under slightly different gradient slopes or acetonitrile content. Adjusting for the former will likely require optimization of polynomial corrections for each individual charge group since the change in salt concentration (gradient) impacts differently charged molecules in different ways. We have demonstrated that modeling variation in acetonitrile concentration requires the reoptimization of retention contributions of individual residues. The vast majority of the tryptic peptides carry 2 or 3 positively charged groups (Figure 4A) and elute in a very narrow range of salt concentration. Therefore, future modeling studies will undoubtedly explore SCX separations using segmented gradients, often applied to make the distribution of peptides across the fractions more uniform.[35,36]

The SSRCalc algorithm has been known for the superior accuracy of RPLC modeling since 2004.[19] In 2017, we expanded our research into peptide separations using CZE,[15] HILIC,[18] and SCX (present work). These examples represented our first experience with these separation techniques, but we achieved the highest prediction accuracy reported for all of them. This leads us to the conclusion that the general principles we use for model optimization have an advantage over other modeling approaches and should be *applicable to any peptide separation technique*. These principles include (1) working with extremely abundant proteomics-derived datasets to achieve at least a 100:1 ratio between number of data points and model variables; (2) acceptance of fraction-based (discrete) retention data from LC-MALDI MS[19] or first dimension of 2D LC-MS/MS[10,18] acquisitions (30−50 fractions); (3) application of peptide retention prediction filtering to improve quality of experimental data; (4) combining findings of high caliber previous studies of separation mechanisms with our own empirical observations; (5) considering peptide secondary structure (helicity), positioning of individual residues relative to peptide ends, peptide orientation relative to the surface, and nearest neighbor effects as the major drivers of sequence-dependent character in peptide separations. We believe that these principles, especially the sequence-specific corrections, should be automatically applied to the modeling of novel peptide separation techniques going forward.

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b03436.

Figure S-1, the accuracy of SSRCalc SCX prior to polynomial correction; Figure S-2, experimental and in silico generated orthogonality plots for SCX-RPLC of optimization dataset (30 482 peptides); Figure S-3, application of SSRCalc SCX to MudPIT-derived data; Table S-2, major parameters of SSRCalc SCX model (PDF)

Table S-1, optimization dataset used in this study (XLS)

■ AUTHOR INFORMATION

Corresponding Author

*Fax: (204) 480 1362. E-mail: oleg.krokhine@umanitoba.ca.

ORCID

Oleg V. Krokhin: 0000-0002-9989-6593

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05963; O.V.K.). The authors thank Dr. A. Alpert for providing the Polysulfoethyl A column and Dr. J. C. Trinidad for sharing the SCX retention data.[28] The authors also thank Dr. D. Court and S. Shuvo for providing *S. cerevisiae* samples.

■ REFERENCES

(1) Di Palma, S.; Hennrich, M. L.; Heck, A. J.; Mohammed, S. *J. Proteomics* **2012**, *75*, 3791−3813.

(2) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *Anal. Chem.* **2005**, *77*, 6426−6434.

(3) Takahashi, N.; Takahashi, Y.; Putnam, F. W. *J. Chromatogr.* **1983**, *266*, 511−522.

(4) Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd. *Anal. Chem.* **2001**, *73*, 5683−5690.

(5) Washburn, M. P.; Wolters, D. A.; Yates, J. R., 3rd. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(6) Tarasova, I. A.; Masselon, C. D.; Gorshkov, A. V.; Gorshkov, M. V. *Analyst* **2016**, *141*, 4816−4832.

(7) Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. *Mol. Syst. Biol.* **2008**, *4*, 222.

(8) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Smith, R. D. *J. Proteome Res.* **2004**, *3*, 760−769.

(9) Yen, C.-Y.; Russell, S.; Mendoza, A. M.; Meyer-Arendt, K.; Sun, S.; Cios, K. J.; Ahn, N. G.; Resing, K. A. *Anal. Chem.* **2006**, *78*, 1071−1084.

(10) Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2008**, *80*, 7036−7042.

(11) Spicer, V.; Ezzati, P.; Neustaeter, H.; Beavis, R. C.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2016**, *88*, 2847−2855.

(12) Krokhin, O. V. *Anal. Chem.* **2006**, *78*, 7785−7795.

