Published in final edited form as:
Anal Chem. 2017 February 07; 89(3): 2000–2008. doi:10.1021/acs.analchem.6b04544.

Predicting electrophoretic mobility of tryptic peptides for high-throughput CZE-MS analysis

Oleg V. Krokhin[*,1,2], **Geoff Anderson**[1], **Vic Spicer**[1], **Liangliang Sun**[3], and **Norman J. Dovichi**[4]

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

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

[3]Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA

[4]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, 46556, USA

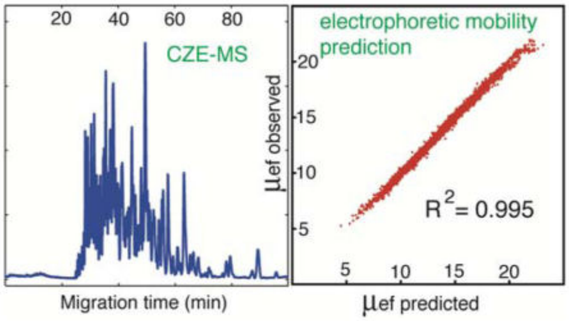
Abstract

A multi-parametric sequence-specific model for predicting peptide electrophoretic mobility has been developed using large-scale bottom-up proteomic CE-MS data (5% acetic acid as background electrolyte). Peptide charge (*Z*) and size (molecular weight, *M*) are the two major factors determining electrophoretic mobility-- in complete agreement with previous studies. The extended size of the dataset (>4000 peptides) permits access to many sequence-specific factors that impact peptide mobility. The presence of acidic residues Asp and Glu near the peptide N-terminus is by far the most the prominent among them. The induction effect of the side chain of N-terminal Asp reduces the basicity of the N-terminal amino group- and as hence its charge- by ~0.27 units, lowering mobility. The correlation of the model (R^2~0.995) indicates that the peptide separation process in CZE is relatively simple and can be predicted to a much higher precision than current RP-HPLC models. Similar to RP-HPLC prediction studies, we anticipate future studies that introduce peptide migration standards, collect larger datasets for modeling through the alignment of multiple CZE-MS acquisitions, and study of the behaviour of peptides carrying post-translational modifications. The increased size of datasets will also permit investigation of the fine-scale effects of peptide secondary structure on peptide mobility. We observed that peptides with higher helical propensity tend to have higher than predicted electrophoretic mobility; the incorporation of these features into CZE migration models will require significantly larger data sets.

Graphical Abstract

[*]corresponding author: Fax: (204) 480 1362, oleg.krokhine@umanitoba.ca.

The authors declare no competing financial interest.



INTRODUCTION

Accelerated developments in mass spectrometry have revolutionized the field of bioanalytical chemistry and protein analysis [1]. High-speed (MS/MS) peptide analysers coupled to various peptide separation devices perform the vast majority of proteomic studies. Different variants of HPLC (predominantly reversed phase but also HILIC and ion-exchange) and capillary zone electrophoresis (CZE) represent the most popular choices of separation techniques; all have been tailored to provide the uniform, highly efficient delivery of separated compounds into the mass spectrometer. It is interesting to note that to a great extent, retention/migration times no longer serve their original purpose of qualitative analysis (identification) of sample components, as the resolving power and speed of MS/MS acquisition in modern mass spectrometers makes the identification of multiple co-eluting components a routine task. However retention/migration times have found their applications in improving the confidence of MS and MS/MS identification by predicting physico-chemical properties of peptides [2, 3], and for guiding quantitative LC-MS procedures such as SRM [4] and SWATH [5].

RP-HPLC, which is the dominant peptide separation technique, has received most of the attention from separation scientists: new retention prediction models were developed using large retention data sets, spanning from a few thousand to hundreds of thousands of peptides [6–9]. These huge datasets provided an enormous advantage over the models developed in 1980s–90s: most of the early models were based on 20–30 peptides for model optimization [10, 11], leaving undiscovered the multitude of sequence-specific factors involved in complex separation mechanisms, based on both hydrophobic and ion-pairing interactions. Early attempts to increase these dataset led to significant disagreement against previously published data (260 synthetic peptides [12]) or inconsistent results due to the uncontrollable oxidation of Met and Trp residues (1738 peptides [13]). New features involved in peptide separation mechanisms were discovered as soon as researchers got access to high quality retention data for a few hundred or more peptides [6, 14–16].

In its own right, prediction of electrophoretic mobility of peptides was a subject of intense studies and debates in late 1990s to the early 2000s [17]. Most of these approaches were based on application of semi-empirical models similar to Offords' [18] equation: $\mu_{ef} = k(Q/M^{2/3})$, where Q and M are peptide charge and mass respectively [19–23]. Several attempts have been made to introduce multi-variable models, which used additional descriptors beyond peptide charge and mass to predict peptide mobility: average mass and charge distribution across peptide sequence [24], peptide width [25], steric interaction effects and molar refractivity [26], and the relative number of oxygen (RNO) and nitrogen (RNN) atoms [27]. Some of these approaches were supplemented by using modern machine learning techniques such as Analytical Neural Networks (ANN) [26, 27]. However, prediction accuracy of these models rarely exceeded 0.98 R^2-value. The only exception from this was a few observations of Kim et al. [21]. The authors reported ~0.99 correlations using simple $\mu_{ef} = k(Q/M^{0.56})$ model applied to sets of 20–30 relatively short proteolytic peptides.

Compared to the development of RP-HPLC prediction models two decades earlier [10–11], many researchers explicitly targeted proteomic techniques as their major application area. For example, Katayama et al. [3] used peptide charge, calculated based on electrophoretic mobility and mass, as an additional parameter in protein identification by peptide mass fingerprinting. Kim et al. [21] observed deviation from the predicted mobility for several peptides carrying post-translational modifications. By 2008, when mobility modeling in CZE was extensively reviewed by Mittermayr et al. [17], the gap in size of optimization datasets between CE and HPLC-MS applications was significant. Petritis et al. [7] reported ANN-based approach using 300,000 peptides identified in RP-HPLC-MS. Krokhin [6] used ~6,000 peptides retention datasets collected under different RP-HPLC conditions. Conversely, the datasets used for developing CE migration models rarely exceeded 100 entries [17].

