Proteomic Data Mining Using Predicted Peptide Chromatographic Retention Times

Megha RenukaJayadev1, Don Blow1, Brian Tripet1, Robert Hodges1 and Krzysztof J. Cios12*
University of Colorado at Denver and Health Sciences Center1 and University of Colorado at Boulder2

ABSTRACT

Correct identification of proteins from peptide fragments is important for proteomic analyses. Peptides are initially separated by reversed-phase high-performance liquid chromatography (RP-HPLC) before mass spectrometry (MS) identification. At the present time, peptide fragment retention (separation) time is not used as a useful scoring filter for identification of the peptide fragments and their parent proteins. In the present paper, we present a new web-based tool for the prediction of peptide fragment retention times and its use in compiling a database of ~133,000 peptide fragments computationally obtained by digestion with trypsin of 4,265 E. coli –K12 proteins. The retention calculation is based on the described formulae and the fragments/protein identification was carried out using a simple search-scoring algorithm.

Key words: Liquid Chromatography (LC), Mass Spectrometry (MS), LC/MS, Reverse-Phase High Performance Liquid Chromatography (RP-HPLC), Retention Time Prediction, Tryptic digest, Mass frequency.

Authors: Megha RenukaJayadev, Don Blow and Krzysztof Cios are at Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center, Denver, CO 80217. Email: Krys.Cios@cu.edu, Office phone: 303-556-6035.
*Corresponding author
Brian Tripet and Robert Hodges are at Department of Biochemistry and Molecular Genetics, University of Colorado at Denver and Health Sciences Center, Aurora, CO 80045. Office phone: 303-724-3253.

INTRODUCTION

Complex peptide samples (such as a tryptic digest of proteins) are typically separated by reversed-phase high-performance liquid chromatography (RP-HPLC) based on subtle differences in the overall hydrophobicity of the peptides. By applying a mobile phase with an increasing organic co-solvent (typically a linear AB gradient where Eluent A is aqueous trifluoroacetic acid (0.1-0.2% TFA) and Eluent B is 0.1-0.2% TFA in acetonitrile) and a C8 or C18 silica based matrix for the stationary phase, the peptides can selectively partition between the stationary and mobile phase at different rates depending on their overall hydrophobicity. The peptides are detected as they elute from the column by mass spectrometry (MS) in the case of LC/MS or LC/MS/MS. Mass spectrometers are used for accurate mass measurements based on elemental compositions for a given peptide. 'Retention time' (defined as the time taken by an individual component to move through the column, from the point of injection to the detector) is a specific and unique parameter of each peptide fragment.

At the present time, although the retention time is recorded during the LC/MS or LC/MS/MS run, the retention time of each fragment is not used in determining or verifying the correct identification of the peptide fragment. This is because prediction of the retention time of many peptide fragments has not been close enough to the observed retention times. Many research groups are now actively investigating methods to predict retention time behavior for proteomic applications (1-4). In this paper we present a new prediction method and a searchable database to evaluate whether this unused parameter can be successfully used to improve the assignment of peptides in LC/MS or LC/MS/MS.

DATABASE AND SEARCH VARIABLES

The database created for analysis was populated with the protein sequences from E. coli-K12. The E. coli K12 proteome was chosen since it is one of the most studied proteomes and the genome is well mapped. 4,265 protein sequences were gathered from the University of Wisconsin at Madison ASAP database. Trypsin, a proteolytic enzyme was considered. Computationally a tryptic digest cleaves after lysine and arginine residues giving 138,291 total fragments of different sizes. Protein sequences are input in text file format to the digest engine. Any number of proteins can be digested with the prescribed format, where each protein is stored in a dynamic array. The name and sequence of the protein are stored in 'protein' object (objects are represented within single quotes throughout the paper). Once the complete text is read into memory, a tryptic digest is carried out. Only a tryptic digest was used at the time of writing this paper. Other proteolytic enzymes with different cleavage specificity may be considered for future enhancement of the project.

