Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents

Oleg V. Krokhin*

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

Continued development of a new sequence−specific algorithm for peptide retention prediction in RP HPLC is reported. Our discovery of the large effect on the apparent hydrophobicity of N-terminal amino acids produced by the ion-pairing retention mechanism has led to the development of sequence-specific retention calculator (SSRCalc) algorithms. These were optimized for a set of ∼2000 tryptic peptides confidently identified by off-line micro-HPLC−MALDI MS (MS/MS) (300-Å pore size C18 sorbent, linear water/acetonitrile gradient, and trifluoroacetic acid as ion-pairing modifier). The latest version of the algorithm takes into account amino acid composition, position of the amino acid residues (N- and C-terminal), peptide length, overall hydrophobicity, pI, nearest-neighbor effect of charged side chains (K, R, H), and propensity to form helical structures. A correlation with R^2∼0.98 was obtained for the 2000-peptide optimization set. A flexible structure for the SSRC programming code allows easy adaptation to different chromatographic conditions. This was demonstrated by adapting the algorithm (∼0.98 R^2 value) for a set of ∼2500 peptides separated on a 100-Å pore size C18 column. The SSRCalc algorithm has also been extensively tested for a number of real samples, providing solid support for protein identification and characterization; correlations in the range of 0.95−0.97 R^2 value have normally been observed.

The analytical chemistry of peptides and proteins has come a long way since introduction of the soft ionization techniques ESI and MALDI.[1,2] Identification of hundreds to thousands of proteins from minute samples is becoming routine thanks to recent advances in genomics, bioinformatics, and mass spectrometers. Proteomics researchers are now able to probe deeper into the dynamics of the proteome by looking for post-translational protein modifications and conducting quantitative analyses.[3]

However, the generation of gigantic data sets of “identified” peptides and proteins causes understandable worries about the quality of these data. False positives are a cause for particular concern, and Carr et al.[4] pointed out the importance of using additional filtering criteria for identification and characterization. Such approaches can be based on the application of various statistical methods to estimate quality of MS/MS data[5−7] or on the measurement or prediction of additional peptide properties such as pI[8] and hydrophobicity.[9]

Although most workers recognize the value of mass spectrometry in such investigations, the role of chromatographic techniques is often relegated to “just a sample preparation procedure”. This is a consequence of its low resolution compared to mass spectrometry and also because the chromatography of peptides and proteins matured long before[10] the appearance of proteomics as a separate field, so it is often taken for granted. In fact the classical chromatographic techniques (reversed-phase, normal-phase, ion-exchange, and immobilized metal ion affinity chromatography) now play a key role in many approaches for protein identification and characterization. In addition to enabling sample fractionation/enrichment prior to mass spectrometry, these techniques provide retention times that can be used as additional parameters to sharpen the accuracy of protein identification.[9,11−13]

Indeed, retention time prediction and study of the mechanism of RP HPLC of peptides have been the subject of considerable

* To whom correspondence should be addressed. Tel: (204) 474 6184, Fax: (204) 474 7622, E-mail: krokhino@cc.umanitoba.ca.

[1] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64−71.
[2] Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299−2301.
[3] MacCoss, M. J.; Matthews, D. E. *Anal. Chem.* **2005**, *77*, 294A−302A.
[4] Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531−533.
[5] Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768 −774.
[6] MacCoss, M. J.; Wu, C. C.; Yates, J. R., 3rd. *Anal. Chem.* **2002**, *74*, 5593−5599.
[7] Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
[8] Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L. *J. Proteome Res.* **2004**, *3*, 112−119.
[9] Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826−5830.
[10] Mant, C. T.; Hodges, R. S. In *HPLC of Biological Macromolecules,* 2nd ed.; Gooding, K. M., Regnier, F. E., Eds.; Marcel Dekker: New York, 2002; pp 433−511.
[11] 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. *Anal. Chem.* **2003**, *75*, 1039−1048.
[12] 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.
[13] Palmblad, M.; Ramström, M.; Bailey, C. G.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. *J. Chromatogr., B* **2004**, *803*, 131−135.

10.1021/ac060777w CCC: $33.50 © 2006 American Chemical Society
Published on Web 10/17/2006

research since the early 1980s. The key part of all prediction algorithms is the representation of the hydrophobicity of a peptide as the sum of the hydrophobicities (retention coefficients) of all the constituent amino acids.[14] A number of similar algorithms have been developed for various sorbents and ion-pairing modifiers[15−18] showing some correlation of the retention time with the calculated hydrophobicity. In spite of these successes, many researchers pointed out that there was much room for improvement; in particular, the algorithms should reflect other peptide properties such as peptide length[19] and presence of secondary structures (α-helix, β-sheets);[20−24] i.e., they should be sequence specific. Houghten and DeGraw[25] used a set of 260 related peptides (13-residue-long cycling through all 20 amino acids in each position) to show clearly that each position within the peptide would require a different set of retention coefficients for accurate retention prediction. However, even these will not be enough for building an adequate model, since apparent hydrophobicity of the residue is affected by various sequence-specific factors. Recently, Kovacs et al.[26] demonstrated the significance of such factors in an effort to completely eliminate them while determining the intrinsic hydrophobicities of amino acid side chains. This was achieved by designing custom synthetic peptides which lack any nearest-neighbor or conformational effects. But working with real digests forces us to study and take into account as many sequence specific features as possible.

Additional problems with prediction accuracy appeared when early algorithms were applied to real proteomic samples, since they were created for the small sets of peptides often with modified N- and C-termini. Proteolytic digests in proteomics, however, contain peptides with free terminal amino and carboxy groups. As we showed recently,[12] this is the main reason for the limited applicability of the early models, since charge distribution effects ion-pair formation and hence retention behavior.

Another important difference between chromatographic and proteomic studies of this problem is how researchers approach the problem. In chromatography, retention prediction in various LC modes (including RP HPLC of peptides) is the subject of fundamental studies to allow fast optimization of chromatographic conditions to improve peak resolution and decrease analysis time. Most of these studies are directed toward the retention prediction of a limited number of sample components under various chromatographic conditions. For example, parameters examined in RP HPLC of peptides included different ion-pairing reagents at various concentrations,[27,28] as well as different gradient slopes and flow rates.[29] In addition, these studies were scattered across a range of different RP sorbents, which also introduced uncertainties. By contrast, chromatographic conditions applied in most proteomics analyses are fairly limited, but proteomics deals with samples carrying a virtually unlimited number of components. Since complete chromatographic resolution of all analytes in such a case is impossible, retention prediction in proteomics is mainly concerned with the question, "assuming constant chromatographic conditions (column, gradient), how will the composition of a given peptide affect its retention time?"

Most proteomic studies use RP HPLC as a final step of sample preparation prior to mass spectrometry, and retention information can be easily extracted following peptide identification. This can provide extensive data sets of peptides for retention prediction optimization. Thus, Petritis et al.[11] reported the collection of 7000 peptides confidently identified by on-line HPLC−ESI MS/MS and used it for algorithm optimization.

We used off-line HPLC-MALDI MS to collect a data set of ~350 peptides and used it for the development of our first version of a sequence-specific retention calculator (SSRCalc) algorithm[12] yielding a correlation R^2 ~0.94. A number of corrections to the additive model were proposed, including compensation for significant changes in the retention coefficients for peptide length, for overall hydrophobicity, and for amino acids located at the N-terminal. A direct consequence of this work was the idea of using separate retention coefficients depending on amino acid position within peptide chain. This required, however, a bigger retention data set. A second version of the algorithm was developed for a set of ~2000 tryptic peptides and was presented at the 2004 ASMS conference in Nashville.[30] It featured separate retention coefficients for amino acids at positions 1, 2, and $n - 1$, corrections for peptide length, pI, overall hydrophobicity, nearest-neighbor effect for charged amino acids (His, Arg, Lys), and the presence of Pro repeats. This yielded correlations with R^2 ~0.96, and Petritis et al.[31] presented similar results for their optimization at the same meeting.