Recent improvements in coupling CZE to MS/MS aimed to bridge the gap between CE and RP-HPLC in combination with tandem mass spectrometry analysis [28–31]. Sun et al. [28] reported the identification of over 10,000 peptides and 2,100 proteins in a single pass analysis of tryptic digest of HeLa cells – almost an order of magnitude improvement from previously reported studies. This was achieved through extending the CZE separation window, along with improvements in sample loading, CZE-MS coupling techniques and mass spectrometer. We have no doubt that the rapid improvement in CZE-MS performance for large-scale peptide identifications will provide significant impact on further advancement in peptide electrophoretic mobility modeling, similar to RP-HPLC in early 2000s.

Our objective is the development of a peptide electrophoretic mobility model, which is based on an extended dataset containing migration data for thousands of peptides. Similar to the successful development of Sequence-Specific Retention Calculator for RP-HPLC [6, 14–16], we based our model on the best approaches reported before, but we have made an attempt to supplement the basic model with an empirical evaluation of sequence-specific factors, which may play a role in altering apparent electrophoretic mobility of peptides. These factors may have been missed in previous prediction studies due to the limited size of their experimental datasets.

EXPERIMENTAL

We used experimental data and identification outputs from Proteome Discoverer 1.4 [28] from CE-MS analysis of tryptic digests from *E. coli*, *S. cerevisiae*, and HeLa whole cell lysates. Raw data from these analyses were converted into MGF (mascot generic format) files and analysed using X!Tandem search engine.

The following X!Tandem search parameters were used: up to one missed cleavage tryptic peptides was considered; constant modification Cys +57.021 Da (cysteine protection with iodoacetamide); variable modifications: Met, Trp +15.995 (oxidation) or +31.989 (double oxidation); Ser, Thr, Tyr +79.966 (phosphorylation), N, Q +0.984 (deamidation); parent mass error: +−20 ppm, fragment mass error: 0.5 Da. Peptide identification with expectation values Log_{10} (e) < −1 were permitted.

Migration times (t_M, min) were determined as time of MS/MS acquisition of the most intense tandem spectra for each unique peptide identification. We assumed that the electro-osmotic flow at 5% acetic acid in the background electrolyte is very low and mapped t_M into electrophoretic mobility (μ_{ef}) using the equation for their experimental conditions (a 90 cm long capillary at 280 volts/cm):

$$\mu_{\text{ef}}=90/(60*t_M*280)\,(\text{units of cm}^2*\text{V}^{-1}*\text{s}^{-1}).$$

Hydrophobicity indexes and helical content for peptides were calculates using on-line versions of SSRCalc [6] and AGADIR models [32], respectively.

RESULTS AND DISCUSSION

Generating high-quality optimization data set using migration data filtering

Table 1 presents the summary of X!Tandem identification outputs for the five CE-MS/MS runs. Note that X!Tandem identification results are somewhat different from those reported by Proteome Discoverer 1.4 [28] due to differences in the peptide identification algorithms. Having an optimization dataset free of false positive identifications and miss-assigned migration/retention time values is critical for correct model optimization. Our own peptide retention prediction project had started with RP-HPLC – MALDI MS analysis of tryptic digest of 17 purified proteins with known sequences [14]. Peptide mapping with high mass accuracy (10 PPM) supplemented by MS/MS confirmation yielded a retention data set of extremely high quality. Later, we switched to high-throughput RP-HPLC – ESI MS/MS approach, but based peptide identification on both MS/MS analysis and peptide retention prediction filtering – often in two LC dimensions [15].

We began this project by using the familiar X!Tandem search engine, which usually yields a false discovery rate ~0.3–0.4% for the datasets of similar complexity acquired with mass spectrometers of comparable characteristics. Upon subsequent improvement of our model, we applied migration prediction filtering, but never excluded peptides identified with high confidence (Log(e) < −3). The identification outputs from Proteome Discoverer represented an additional source of information, which was used to confirm questionable identifications/ migration time assignments. The following discussion describes model development using X!Tandem identifications of 4463 tryptic peptides in first CE-MS/MS replicate of yeast digest. Peptides were 6–51 residues in length (17 on average), carrying 1–5 positively charged residues. Identifications were filtered through mobility prediction and the cross correlation with results reported from Proteome Discoverer. The remaining datasets were used as prediction algorithm test sets.

Application of the classical semi-empirical models showed different accuracy for the dataset under consideration (Table 2). We concluded that the Cifuentes and Poppe [23] model provided the best correlation of R^2~0.974 (Figure 1a). Molecular weight (*M*) and charge (*Z*) equal to the number of positively charged residues at pH of 5% acetic acid in the background electrolyte (Arg, Lys, His and N-terminus) were used as parameters in this model. We used this simplified version of charge calculation due to the well-recognized concern about the accuracy of estimation of *Z* value based on known p*Ka* values of ionizable functional groups [17]. There is no doubt that these values are affected by steric and electrostatic interaction, leading to discrepancies in charge calculation and uncontrollable deviations from the model. Our rationale was that leaving the charge calculation in its most fundamental form would hopefully allow establishing the most important sequence-related factors affecting p*Ka* values and peptide charge.

Model optimization

Finding sequence specific factors, which affect peptide physico-chemical properties (electrophoretic mobility in our case), is a key process applied in our modeling approach [6, 14, 16]. It is based on repetitive application of following steps:

i. the visual inspection of peptides with the most significant deviations from current version of prediction model (both positive and negative) and finding a common features between these peptides;

ii. the introduction of corrections to the model to improve prediction accuracy for these outlier peptides;

iii. the modified version of the model with improved prediction accuracy (R^2-value) is applied in the next cycle of optimization in pp. (i).

