Proteomic data mining using predicted peptide chromatographic retention times

Brian Tripet

Department of Biochemistry and Molecular Genetics,
University of Colorado at Denver and Health Sciences Center,
Aurora, CO 80045, USA
E-mail: brian.tripet@uchsc.edu

Megha Renuka Jayadev, Don Blow and Cao Nguyen

Department of Computer Science and Engineering,
University of Colorado at Denver and Health Sciences Center,
Denver, CO 80217, USA
E-mail: mrenukaj@ouray.cudenver.edu
E-mail: ddbcrip@yahoo.com
E-mail: cdnguyen@vcu.edu

Robert S. Hodges

Department of Biochemistry and Molecular Genetics,
University of Colorado at Denver and Health Sciences Center,
Aurora, CO 80045, USA
E-mail: robert.hodges@uchsc.edu

Krzysztof J. Cios*

Department of Computer Science,
Virginia Commonwealth University,
Richmond, VA 23284, USA
E-mail: kcios@vcu.edu
*Corresponding author

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.

Keywords: liquid chromatography; LC; mass spectrometry; MS; LC/MS; reversed-phase high performance liquid chromatography; RP-HPLC; retention time prediction; tryptic digest; mass frequency.

Reference to this paper should be made as follows: Tripet, B., Renuka Jayadev, M., Blow, D., Nguyen, C., Hodges, R.S. and Cios, K.J. (2007) 'Proteomic data mining using predicted peptide chromatographic retention times', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Brian Tripet is Manager of the Peptide and Protein core facility at the University of Colorado at Denver and Health Sciences Center. His current research interests include peptide synthesis, peptide modifications, peptide mapping using mass spectrometry, HPLC methodology development, understanding protein folding and stability and the development of a SARS-CoV vaccine. He received his PhD in Biochemistry from the University of Alberta. He has published over 30 papers.

Megha Renuka Jayadev is currently finishing her MS Degree in Computer Science with the option in Computational Biology at the UCDHSC. She graduated from Vani High School in Bangalore, India and holds a Degree in Computer Science (Cum Laude) from M.S. Ramiah Institute of Technology, Bangalore, India. In addition to her roles of an Engineer, Programmer and Systems Analyst, and she is also active in community service: contributing towards welfare of under privileged children and volunteering for social service. She loves sports and has won many awards in field and track events.

Don Blow graduated from UCDHSC in 2005, attaining his MS in Computer Science with the option in Computational Biology. He is currently employed as a Software Engineer for Lockheed Martin – one of the world's leading rocket builders. In early 1993, he broke his neck at the C5-6 level, leaving him paralysed from the mid-chest down. With the help of his trusty service-dog, 'Rudder', he has become a Well-respected Engineer amongst his peers.

Cao Nguyen is a PhD student in Computer Science and Information Systems program, with the option in Computational Biology, at the UCDHSC. He conducts research with Dr. Cios in the area of mathematical modelling (hidden Markov models, clustering, fuzzy cognitive maps) for prediction of protein-protein interfaces and protein functions. He holds an MS Degree in Computer Science from the Vietnam National University and has been awarded full scholarship for his study in the USA.

Robert S. Hodges research interests include development of synthetic peptide vaccines, antimicrobial peptides, peptide/antibody inhibitors of SARS-coronavirus, development of new HPLC and CE methodology for separation of peptides/proteins. He received PhD in Biochemistry from the University of Alberta. He is currently Professor of Biochemistry and Molecular Genetics, Director of the Program in Biomolecular Structure and John Stewart Endowed Chair in Peptide Chemistry at UCDHSC. Awards include the Boehringer-Mannheim Award for outstanding research achievements in the field of Biochemistry, the Vincent Du Vigneaud Award from the American Peptide Society for outstanding achievements in peptide research. He has published over 485 papers.

Krzysztof J. Cios research is in the areas of data mining, biomedical informatics, and machine learning. He published three books, five special issues of journals, and over 150 peer-reviewed papers. He serves on several

journals editorial boards and has been the recipient of the Norbert Wiener Outstanding Paper Award, Neurocomputing Best Paper Award and Fulbright Senior Scholar Award. He received MS and PhD Degrees from the AGH University of Science and Technology, Krakow, MBA from the University of Toledo, and DSc from the Polish Academy of Sciences. He is a foreign member of the Polish Academy of Arts and Sciences.