For efficient use of calculation time, during the digest process the data length of each fragment is also calculated. Each 'fragment' object contains sequence, mass, retention time, net charge, chain length, fragment position and protein name. To obtain the calculation time, the number of individual amino acids from the data set (n) is used rather than the number of fragments formed by the digest. This is simply due to fragment
numbers being variable on the digest method and protein/peptide composition, whereas the amino acid count is fixed. Since we read each amino acid residue in the protein sequence, cleave the sequence as we read, and sum up the fragment retention time up to each cleavage site, the calculation time is O(n).

Once the tryptic digest is complete, analysis involves data mining and formatting the results so that the output is easy to read. In the case of identifying an unknown peptide, mass (accurate to 5 decimal places) and retention time along with its error variance (if any) are entered into the peptide identifier and the database is searched.

ANALYSIS OF PEPTIDE FREQUENCY

In our initial attempt at populating the database, only whole number fragment masses were considered. It was seen that the frequency of fragments between mass 500-1000 was in the range of 80-160 per mass unit (Figure-1). It was understood that reducing the frequency would improve the accuracy for peptide identification. So the masses of fragments with 8 significant figures were stored in the database. This made a noticeable improvement in the frequency of peptides with the same mass. For example, with one decimal place added to the mass the frequency was reduced to 8-10 fragments with the same mass.

Since the peptide retention time search variable will depend on the amino acid composition of the peptide fragments, we also examined the frequency of peptides with the same compositions. Interestingly, in the entire genome there were only 371 specific cases of peptide fragment sequences having the same amino acid composition (mass 500 or greater). The total sum of peptides was 1016 (only 0.74% of the proteome studied). Thus, peptide retention times if predicted accurately enough can indeed be a useful search criteria for peptide fragment identification. In another analysis, it was noted that this ambiguity decreased as the number of residues in the peptide increased. Specifically, the ratios of the number of residues to the number of cases with the same amino acid composition were found to be - 5/197, 6/96, 7/44, 8/22, 9/8, 11/2, 17/1, and 23/1, fragment length/number of cases. Thus, the longer the sequence the lower the number of cases of peptide fragments having the same amino acid composition. At peptide lengths of 7 residues or greater the number of cases is only 78 and the total sum of peptides reduces dramatically to 164 (only 0.12% of the proteome).

RETENTION TIME CALCULATION

Using the digest engine discussed in the previous section, retention time predictions can be calculated for individual proteins or peptides entered and defined by a user. Previously, investigators have been using one set of side-chain coefficients for all positions in the sequence (5). However, currently we are using three different retention time predictions in the program output. The first of these — Internal Coefficients Only” uses side-chain coefficients for all 20 amino acids derived experimentally from an internal region of a synthetic peptide. These coefficients are used for all positions of a peptide sequence to predict the retention time independent of amino acid position in sequence. The second type of prediction of retention time, labeled “N-Term + Internal Coefficients”, predicts retention time using two sets of experimentally derived coefficients, one set for the N-terminal amino acid residue and another set for the internal coefficients which are used for all the remaining amino acid residues in the sequence. The third type of prediction of retention time, labeled “N-Term + C-Term + internal coefficients” predicts retention time using three sets of experimentally derived coefficients, one set for the N-terminal amino acid residue, one set for the C-terminal amino acid residue and the other for the internal coefficients which are used for all the remaining amino acid residues in the sequence. We have shown that the terminal amino acid side-chain coefficients with varying end-groups (N-terminal acetyl vs. amino or C-terminal carboxyl vs. amide) vary dramatically from each other. For example, terminal coefficients when compared to internal coefficients can vary as much as a factor of two (6).