Table 3 and Figure 1 shows the three cycles of this optimization process.

Cycle 1 of optimization started with a visual inspection of the prediction errors from the Cifuentes and Poppe [23] model shown in Figure 1a, and revealed following features:

1. Very large negative deviations from predicted mobility (peptide migrate slower than expected) were characteristic for peptides carrying acidic amino acids Asp (predominantly) and Glu in N-terminal positions;
2. Positive deviations were observed for peptides with relatively high content of Asn and possibly Gln;
3. Negative deviations – for peptides carrying multiple internal Asp and Glu.

These findings are not surprizing considering the possible decrease/increase of peptide basicity because of presence of acidic/basic residues. This induction effect is particularly profound for N-terminal amino group of Asp, due to presence of a carboxy group in close proximity of the N-terminal amino group. A three-fold larger induction effect of Asp compared to Glu is observed due to the shorter (by one methylene group, Figure 1b) side chain.

One of the multi-variable models developed in the past used the relative number of oxygen (RNO) and nitrogen (RNN) to supplement Offord's model and found it applicable for one of the datasets under investigation [27]. Our findings on the effect of Asp and Glu confirm this observation, but also show that there is a significant difference between carboxy and hydroxy groups in this regard. Thus, the oxygen in hydroxy groups of N-terminal Ser and Thr is located closer to a terminal amino group, but showed no effect on peptide mobility. Based on this we introduced the notion of corrected charge *Zc*:

$$Zc = Z + \sum \left({C_{\text{res}}}^{\text{pos}} \right);$$

Where ${C_{\text{res}}}^{\text{pos}}$ represent correction coefficients for Asp, Glu, Asn, Gln in a position dependent manner: each residue had separate coefficient value for positions 1, 2, 3, a general "internal" position, and well as N-2 and N-1 (since most of analytes had N-terminal Arg or Lys). Figure 1b shows optimized values for these correction coefficients. An N-terminal Asp residue reduces the apparent charge value by ~0.27, with diminished effect for the internal positions. The contribution of Glu is also negative but is three times lower in magnitude. Asn shows a small positive, position-independent effect. Therefore, 0.016 average corrections were applied for all Asn regardless of their position (Figure 1b). Gln show no effect, so these corrections were removed from the model.

Following the same logic, we then explored possible influence of basic residues Arg, Lys, His on basicity of N-terminal amino group and found no effect. The next step tested a possible charge correction for a "nearest neighbour" effect that Asp and Glu might have on internal basic residues (Arg, Lys, His). We found small a decrease in the apparent charge of basic residues for this situations: −0.026 and −0.007 for Asp and Glu, respectively. This effect is 10-times smaller compared to the N-terminal amino group and also tracks the ~3-fold difference between Asp and Glu. Introduction of these corrections to peptide charge significantly improved model accuracy: from 0.974 to 0.991 (Figure 1c). All together charge correction had 13 residue and position dependent coefficients: 5 for Asp and Glu each, one for Asn and two for nearest neighbour effect of Asp and Glu on internal basic residues.

In cycle 2 of optimization we introduced a correction related to peptide mass associated with the average size of the residues in a particular peptide. We observed that peptides composed of heavier residues tend to have positive prediction errors (migrating faster than predicted) and vice versa – peptides composed of light residues migrate slower than predicted. Indeed - two peptides of the same mass and charge may have different number of residues (peptide length, *N*), which will result in higher mobility of sorter peptide (composed of heavier residues).

We introduced a corrected mass value as:

$$Mc = (0.66 * M + 0.34 * N * 110.9);$$

where N is the peptide length, 110.9 the mass of average residue in our dataset, and 0.66 and 0.34 are empirically optimized coefficients.

The larger contribution of M (2/3 of calculated value) compared to peptide length N confirmed the common features of previous models, which usually use molecular weight as one of the descriptors. ($0.34*N*110.9$) term was introduced to correct for the difference in average size of the residues in a particular peptide and further improved observed correlation up to 0.992 (Figure 1d).

Cycle 3 of the optimization followed from a visual inspection of correlation in Figure 1d showing significant deviation for slowest (positive) and fastest (negative errors) peptides. We believe that these two extreme cases are different in nature, but can be corrected using the same polynomial function with (Zc/N) ratio as its argument.

Singly charged peptides should exhibit different geometry while migrating towards cathode. Because they have only one charged residue they should assume orientation parallel to capillary axis, which provides smaller frictional force and higher mobility. We found that the vast majority of tryptic peptides from protein C-termini (those having only N-terminal charged group) migrate faster than predicted. The symmetry of distribution of charged residues was considered as one of the parameters affecting peptide mobility [24]. Figure 1e shows the correlation between prediction error $\Delta\mu = \mu_{\text{ef observed}} - \mu_{\text{ef predicted}}$ and the Zc/N ratio. Singly charged peptides with smallest Zc/N showed large positive prediction errors. These are the heaviest singly charged peptides, which mobility benefits the most from this (axial) orientation of the molecule.

Peptides with large ratios ($Zc/N > 0.4$) are relatively short in length but with a large number of basic functional groups. These molecules have extremely high charge density, which lead to excessive peptide hydration (relative to its size) and a corresponding decrease in observed mobility. For example, YHLEHHYK and QQEQYGNSNFGGAPQGGHNNHHR have the same number of closely spaced internal His residues, but differ by ~3-fold in length. The former showed extremely large negative prediction error (−2.24) in Figure 1d and the latter – small positive (0.24). We believe that the difference in (Zc/N) values for these peptides--0.60 vs. 0.21, explains this behaviour.

We used a fifth-order polynomial correction function in Figure 1e to calculate offset for predicted mobility value. Resulting gains were impressive with a final model fit of R^2-value 0.995 (Figure 1f). It should be noted that when done correctly, this polynomial offset function should be introduced as a peptide size correction to Mc. However, at this point we do not sufficient theoretical or experimental background to predict how peptide orientation and hydration status will impact the peptide size.