1 Introduction

Complex peptide samples (such as a tryptic digest of proteins) are typically separated by 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 behaviour for proteomic applications (Palmbald et al., 2002, 2004; Petritis et al., 2003; Krokhin et al., 2004). The basic premise for prediction of peptide retention time is the assumption that unless a peptide is subject to conformational restraints, its chromatographic behaviour in RP-HPLC can be correlated with its amino acid composition. Thus, the first requirement for prediction of peptide retention time is to have a set of hydrophilicity/hydrophobicity coefficients for the 20 amino acids found in proteins. The most systematic study for determining the contribution of individual amino acid residues to RP-HPLC retention behaviour was carried out by Guo and coworkers where amino acid substitutions were made in a model synthetic peptide, Ac-Gly-X-X-$(Leu)_3$-$(Lys)_2$-amide (Guo et al., 1986a, 1986b). The advantage of this approach is that the frequency of occurrence of each amino acid in the 20 synthetic peptides is the same. This is not the case when a random collection of peptides are used to calculate amino acid retention time coefficients. Amino acid coefficients generated from observed RP-HPLC retention times of these peptides were used to show good predictive accuracy (correlation coefficient of 0.98 and an average error of 1.29 min using a linear gradient of 1% acetonitrile per min.) for a wide range of peptides varying in size from 2–16 residues and composition (Guo et al., 1986b). Recently Krokhin et al. (2004) introduced a predictive algorithm using the coefficients of Guo et al. (1986a, 1986b) to predict retention times of 346 tryptic peptides in the 560 to 4,000 dalton mass range from a mixture of 17 protein digests. These authors noted that

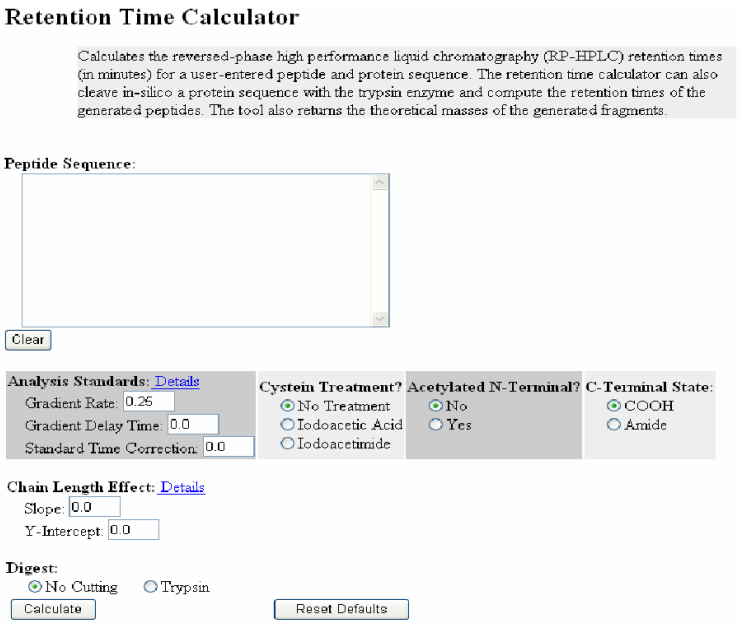
their predictions could be improved further if adjustments were made to the N-terminal coefficients (containing a free N-terminal amino group). Their results suggested that we should investigate the hydrophilicity/hydrophobicity of side-chains at the N and C-termini of peptides while varying the functional end-groups at the termini. Thus, we substituted all 20 naturally occurring amino acids at the termini (position X) where the functional end-groups at the N-terminus were N^{α}-acetyl-X- and N^{α}-amino-X and at the C-terminus, -X-C^{α}-carboxyl and -X-C^{α}-amide. These coefficients were compared to internal coefficients determined in the centre of the polypeptide chain. These results clearly showed that if you are going to predict retention times of peptides the sum of the retention coefficients ($\sum R_c$) must include internal coefficients and terminal coefficients (Tripet et al., 2007). In this paper we present a new prediction method and a searchable database to evaluate whether this unused parameter (retention time) can be successfully used to improve the assignment of peptides in LC/MS or LC/MS/MS.

