Prediction of Peptide Retention Times in Hydrophilic Interaction Liquid Chromatography (HILIC) Based on Amino Acid Composition

by Majors J. Badgett[1], Barry Boyes[1,2], Ron Orlando[1]
[1]Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602 USA
[2]Advanced Materials Technology, Wilmington, DE 19810 USA

A retention prediction model for peptides was created for hydrophilic interaction liquid chromatography (HILIC). This model predicts coefficients for each amino acid that can be summed to predict the retention time of peptides. The correlation coefficient (R^2 = 0.960) is similar to previous reverse-phase (RP) and HILIC peptide retention prediction models. This model was developed using gradient elution on a HALO Penta-HILIC column and can predict the retention times of peptides based on amino acid composition with a site-specific correction for hydrophobic residues at the N-terminus.

Introduction

Recent developments have shown that HILIC is an incredibly useful tool for the analysis of proteins and peptides, and is complimentary to reversed-phase (RP) chromatography, which has been the preferred analytical method for these analytes due to their large hydrophobicity and low polarity. [1,2] Since 1979, when O'Hare and Nice noted that small peptide retention on RP columns was directly related to the sum of the hydrophobicity of the amino acids within the peptide, many researchers have created models that accurately predict the retention of peptides by the summation of amino acid coefficients. [3-14] These coefficients can be derived a number of ways, from linear regression analysis to the use of MATLAB® or even the substitution of amino acids on a synthetic peptide. [9,10,14] Although the majority of peptide retention prediction models available use RP chromatography, there have been some attempts to create similar models using HILIC, especially since the types of available HILIC columns has steadily increased through the years. Yoshida was the first to do so in 1998 on a TSK Amide-80 column, then in 2011 Gilar et al. created coefficients for three different HILIC stationary phases: bare silica, bridge-ethyl hybrid silica, and an amide modified bridge-ethyl hybrid silica. [1,2] These models have high correlation coefficients in the range of 0.92-0.97, illustrating that the prediction of peptides with these columns can be very accurate. These models have also shown that amino acid coefficients change with different HILIC stationary phases, and are dependent on operating conditions (for example, pH). Thus, amino acid coefficients need to be created for specific mobile phase and stationary phase operation. This does not necessarily limit the usefulness of these models, but rather requires an understanding of the separation methods and conditions that are needed for specific purposes.

Retention prediction models are useful for many different reasons, including being able to improve the confidence in identifying proteins as well as eliminating false positives when MS2 data is insufficient in confidently identifying a peptide. Accurately predicting where peptides will elute can help further the characterisation process and lead to more confident and accurate identifications when paired with database searching. [13,15] Accurate mass and time (AMT) tagging technology has been used frequently to quickly identify peptides based off of their mass to charge ratio and retention times. [16] However, as the type and complexity of chromatographic columns increases, so must the number of models specifically made for those columns that are able to predict retention.

The model that is presented here can predict peptide retention using a HILIC column with gradient elution, and uses dextran as a retention time calibrant. Coefficients for all the amino acids have been derived using linear regression from a data set of tryptic peptides that resulted in a high correlation coefficient (0.960). We introduce specific criteria for peptide selection as well as optimised coefficients for hydrophobic residues at the N-terminus of a peptide. This model is incredibly useful by not only predicting peptide retention, but also heightening protein confidence and decreasing the length of the identification process.

Materials and Methods

Protein Digestion

Myoglobin, transferrin, concanavalin A, fetuin, cytochrome C, lysozyme, ribonuclease B, carbonic anhydride, and dextran were purchased from Sigma-Aldrich (St. Louis, MO, USA). Bovine serum albumin was purchased from Waters (Milford, MA, USA). These proteins were reduced using 10-mM dithiothreitol (DTT) and then alkylated using 55-mM iodoacetamide (IDA), which were both purchased from Sigma Aldrich (St. Louis, MO, USA). Sequencing-grade trypsin or chymotrypsin purchased from Promega (San Luis Obispo, CA, USA) was added (50:1, w/w, protein/trypsin) and samples were incubated at 40°C overnight.