Final version of the model:

$$\mu_{\text{ef}}=3.069+386*(\ln(1+0.35*Zc)/Mc^{0.411})+\text{OFFSET}(Zc/N);$$

where 3.069 and 386 are coefficients applied to align output with experimentally measured values (slope 1 and intercept 0 in Figure 1f). Zc was calculated as a number of positively charges groups at pH 2.4, corrected using 13 residue and sequence specific coefficients. $Mc = (0.66{*}M + 0.34{*}N{*}110.9)$, where M and N are peptide mass and length, respectively; 0.66 and 0.34 are empirically optimized coefficients. OFFSET is a function of Zc/N: OFFSET = −783*x5 + 1380*x4 − 902*x3 + 256*x2 − 29.7*x +1.07 from Figure 1e.

Please note, the number of empirical correction coefficients is significantly lower than the number of peptides in the data set, which is a key condition to prevent over-fitting of the model. Multi-parametric models benefit from repetitive optimizations of the model parameters. Each cycle of optimization procedure will affect optimal parameter's values from previous cycles. In our case we did not attempt additional optimization cycles: further improvements in model accuracy would be marginal and possibly dataset specific. However repetitive optimization would be needed when similar models will be developed for larger data sets using larges number of composition and sequence-specific variables.

Peptide hydrophobicity and secondary structure have been proposed as some of the parameters affecting accuracy of mobility prediction [17]. Cross-correlations between prediction errors in Figure 1f with calculated peptide hydrophobicity (SSRCalc [6]) and propensity to form helical structures (AGADIR [32]) are shown in Figure 2. There is no correlation between prediction error and peptide hydrophobicity (Figure 2a). At the same time, peptides with high helical propensity tend to have higher than predicted mobility. This is not surprizing considering the more compact structure of peptides in a helical conformation. However, a comprehensive evaluation of impact of peptide helicity on electrophoretic mobility will need a significantly larger dataset. For example, we needed a ~300,000-peptides retention dataset to establish rules for N-capping helix stabilization in RP-HPLC [16] from a sub-population of ~ 5,000 amphipathic helical peptides.

Model testing using additional data

Examining correlations between observed and predicted values ($\mu_{ef\ obs}$ vs. $\mu_{ef\ pred}$) for test data provides real estimation of model's accuracy. Figure 3a–c shows application of final model to outputs of three additional CE-MS/MS runs from Table 1. All non-modified tryptic peptides with Log (e) < −1 are shown. Clearly, the accuracy of prediction remains the same (counting the impact of obvious false-positive ID outliers), except for the HeLa digest. The latter shows a systematic deviation for peptides with low mobility. This feature could be the result of a sudden change in separation conditions observed in this run. Similar to nano-RP-HPLC, such variations could be corrected by using a set of standard synthetic peptides that cover a wide range of electrophoretic mobility. Application of such peptide mixtures was introduced in proteomic RP-HPLC by Krokhin & Spicer [33] and reagents have been provided by several LC-MS vendors [34].

Figure 3d shows CE mobility prediction for the whole set of identified peptides in yeast digest, including PTM carrying species. We find that most PTMs lead to reducing mobility of peptides. There are distinct trend lines in the plots shown in Figure 3d, each representing a different group of peptides. The most dramatic shifts are observed for N-terminal cyclization (N-terminal Gln and Cys alkylated with iodoacetamide [35]) and N-terminally acetylated peptides. These modifications eradicate the N-terminal amino group and reduce the *Z* value by 1. There are at least two distinct trend lines corresponding to these modifications: the lower group contains peptides with no internal basic residues (charge reduction from 2 to 1). The upper line represents species with one internal charge (charge reduction from 3 to 2). Significant negative migration shifts are expected for phosphorylated peptides due to extremely acidic character of phospho- group [21]. Asparagine deamidation is another PTM, which will lead to formation of slower migrating (at acidic pHs) species. Deamidation of Asn yields two different products: Asp and β-Asp residues. We expect that the latter will possess lower electrophoretic mobility because the carboxy group of β-Asp is located closer to the peptide backbone than for Asp. Met oxidations represent another PTM contributing to migration shifts observed in Figure 3d. Our preliminary assessment indicates that oxidation leads to decreased electrophoretic mobility as well. Recent contribution by Barroso et al. [36] showed a significant decrease of peptide mobility due to glycosylation. Taken all together, studying the details of PTMs impact on electrophoretic mobility will be another exciting addition to CE prediction modeling – similar to RP-HPLC studies [37, 38].

CE vs. RP-HPLC: pros and cons

Having the capability to predict both peptide's hydrophobicity and electrophoretic mobility allowed us to compare CE and RP-HPLC from the point of view of the uniformity of the peptide distribution in separation space – a key feature in determining overall output of the analysis. Figure 4a, b show theoretical distribution of these parameters for *in-silico* digestion of whole yeast proteome with one allowed tryptic missed cleavage (more than 4 amino acids peptide length). Being transformed into migration/retention times for CE conditions used in this paper and typical LC-MS acquisition with 0.5% acetonitrile per minute gradient and trap column injection (0.1% formic acid ion-pairing modifier), it provides a realistic distribution of tryptic peptides from complex mixtures in typical runs (Figure 4c, d). The first finding indicates the clear advantage of CZE: peptides with SSRCalc HI < 0 will not retain on a C18 trap column and will be lost during sample loading. But CZE has access to a whole population of peptides, compared to RP-HPLC. This clearly suggests that combination of RP-HPLC and CZE will provide more comprehensive peptide identifications from complex proteome samples, thus leading to better protein sequence coverage and better characterization of protein isoforms. Note, that hydrophilic peptides will have a chance to be detected in RP-HPLC-MS if direct injection is used, but will elute along with all other hydrophilic species in the beginning of the chromatogram.