2 Methods

2.1 Graphical user interface

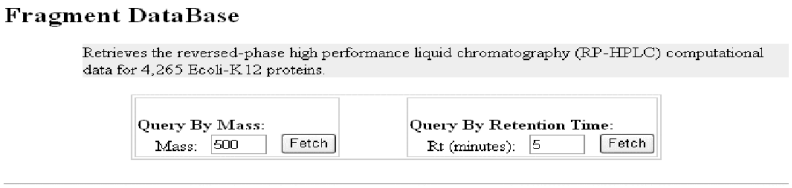
A web page was set up at the following http address: http://isl.cudenver.edu/RetenMassPrediction/. The web page displays four tabbed windows. The 'Main' window describes the purpose of the site for new users. The 'Calculator' window (Figure 1(A)) allows the user to calculate the predicted retention time of any entered sequence. The 'Fragment' and 'Protein' windows (Figure 1(B) and (C)) allow the user to search the *E. coli* database (described below) using a number of search query variables.

Figure 1 Retention time calculator, fragments search and the protein identifier

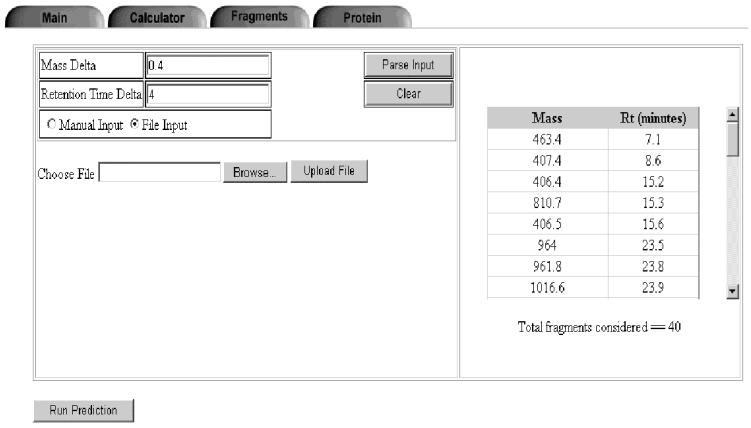


(A)

Figure 1 Retention time calculator, fragments search and the protein identifier (continued)



(B)



(C)

2.2 Retention time calculation

Retention time predictions can be calculated for individual proteins or peptides entered and defined by a user in the 'Calculator' window. Previously, investigators have been using one set of side-chain coefficients for all positions in the peptide sequences. To compare and contrast the predicted retention times of this older method with that of our newer method, three different retention time predictions are output from the program. 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, labelled "N-Term + Internal Coefficients", predicts retention time using two sets of experimentally derived coefficients, one set for the N-terminal amino acid residue containing a N^{α}-amino group 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, labelled "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 containing the N^{α}-amino group, one set for the C-terminal amino acid residue containing the C^{α}-carboxyl group 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^{α}-acetyl vs. amino or C^{α}-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 (Tripet et al., 2007).

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 NH_2 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 (Mant et al., 1989).

Table 1 shows the three sets of coefficients used principally in this study to predict peptide retention time of tryptic peptides, which contain N^{α}-amino groups and C^{α}-carboxyl groups.

Table 1 Retention time coefficients used in this study

Amino acid	*C-terminal*[a] (*-G-X-OH*) Δt_R *Gly* (*min*)	*N-terminal*[a] (*NH_2-X-G-*) Δt_R *Gly* (*min*)	*Internal*[a] (*-G-X-G-*) Δt_R *Gly* (*min*)
Trp	40.0	27.9	22.9
Phe	37.0	22.3	20.6
Leu	32.2	15.8	16.8
Ile	30.5	14.2	15.3
Met	21.2	11.8	11.2
Tyr	18.9	12.8	8.2
Val	20.0	8.1	8.6
Pro	12.2	4.5	3.6
Cys	10.8	4.3	6.0
Ala	5.0	1.5	2.8
Glu	2.1	1.4	2.3
Thr	3.6	1.9	1.5
Arg	2.5	3.0	–1.1
Asp	1.4	1.4	1.5
Gln	0.0	1.4	0.8
Gly	0.0	0.0	0.0
His	0.0	1.4	–2.4
Ser	–0.8	0.0	0.6
Lys	–1.0	1.3	–2.3
Asn	–2.3	0.0	–0.5

[a]The peptides used to generate these coefficients and their design are discussed in detail in reference (Tripet et al., 2007).