LC-MS/MS Settings and Instrumentation

Data were acquired using a Finnegan

LTQ (Thermo-Fisher, San Jose, CA, USA) and an 1100 Series Capillary LC system (Agilent Technologies, Palo Alto, CA, USA) with an ESI source that used spray tips made in-house. Samples were dissolved in 25% H_2O, 75% ACN and 0.1% formic acid (Sigma-Aldrich, St. Louis, MO, USA) prior to injection, and 6 µL of each sample were directly injected into the LC. Peptides were separated using a 200 µm x 150 mm HALO Penta-HILIC column that has five hydroxyl groups on the bonded ligand and was packed with 2.7-µm diameter superficially porous particles (Advanced Materials Technology, Wilmington, DE, USA). The gradient used for each sample was 95-30% ACN over 90 minutes at a 2 µL/min flow rate. The mobile phase contained 0.1% v/v formic acid (Sigma Aldrich, St. Louis, MO, USA) and the aqueous solvent contained 50 mM ammonium formate (Thermo-Fisher, San Jose, CA, USA).

To evaluate the general applicability of this model, some of the same digested proteins were run on a 4000 Q Trap (AB Science, Chatham, NJ, USA). Peptides were separated by a 2.1 mm x 15 cm HALO Penta-HILIC column packed with 2.7-µm diameter superficially porous particles using a Nexera UFLC (Shimadzu, Columbia, MD, USA). The gradient used for each sample was 78-48% v/v ACN over 80 minutes at a 0.4 mL/min flow rate. Spectra were obtained using an ESI source.

Database Search Parameters

The resulting RAW files were converted using Trans-Proteomic Pipeline (Seattle Proteome Center, Seattle, WA, USA), then the MS/MS spectra of each sample were searched using Mascot (Matrix Scientific, Boston, MA, USA) against corresponding protein databases of theoretical MS/MS spectra. Mascot is versatile software that identifies and characterises proteins based on mass spectrometry data. The following parameters were utilised in Mascot: a peptide tolerance of 1000 ppm, a fragment tolerance of 0.6 Da, two max missed cleavages of trypsin, and a fixed modification of carbamidomethylation (C).

Selection of Peptides for Prediction Model and Post-Run Data Analysis

All peptides that had a higher Mascot score than 10 were considered. Peptide retention times were found by hand from .RAW files from the apex of the peaks using Xcalibur software (Thermo-Fisher, San Jose, CA, USA), and resulting MS/MS data were visually inspected to verify the peptide assignments. Chromatographic peaks for each peptide had to have a peak asymmetry value of between 0.25 - 4, and peptides exhibiting peak widths greater than 5.5 minutes were excluded from analysis. Peptides had to be fewer than 15 amino acids in length. Peptide retention times in minutes were converted to glucose units based on dextran samples that were run immediately before. Linear regression analysis using StatPlus (AnalystSoft, Walnut, CA, USA) was used to find the coefficients for each amino acid. One hundred and eighteen peptides met these criteria and were used in this study.

Results

Amino Acid Coefficients

Table 1 shows amino acid coefficients that were derived using linear regression analysis of peptide retention times and their corresponding amounts of each amino acid residue. Amino acids with positively charged side chains (arginine, histidine, and lysine) had the strongest positive effect on retention time and the strongest effect overall. Negatively charged side chains (aspartic acid and glutamic acid) also had a large positive effect on retention time. All amino acids with aromatic side chains (phenylalanine, tyrosine and tryptophan) and some aliphatic amino acids (leucine and isoleucine) had a negative impact to peptide retention. All other amino acids did not affect retention time to the same degree and were statistically insignificant according to their p-values (calculated probabilities) from the regression analysis. Predicted retention times of peptides, R_T, can be calculated by using Equation 1 shown below, where L_i is the amount of residue i in the peptide, AA_i is the amino acid coefficient of residue i, and b_0 is the intercept of the model:

$$R_T = \sum(L_i AA_i) + b_0 \quad (1)$$

When the predicted times of the 118 peptides used in this model were plotted against their actual times in Figure 1, there is a high correlation coefficient that expresses the accuracy of the amino acid coefficients. This value (0.960) is on the higher end of previous RP and HILIC peptide retention prediction models. [1-14]

In order to make this model capable of being used on any LC-MS system, all coefficients are expressed in glucose units (GU) from procainamide-labelled dextran ladder samples that were run immediately before the standard digests. These dextran samples elute in a logarithmic fashion in order of increasing monosaccharide linkage and provide reference for peptide retention times. A set of peptide standards run after the dextran samples was used over the course of a month on multiple LC-MS systems to make sure that dextran was a suitable retention time calibrant for our purposes.