The obvious advantage of RP-HPLC is a more uniform distribution of peptides throughout the separation space. CZE separation will deliver close to 50,000 unique tryptic peptides (*in-silico* yeast digest) in a 2-minute wide bin of separation space at the densest portion of electropherogram (Figure 4c). RP-HPLC has maximum of ~18,000 peptides per 2 minute wide separation bin. Another operational advantage of RP-HPLC is the relative ease of manipulation of elution speed. Gradient slope can be adjusted (increased) to make overall elution speed (peptides/time bin) more uniform at the end of chromatogram. This feature is widely used by proteomic practitioners to minimize MS instrument time without significant impact on ID output.

The differences outlined here will dictate the choice of separation technique depending on the sample complexity and purpose of the analysis. CZE is preferable for relatively simple samples, when complete coverage for peptide mapping and fast analysis are needed. RP-HPLC will be more useful for extremely complex samples targeting protein identification. But we also need to note that RP-HPLC is prone to significant sample loss during sample loading due to use of injection valves and suffers from significant peak diffusion due to larger dead volume of the entire separation system. CZE has much lower sample loss due to the direct sample injection from sample vial to capillary and its open-tubular design. Accordingly, CZE has advantages for analysis of mass-limited samples [39].

Separation scientists, who are heavily invested in studying separation mechanisms of peptides, will find many striking differences and analogies between CZE and RP-HPLC modeling. Our data indicates a dramatic gap between the classical additive models for RP-HPLC [10, 11] and semi-empirical models for CZE (R^2~0.974) shown here. The simplest 20-parameter additive models with peptide length correction produce at best a R^2~0.93 correlation for the common RP-HPLC separation proteomic data set of tryptic peptides [40]. This performance indicates a significant and fundamental difference in complexity of separation mechanisms between these two techniques. RP-HPLC retention is affected by hydrophobic, ion-pairing, and hydrogen-bond interactions. It occurs on hydrophobic surface in the environment with constantly changing proportion of organic solvent to water. Modeling these interactions presents an extremely complicated task. Application of the most advanced models available today will rarely produce correlations above R^2 0.965 for a typical set of tryptic peptides [41]. In this study we showed that R^2 0.99+ correlations for predicting peptide electrophoretic mobility could be achieved. Electrophoretic mobility depends on the charge of the particle and its Stoke’s radius. The former is affected by p*Ka* values of all ionisable groups in a peptide, and the pH of the CZE background electrolyte. Peptide size (Stoke’s radius) will also depend on the conformation of the peptide in solution determined through ion-pairing, hydrogen bonds and hydrophobic interactions. However, these effects in free solution combined with the constant polarity of the CZE background electrolyte are much more subtle.

The similarities are notable too, and are observed due to distinct chemical properties of peptides at acidic pHs. The first sequence-specific effect we found for RP-HPLC modeling included N-terminal amino group as well [14]. We showed that hydrophobic contribution of N-terminal residues into peptide retention is much smaller due to ion-pair formation and effective “shielding” of the hydrophobic residues by more hydrophilic counter-ions (formate, trifluoroacetate). In case of CE – this is a direct induction effect of acidic residues on the apparent basicity of N- terminal amino group. Correct prediction of peptide helicity is another key development, which will improve prediction modeling for both RP-HPLC [16] and CE.

The historical time differential (~10–15 years) between developing RP-HPC and CZE predictive tools – both from the point of view of modeling and instrument throughput – allows us to predict future progress/developments in CZE modeling studies. These will include:

i. The introduction and wider application of peptide migration standards, similar to HPLC peptide retention standards, which will facilitate addressing separation reproducibility issues. Both nano-LC and CZE suffer from this problem due to their extremely low flow/mobility rates, and the dependence of separation on the sample load. Peptide retention/migration standards, which cover entire separation range, represent the only viable solution of this problem.

ii. The accumulation of significantly larger datasets through aligning multiple CZE-MS/MS runs.

iii. The incorporation of PTMs into migration modeling, similar to RP-HPLC studies.

iv. The guided development of optimal 2D combinations of HPLC-CZE separation techniques using peptide retention/migration prediction algorithms.

We have developed the first sequence specific model to predict peptide electrophoretic mobility for an extended set of tryptic peptides (more than 4,000) identified by CZE-MS/MS. Despite a significant increase in dataset size and the diversity of analytes, our model accuracy has exceeded all previously reported approaches: an R^2-value of 0.995 vs. 0.98. We attribute this improvement to our ability to select major sequence-specific features that alter apparent peptide charge at acidic conditions. This approach was previously unavailable to researchers studying electrophoretic mobility of peptides due to the limited dataset sizes. The presence of acidic residues (Asp and Glu) near the peptide N-terminus is by far the major the factor causing deviation of experimental mobility values from values predicted via the classical approaches. The superior accuracy of our model suggests that it may be used for migration time filtering of CZE-MS/MS analysis outputs, and the *a-priori* selection of separation conditions for targeted analysis. Further developments in modeling CZE separation process will undoubtedly include use of significantly larger datasets, which will allow the elucidation of additional sequence-specific features affecting electrophoretic mobility.

Acknowledgments

This work was supported by grant from the Natural Sciences and Engineering Research Council of Canada (OVK) and the National Institutes of Health (R01GM096767 –NJD).