In addition to the above user-entered values, 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' (t_c), which includes the elapsed time from when the HPLC pump starts to deliver the gradient, for the gradient to travel through the instrument to the top of the column (instrument dependent), through the column (column dependent) and the beginning of gradient linearity at the detector (t_c is also dependent on the specific flow rate used) and the "Peptide Standard Correction time" (t_s) which allows the researcher to use any HPLC instrumentation, reversed-phase columns of any length and diameter, reversed-phase packing of any *n*-alkyl chain length and ligand density, and counter ion concentration differences from those which the retention time coefficients were derived. The predicted retention time (τ) for any linear gradient rate using retention time coefficients determined at a gradient rate of 0.25% acetonitrile/min is given by the equation:

$$\tau^{\mathrm{GR}} = \sum R_c(0.25/\mathrm{GR}) + t_c + t_s$$

where $\sum R_c$ is the sum of the amino acid coefficients in the peptide. The peptide standard used in this study has the sequence G A G A G V G L G G with an N^{α}-amino group and a C^{α}-carboxyl group.

$$t_s = t_{\mathrm{std}}^{\mathrm{obs}} - \sum R_c^{\mathrm{std}}(0.25/\mathrm{GR}) - t_c$$

where $t_{\mathrm{std}}^{\mathrm{obs}}$ is the observed retention time of the standard peptide and $\sum R_c^{\mathrm{std}}$ is the sum of the retention time coefficients for the peptide standard.

2.3 Database

The database created for testing 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 about each fragment is also calculated. Each 'fragment' object contains sequence, mass, predicted 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 $O(n)$.

Since mass spectrometers used routinely for LC/MS or LC/MS/MS analysis are only able to mass peptide fragments within the mass range 150–4000 Da. and peptide masses in the range 200–500 Da were considered of little informational value, only peptide fragments within the mass range 500–4000 Da were retained in the database.

2.4 Database queries

Database queries are carried out in the 'Fragment' window (Figure 1(C)). The database stores all the fragment information gathered from *E. coli-K12* proteome trypsin digest. By storing information about each fragment like mass, predicted retention time and the protein it belongs to, complex queries can be constructed for different analysis. For example, one can query by mass, predicted 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, Δrt, see description below), protein prediction for the fragments can be carried out easily (Figure 1).

2.5 Data mining of fragments

Data mining of fragments is carried out in the 'Protein' window, and is based on mass $\pm\Delta m$ and retention time $\pm\Delta rt$ (see below for a description of Δm *and* Δrt). If no value is entered for Δrt 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 one for every hit when mass and retention time match. The algorithm adds a score of bonus one 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. Mascot Generic Format (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 Δrt, retention time correction. This predicts the protein for the given set of values.

2.6 Δm and Δrt

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 $\pm\Delta m$ and $\pm\Delta rt$, respectively). This then allows one to search for a mass range for the fragment mass to match to and a corresponding retention time range.

2.7 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 $\pm\Delta m$ and $\pm\Delta rt$, 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'.

3 Analysis

3.1 Predicted retention time vs. observed retention time

To validate the program we randomly selected four tryptic peptides from one *E. coli* protein varying in length from 7–11 residues as shown below:

P1 L S D E E L K
P2 S E L V S N E L T K
P3 Y E V I S T L S K
P4 I L A Q S I E V Y Q R.

The retention times of these four peptides were predicted at three different gradient rates (0.25%, 0.50% and 1.0% acetonitrile/min on two columns containing different *n*-alkyl chain length packings (C8 and C18) and compared to the observed retention times (Table 2). This limited analysis does show that the error in prediction is linearly related to the gradient rate. At 0.25% acetonitrile/min the errors in prediction are proportional but greater than at 1% acetonitrile/min. The only reason for using shallow gradients is to enhance peak capacity and resolution for digests containing large numbers of peptides. Nevertheless, this limited study suggests errors of ±6 min, ±4 min and ±2 min for gradient rates 0.25%, 0.5% and 1%, respectively.

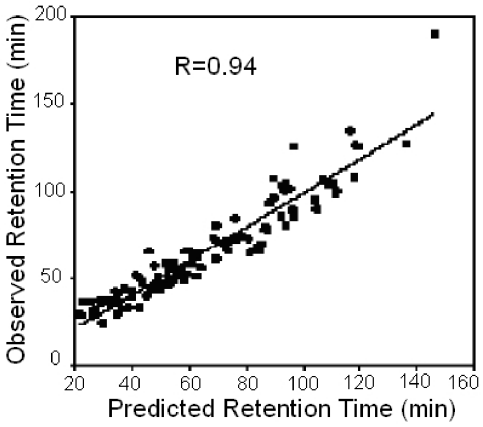
Table 2 Predicted and observed retention times of four peptides