Optimised Coefficients for Hydrophobic Residues at the N-Terminus

Site-specific trends in the peptide dataset were investigated and it was found that 19 out of 30 peptides with hydrophobic amino acids located at the N-terminus had actual retention times that were greater than their predicted retention times. Table 2 shows optimised coefficients that account for this trend. Using an iterative process that maximised the R^2 value, a 15% increase in the original hydrophobic coefficients was found to have the best fit. The deviation between actual and predicted retention times decreased from .283 GU to .204 GU using these coefficients, indicating an increase in prediction accuracy. These optimized coefficients are only to be used for the first hydrophobic residue at the N-terminus and no others. For unknown peptides, MS2 data needs to be utilised to identify a peptide with a hydrophobic residue at the N-terminus so that these coefficients can be used to predict retention.

Test Peptides

Helicobacter pylori protein digests were run on the same LC-MS setup as the 118 peptides used to create the model so that the model's accuracy could be tested. From these digests, 18 peptides fit the selection criteria and their actual retention times plotted against their predicted retention times yielded a correlation coefficient of 0.949. The relatively high correlation coefficient indicates that the model was suitable for predicting the retention time of these peptides. Table 3 shows the actual retention times and the predicted retention times for the 18 peptides as well as their deviations, with the average deviation being 1.62 minutes. Eight of the 18 test peptides had larger actual retention times that their predicted ones indicating that there was no trend, and all predicted retention times were calculated by using Equation 1.

BSA and carbonic anhydrase were tested

on another LC-MS system, a 4000 Q Trap with a Nexera UFLC, to make sure that the model was universal. Although the LC-MS system, gradient, column size, and flow rate differed, peptides from BSA and carbonic anhydrase that were identified using both LC-MS systems differed only by an average of 2.29 minutes and their retention times were within 3.73% of each other.

Discussion

In order to be able to predict peptide retention with the Penta-HILIC column, a new peptide retention model required calculating. This is because HILIC stationary phases exhibit different selectivites from one another and models made using these columns will produce different amino acid coefficients. [2] It was widely known that amino acid composition is the main characteristic that influences peptide retention, but it was demonstrated that location has an affect as well.

The amino acids that have the strongest effect on retention are histidine, lysine and arginine, and this is evident in other studies. [2,17] Because these residues have positively charged side chains, they interact with the stationary phase to a greater extent than other hydrophilic amino acids and increase peptide retention. These amino acid coefficients, as well as many others, matched up to the inverse of reverse phase coefficients from other models. This finding was expected, however Gilar, et. al. showed that it is not necessarily a linear correlation, illustrating that HILIC and RP can be used in multidimensional HPLC for more complex separations. [2]

While most models attribute retention time solely to amino acid composition, other models have indicated that the length of the peptide and the position of the amino acids have an affect on retention time as well. [13,18-20] Mant et al. concluded that the retention times of longer peptides (over 15 residues) deviate more than expected and cannot be overlooked. [19,20] Since peptides over 15 residues tend to be non-polar due to their large size, most of them would not be retained well on HILIC columns and would elute very early. This consideration was applied to this study, and the peptides in our study were limited to a max of 15 amino acids in length.

The Effect of Amino Acid Location

Krokhin, et al. reported that amino acid location in a peptide influences retention time in RP chromatography and created optimised coefficients to account for position. [13] This is also evident in HILIC, as it was found in this work that most peptides with hydrophobic amino acids located at the N-terminus eluted later than expected. Optimised coefficients were created to account for this difference between expected and actual retention times and they were shown to increase the correlation coefficient and improve predictions. Hydrophilic amino acids at the N-terminus and both hydrophilic and hydrophobic amino acids at the C-terminus were also examined, but the location of these residues appeared to have a negligible affect on retention and there were no detected trends in deviation from expected and actual retention times. Some previous models have incorporated optimised coefficients based on the distance of a specific residue from one of the termini, but no trends were identified that suggested that doing the same would help improve the accuracy of this model. [2, 15]

Summary

A peptide retention prediction model using a HALO Penta-HILIC column and gradient elution was created using LC-MS data from tryptic digests of standard proteins. This model produced a high correlation coefficient (0.960) and contains coefficients for each amino acid that can be used to predict peptide retention times by using Equation 1. Dextran was shown to be a suitable retention time calibrant and we showed that it was able to make this model capable of peptide prediction on two completely different LC-MS systems.