The next step in updating our model included introduction of separate retention coefficients for the short peptides and took into account the peptide's propensity to form helical structures. The optimization of the third version of SSRCalc with correlation ~0.98 was accomplished in early 2005. It should be noted that the current versions of the algorithm have been available on the Internet (http://hs2.proteome.ca/SSRCalc/SSRCalc.html) since May 2004 and have received a number of positive references.

We spent the year 2005 testing version 3 of the algorithm for a 300-Å pore size sorbent and collecting the data for a 100-Å pore size C18 column. The flexible structure of the SSRCalc allowed

(14) Meek, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632−1636.

(15) Browne, C. A.; Bennett, H. P. J.; Solomon, S. *Anal. Biochem.* **1982**, *124*, 201−208.

(16) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M. R.; Hodges, R. S. *J. Chromatogr.* **1986**, *359*, 499−517.

(17) Su, S. J.; Grego, B.; Niven, B.; Hearn, M. T. W. *J. Liq. Chromatogr.* **1981**, *4*, 1745−1753.

(18) Sasagawa, T.; Okuyama, T.; Teller, D. C. *J. Chromatogr.* **1982**, *240*, 329−340.

(19) Mant, C. T.; Burke, T. W. L.; Black, J. A.; Hodges, R. S. *J. Chromatogr.* **1988**, *458*, 193−205.

(20) Zhou, N. E.; Mant, C. T.; Hodges, R. S. *Pept. Res.* **1990**, *3*, 8−20.

(21) Blondelle, S. E.; Ostresh, J. M.; Houghten, R. A.; Perez-Paya, E. *Biophys. J.* **1995**, *68*, 351−359.

(22) Purcell, A. W.; Aguilar, M. I.; Hearn, M. T. W. *Anal. Chem.* **1993**, *65*, 3038−3047.

(23) Sereda, T. J.; Mant, C. T.; Hodges, R. S. *J. Chromatogr.* **1995**, *695*, 205−221.

(24) Lazoura, E.; Maidonis, I.; Bayer, E.; Hearn, M. T. W.; Aguilar, M. I. *Biophys. J.* **1997**, *72*, 238−246.

(25) Houghten, R. A.; DeGraw, S. T. *J. Chromatogr.* **1987**, *386*, 223−228.

(26) Kovacs, J. M.; Mant, C. T.; Hodges, R. S. *Biopolymers* **2006**, *84*, 283−297.

(27) Guo, D.; Mant, C. T.; Hodges, R. S. *J. Chromatogr.* **1987**, *386*, 205−222.

(28) Shibue, M.; Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **2005**, *1080*, 86−75.

(29) Mant, C. T.; Burke, T. W. L.; Hodges, R. S. *LC GC* **1994**, *12*, 396.

(30) Krokhin, O. V.; Ying, S.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *52nd ASMS Conference on Mass Spectrometry and Allied Topics*, Nashville, TN, Poster 2004.

(31) Petritis, K.; Kangas, L. J.; Yan, B.; Strittmatter, E. F.; Camp II, D. G.; Lipton, M. S.; Xu, Y.; Smith, R. D. *52nd ASMS Conference on Mass Spectrometry and Allied Topics*, Nashville, TN, Poster 2004.

easy adaptation to different chromatographic conditions, once the key difference between the 100- and 300-Å pore size was established. The optimization for the data set of ~2500 tryptic peptides separated on a 100-Å pore size C18 column resulted in a similar (~0.98 R^2 value) retention time versus hydrophobicity correlation. This is to our knowledge the first demonstration of the difference between RP sorbents of various pore sizes.

This paper describes the data collection, approach to algorithm optimization, and general structure of SSRCalc for both 100- and 300-Å pore size C18 columns.

EXPERIMENTAL SECTION

Reagents. Dithiothreitol, iodoacetamide, trifluoroacetic acid (TFA), and 2,5-dihydroxybenzoic acid (DHB) were obtained from Sigma Chemicals (St. Louis, MO). Sequencing grade modified trypsin (Promega, Madison, WI) was used for digestion. A number of commercially available proteins and a few well-characterized real samples (affinity-purified α5β1, αVβ3 human integrins,[32] and mammalian reovirus[33]) were used to generate mixtures of peptides for subsequent HPLC-MALDI analysis. A list of the proteins is provided in the Supporting Information. Three different protein mixtures used for model testing in this paper included the following: pea protein mixture (Parrheim Foods, Portage la Prairie, Manitoba, Canada) containing 10 major protein components; mixture of 49 human proteins provided by the ABRF Proteomics Standards Research Group as sPRG06 sample; protein mixture obtained from human NK like cell line, YTS (ATCC) by immunoprecipitation with anti-IQGAP1 rabbit antibodies (Dr. Xiaobo Meng, unpublished results).

Preparation of Digests for HPLC−MS Analysis. Sample preparation included reduction (10 mM dithiothreitol, 30 min, 57 °C), alkylation (50 mM iodoacetamide, 30 min in the dark at room temperature), dialysis (100 mM NH_4HCO_3, 6 h, 7 kDa molecular mass cutoff, Pierce), and trypsin digestion (1/50 enzyme/substrate weight ratio, 12 h, 37 °C). Mixtures of these proteins (~0.4 pmol/μL of each protein) were carefully composed prior to digestion in order to provide 300−400 or 100−150 peptide samples for 300- or 100-Å sorbent data acquisition, respectively. In this way, a data set of ~2000 peptides for a 300-Å column was collected using six separations (~2500 peptides in 20 chromatograms for the 100-Å one). Samples were acidified with 0.5% TFA and injected directly (5 μL, ~2 pmol of each protein) into the μ-HPLC system. The pea protein digest and the IQGAP1 immunoprecipitate (200−300 ng/μL protein final concentration, 1000−1500 ng/injection) were prepared in the same way. One-third of the sPRG06 sample (~1600 ng of protein) was used for digestion and HPLC−MALDI MS analysis.

Each sample was spiked with the digest of a standard protein for calibration/data alignment purposes (see later discussion). Human transferrin and horse myoglobin were used as calibrating proteins for 300- or 100-Å sorbent data acquisition, respectively.

Chromatography and Fraction Collection. These have been described in detail elsewhere.[12] In brief, samples were fractionated using linear water/acetonitrile gradients (1% acetonitrile starting conditions) on a micro-Agilent 1100 Series system (Agilent Technologies, Wilmington, DE). The column effluent (3 or 4 μL/min) was mixed on-line with DBH MALDI matrix solution (0.5 μL/min, 150 mg/mL DHB in water/acetonitrile 1:1) and deposited by a computer-controlled robot onto a movable metal target at 0.5-min intervals. Both eluents contained 0.1% TFA as an ion-pairing modifier. Samples (5 μL) were injected directly onto a 150 μm × 150 mm column (Vydac 218 TP C18, 5 μm; Grace Vydac, Hesperia, CA) or a 300 μm × 150 mm column (PepMap100, 3 μm; LC Packings-Dionex, Sunnyvale, CA). The gradients used were 0.66% acetonitrile/min at 4 μL/min (or later 0.8% acetonitrile/min at 3 μL/min) and 0.75% acetonitrile/min at 3 μL/min for the two columns, respectively.

TOF Mass Spectrometry. Spots of the chromatographic fractions were analyzed by single MS with *m/z* range 550−5000 and by tandem mass spectrometry (low-energy CID, MS/MS) in the Manitoba/Sciex prototype MALDI quadrupole/TOF (QqTOF) mass spectrometer.[34]

Peak Assignments, MS/MS Identification of Peptides, and Software. "M/z" and "Global Proteome Machine" (GPM)[35] programs (Manitoba Centre for Proteomics and Systems Biology, www.proteome.ca) were used for peak assignment and MS/MS identification of peptides, respectively.

Combined peak lists for HPLC−MALDI MS runs were created for each separation by concatenating peak lists from individual fractions. Fraction number was used as a measure of peptide's retention time. If the full intensity of a peak was contained in a single fraction, the peak was assigned a retention time equal to the fraction number. However, if that peak's signal was distributed between two (three) consecutive fractions, the retention time assigned was the intensity-weighted average of the fraction numbers. Thus, only one *m/z* value (for the fraction containing the maximum intensity) remained in the combined peak list.