		Gradient rate								
		0.25%			0.5%			1%		
Column	*Peptide*	*Pred.*	*Obs.*	Δ	*Pred.*	*Obs.*	Δ	*Pred.*	*Obs.*	Δ
	P1	31.8	30.7	+1.1	21.6	21.1	+0.5	14.9	14.6	+0.3
C8	P2	40.9	47.6	–6.7	26.1	29.5	–3.4	17.2	18.8	–1.6
	P3	51.0	56.6	–5.6	31.2	34.1	–2.9	19.7	21.2	–1.5
	P4	66.1	67.3	–1.2	38.7	39.2	–0.5	23.4	23.6	–0.2
				Av[a] = 3.7			*Av[a] = 1.8*			*Av[a] = 0.9*
	P1	46.8	44.8	+2.0	29.0	27.7	+1.3	18.6	17.7	+0.9
C18	P2	55.9	60.7	–4.8	33.5	35.6	–2.1	20.8	21.6	–0.8
	P3	66.0	69.0	–3.0	38.6	39.8	–1.2	23.4	23.8	–0.4
	P4	81.1	75.9	+5.2	46.1	43.0	+3.1	27.2	25.2	+2.0
				Av[a] = 3.8			*Av[a] = 1.9*			*Av[a] = 1.0*

[a]Av is the absolute average error in min.

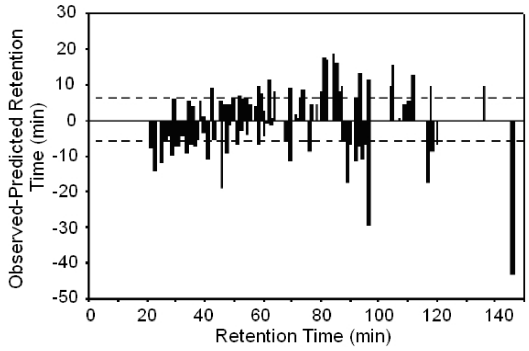
Analysis of the predicted retention time vs. observed retention time of 108 tryptic peptide fragments from ten proteins shows that there is a strong correlation (0.94) between predicted and observed times (Figure 2(A)). Further, plotting of the differences between predicted and observed retention times (about the mean) shows that approximately 80% of the peptides are within the range of ±10 min (at this very shallow gradient of 0.25% acetonitrile/min) and 50% are within the range of ±6 (Figure 2(B)). This suggests that there are other variables such as chain length, clustering of hydrophobes and conformational effects, which still need to be included in 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 re-calculated with these corrected values/changes.

Figure 2 (A) A correlation between the observed and predicted retention time values in minutes of 108 tryptic peptide fragments and (B) the differences between predicted and observed values in minutes at a gradient rate of 0.25% acetonitrile/min



(A)

Figure 2 (A) A correlation between the observed and predicted retention time values in minutes of 108 tryptic peptide fragments and (B) the differences between predicted and observed values in minutes at a gradient rate of 0.25% acetonitrile/min (continued)

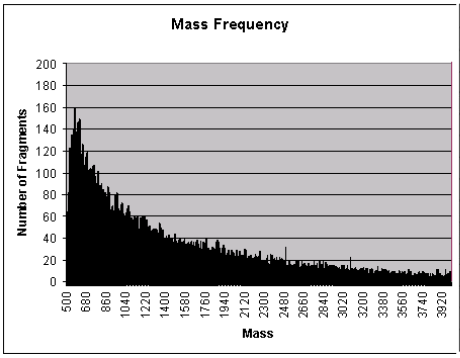


(B)

3.2 Database analysis

Plotting the frequency of peptide fragments in the database showed that the greatest numbers of fragments were observed at the lower mass range (Figure 3). Further, the number of fragments at a given fragment size decreases rapidly with increasing fragment size. For example, there are 9,393 fragments at five residues, 2,777 fragments at 15 residues and 655 fragments at 30 residues.