User input determines which retention time values to apply to each amino acid in a given peptide. A user enters a peptide/protein sequence and selects the type of processing – “No Cutting” and “Trypsin”. An alteration is made to this if the user selects “Yes” for the “Acetylated N-Terminal” question. In this case, acetylated retention time coefficients are used instead of the NH2 retention time coefficients. Similarly, the C-terminal amino acid is processed according to fragment position and whether the user indicates the C-terminal amino acid contains a carboxyl or -amide group. The user can make the appropriate selection to increase the accuracy of the prediction when cysteine is treated with iodoacetamide or iodoacetic acid. We plan to incorporate a time correction when the peptide chain contains more than ten residues since chain length affects retention time predictions (7).

The retention time prediction algorithm is based on three HPLC variables that allow the user to select settings specific to their instrumentation. The “Gradient Rate” (GR) which is the rate of acetonitrile/minute during a linear gradient, the “Gradient Correction Factor” (Tc) which includes the elapsed time from when the HPLC pump starts to deliver a set GR and when it is detected at the back end of the column at the detector at a specific flow rate and the “Peptide Standard Correction Time” (Tw) which allows the researcher to use any HPLC instrumentation, reversed phase columns of any length or diameter, reversed phase packings of any n-alkyl chain length and ligand density, and counter ion concentration differences from those which the retention time coefficients where derived. The following retention time (Rt) prediction formula for any linear gradient rate using coefficients determined at a gradient rate of 0.25% acetonitrile/min (6):

\[ Rt = \text{internal coefficients} + \text{N-Term} \times Tc \]
\[ R_i = \sum (AA_i) \times 0.25/GR + T_e + T_s \]

Current analysis of the predicted retention time versus observed retention time of several tryptic peptide fragments of several proteins shows that there is a strong correlation (0.94) between predicted and observed times (Figure 3). Further, plotting of the differences between predicted and observed retention times (about the mean) shows that approximately 50% of the peptides are outside the range of ± 6 min (at this very shallow gradient of 0.25% acetonitrile/min) (Figure 3). This suggests that there are other variables such as chain length, clustering of hydrophobes and conformational effects which still need to be included on the retention time equation. Current experimentation is focusing on determining and/or accounting for these variables. Once the predicted retention times are within the targeted error range, the database will be recalculated with these corrected values/changes.

**DATABASE QUERIES**

The database stores all the fragment information gathered from E. coli K12 proteome trypsin digest. By storing information about each fragment like mass, retention time and the protein it belongs to, complex queries can be constructed for different analysis. One can query by mass, retention time, mass/retention time and protein prediction. Upon querying, fragment information from the entire proteome is listed in table format along with the parent chain (original protein sequence).

Since the database, calculator and protein identifier are on the same site, it is easy enough for the user to copy the protein into the calculator and digest it with settings for his/her instrument. Giving the set of masses and retention times with or without error corrections (Δm, Δt, see description below), protein prediction for the fragments can be carried out easily (Figure 4).

**Δm and Δt**

Since it is well known that different mass spectrometers will have different levels of accuracy for obtaining the mass of peptide fragments, and different HPLC chromatography units will have run to run retention time variation, it was important to also include into the program variance values (error values) for the mass and the retention time values (denoted ± Δm and ± Δt, respectively). This then allows one to search for a mass range for the fragment mass to match to and a corresponding retention time range.

**Data Mining of Fragments**

Data mining of fragments is carried out based on mass ± Δm and retention time ± Δt. If no value is entered for Δt and Δm then they are set to default values of 4 and 0.4, respectively. The default Δ values have been determined at the Hodges Laboratory. Peptide/Fragment hits help in identifying their respective Proteins, which are stored in a dynamic array of 'Protein Identified' object. It stores Protein ID, Sequence, Hit score, Fragments hit, the Observed and Predicted masses and retention time, when the prediction is run. Each mass and retention time is seen to have number of peptide hits. The protein to which the peptide belongs is given a score as it is identified. The scoring algorithm assigns the identified protein a score of 1 for every hit when mass and retention time match. The algorithm adds a score of bonus 1 when the protein identified is already present in the identified list, as it was seen to have occurred a number of times. Ambiguity of the peptides is further reduced by considering the protein only once and giving the bonus points. For example, a peptide maybe repeatedly found in a protein; in such case the protein is considered only once. The input is given an option to enter manually all the LC/MS masses and retention with charges for each or upload a file-.mgf having all the input data. The actual mass and retention time is calculated based on the formula including the charges. The actual mass calculated by the formula with charge into consideration will be displayed and the prediction is run with the Δm, mass correction and Δt, retention time correction. This predicts the protein for the given set of values.