Peptide identification for subsequent inclusion in the optimization data set consisted of two steps: peptide mass mapping using known protein sequences (allowing one missed cleavage, 10 ppm mass accuracy) and MS/MS confirmation on preselected parent ions. MS/MS identification of small (less than 8−9 amino acids) peptides rarely produced reliable results from search tools; in these cases, the MS/MS spectra were processed manually to confirm peptide identity.

The SSRCalc algorithm was written in Perl 5.8.6, a core component of the Mac-OS 10.4.3 operating system (Apple Computer, Curpitino, CA). The optimization user interface was written in RealBasic 5.5.5 (Real Software, Austin, TX) and presents the operator with a retention time/hydrophobicity scatter plot and R^2 value as well as information about which peptides' prediction values have improved or been made worse by the current iteration's table values. Processing time for the algorithm to evaluate 2000 peptides is ~30 s on an iMac-G5−2.GHz desktop computer.

RESULTS AND DISCUSSION

Data Collection and Alignment. A critical point for the development of the retention prediction algorithm is the collection of the optimization data set. It should contain confidently identified peptides within a wide mass range (550−5000 Da in our case)

[32] Wilkins, J. A.; Li, A.; Ni, H.; Stupack, D. G.; Shen, C. *J. Biol. Chem.* **1996**, *271*, 3046−3051.

[33] Hadzisejdic, I.; Cheng, K.; Wilkins, J. A.; Coombs, K. M. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 438−446.

[34] Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1047−1057.

[35] Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.

and represent as fully as possible typical samples to which the model will be applied, i.e., proteomic samples. Current techniques for peptide separation/identification provide a variety of approaches to collect such a data set. We have chosen to work with an off-line HPLC−MALDI MS combination and used composed samples of purified proteins instead of real samples.

(a) Off-Line HPLC−MALDI MS versus On-Line HPLC−ESI-MS/MS. The on-line combination of nanoHPLC and mass spectrometry is the method of choice for the majority of proteomics research groups. It provides high-throughput analysis, and respective retention/identification data can be easily extracted. ESI combination has a definite advantage in this component. Off-line micro-HPLC−MALDI combination gained popularity in recent years[36−39] but still remains relatively unknown to the proteomics community. We have chosen this method to collect our data set first due to historical reasons: it was complementary to the HPLC-ESI system in Manitoba Centre for Proteomics and Systems Biology for several years and provided exceptionally good results.[40−42] Second, advantage is provided by the size of the columns and flow rates (2−5 μL/min) typically used in micro-HPLC−MALDI MS. This enables more robust chromatographic separation compared to hundreds of nanoliters per minute flow rates in nanoversion. Nanocolumns are also more prone to "overload", which leads to retention time shifts and consequent deterioration in the quality the of optimization data set. Third, off-line combination of HPLC and MALDI MS was chosen due to its unique archiving capability: one can collect chromatographic fractions and store them for further analysis. The separation is completely decoupled from the mass spectrometry part of the analysis, so the two techniques can be optimized separately. This provides time for a detailed analysis of separated species (by MS/MS) and leaves no room for false-positive identifications. The final reason is the compatibility of the HPLC-MALDI technique with a wider range of mobile phases (ion-pairing modifiers). Thus, ESI MS is typically employed with formic acid-based eluents, while MALDI can be used with any acidic ion-pairing modifiers and even tolerates some amounts of salts/detergents in the mobile phase. All these establish, in our opinion, the off-line micro-HPLC−MALDI MS combination as a method of choice for collection of the retention data sets under the variable chromatographic conditions.

(b) Real Proteomics Samples versus Artificially Composed. There is no difference in principle between digests of real proteomic samples and purified proteins. Resulting peptides will behave similarly independent from their source (purified protein or real mixture) or organism of origin. We use about the same protein concentration (~2 pmol/injection) for preparation of our peptide mixtures. This alleviates problems with column overloading and the associated retention shifts. An important advantage in working with a digest of known protein is additional confidence in peptide identification. Separated peptides were first assigned by peptide mass fingerprint using database sequences and then confirmed by MS/MS.

(c) Data Alignment. Preferably all peptides for the optimization set should be separated under the same chromatographic conditions. Column temperature should be controlled as well, as it may affect peptide retention through changing the enthalpic/entropic contributions of analyte interactions with the stationary phase. We choose to maintain a constant 30 °C column temperature throughout all experiments,[12] since introducing this variable into the algorithm represents a formidable task. Keeping chromatographic conditions constant removes many problems with retention time alignment in the peptide database. Our first set of ~2000 peptides for a 300-Å column was collected using the same column and identical chromatographic conditions. Each of the six separations contained a digest of the calibrating protein, human transferrin. By identifying the same ~50 human transfrerrin peptides and plotting retention time versus hydrophobicity, we were able to compensate small differences in the intercepts observed for these chromatograms. We used horse myoglobin as a calibrating protein for the 100-Å pore size column; the nine most abundant peptides (Table 1) from a horse myoglobin digest were used to provide internal calibration for each of 20 chromatograms acquired in this case.

This approach can also be used for the alignment of data collected under slightly different chromatographic conditions (gradient slope, column size, flow rate). Thus, the original data set of ~2000 peptides for the 300-Å column was collected for the 0.66% acetonitrile/min gradient and 4 μL/min flow rate. We continue to populate this database using a chromatographic column of the same size but at 0.8% gradient and 3 μL/min flow rate. To make this transformation more precise, we calculated the "ideal" hydrophobicity of 9 peptides from horse myoglobin by extrapolating their actual retention times onto a resulting fitting curve of the data set of 2000 peptides. The ideal hydrophobicity values for nine calibrating peptides are presented in Table 1 for C18 columns of both pore sizes. Retention time correction from current into standard data set conditions involves the following: (i) the retention time assignment for nine horse myoglobin peptides; (ii) plotting a calibration line using ideal hydrophobicity values; (iii) calculation using the formula (Figure 1a):

$$T_{\mathrm{Rcorr}} = ((1.8727(T_{\mathrm{R}} - 12.507))/1.5066) + 12.677$$

A better solution for an alignment problem will be the use of a set of standard synthetic peptides spanning wide range of hydrophobicities. Calibration with real digests (such as myoglobin, described here) is less preferable since it generates more than nine abundant peaks shown in Table 1 and can affect MS (MS/MS) analysis. However, even in this form, this approach provides a solid background for intercolumn and even interlaboratory correlation and data collection.

Empirical Multiparametric Algorithm Optimization: Starting Point and Work Flow. As mentioned above, application of

[36] Hsieh, S.; Dreisewerd, K.; van der Schors, R. C.; Jimenez, C. R.; Stahl-Zeng, J.; Hillenkamp, F.; Jorgenson, J. W.; Geraerts, W. P. M.; Li, K. W. *Anal. Chem.* **1998**, *70*, 1847−1852.

[37] Miliotis, T.; Kjellstrom, S.; Nilsson, J.; Laurell, T.; Edholm, L. E.; Marko-Varga, G. *J. Mass Spectrom.* **2000**, *35*, 369−377.

[38] Chen, H. S.; Rejtar, T.; Andreev, V.; Moskovets, E.; Karger, B. L. *Anal. Chem.* **2001**, *73*, 2323−2331.

[39] Zhang, B.; McDonald, C.; Li, L. *Anal. Chem.* **2004**, *76*, 992−1001.

[40] Krokhin, O.; Li, Y.; Andonov, A.; Feldmann, H.; Flick, R.; Jones, S.; Stroeher, U.; Bastien, N.; Dasuri, K. V. N.; Cheng, K.; Simonsen, J. N.; Perreault, H.; Wilkins, J.; Ens, W.; Plummer, F.; Standing, K. G. *Mol. Cell. Proteomics* **2003**, *2*, 346−356.

[41] Krokhin, O.; Cheng. K.; Sousa, S.; Ens, W.; Standing, K. G.; Wilkins, J. A. *Biochemistry* **2003**, *42*, 12950−12959.

