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

Majors Badgett[1], Barry Boyes[2], Ron Orlando[1]

[1]Complex Carbohydrate Research Center, University of Georgia, Athens, GA, [2]Advanced Materials Technology, Wilmington, DE

The University of Georgia

Introduction

There have been many peptide retention prediction models for reverse phase chromatography experiments, but very few so far for hydrophilic interaction liquid chromatography (HILIC). These models predict retention times of specific analytes by amassing data from LC-MS runs and using the data to predict the elution of similar analytes in future runs based on amino acid composition. Although reverse-phase chromatography is the go-to chromatography method for the analysis of proteins and peptides, HILIC has shown to be very useful and complimentary. HILIC is particularly useful in the analysis of peptides containing hydrophilic post-translational modifications, such as glycosylation. Presented here is a prediction model employing gradient elution on a HILIC column that can predict the retention times of peptides based on amino acid composition.

Methods

Protein Digestion

Bovine serum albumin, myoglobin, transferrin, concanavalin A, fetuin, cytochrome C, lysozyme, ribonuclease B, and carbonic anhydride were reduced using 10-mM DTT and alkylated using 55-mM IDA. Trypsin was added (50:1, w/w, protein/trypsin) and samples were incubated overnight.

LC-MS/MS Settings

For LC-MS/MS analysis using a Finnegan LTQ (*Thermo Scientific*), samples were suspended in 11 µL of 25% H_2O, 75% ACN and 0.1% FA, and 8 µL of each sample was injected into an Agilent 1100 Series LC. Peptides were separated by a Halo penta-HILIC column (200 µm x 150 mm, 2.7-µ particle size). The gradient used for each sample was 95-30% ACN in 0.1% formic acid/50mM ammonium formate/water gradient for 90 minutes at a 2-µL/min flow rate. Spectra were obtained using an ESI source with spray tips made in-house.

To make sure that this model would be universal, some of the same digested proteins suspended in 11 µL of 25% H_2O, 75% ACN and 0.1% FA were run on a 4000 Q Trap (*AB Sciex)*. Peptides were separated by a Halo penta-HILIC column (2.1 mm x 15 cm, 2.7-µ particle size) using a Nexera UFLC (*Shimadzu*). The gradient used for each sample was 78-48% ACN in 0.1% formic acid/50mM ammonium formate/water gradient for 80 minutes at a 0.4-mL/min flow rate. Spectra were obtained using an ESI source.

Database Search Parameters

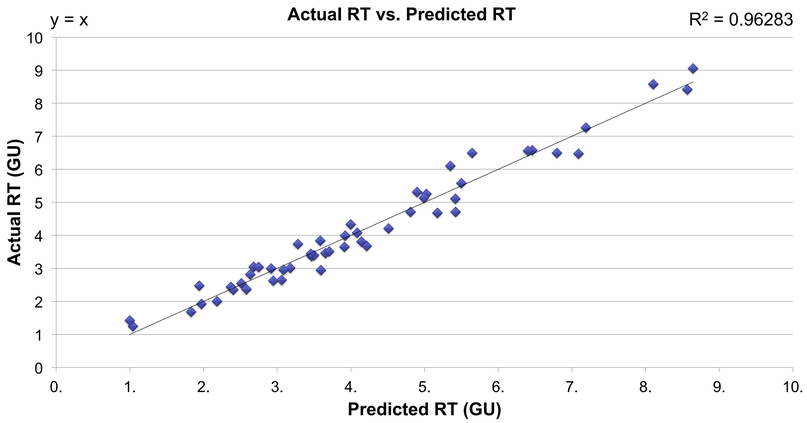
The resulting RAW files were converted, then the MS/MS spectra of each sample were searched using Mascot Daemon against corresponding protein databases of theoretical MS/MS spectra. The following parameters were utilized in Mascot Daemon: a peptide tolerance of 1000 ppm, a fragment tolerance of 0.6 Da, two max missed cleavages of trypsin, and a fixed modification of carbamidomethyl (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, and resulting MS/MS data were visually inspected to verify the peptides. Peptides 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 under 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 was used to find the coefficients for each amino acid and fifty peptides were used in this study.

Results

Actual retention times plotted against predicted retention times for the fifty peptides that matched the criteria are shown below:



The r-squared value (0.96283) indicates that this model is very accurate for predicting the RTs for peptides. The average deviation from predicted and actual RT was 1.45 minutes.