(13) 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. *Anal. Chem.* **2006**, *78*, 5026−5039.

(14) Moruz, L.; Tomazela, D.; Kall, L. *J. Proteome Res.* **2010**, *9*, 5209−5216.

(15) Krokhin, O. V.; Anderson, G.; Spicer, V.; Sun, L.; Dovichi, N. J. *Anal. Chem.* **2017**, *89*, 2000−2008.

(16) Gilar, M.; Jaworski, A. *J. Chromatogr A* **2011**, *1218*, 8890−8896.

(17) Badgett, M. J.; Boyes, B.; Orlando, R. Prediction of Peptide Retention Times in Hydrophilic Interaction Liquid Chromatography (HILIC) Based on Amino Acid Composition. In *Chromatography Today*; International Labmate Limited: St. Albans, Hertfordshire, 2015; (Nov-Dec), pp 39−42.

(18) Krokhin, O. V.; Ezzati, P.; Spicer, V. *Anal. Chem.* **2017**, *89*, 5526−5533.

(19) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908−919.

(20) Krokhin, O. V.; Spicer, V. *Anal. Chem.* **2009**, *81*, 9522−9530.

(21) Isobe, T.; Takayasu, T.; Takai, N.; Okuyama, T. *Anal. Biochem.* **1982**, *122*, 417−425.

(22) Cachia, P. J.; Vaneyk, J.; Chong, P. C.; Taneja, A.; Hodges, R. S. *J. Chromatogr.* **1983**, *266*, 651−659.

(23) Alpert, A. J.; Andrews, P. C. *J. Chromatogr.* **1988**, *443*, 85−96.

(24) Burke, T. W.; Mant, C. T.; Black, J. A.; Hodges, R. S. *J. Chromatogr.* **1989**, *476*, 377−389.

(25) Kopaciewicz, W.; Rounds, M. A.; Fausnaugh, J.; Regnier, F. E. *J. Chromatogr.* **1983**, *266*, 3−21.

(26) Hodges, R. S.; Parker, J. M.; Mant, C. T.; Sharma, R. R. *J. Chromatogr.* **1988**, *458*, 147−167.

(27) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. *Anal. Chem.* **2004**, *76*, 3556−3568.

(28) Trinidad, J. C.; Specht, C. G.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. *Mol. Cell. Proteomics* **2006**, *5*, 914−922.

(29) Ballif, B. A.; Villén, J.; Beausoleil, S. A.; Schwartz, D.; Gygi, S. P. *Mol. Cell. Proteomics* **2004**, *3*, 1093−1101.

(30) Petritis, K.; Kangas, L. J.; Jaitly, N.; Monroe, M.; Lopez-Ferrer, D.; Maxwell, R. A.; Mayampurath, A. M.; Petritis, B. O.; Mottaz, H. M.; Lipton, M. S.; Camp, D. G.; Smith, R. D. Strong cation exchange LC peptide retention time prediction and its application in proteomics, Poster# WP 591, *56th ASMS Conference on Mass Spectrometry and Allied Topics*, Denver, CO, June, 2008.

(31) Alpert, A. J.; Petritis, K.; Kangas, L.; Smith, R. D.; Mechtler, K.; Mitulović, G.; Mohammed, S.; Heck, A. J. *Anal. Chem.* **2010**, *82*, 5253−5259.

(32) Mant, C. T.; Litowski, J. R.; Hodges, R. S. *J. Chromatogr A* **1998**, *816*, 65−78.

(33) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. *Nat. Methods* **2009**, *6*, 359−362.

(34) Webb, K. J.; Xu, T.; Park, S. K.; Yates, J. R., 3rd *J. Proteome Res.* **2013**, *12*, 2177−2184.

(35) Karsan, A.; Pollet, I.; Yu, L. R.; Chan, K. C.; Conrads, T. P.; Lucas, D. A.; Andersen, R.; Veenstra, T. *Mol. Cell. Proteomics* **2005**, *4*, 191−204.

(36) Jacobs, J. M.; Yang, X.; Luft, B. J.; Dunn, J. J.; Camp, D. G., 2nd; Smith, R. D. *Proteomics* **2005**, *5*, 1446−1453.

DOI: 10.1021/acs.analchem.7b03436
Anal. Chem. XXXX, XXX, XXX−XXX