[42] Ghosh, D.; Krokhin, O.; Antonovici, M.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *J. Proteome Res.* **2004**, *3*, 841−850.

Table 1. Calculated "Ideal" Hydrophobicities and Retention Times for Nine of the Most Abundant Peptides from Horse Heart Myoglobin Used for Internal Chromatogram Calibration and Data Alignment (Figure 1a)

		300-Å sorbent			100-Å sorbent	
peptide mass	peptide sequence	H_{ideal}	T_R[a]	T_R[b]	H_{ideal}	T_R[c]
649.3071	ELGFQG	16.2989	43.2	38	20.2459	68
747.4279	ALELFR	27.9933	65.1	55.5	29.3991	82.7
1270.6558	LFTGHPETLEK	21.532	53	45	19.1252	66.2
1377.8344	HGTVVLTALGGILK	40.4352	88.4	73	35.3767	92.3
1501.662	HPGDFGADAQGAMTK	18.8621	48	40.6	16.1364	61.4
1605.8475	VEADIAGHGQEVLIR	25.4835	60.4	50	22.0517	70.9
1814.8952	GLSDGEWQQVLNVWGK	41.9304	91.2	75.6	39.3618	98.7
1852.9544	GHHEAELKPLAQSHATK	18.5951	47.5	40.1	13.957	57.9
1884.0145	YLEFISDAIIHVLHSK	52.1295	110.3	91.4	46.3978	110

[a] Retention times (fraction) for 0.66% acetonitrile/min gradient, $T_R = 1.8727H + 12.677$ (R^2 value = 1). [b] Retention times for 0.8% acetonitrile/min gradient, $T_R = 1.5066H + 12.507$ (R^2 value = 0.999). [c] Retention times for 0.75% acetonitrile/min gradient, $T_R = 1.606H + 35.485$ (R^2 value = 1).

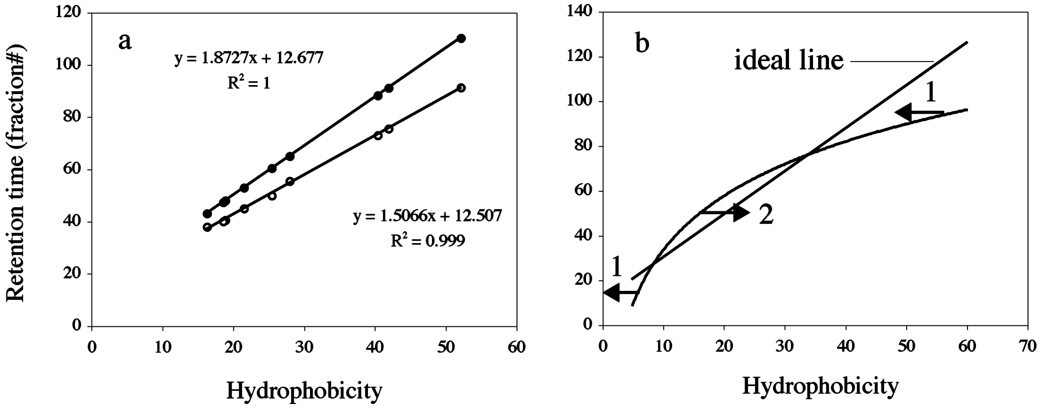

Figure 1. (a) Data alignment using nine peptides from horse myoglobin (Table 1) for separations at different chromatographic conditions: $y = 1.8727x + 12.677$ for 150 μm × 150 mm Vydac 218 TP C18 column, 0.66% acetonitrile/min at 4 μL/min; $y = 1.5066x + 12.507$, 0.8% acetonitrile/min at 3 μL/min. Note slightly lower intercept at the latter plot due to efforts to minimize pre- and postcolumn dead volumes for 3 μL/min flow rate. (b) Different ways of "steering" of concave retention time vs hydrophobicity plots.

additive algorithms to real proteomics samples provides disappointing results. We found that retention coefficient optimization for the models based solely on summation of the retention coefficients of 20 amino acids (or additive models) cannot provide correlation better than 0.85−0.9 for a data set of reasonable size.[12] One of the main reasons behind this is the neglecting of the ion-pairing mechanism of peptide separation in RP HPLC. Despite years of studies and data collection, this important feature was not explored enough to provide any impact on the accuracy of retention prediction algorithms. Therefore, even keeping in mind all the experience and data about peptide RP HPLC collected over the years, we decided intentionally avoid using it and approach optimization empirically as "nonexpert" users. At the end, this could provide a good chance to compare our findings and confirm it to the "state-of-art" chromatographic knowledge in this field.

(a) Starting point. The model proposed by Guo et al.[16] was chosen as a starting point for our optimization. They determined retention coefficients using a collection of synthetic peptides with designed amino acid substitutions, which in our opinion represents a more general approach to the study of chromatographic retention of peptides. We used the R^2 values as optimization criteria. Each time a new correction introduced, it was accepted only if increase in R^2 value was observed. Only corrections providing improvement in retention time prediction are described in this paper.

(b) Empirical Approach. Following simplistic representation of the ion-pairing mechanism, we proposed to correct the shielding effect that counterions exert on the apparent hydrophobicity of N-terminal residues at positions 1, 2, and 3.[12] We were able to make such an observation by manually inspecting the list of peptides and trying to propose the reason for big (positive or negative) deviations from predicted retention time. Once the usefulness of this empirical approach was established, we applied it at the beginning of each optimization round. The feature exhibiting the most abundant effect on retention time deviation was chosen each time, and a corresponding correction was introduced to compensate for it.

Apart from the individual approach to each peptide, it is always important to observe the character of the whole retention time versus hydrophobicity plot during the optimization. Thus, if we apply an optimized additive model to the separation with linear water/acetonitrile gradient, the fitting curve will have a concave shape (Figure 1b). One can fix it by decreasing calculated hydrophobicities of "too hydrophobic" and "too hydrophilic" peptides or by increasing it for medium ones. This requires introduction of correction coefficients for total calculated hydrophobicity or peptide lengths[19] or both.[12]

(c) Multiparametric Optimization. Introduction of new corrections requires "reoptimization" of retention coefficients for all 20 amino acids. All this makes the algorithm development a

Table 2. Optimized Retention Coefficients for 20 Naturally Occurring Amino Acids[a]

residue	R_c	R_{c1}	R_{c2}	R_n	R_{n-1}	R_c (100 Å)
W	12.25	11.1	11.8	12.25[b]	12.1	13.35
F	10.9	7.5	9.5	10.9[b]	10.3	11.67
L	9.3	5.55	7.4	9.3[b]	9.3	9.4
I	8.0	5.2	6.6	8.0[b]	7.7	7.95
M	6.2	4.4	5.7	6.2[b]	6.0	6.25
V	5.0	2.9	3.4	5.0[b]	4.2	4.7
Y	4.85	3.7	4.5	4.85[b]	4.4	5.35
C[c]	0.45	0.9	0.2	0.45[b]	−0.5	0.1
P	2.1	2.1[d]	2.1	2.1[b]	2.1	1.85
A	1.1	0.35	0.5	1.1[b]	−0.1	1.0
E	0.95	1.0	0.0	0.95[b]	−0.1	1.0
T	0.65	0.8	0.6	0.65[b]	0.0	0.65
D	0.15	0.5	0.4	0.15[b]	−0.5	0.15
Q	−0.4	−0.7	−0.2	−0.4[b]	−1.1	−0.6
S	−0.15	0.8	−0.1	−0.15[b]	−1.2	−0.15
G	−0.35	0.2	0.15	−0.35[b]	−0.7	−0.35
R	−1.4	0.5	−1.1	−1.3	−1.1	−2.55
N	−0.85	0.2	−0.2	−0.85[b]	−1.1	−0.95
H	−1.45	−0.1	−0.2	−1.45[b]	−1.7	−3.0
K	−2.05	−0.6	−1.5	−1.9	−1.45	−3.4

[a] Complete list of values for small peptides and for 100-Å sorbent is provided in Supporting Information. [b] Retention coefficients for position n (R_n) were assigned equal to R_c. [c] Retention coefficients for carbamidomethylated Cys. [d] Retention coefficient for Pro at N-terminal was assigned equal to R_c due to lack of N-terminal prolines in our data set.

slow and tedious procedure. We did not apply linear regression analysis or any other computerized approach at this point since manual optimization provides better control of the process and helps to avoid falling into local extrema. During retention coefficients optimization, we followed two rules:

(1) Retention coefficients of individual amino acids were optimized according to the order of their abundance—starting with Leu and finishing with Trp.