**Protein Identification**

Protein identification is carried out by matching different masses and respective retention times of different peptide/fragments in the database. 'Peptide' object is created for the initial given mass and retention time. For each mass and retention time the protein hits are compared with the existing proteins in the 'Protein Identified' object and the scoring algorithm scores on the number of times a protein was identified. This process is continued for all the given masses and retention time (with ± Δm and ± Δt, if mentioned otherwise with the default values) that the LC/MS specifies when a sample is run. The highest scored proteins are displayed as the predicted proteins along with their "Hit scores", "Hit fragments" for the protein, "Observed and Predicted masses," "Retention times" and "Protein sequence".

At the time of writing this paper, only a limited number of LC/MS files were tested. We are presently purifying known E. coli proteins to accurately test the algorithm and to determine the optimal values for the mass and retention time error values (Δm and Δt). For the few samples which we have run, we can see that there can be more than one protein hit predicted by the tool for the sample data. The total number of proteins identified depends on the accuracy of the mass, retention time and the error values (Δm and Δt) provided. For example, one E. coli sample protein showed 18 hit protein hits with default Δ values (Figure 5). Decreasing the Δ value in this case reduced the false positive hits. However, a detailed analysis of the optimal Δ value can be performed only when E.coli samples will become available.
FUTURE WORK

Future work planned include enhancement to the digest engine and calculator by implementing different enzyme digests. Another planned enhancement is to make a graphical user interface (GUI) for the digest engine. Implementing a GUI will make the application more user-friendly and will give us the option to display dynamic charts instead of using a secondary spreadsheet application. Specifically referring to the calculator, more variables like temperature may be added to improve prediction accuracy.

More work is being carried out in Robert Hodges Laboratory to determine an ideal Δ values for the mass and retention time. However, presently the default values are in use.

As for the database, support for various species will be added. Of course, more queries will be made available too. Two of the planned queries will be same fragment combination with varying sequences and chain-length/mass/retention time. Coloring the identified fragments when the whole protein sequence is shown is also planned. This would give easy visual interpretation of the fragment coverage of a protein sequence.

The retention time calculator based on current retention time values for each amino acid, fragment identifier and protein identifier with the file or manual input for a sample run on LC/MS is available online at: http://isl.cudenver.edu/RetenMassPrediction/frameset1.htm (This web tool works best using Internet Explorer 5.0 or higher).

FIGURES LEGEND

Figure 1 – Number of fragments at each mass in the 500-4000 range, with no decimal values considered in the database. The chart shows ambiguous peptides with mass range 700-705. It specifies first peptide, Chain 1 and second peptide, Chain 2; the first and second occurrence of peptide fragments with the same mass and same composition but a different sequence.

Figure 2 – Number of fragments for each peptide length. The chart shows analysis done on retention times. It shows the retention times (in minutes) of sequences sorted by time range and the ambiguity information for the peptide fragments of the same mass and composition but different sequence (shaded).

Figure 3 – A correlation between the observed and Predicted Retention time values in minutes of 108 tryptic peptide fragments (left) and the differences between predicted and observed values in minutes at a gradient rate of 0.25% acetonitrile/min.

Figure 4 – Retention calculator, fragments search, and the Protein Identifier.

Figure 5 – Display of the protein ID’s detected while changing the delta value. Manual sample and an E.coli sample was run at the Hodges Lab.

ACKNOWLEDGEMENTS

The authors acknowledge the use of the E.coli Genome Project (ASAP Database) at the University of Wisconsin, Madison, http://www.genome.wisc.edu/tools/asap.htm.

SELECTED REFERENCES