References

1. Bensimon A, Heck AJ, Aebersold R. Annu Rev Biochem. 2012; 81:379–405. [PubMed: 22439968]
2. Strittmatter EF, Kangas LJ, Petritis K, Mottaz HM, Anderson GA, Shen Y, Jacobs JM, Camp DG 2nd, Smith RD. J Proteome Res. 2004; 3:760–769. [PubMed: 15359729]
3. Katayama H, Ishihama Y, Oda Y, Asakawa N. Rapid Commun Mass Spectrom. 2004; 14:1167–1178.
4. Lange V, Picotti P, Domon B, Aebersold R. Mol Syst Biol. 2008; 4:222. [PubMed: 18854821]
5. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R. Mol Cell Proteomics. 2012; 11:O111016717.
6. Krokhin OV. Anal Chem. 2006; 78:7785–7795. [PubMed: 17105172]
7. Petritis K, Kangas LJ, Yan B, Monroe ME, Strittmatter EF, Qian WJ, Adkins JN, Moore RJ, Xu Y, Lipton MS, Camp DG 2nd, Smith RD. Anal Chem. 2006; 78:5026–5039. [PubMed: 16841926]

8. Shinoda K, Sugimoto M, Yachie N, Sugiyama N, Masuda T, Robert M, Soga T, Tomita M. J Proteome Res. 2006; 5:3312–3317. [PubMed: 17137332]

9. Moruz L, Tomazela D, Kall L. J Proteome Res. 2010; 9:5209–5216. [PubMed: 20735070]

10. Meek JL. Proc Natl Acad Sci USA. 1980; 77:1632–1636. [PubMed: 6929513]

11. Guo D, Mant CT, Taneja AK, Parker JMR, Hodges RS. J Chromatogr. 1986; 359:499–517.

12. Houghten RA, DeGraw ST. J Chromatogr. 1987; 386:223–228. [PubMed: 3558605]

13. Wilce MCJ, Aguilar MI, Hearn MTW. Anal Chem. 1995; 67:1210–1219.

14. Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA. Mol Cell Proteomics. 2004; 3:908–919. [PubMed: 15238601]

15. Dwivedi RC, Spicer V, Harder M, Antonovici M, Ens W, Standing KG, Wilkins JA, Krokhin OV. Anal Chem. 2008; 80:7036–7042. [PubMed: 18686972]

16. Spicer V, Lao YW, Shamshurin D, Ezzati P, Wilkins JA, Krokhin OV. Anal Chem. 2014; 86:11498–11502. [PubMed: 25372782]

17. Mittermayr S, Olajos M, Chovan T, Bonn GK, Guttman A. Trends in Analytical Chemistry. 2008; 27:407–417.

18. Offord RE. Nature (London). 1966; 211:591. [PubMed: 5968723]

19. Adamson NJ, Reynolds EC. J Chromatogr B. 1997; 699:133–147.

20. Tanford, C. Physical Chemistry of Macromolecules. Wiley; New York, USA: 1961.

21. Kim J, Zand R, Lubman DM. Electrophoresis. 2003; 24:782–793. [PubMed: 12627438]

22. Grossman PD, Colburn JC, Lauer HH. Anal Biochem. 1989; 179:28–33. [PubMed: 2757198]

23. Cifuentes A, Poppe H. J Chromatogr A. 1994; 680:321–340. [PubMed: 7952009]

24. Metral CJ, Janini GM, Muschik GM, Issaq HJ. J Sep Sci. 1999; 22:373–378.

25. Janini GM, Metral CJ, Issaq HJ. J Chromatogr A. 2001; 924:291–306. [PubMed: 11521876]

26. Jalali-Heravi M, Shen Y, Hassanisadi M, Khaledi MG. Electrophoresis. 2005; 26:1874–1885. [PubMed: 15825217]

27. Ma W, Luan F, Zhang H, Zhang X, Liu M, Hu Z, Fan B. Analyst. 2006; 131:1254–1260. [PubMed: 17066195]

28. Sun L, Hebert AS, Yan X, Zhao Y, Westphall MS, Rush MJP, Zhu G, Champion MM, Coon JJ, Dovichi NJ. Angew Chem Int Ed. 2014; 53:13931–13933.

29. Faserl K, Kremser L, Müller M, Teis D, Lindner HH. Anal Chem. 2015; 87:4633–4640. [PubMed: 25839223]

30. Yan X, Sun L, Zhu G, Cox OF, Dovichi NJ. Proteomics. 2016; doi: 10.1002/pmic.201600262

31. Sun L, Zhu G, Yan X, Zhang Z, Wojcik R, Champion MM, Dovichi NJ. Proteomics. 2016; 16:188–196. [PubMed: 26508368]

32. Lacroix E, Viguera AR, Serrano L. J Mol Biol. 1998; 284:173–191. [PubMed: 9811549]

33. Krokhin OV, Spicer V. Anal Chem. 2009; 81:9522–9530. [PubMed: 19848410]

34. Grigoryan M, Shamshurin D, Spicer V, Krokhin OV. Anal Chem. 2013; 85(22):10878–10886. [PubMed: 24127634]

35. Krokhin OV, Ens W, Standing KG. Rapid Commun Mass Spectrom. 2003; 17:2528–2534. [PubMed: 14608624]

36. Barroso A, Gimenez E, Benavente F, Barbosa J, Sanz-Nebot V. Anal Chim Acta. 2015; 854:169–177. [PubMed: 25479881]

37. Reimer J, Shamshurin D, Harder M, Yamchuk A, Spicer V, Krokhin OV. J Chromatogr A. 2011; 1218:5101–5107. [PubMed: 21665210]

38. Lao YW, Gungormusler-Yilmaz M, Shuvo S, Verbeke T, Spicer V, Krokhin OV. J Proteomics. 2015; 125:131–139. [PubMed: 26025879]

39. Wang Y, Fonslow BR, Wong CC, Nakorchevsky A, Yates JR 3rd. Anal Chem. 2012; 84:8505–8513. [PubMed: 23004022]

40. Shamshurin D, Spicer V, Krokhin OV. J Chromatogr A. 2011; 1218:6348–6355. [PubMed: 21798546]

41. Spicer V, Ezzati P, Neustaeter H, Beavis R, Wilkins JA, Krokhin OV. Anal Chem. 2016; 88:2847–2855. [PubMed: 26849966]