(2) If prior information about anomalous behavior of particular residues was available (as in the case of His, Arg, or Lys on 100-Å sorbent; see later discussion) from the manual inspection of peptide list, these resides were considered for optimization first.

Algorithm Structure. Factors Affecting Peptide Retention in Ion-Pair RP HPLC and Their Incorporation into the Retention Prediction Algorithm. (1) Retention Coefficients for Individual Amino Acids. Our algorithm starts with summation of individual retention coefficients for all amino acids represented in a given peptide. The natural consequence of our early work[12] was to introduce separate retention coefficients depending on the amino acid position inside the peptide chain. Thus, the second version of the algorithm featured five different coefficients (R_{c1}, R_{c2}, R_{cn}, R_{cn-1} and R_c; Table 2) for each residue: positions 1, 2, n, $n - 1$, and for internal amino acids, respectively. Position n is occupied by Arg and Lys in tryptic peptides. At this time, we do not have enough retention data for non-Arg/Lys-terminated peptides. These values will be optimized as relevant data sets of nontryptic peptides become available. While the effect of positions 1 and 2 was evident, the idea of correcting position $n - 1$ originated also from the ion-pairing nature of the separation mechanism. Arg and Lys are carrying positively charged side chains at pH 2. Therefore a "cloud" of counterions will affect the hydrophobicities of the surrounding amino acids as well. We found, however, this effect not as important as for N-terminal, α-amino group (compare R_{c1}, R_{cn-1}, and R_c in Table 2).

(2) Nearest-Neighbor Effect. This effect has the same nature and a similar method of correction. All amino acids adjacent to His, Arg, or Lys inside the peptide chain will be affected by trifluoroacetate anions associated with positively charged side chains. This exerts the biggest effect on the most hydrophobic amino acids: W, F, L, I, V, and Y. The individual corrections have been assigned for these amino acids: 0.15, 0.1, 0.3, 0.15, 0.2, and 0.05, respectively. In addition, this correction as well as all others have specific weighing factors, which allows more flexible bulk optimization of corrections. For example, if an Arg residue is followed by Leu, the total peptide hydrophobicity will be decreased by 0.3 × 1.0 (weighting factor) = 0.3.

(3) Clusters of Hydrophobic Amino Acids. These will decrease the apparent hydrophobicity of the peptide. This observation can be explained from the point of view of steric hindrance that prevents neighboring amino acids from simultaneous interaction with the C18 phase. To incorporate this correction in the algorithm, all hydrophobic residues were represented as binary symbols X and Y corresponding to highly hydrophobic (WFLIY) and hydrophobic (MVA) residues, respectively. Individual correction factors were assigned to each possible combination. For example an "XXXXXX" combination corresponding to the stretch of 6 very hydrophobic residues in a row decreases hydrophobicity by 4.5 (correction) × 0.4 (weighting factor) = 1.8.

(4) Specific Role of Pro Residues. Proline is the most rigid of the 20 naturally occurring amino acids since its side chain is

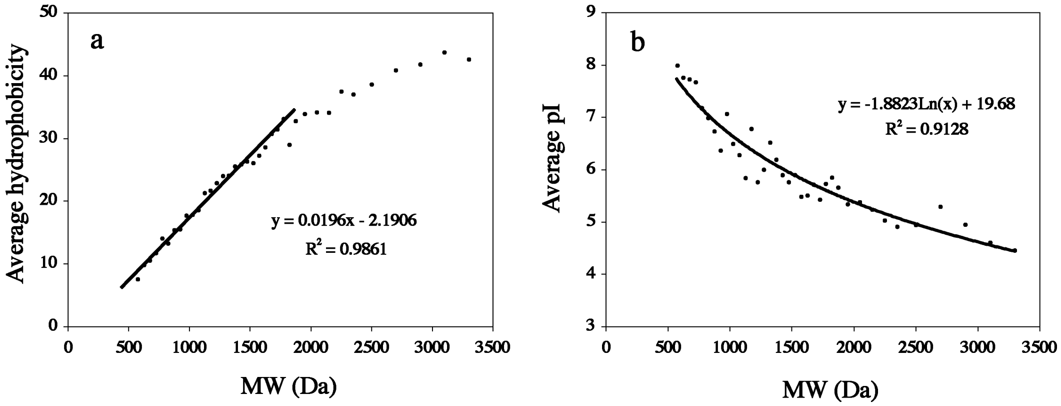

Figure 2. Dependence of average bulk properties of peptides from molecular weight for optimization data set. Average hydrophobicity in 550−1700-Da range shows linear dependence with correlation ~0.986 R^2 value (a).

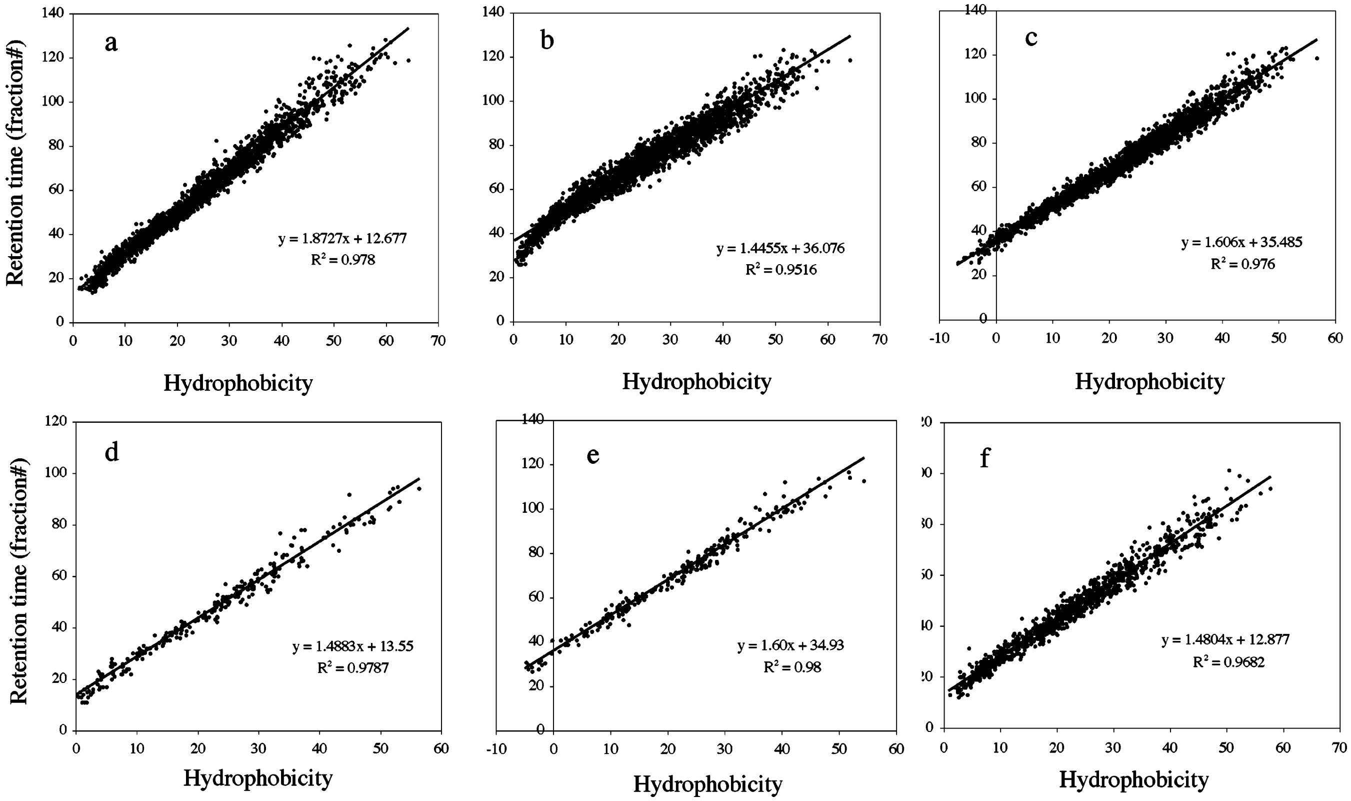

Figure 3. Retention time vs calculated hydrophobicity correlation for various data sets and conditions (details in Table 3). (a) Optimization set ∼2000 peptides, 300-Å sorbent; (b) optimization set ∼2500 peptides, 100-Å sorbent using calculated hydrophobicities for 300-Å version; (c) optimized conditions for 100-Å sorbent for the same ∼2500 peptides; (d, e) pea proteins digest on 300- and 100-Å columns, respectively; (f) digest of IQGAP1 associated proteins on 300-Å column.