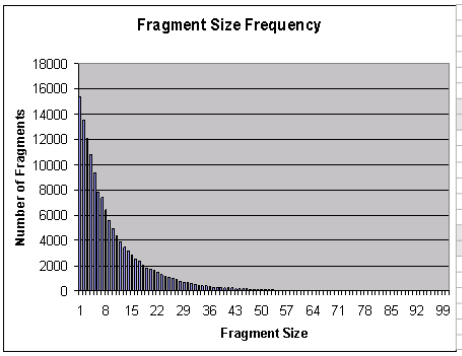
Figure 3 Number of fragments at each mass in the 500–4000 range. The chart shows ambiguous peptides with mass range 700–705



Mass	Chain1_Name	Chain1	Chain2_Name	Chain2
700	MG1655_m56_10.24	PETAQR	MG1655_m56_8478.27	QPEATR
700	MG1655_m56_1062.21	AAEAAIR	MG1655_m56_11385.27	AAIEAAR
701	MG1655_m56_58.3	NAGALTR	MG1655_m56_11729.34	NTLAGAR
701	MG1655_m56_370.8	AELLTR	MG1655_m56_7003.111	LETALR
701	MG1655_m56_370.8	AELLTR	MG1655_m56_8263.48	LAETLR
702	MG1655_m56_10.27	LDTAGAR	MG1655_m56_8380.95	LAATGDR
702	MG1655_m56_608.7	ASELQR	MG1655_m56_1000.2	SALQER
702	MG1655_m56_608.7	ASELQR	MG1655_m56_1613.39	ALQSER
702	MG1655_m56_608.7	ASELQR	MG1655_m56_2793.17	LAQESR
702	MG1655_m56_608.7	ASELQR	MG1655_m56_11362.20	SEQLAR
703	MG1655_m56_91.41	DELTAR	MG1655_m56_9061.3	LTDAER
704	MG1655_m56_198.94	NEADEK	MG1655_m56_8806.29	ENADEK
705	MG1655_m56_789.48	TWDER	MG1655_m56_12494.2	DWETR
705	MG1655_m56_789.48	TWDER	MG1655_m56_13262.25	DTWER
705	MG1655_m56_2451.16	EIAAFR	MG1655_m56_9038.29	AFEIAR

The number of fragments per mass unit was also analysed. The number of fragments within one mass unit is within the range of 80–160 Da between masses 500–1000 Da. (Figure 4) indicating a large number of possibilities for a single mass.

Figure 4 Number of fragments for each peptide length. The chart shows the retention times of sequences and the ambiguity information for the peptide fragments of the same mass and composition but different sequence (shaded)



TimeRange	Sequence	Netcharge	RetentionTime	Sequence_Length	Mass
-10_to_0	NNNGR	1	-1	5	573.26192
-10_to_0	NSSNSNR	1	-1	7	777.33655
-10_to_0	GSNNR	1	-0.5	5	546.25102
0.1_to_10	DNEEDR	-3	9.2	6	776.29366
0.1_to_10	SDDDSDK	-3	9.7	7	780.27734
0.1_to_10	DDEER	-3	9.8	5	662.25073
0.1_to_10	EEDDR	-3	9.8	5	662.25073
0.1_to_10	EGDNGQ	-2	4.7	6	618.22452
0.1_to_10	DDDGSR	-2	5.7	6	663.24598
0.1_to_10	SDADGG	-2	6.3	6	520.1765
0.1_to_10	NDDER	-2	6.4	5	647.25107
0.1_to_10	DGDNER	-2	6.4	6	704.27253
0.1_to_10	ESDNDR	-2	6.4	6	734.2831
0.1_to_10	DDEGR	-2	6.9	5	590.2296
0.1_to_10	GDEDR	-2	6.9	5	590.2296
0.1_to_10	DSDER	-2	6.9	5	620.24017
0.1_to_10	GDNEENR	-2	7	7	832.33111
0.1_to_10	DDDNK	-2	7.4	5	605.22927
0.1_to_10	QDEDR	-2	7.5	5	661.26672
0.1_to_10	SGQAQDD	-2	7.5	7	719.2722
0.1_to_10	EDENR	-2	7.6	5	661.26672

Additionally, 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 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 seven 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).

3.3 Added value of retention time data

To evaluate the degree to which the retention time data can reduce the complexity of proteins identified (No. of hits), we randomly selected a single mass (e.g., 2001 Da) and a retention time (110 min) and varied the Δ*rt* window size from ±20 min to ±1 min. The number of proteins identified from a single fragment mass and retention time decreased from 30 to 1 as the Δ*rt* window size was reduced (Table 3). At a Δ*rt* size of ±6 min (our observed retention time variance) the number of proteins identified was reduced 66% (e.g., 30–12 possibilities). Thus, retention time data can indeed be useful in reducing the possibilities in identifying a given protein.