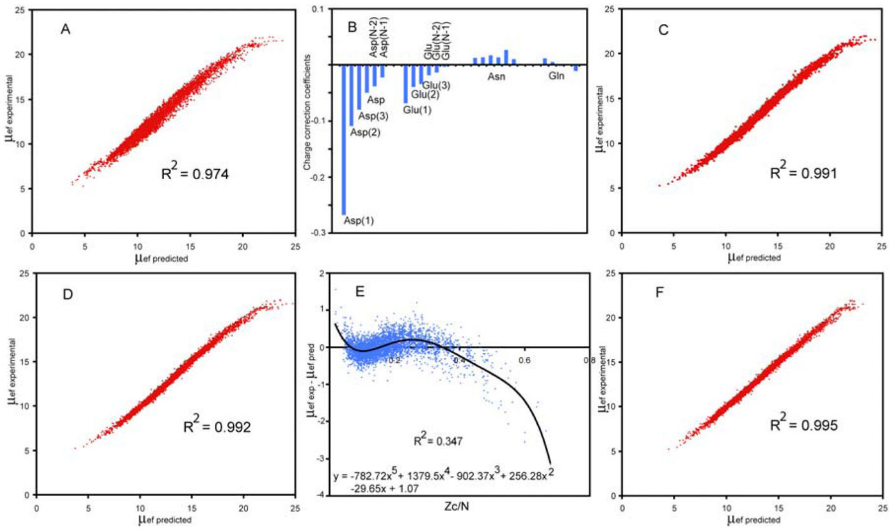


Figure 1.

Step-by-step optimization of mobility prediction model (see Table 3 for details). A) initial correlation using Cifuentes and Poppe [23] model; B) sequence specific charge correction coefficients: Asp(N1), Asp(N2), Asp(N3), Asp, Asp(N-2), Asp(N-1) correspond to correction coefficients of Asp in position 1, 2, 3, internal, third to last and second to last in peptide sequence, respectively; C) correlation after incorporating charge correction; D) correlation after incorporating mass correction; E) correlation between prediction error and *Zc/N* ratio approximated by polynomial function; F) final correlation after introducing polynomial offset function. Electrophoretic mobility μ_{ef} x 10^5 ($cm^2 * V^{-1} * s^{-1}$) is shown.

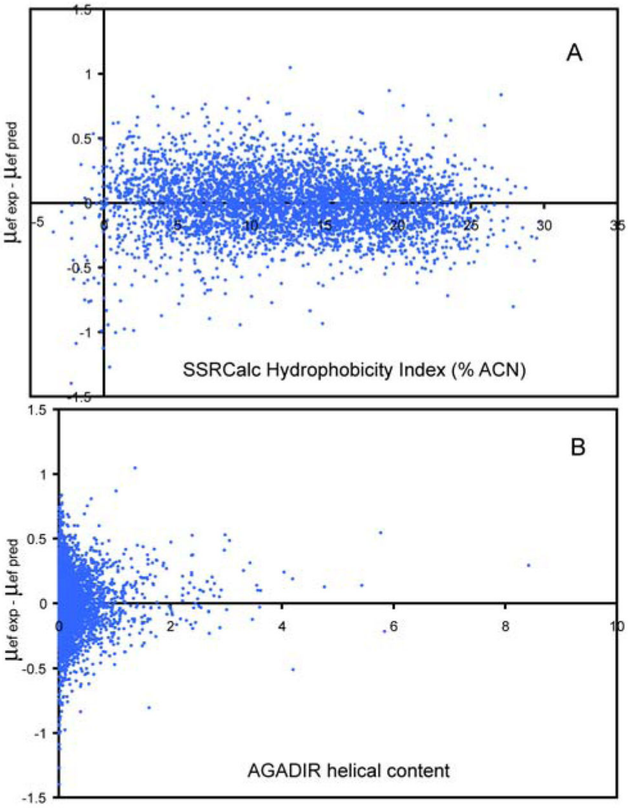


Figure 2.

Correlation of mobility prediction errors vs. peptide hydrophobicity (A) and helical content (B).