covalently linked with nitrogen in the peptide chain. This, to some extent, prevents peptide folding, makes the rest of amino acids available for hydrophobic interactions, and likely provides the main contribution to the Pro retention coefficient (2.1). We found that peptides carrying proline repeats typically show negative deviation from the predicted retention time. Therefore, the hydrophobicity of each peptide carrying PP, PPP, and PPPP sequences was decreased by 1.2, 3.5, and 5.0, respectively.

(5) Influence of Isoelectric Point of a Peptide. This and its incorporation into the algorithm is still under investigation and therefore is not reported here in detail. The importance of this parameter was established yet again by manual inspection of the peptide list. Thus, we found that lower pI favors retention of relatively hydrophobic peptides and decreases retention for hydrophilic ones. The reverse is also true; high pI is the reason for preferential retention of relatively hydrophilic compared to hydrophobic peptides. Such a phenomenon can be explained from the point of view of ion-pair formation as well. Counterions of TFA provide a shielding effect for a whole peptide and its influence will depend on the relative peptide/counterion hydrophobicity. This also will explain the differences in retention behavior in the systems with different ion-pairing modifiers, which vary in its own hydrophobic properties.

Taking into account bulk properties of peptides such as pI and hydrophobicity may be difficult since the properties of individual amino acids that provide the highest impact on these parameters were already considered at the level of retention coefficients' optimization. Our attempts to reflect these properties in the algorithm are based on calculating average bulk properties of peptides depending on molecular weight. Figure 2 shows how average peptide hydrophobicity (a) and average peptide pI (b) depend on molecular weight. Following these considerations, each peptide can be assigned to the categories of too hydrophobic, too hydrophilic, too acidic, or too basic for its mass and respective corrections can be introduced.

(6) Separate Set of Retention Coefficients for Short Peptides (N < 9). The separate set was introduced during optimization of the third version of the SSRCalc algorithm. A smaller size of peptide makes all its residues readily available for interaction with a stationary phase. In addition, this could help to correct the influence of pI discussed in the previous paragraph, because short tryptic peptides possess high isoelectric point.

All parameters and correction (1−6) listed above are used for calculation of "base hydrophobicity" and reflect local (sequence-specific) properties of peptides. The following corrections (7−9) are applied last to take into account peptides bulk properties.

(7) Corrections for Peptide Length. The corrections remain the same as described previously.[12] Following calculation of base peptide hydrophobicity peptide length, a weighting factor K_L is applied as follows:

$$\text{if } N < 8, \quad K_L = 1 - 0.055(8 - N)$$

$$\text{if } N > 20, \quad K_L = 1/1 + 0.027(N - 20);$$

$$\text{otherwise } K_L = 1$$

(8) Correction for Overall Peptide Hydrophobicity. This correction in the first version of the SSRCalc algorithm was applied for the peptides with hydrophobicity values higher than 38.[12] The second and third versions feature more detailed separation of peptides into the groups depending on their base hydrophobicity:

$$\begin{aligned}
&\text{if } H < 20, \quad H_{\text{final}} = H \\
&\text{if } 20 < H < 30, \quad H_{\text{final}} = H - 0.27(H - 18) \\
&\text{if } 30 < H < 40, \quad H_{\text{final}} = H - 0.33(H - 18) \\
&\text{if } 40 < H < 50, \quad H_{\text{final}} = H - 0.38(H - 18) \\
&\text{if } 50 < H, \quad H_{\text{final}} = H - 0.447(H - 18)
\end{aligned}$$

These calculations keep peptide hydrophobicity unchanged if it is less than 20 and penalize the rest of peptides for being too hydrophobic.

(9) Influence of Peptide Propensity To Form Helical Structures. The influence of this parameter was established by monitoring peptides with substantial positive deviations from the predicted retention time. We found that the main contribution to these deviations was provided by short stretches (and repeats of such stretches) of 4−6 amino acids with the following patterns: XXOX, XXOXX, XXOOXX, where X represent hydrophobic amino acids and O all others. To introduce a correction for that effect, we needed to generate a table of corrections for each combination of 4−6 residues mentioned before out of 20 amino acids. This was impossible for our set of 2000 peptides as only 1080 of them featured these "short helical structures", which is below the number of possible combinations. We addressed this issue by separating amino acids into four different groups: X, strongly hydrophobic (WFLI); Z, hydrophobic (YMVA); O, acidic (ED); U, all others (GCHRKPNQST). Each possible combination was assigned with an optimized value of a correction factor. For example, the highest correction was assigned to XXOXX sequences: 3.75(correction) × 1.6 (weighting factor) = 6.0. This value was added to the H_{final} to obtain the calculated hydrophobicity. The way of incorporating corrections in the case of multiturn helixes is still under development and is not described here in detail.

Adaptation of SSRCalc Algorithm to 100-Å Pore Size C18 Sorbent. Optimization of all parameters described above resulted in the ~0.98 R^2 value retention time versus calculated the hydrophobicity correlation for the 300-Å pore size C18 column (Figure 3a). We are working on extending the applicability of the algorithm to a wider range of chromatographic conditions. The 100-Å pore size C18 columns were the first candidates for the study since these are the most popular for peptide fractionation prior to mass spectrometry. We composed the data set of ~2500 tryptic peptides using essentially the same set of proteins and a PepMap100 column. Direct application of the 300-Å pore size version of algorithm to these data showed a decrease in correlation from ~0.98 to ~0.95 R^2 value (Figure 3b).

As mentioned before, inspection of the peptide list revealed anomalous behavior of positively charged residues on the 100-Å column: peptides carrying internal His, Arg, or Lys residues showed mostly negative deviations from predicted retention times. Due to ion-pairing formation, their effective size under chromatographic conditions at pH 2 is essentially larger than for bare residues. Therefore, an additional size exclusion barrier appears for His-, Arg-, or Lys-carrying peptides and results in a decrease of optimal R_c values for these amino acids (Table 2).

Smaller pore size prevents big peptides from penetration inside the bead and, consequently, decreases their retention. On the other hand, increased retention was observed for small (especially highly hydrophobic) peptides. This all explains changes in the character of the fitting curve seen from comparison of Figure 3a and b. To explore possible bulk corrections for the peptides of small hydrophobicity (i.e., small size), we look to determine how deviation from average hydrophobicity for a peptide of particular mass influences retention prediction. All peptides were grouped according to their masses (550−700, 700−800, ..., 1400−1500 Da), and respective ($T_{\text{Rmeasured}} - T_{\text{Rpredicted}}$) versus ($H_{\text{calculated}} - H_{\text{average}}$) and graphs were plotted for all mass ranges. Figure 4 shows such dependencies for three mass ranges: 550−700, 900−1000, and 1400−1500 Da. The average hydrophobicity for each particular mass was calculated using the equation, $H = 0.0196\text{MW} - 2.1906$ (Figure 2a). As can be seen (Figures 3b and 4), small peptides (especially hydrophobic) show mostly positive deviations from predicted retention time and the degree of this influence decreases with the mass. The correction for this effect was introduced directly into the SSRCalc code as very last using equations for fitting curves at Figure 4. For example

$$\text{if MW} < 700, \delta H = H - (0.0196\text{MW} - 2.19)$$

$$\delta T_{\text{R}} \text{ (fractions)} = 0.0018\delta H^3 - 0.0544\delta H^2 + 0.6434\delta H + 3.2$$

$$H_{\text{final}} = H + \delta T_{\text{R}}/1.5$$

Addition of this final correction together with optimization of all parameters listed above for the 100-Å pore size column resulted in similar R^2 value ~0.98 correlation (Figure 3c). Note that the significantly larger intercept of T_{R} versus hydrophobicity plot for the PepMap100 column is a consequence of its larger dead volume.