We hope to investigate the effect that some post-translational modifications have on retention (such as oxidation, glycation, deamidation, and glycosylation) and create coefficients that account for them to expand this model. We also hope to investigate peptide size to a greater extent so that we can predict peptides that are longer than 15 amino acids with high accuracy using HILIC. Our group is currently researching a model that predicts glycan retention with the same HILIC column, and eventually we would like to create a glycopeptide retention prediction model that would combine this peptide model with the glycan model.

Acknowledgments

Support for this work comes from NIH Grant R01AI 055624 to Dr Judith H. Willis and NIH Grant GM0 93747 to Dr Barry Boyes.

We would like to thank Rudradatt Persaud for his help with the early visualisation of the project and both Dr. Mary Elizabeth Thompson and Dr T. Colin Campbell for their help with the data sorting and regression analysis.

References

[1] T. Yoshida, *J. Chromatogr. A* 811 (1998) 61-67.

[2] M. Gilar, A. Jaworski, *J. Chromatogr. A* 1218 (2011) 8890-8896.

[3] M.J. O'Hare, E.C. Nice, *J. Chromatogr.* 171 (1979) 209-226.

[4] J.L. Meek, *Proc. Natl. Acad. Sci. USA* 77 (1980) 1632-1636.

[5] S.J. Su, B. Grego, B. Niven, M.T.W. Hearn, *J. Liq. Chromatogr.* 4 (1981) 1745-1764.

[6] K.J. Wilson, A. Honegger, R.P. Stotzel, G.J. Hughes, *Biochem. J.* 199 (1981) 31-41.

[7] T. Sasagawa, T. Okuyama, D.C. Teller, *J. Chromatogr.* 240 (1982) 329-340.

[8] C.A. Browne, H.P.J. Bennet, S. Solomon, *Anal. Biochem.* 124 (1982) 201-208.

[9] D. Guo, C.T. Mant, A.K. Taneja, J.M.R. Parker, R.S. Hodges, *J. Chromatogr.* 359 (1986) 499-517.

[10] J.M.R. Parker, D. Guo, R.S. Hodges, *Biochem.* 25 (1986) 5425-5432.

[11] Y. Sakamoto, N. Kawakami, T. Sasagawa, *J. Chromatogr.* 442 (1988) 69-79.

[12] M. Palmblad, M. Ramstrom, K.E. Markides, P. Hakansson, J. Bergquist, *Anal. Chem.* 74 (2002) 5826-5830.

[13] O.V. Krokhin, R. Craig, V. Spicer, W. Ens, K.G. Standing, R.C. Beavis, J.A. Wilkins, *Mol. Cell. Proteomics* 3.9 (2004) 908-919.

[14] E. Tyteca, A. Periat, S. Rudaz, G. Desmet, D. Guillarme, *J. Chromatogr. A* 1337 (2014) 116-127.

[15] B. Tripet, D. Cepeniene, J.M. Kovacs, C.T. Mant, O.V. Krokhin, R.S. Hodges, *J. Chromatogr. A* 1141 (2007) 212-225.

[16] L. Pasa-Tolic, C. Masselon, R.C. Barry, Y. Shen, R.D. Smith, *BioTechniques* 37 (2004) 627-639.

[17] T. Yoshida, *J. Chromatogr.* A 808 (1998) 105-112.

[18] J.L. Meek, Z.L. Rossetti, *J. Chromatogr.* 211 (1981) 15-28.

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

[20] C.T. Mant, N.E. Zhou, R.S. Hodges, *J. Chromatogr.* 476 (1989) 363-375.