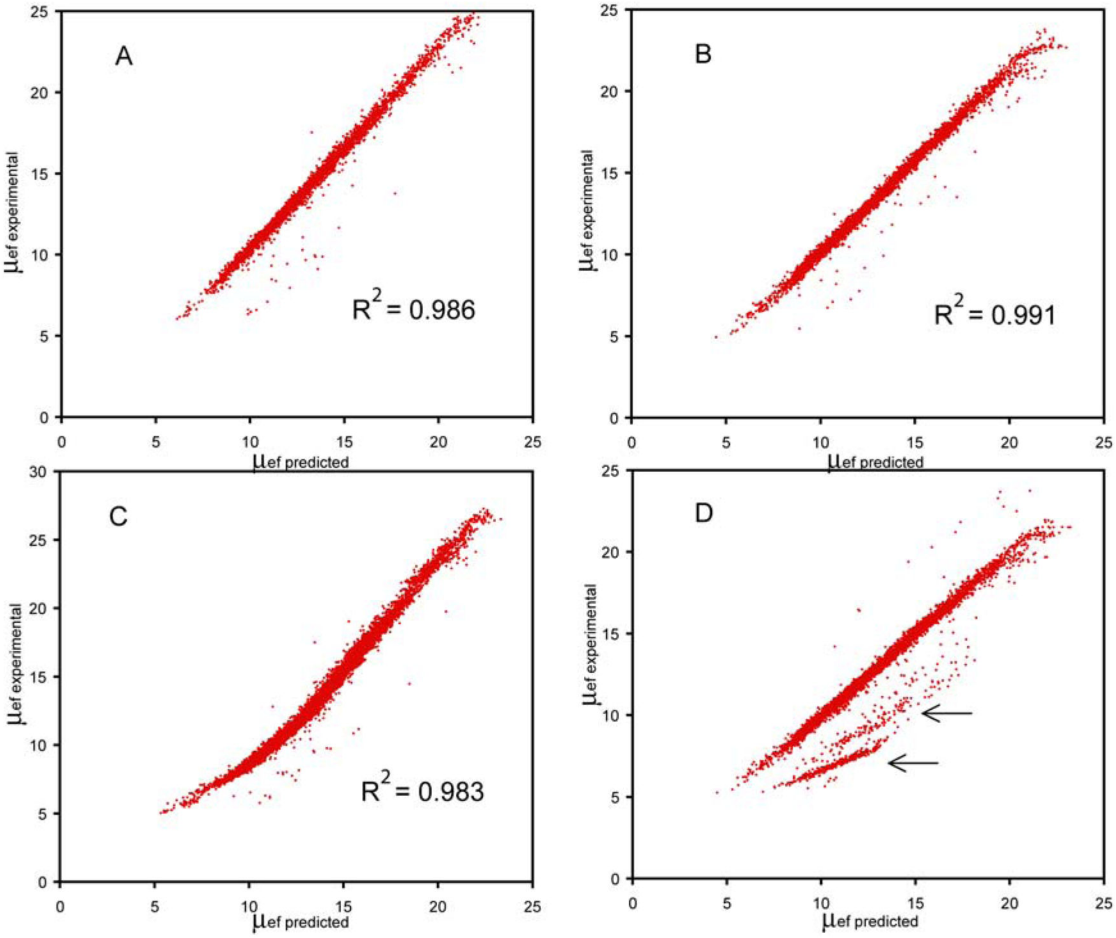


Figure 3.

Predicting electrophoretic mobility for test sets of peptides A) *E. coli* tryptic digest (3290 non-modified tryptic peptides); B) replicate #2 of yeast digest (4663); C) HeLa cells tryptic digest (7979); D) replicate #2 of yeast digest (5164) including peptides with PTM. Two distinctive groups of peptides (N-terminal cyclization and acetylation) are marked with arrows. Note, that all plots contain false positive identifications, which reduce accuracy of the model (R^2-values). Electrophoretic mobility μ_{ef} x 10^5 ($cm^2*V^{-1}*s^{-1}$) is shown.

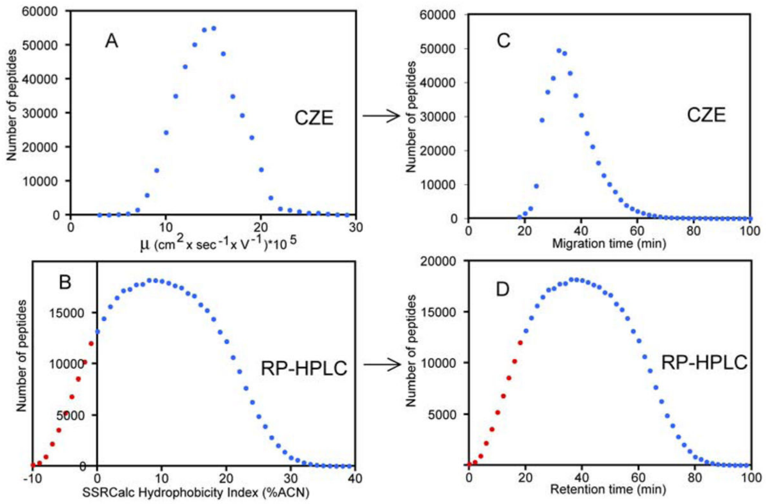


Figure 4.

Predicted distributions of μ_{ef} (A), peptide hydrophobicity (B), CE migration times (C) and retention times for typical LC-MS with trap column injection and 0.5% per minute acetonitrile gradient (D). 441,664 tryptic peptides from yeast proteome (> 4 residues, 1 missed cleavage allowed) were used for these *in-silico* calculations. Note, that hydrophilic peptides with SSRCalc HI < 0 (%ACN) will not retain in RP-HPLC system (labeled with red dots) with 0.1 % formic acid as ion pairing modifier.

Table 1

The summary of X!Tandem CZE-MS/MS identification output for the analyses of three different tryptic digests.

Sample (replicate)	**# of MS/MS**	**# of peptide IDs**	**# of unique peptide IDs**	**# of protein IDs**
E.Coli (#1)	41,982	8,539	3,841	830
E.Coli (#2)	40,585	9,002	4,109	872
Yeast (#1)	61,923	11,549	5,164	1,428
Yeast (#2)	51,287	11,002	5,256	1,497
HeLa (#1)	49,228	15,063	9,022	2,119

Table 2

Performance of previously reported semi-empirical models* for 4463 peptides training data set.

Model	R^2-value correlation
$\mu_{ef} = k(Z/M^{1/3})$ [19]	0.909
$\mu_{ef} = k(Z/M^{2/3})$ [18]	0.953
$\mu_{ef} = k(Z/M^{1/2})$ [20]	0.959
$\mu_{ef} = k(Z/M^{0.56})$ [21]	0.962
$\mu_{ef} = k(\ln(1+Z)/M^{0.43})$ [22]	0.945
$\mu_{ef} = k_1(\ln(1+k_2Z)/M^{0.411})$ [23]	0.974**

*charge *Z* was calculated as a number of positively charged residues (Arg, Lys, His, N-terminus);

**best fit model from Cifuentes and Poppe [23] $\mu_{ef} = 900*(\ln(1+0.35*Z)/M^{0.411})$

Table 3

Step-wise optimization process and model accuracy.

Optimization cycle	**Corrections introduced**	**R^2-value**
Starting point μ_{ef} = 900*(ln(1+0.35*Z)/$M^{0.411}$)	-	0.974 (Figure 1a)
1 (sequence specific Z corrections)	Position specific charge corrections for Asn, Glu, Asn; nearest neighbour effect for Asn, Glu	0.991 (Figure 1b, c)
2 (mass correction)	Correction related to relative size of the residues in peptide sequence	0.992 (Figure 1d)
3 (polynomial correction)	Polynomial offset – function of Zc/N	0.995 (Figure 1e, f)