Testing the Model and Its Applicability. The SSRCalc algorithm is now routinely applied in our laboratory for protein identification and characterization tasks. Thus, we are working on development of a search engine that will use both peptide mass and retention time for protein identification by peptide mass fingerprint in conjunction with HPLC−MALDI MS analysis. Peptide mixtures being analyzed with this approach provide additional information for inclusion in our retention database. Simultaneously, the correlations observed in these studies indicate applicability of SSRCalc to a variety of real proteomic samples. Retention times for linear gradient of 0.66% acetonitrile/min can be predicted with an accuracy of ±4, ±2, and ±1 min for 97, 78, and 49% of the peptides, respectively (Figure 3a). Analyses of real samples typically provide correlation in the range 0.95−0.97 compared to ~0.98 for an optimization set (although a number of cases with 0.98−0.99 have been observed). Table 3 shows a summary of the results obtained for some recently analyzed samples. These include a mixture of pea protein peptides (their retention times for both columns are provided in the Supporting Information), the digest of the sPRG06 sample, and the digest of the IQGAP1 immunoprecipitate (see Experimental Section). As

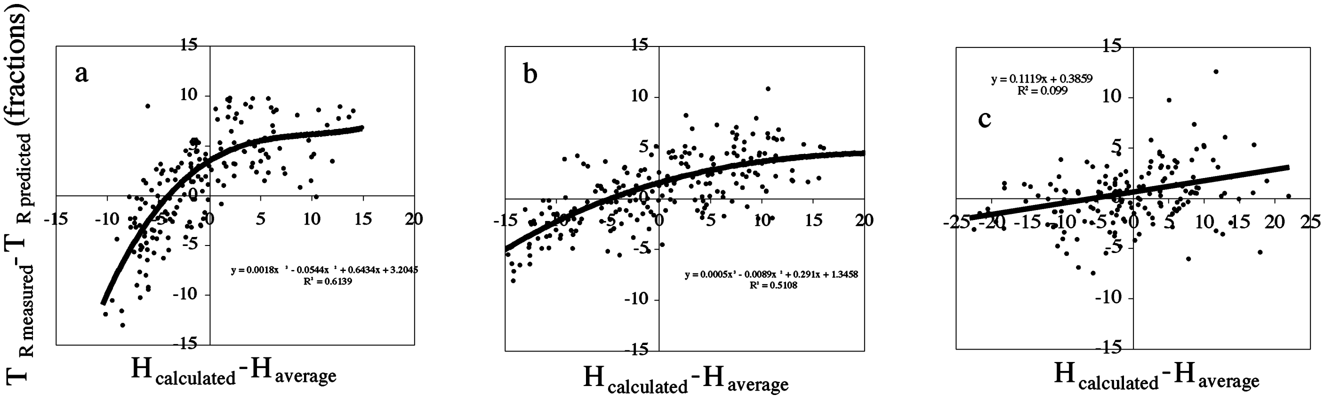

Figure 4. Influence of small peptide's hydrophobicities on the accuracy of retention time prediction for 100-Å pore size sorbent. Average hydrophobicity for a peptide of particular mass was calculated using the approximation from Figure 2a. All peptides were divided into groups depending on their mass: (a) 550−700, (b) 900−1000, and (c) 1400−1500 Da.

Table 3. SSRCalc Algorithm Performance for Various Conditions and Data Sets

data set, conditions	plot parameters, correlation
optimization set ∼2000 peptides, 300-Å sorbent, 0.66% acetonitrile/min gradient	$y = 1.8727x + 12.677$, 0.978
optimization set ∼2500 peptides, 100-Å sorbent, 0.75%/min gradient	$y = 1.606x + 35.485$, 0.976
test data set (pea proteins), 267 peptides, 300-Å sorbent, 0.8%/min gradient	$y = 1.4883x + 13.55$, 0.9787
test data set (pea proteins), 255 peptides, 100-Å sorbent, 0.75%/min gradient	$y = 1.60x + 34.93$, 0.98
test data set (IQGAP1 associated proteins), 1062 peptides, 300-Å sorbent, 0.8%/min gradient	$y = 1.4804x + 12.877$, 0.9682
test data set (IQGAP1 associated proteins), 827 peptides, 100-Å sorbent, 0.75%/min gradient	$y = 1.5831x + 34.061$, 0.9673
test data set (sPRG06 sample), 576 peptides, 300-Å sorbent, 0.8%/min gradient	$y = 1.4778x + 13.35$, 0.9685

can be seen from Table 3, both 300- and 100-Å pore size columns produce results consistent with correlations for the optimization set of peptides. It should be noted that all peptides confirmed by MS/MS in relevant samples were used to plot correlations for Table 3.

Model Applicability. It is very important to know model limitations and to apply correct chromatographic conditions providing the best results.

(1) SSRCalc was developed for linear water/acetonitrile gradients with 1% acetonitrile initial conditions. Changing starting acetonitrile concentration will result in isocratic elution of some hydrophilic peptides and, consequently, an inability to predict their retention. Application of alternative gradients (exponential, etc.) will require more careful internal calibration of the chromatogram, since nonlinear dependence of retention time versus hydrophobicity will be observed in this case.

(2) The stationary phase should have same pore size and similar C18 and end-capping chemistry. Small differences in end-capping treatment and even in purity of the silica used for sorbent preparation can affect the accuracy of retention time prediction. For example, hydrophilic end capping or application of phenyl stationary phases will likely introduce additional interactions that will deteriorate prediction accuracy.

(3) The use of TFA as an ion-pairing modifier will also ensure better algorithm accuracy. Since ion-pairing interactions provide a large effect on peptide retention, any change in counterion hydrophobicity will affect separation selectivity. Such an influence was the subject of intensive studies and widely represented in the chromatographic literature.[27,28] Upgrading the algorithm for use with another modifier will require collection of a new data set and optimization.

(4) The model was developed for Cys residues alkylated with iodoacetamide. Therefore, its application to the peptides with unmodified cysteines or cysteines with alternative alkylation chemistry will provide unsatisfactory results.

(5) The best prediction accuracy will be obtained for sets of peptides containing preferably small ($N < 15$) peptides and peptides with small tendency to form helical structures. For example, a data set of 2000 peptides (Vydac TP218, 300 Å) contains 920 peptides with no short helical stretches (as discussed above). The average absolute deviation from predicted retention time of these peptides was found to be 1.14 min, while a 1.46-min value was observed for the rest of the sequences. This indicates lower prediction accuracy for the species with higher helical propensity and that the potential for model improvement is not explored completely in this aspect. Higher or lower abundance of helical peptides will affect accuracy of prediction for real samples and may explain some discrepancies in correlation values shown in Table 3. For example, (103−118) YLEFISDAIIHVLHSK peptide from horse myoglobin (P68082) contains a continuous stretch of three "short helical structures" as discussed above, which likely contributes to the large positive (∼5.6 min) deviation from predicted retention time. The model was optimized for "average" peptides, while anomalous physicochemical properties of some separated species may cause significant deviations. An example of such case is (230−255) ESTVFEDLSDEAERDEYELLCPDNTR (∼6.7-min positive deviation) from human lactotransferrin (P02788), which carries 10 Asp/Glu residues and exhibits an extreme pI value.