Table 3 Effect of Δ*Rt* on the number of proteins identified from a single mass fragment

Δ *mass*	Δ *Rt*[a]	*Mass 2001 Rt 110 min no. proteins ID*	*Mass 1001 Rt 70 min no. proteins ID*	*Mass 1001 Rt 60 min no. proteins ID*	*Mass 501 Rt 30 min no. proteins ID*
1	20	30	62	86	42
1	10	17	28	51	30
1	8	14	22	46	28
1	6[b]	12	13	37	25
1	4	10	7	26	13
1	2	3	4	14	7
1	1	1	2	5	2

[a]Gradient rate 0.25% acetonitrile/min.

[b]With a Δ*Rt* of ±6 min the average reduction in proteins identified was 66% of the total found at ±20 min.

3.4 Mining with peptide fragments P1–P4

To test whether the scoring algorithm can accurately identify the correct protein from the *E. coli* database, we mined the database using the four peptide fragments P1–P4 (Figure 5). Using the four peptides and Δ*rt* = 6 min, the correct protein was identified with the highest hit score of seven. Increasing the Δ*rt* value to 20 min (largely excluding retention time selection) also showed the correct protein identified with a hit score of seven. Thus, a large number of peptide fragment masses (>4) alone are sufficient to identify the correct protein largely independent of retention time. However, when the sample size is reduced to two peptide fragment masses (e.g., P1 and P3), the value of the added retention time data is observed. For example, as the retention time error is reduced from ±20 min, ±10 min, ±8 min, to ±6 min at a gradient rate of 0.25% acetonitrile/min the number of proteins identified having a hit score of three was 4, 3, 2 and 1 protein, respectively.

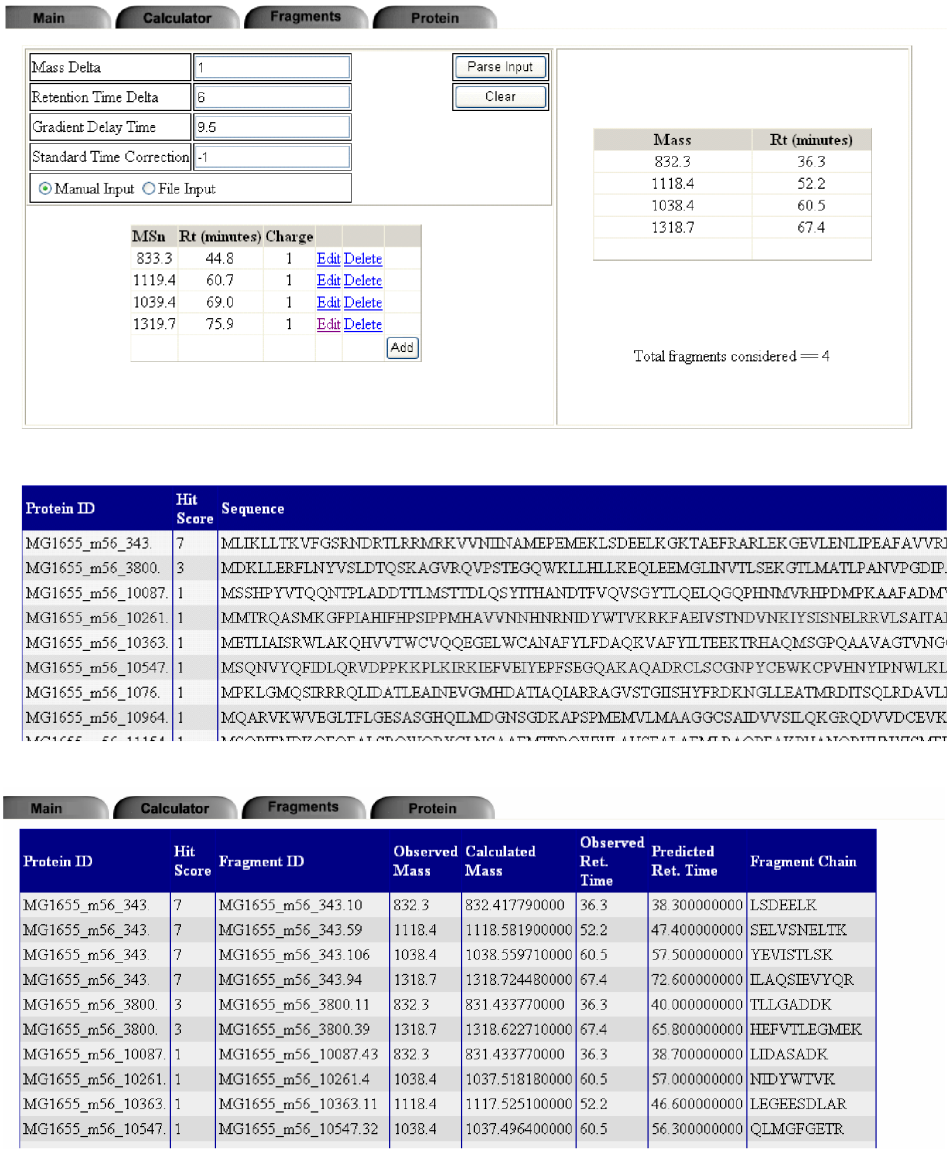
Further, sampling of different Δ*rt* sizes indicates that a useful working value appears to be between 6 min and 8 min at this very shallow gradient rate of 0.25% acetonitrile/min. In some cases this may exclude some data from scoring but the added score given to scoring more reliable data will more than offset the loss in data.