Peptide Prediction Model Coefficients

Amino Acid	Coefficient
Alanine (A)	0.32562
Cysteine (C)*	1.17524
Aspartic Acid (D)	1.06439
Glutamic Acid (E)	1.31496
Phenylalanine (F)	-0.91699
Glycine (G)	0.41634
Histidine (H)	2.29215
Isoleucine (I)	-0.66694
Lysine (K)	2.36468
Leucine (L)	-1.06970
Methionine (M)	-0.89730
Asparagine (N)	0.42535
Proline (P)	-0.12911
Glutamine (Q)	0.83631
Arginine (R)	2.16469
Serine (S)	0.60117
Threonine (T)	0.50174
Valine (V)	-0.36148
Tryptophan (W)	-1.43865
Tyrosine (Y)	-0.43965
Intercept	0.75100
R-Squared Value	0.96283

KEY

Most Hydrophilic Amino Acids:
Cysteine (C), Aspartic Acid (D), Glutamic Acid (E), Histidine (H), Lysine (K), Glutamine (Q), Arginine (R) and Serine (S)

Most Hydrophobic Amino Acids:
Phenylalanine (F), Isoleucine (I), Leucine (L), Methionine (M), Tryptophan (W), and Tyrosine (Y)

Amino Acids that are Statistically Insignificant:
Alanine (A), Glycine (G), Asparagine (N), Proline (P), Threonine (T), and Valine (V)

*Carboxamidomethylated Cysteine

Support for this work comes from NIH-R01AI 055624 to JHW and NIH Grant GM093747 to BEB

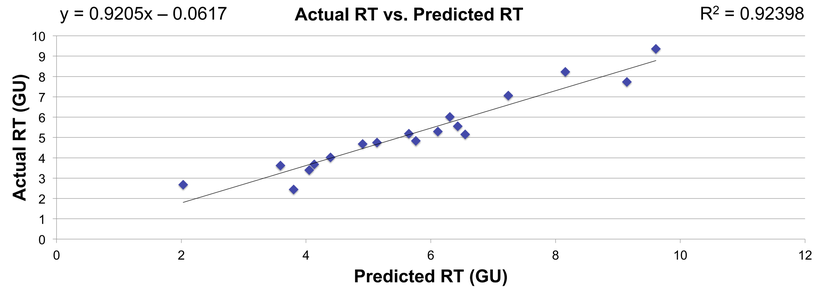
Effect of Peptide Length

It has been reported by several researchers (Mant, et. al., Meek, et. al., Krokhin, et. al.) that the length of a peptide has an important effect to retention, and that peptides longer than 15-20 amino acids deviate more from expected retention times than smaller ones in reverse phase methods. To test this, predicted and actual retention times of peptides from the standard digests that were not used in our prediction model due to length were examined. It was generally found that longer peptides had higher deviations from predicted retention times as shown below (negative deviations indicate the peptide was actually retained less than predicted):

Peptide	Length	Deviation (min)
RPCFSALTPDETYVPK	16	-1.728224739
NTDGSTDYGILQINSR	16	1.958565790
SPDSHPADGIAFFISNIDSSIPSGSTGR	28	-2.913638120
GLSDGEWQQVLNVWGK	16	3.288687946
YGDFGTAAQQPDGLAVVGVFLK	22	-4.564985493
AVVQDPALKPLALVYGEATSR	21	-4.786889756
GLVLIAFSQYLQQCPFDEHVK	21	-5.186307767
CKPVNTFVHESLADVQAVCSQK	22	5.232988641
DLILQGDATTGTDGNLELTR	20	6.146492993
HIIVACEGNPYVPVHFDASV	20	-6.387981695
DAIPENLPPLTADFAEDK	18	7.941109191
NLCNIPCSALLSSDITASVNCAK	23	-8.341706660
AQFVPLPVSVSVEFAVAATDCIAK	24	-13.264748300
ASEDLKKHGTVVLTALGGILK	21	-22.673794270

Test Peptides

Helicobacter pylori protein digests were run on the same LC-MS setup as the 50 peptides used to create the model so that the model's accuracy in prediction could be tested. From these digests, eighteen peptides fit the selection criteria and were examined as shown below:



The relatively high r-squared value indicates that the model was suitable for predicting the RT of these peptides. The average deviation between actual and predicted retention times for the peptides was 3.23 minutes with the largest deviation being 5.6 minutes. Fifteen of the eighteen test peptides had larger actual retention times than their predicted ones.

BSA and carbonic anhydrase were analyzed on another LC-MS system to makes sure that model was universal. All peptides identified from both systems eluted within 3.73% of the expected retention times of each other, indicating that this model can be used regardless of system as long as each peptide RT is converted to GU by using dextran.

Conclusions

- A model that can predict peptide retention times from amino acid composition was created, and tests indicate that it is very accurate.
- The position of amino acids residues is shown to affect retention.
  - 13 of 18 peptides with a hydrophobic residue at the N-terminus eluted earlier than predicted.
- Longer peptides tend to have larger deviations from expected retention times.