(6) As mentioned above, the model was developed for tryptic (i.e., carrying Lys or Arg at N-terminal) peptides. Extension of the model to the peptides of arbitrary composition requires collection of the respective retention data set.

Retention Mechanism and Future Development. Correct prediction of retention times and of the influence of various mobile/stationary-phase parameters is impossible without a clear understanding of the separation mechanism. Our understanding of retention processes generally coincides with the description provided by Mant and Hodges.[10] Peptides in random coil conformation interact with C18 phase mostly due to contribution of hydrophobic amino acids. The presence of a preferred binding domain such as a hydrophobic face of an amphiphatic α-helix results in preferential interaction of this part of the peptide and a consequent increase in retention. This process provides the most significant sequence-specific contribution into deviation from additive models. Zhou et al.[20] showed that, in the absence of a hydrophobic binding domain, other interactions (electrostatic, etc.), which could stabilize α-helical conformation, do not have significant impact on retention. We confirmed that by unsuccessful application of the AGADIR algorithm to our set of peptides at different stages of optimization (result not shown here). This program was developed for prediction of conformational behavior of monomeric helical peptides in solution and provided good prediction of experimentally determined helical content.[43] This suggests once again that only α-helical sequences stabilized by hydrophobic interaction play an important role in peptide RP retention.

Our findings also indicate the second most important sequence-specific effect provided by ion-pairing involving a free α-amino group at the peptide N-terminus. For example, transfer of the internal Leu residue to the N-terminal of a peptide will result in decrease of its retention coefficient by 3.75 units, which corresponds to 3.49 min in retention time for a 0.66% gradient (Table 2, Figure 3a). Introduction of separate retention coefficients for amino acids at different positions confirmed our initial conclusions about the nature of this effect.[12] The same mechanism explains the nearest-neighbor effect of Lys, Arg, and His residues with positively charged side chains. The most intriguing part of the ion-pairing influence is its selective character, which can provide positive/negative deviations depending on the relative hydrophobicity of the residue/peptide and the counterion.

It should be noted that the sequence-dependent effect of the free α-amino group on hydrophobicity of N-terminal amino acid residues was described more than 10 years ago by Sereda et al.[44] The authors, however, provided an explanation based on the influence of the α-amino group itself, rather than the associated counterions. This shows once again that most likely proteomic researchers will "rediscover" many features described in the past as we found[12] this effect independently. It is interesting that changes in our retention coefficients due to transfer of an amino acid from the N-terminal (R_{c1}) inside a peptide (R_c) match very closely the retention shifts observed for N-acetylated/non-acetylated synthetic peptides (Table 4).[44]

Overall, peptide sequence affects retention through differences in distribution of hydrophobic and charged amino acids. Development of an optimal prediction algorithm will require efforts for very precise introduction of parameters reflecting these properties into the programming code. This paper first demonstrates an attempt to incorporate a number of sequence-specific features into the structure of the prediction algorithm. They definitely require further exploration and corrections if needed. For example, correction for the peptide propensity to form helical structures is applied last, rather than together with other sequence-specific corrections (positions 1−7 and 9 in algorithm structure section). This was done intentionally, since interaction with the hydrophobic face of amphiphatic helixes affects the overall peptide hydrophobicity by preferential binding through this domain and against discriminating all others. Other ways of making this correction may provide different results.

Table 4. Comparison of Measured[44] Effect of N-Terminal Acetylation and Calculated by SSRCalc Effect from Removing of N-Terminal Correction for Ion-Pairing Formation

residue	measd value of retention shift by Sereda et al.[42]	calcd value, present work
Trp	−0.8	−0.71
Phe	−3.3	−2.1
Leu	−2.9	−2.31
Ile	−2.8	−1.72
Met	−1.8	−1.1
Val	−2.2	−1.29
Tyr	−1.4	−0.71
Ala	−0.1	−0.46
Thr	1.1	0.09
Pro	−2.0	*a*
Glu	0.2	0.03
Asp	0.6	0.22
Cys[b]	−0.5[b]	0.34[b]
Ser	0.7	0.58
Gln	0.9	0.18
Gly	0.0	0.34
Asn	0.9	0.65
Arg	1.0	1.17
His	1.8	0.83
Lys	1.9	0.89

[a] The value is not provided due to lack of N-terminal Pro in the data set. [b] The value for carbamidomethylated Cys is provided in this study.

Despite a definite increase in the accuracy of retention time prediction, we believe that the potential for useful improvement to the SSRCalc algorithm has been exhausted in the present form. Studying and introducing new sequence-specific correction factors will provide further progress, but examining these smaller factors will first require that we deal effectively with the major contributor of sequence-specific retention features—the formation of α-helixes. It is important to note, however, that this is unlikely that any predictive approach will overcome all problems associated with sequence-specific effects in the foreseeable future.

We plan to continue development of the SSRCalc algorithm through searching new sequence-specific features and further enrichment of our peptide retention database. We plan the next optimization round when the data sets for both columns will reach the ∼10 000-peptide mark. The collection of relevant data sets is important but not a primary goal. Detailed studies of each sequence-specific correction and the correct way of to incorporate it in the algorithm will provide the most significant impact on prediction accuracy. As mentioned above, corrections for peptides helicity are the first candidates for revisiting. A growing database will provide more material to work in this direction. However, an approach that uses designed synthetic peptides[16,23,26] may provide faster and more precise answers to these questions. For example, one of the most interesting questions in this regard is how ion-pairing formation will effect interaction of α-helixes with the C18

(43) Lacroix, E.; Viguera, A. R.; Serrano, L. *J. Mol. Biol.* **1998**, *284*, 173−191.
(44) Sereda, T. J.; Mant, C. T.; Quinn, A. M.; Hodges, R. S. *J. Chromatogr.* **1993**, *646*, 17−30.

phase. Due to the low abundance of His and the fact that mostly -Arg-Pro- and -Lys-Pro- sequences represent internal Arg and Lys in tryptic digests, the α-helical structures with charged residues are definitely underrepresented in our data set. Design and synthesis of relevant sets of peptides is a preferable solution here, rather than data collection with real digests.

Further developments will also be directed to widen the utilization of the SSRCalc application by the proteomics community. This work will include the following: (i) updating the Web-based version of SSRCalc; (ii) studying the algorithm applicability for different RP phases; (iii) optimization of the algorithm for wider range of mobile and stationary phases; (iiii) study and introduction of retention coefficients for alternative Cys alkylation, post-translationally and chemically modified residues (including the most popular stable isotope tags).

CONCLUSIONS

A new sequence-specific retention calculator algorithm was developed for the accurate peptide retention prediction in RP HPLC. Correlation up to ~0.98 R^2 value has been obtained for both 300- and 100-Å pore size C18 sorbents. To our knowledge, this is one of the best algorithms developed to date and the only one developed in conjunction with off-line HPLC−MALDI MS. The differences between 300- and 100-Å pore size prediction models were demonstrated for the first time as well. We showed that collection of a high-quality data set of reasonable size (2000−2500 peptides) together with a delicate approach to the introduction of sequence-specific corrections into the algorithm provides superior accuracy of retention prediction. Acquisition of a similar pool of the data is a matter of hours or days for LC-ESI or LC-MALDI techniques, respectively. This opens up a possibility for the development of custom algorithms virtually for any RP HPLC conditions applied in different proteomics research groups. We hope that publishing these data and making the algorithm available will promote future developments, which will benefit studies in both proteomics and classical HPLC fields.

ACKNOWLEDGMENT

The author thanks Victor Spicer and John Cortens for exceptional computer support of this work. Sample deposition station was designed and manufactured by James McNabb. Constant discussions and support of Dr. John Wilkins, Dr. Kenneth Standing, Dr. Ronald Beavis, and Dr. Werner Ens is greatly appreciated. The author also thanks to Dr. Robert Hodges for reading the manuscript and providing valuable corrections. This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (K.G.S., W.E.), Genome Canada (W.E.), Canadian Institute for Health Research (J.A.W.), and the U.S. National Institutes of Health (GM 59240, K.G.S.).

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review April 26, 2006. Accepted September 11, 2006.

AC060777W