4 Discussion

The work presented in this paper demonstrates that protein identification from peptide fragments can be enhanced by utilising both mass and retention time data. Despite the seemingly large retention time delta value (±6 min) presently used the addition of this data is still sufficient to reduce the total number of protein possibilities by at least 50% based on two fragment identification.

With a larger number of fragments, the mass values alone are sufficient to discern the correct protein and thus the added retention time data does not appear to aid in the prediction.

Figure 5 Display of the proteins identified and fragment hits identified when searching peptides fragments P1–P4



Finally, in order to fully asses the benefits of protein identification based on mass and retention time values, we will need to create more complicated *E. coli* proteome samples which are presently in progress.

5 Conclusions

A database of tryptic fragments of the *E. coli* proteome with masses in the range of 500–4000 was generated. We have shown that with a minimum of two observed masses (±1 mass unit) and retention time (±6 min at a gradient rate of 0.25% acetonotrile/min) a single protein can be identified with high probability.

Acknowledgements

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

This research was supported by a grant from the National Institutes of Health R01 GM 61855 to R.S.H.

References

Guo, D., Mant, C.T., Taneja, A.K. and Hodges, R.S. (1986b) 'Prediction of peptide retention times in reversed-phase high-performance liquid chromatography II. Correlation of derived and predicted retention times of peptides', *J. Chromatography*, Vol. 359, pp.519–532.

Guo, D., Mant, C.T., Taneja, A.K., Parker, J.M.R. and Hodges, R.S. (1986a) 'Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides', *J. Chromatography*, Vol. 359, pp.499–518.

Krokhin, O., Craig, R., Spicer, V., Ens, W., Standing, K., Beavis, R. and Wilkins, J. (2004) 'An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC', *Molecular and Cellular Proteomics*, Vol. 3, pp.908–919.

Mant, C., Zhou, N. and Hodges, R.S. (1989) 'Correlation of protein retention times in reversed-phase chromatography with polypeptide chain length and hydrophobicity', *Journal of Chromatography*, Vol. 479, pp.363–375.

Palmblad, M., Ramstrom, M., Gailey, C., McCutchen-Maloney, S., Bergquist, J. and Zeller, L. (2004) 'Protein identification by liquid chromatography-mass spectrometry using retention time prediction', *Journal of Chromatography B*, Vol. 803, pp.131–135.

Palmblad, M., Ramstrom, M., Markides, K., Hakansson, P. and Bergquist, J. (2002) 'Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry', *Anal. Chem.*, Vol. 74, pp.5826–5830.

Petritis, K., Kangas, L., Ferguson, P., Anderson, G., Pasa-Tolic, L., Lipton, M., Auberry, K., Strittmatter, E., Shen, Y., Zhao, R. and Smith, R.D. (2003) 'Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses', *Anal. Chem.*, Vol. 75, pp.1039–1048.

Tripet, B., Cepeniene, D., Kovacs, J.M., Mant, C.T., Krokhin, O.V. and Hodges, R.S. (2007) 'Requirements for prediction of peptide retention time in reversed-phase HPLC: hydrophobicity/hydrophilicity of side-chains at the N and C-termini of peptides are dramatically affected by the end-groups and location', *J. Chromatography A*, Vol. 1141, pp.212–225.