Appendix

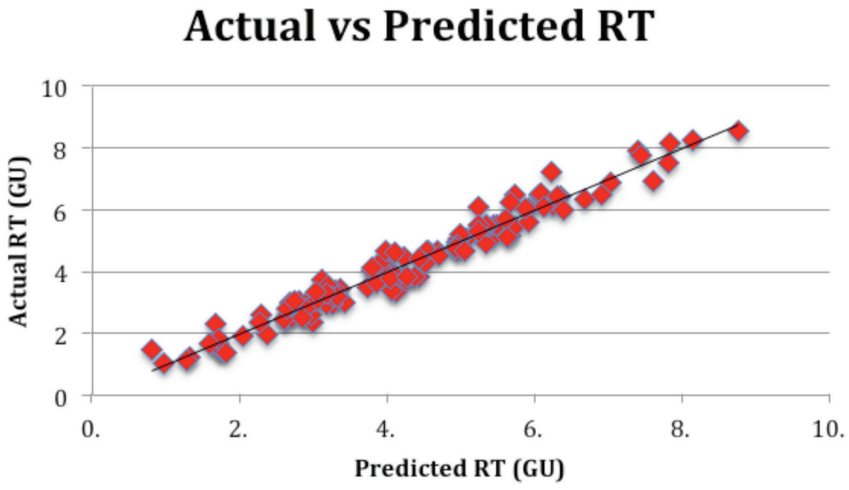
Table 1: Coefficients for all 20 amino acids as found by regression analysis. Proteins were reduced with DTT, alkylated with IDA, and digested with trypsin before being subject to HILIC-MS analysis. Negative deviations indicate the peptide was retained less than predicted and positive deviations indicate the peptide was retained more than predicted.

Amino Acid	Coefficient
Alanine (A)	0.20957
Cysteine (C)*	0.40773
Aspartic Acid (D)	0.67119
Glutamic Acid (E)	0.67791
Phenylalanine (F)	-0.7876
Glycine (G)	0.27677
Histidine (H)	1.50711
Isoleucine (I)	-0.40456
Lysine (K)	2.08285
Leucine (L)	-0.79306
Methionine (M)	-0.46693
Asparagine (N)	0.57851
Proline (P)	0.06800
Glutamine (Q)	0.70475
Arginine (R)	1.85008
Serine (S)	0.32276
Threonine (T)	0.40579
Valine (V)	-0.35101
Tryptophan (W)	-0.97668
Tyrosine (Y)	-0.42613
Intercept	1.36245

**Carbomidomethylated cysteine*

Table 2: Optimized coefficients for all of the hydrophobic amino acids.

Amino Acid	Coefficient
Phenylalanine (F)	-0.90574
Isoleucine (I)	-0.46525
Leucine (L)	-0.91201
Methionine (M)	-0.53697
Tryptophan (W)	-1.12318
Tyrosine (Y)	-0.49005



Equation of the line: y = 0.9948x + 0.0447. The R2 value for the trend line is 0.960.

Figure 1: Predicted vs. actual retention times of peptides used in the prediction model

Table 3: Predicted and actual retention times of helicobacter pylori peptides

Peptide	Actual RT (min.)	Predicted RT (min.)	Deviation (min.)
ADIGIK	55.27	56.24	-0.97
AILEMRLQRLTGLER	62.50	64.92	-2.42
DYDVLFEEAIALR	47.15	49.90	-2.75
EELGLER	60.83	56.99	-3.84
EVTSKPANK	71.95	69.89	2.06
FEPGEEK	63.95	61.98	1.97
GFHGAK	62.58	61.54	1.04
LDIASGTAVR	54.03	56.06	-2.03
LVTVHTPIEANGK	65.56	64.40	1.16
NEDITINEGK	68.82	67.16	1.66
NEDITINEGKK	74.56	72.80	1.76
QVLPVK	49.19	49.52	-0.33
SIKEDVQFADSR	70.68	69.60	1.08
SVELIDIGGNR	57.47	57.63	-0.16
SVELIDIGGNRR	62.47	65.24	-2.77
TWQTADK	60.53	61.29	-0.76
VNDIIADSLTR	55.63	57.90	-2.27
YDANITFVSQA-AYDK	61.14	61.97	-0.83