Published in final edited form as:
J Proteome Res. 2011 May 6; 10(5): 2318–2329. doi:10.1021/pr1011312.

Using Ion Mobility Data to Improve Peptide Identification: Intrinsic Amino Acid Size Parameters

Stephen J. Valentine[1], **Michael A. Ewing**[1], **Jonathan M. Dilger**[1], **Matthew S. Glover**[1], **Scott Geromanos**[2], **Chris Hughes**[3], and **David E. Clemmer**[1]

[1] Department of Chemistry, Indiana University, Bloomington, IN 47405

[2] Waters Corporation, Milford, MA 01757

[3] Waters Corporation, Manchester, UK

Abstract

A new method for enhancing peptide ion identification in proteomics analyses using ion mobility data is presented. Ideally, direct comparisons of experimental drift times (t_D) with a standard mobility database could be used to rank candidate peptide sequence assignments. Such a database would represent only a fraction of sequences in protein databases and significant difficulties associated with the verification of data for constituent peptide ions would exist. A method that employs intrinsic amino acid size parameters to obtain ion mobility predictions that can be used to rank candidate peptide ion assignments is proposed. Intrinsic amino acid size parameters have been determined for doubly-charged peptide ions from an annotated yeast proteome. Predictions of ion mobilities using the intrinsic size parameters are more accurate than those obtained from a polynomial fit to t_D versus molecular weight data. More than a two-fold improvement in prediction accuracy has been observed for a group of arginine-terminated peptide ions twelve residues in length. The use of this predictive enhancement as a means to aid peptide ion identification is discussed and a simple peptide ion scoring scheme is presented.

Introduction

Since the inception of methods to identify peptide ions by tandem mass spectrometry (MS/MS) techniques,[1–3] there has been a rapid advance in mass spectrometric instrumentation development. These advances are in large part spurred by the need to increase the overall numbers of identified peptides and proteins in proteomics experiments in order to provide the necessary increased protein complement coverage for accurate and relevant comparative analyses. Over the last 15 years, improvements in mass spectrometry (MS) instrumentation have resulted in increased numbers of assigned peptide ions obtained from liquid chromatography (LC)-MS/MS experiments for complex proteomics samples;[4–6] in the characterization of human plasma digests,[7–12] numbers of assigned peptide ions in a given experiment have increased by nearly 2 orders of magnitude over this time period.

Although improvements in instrumentation sensitivity and speed have enabled increased numbers of peptides to be identified, a problem of false identification has persisted. The problem is so pervasive in the field that there has been a push to standardize proteomics reporting consisting of the establishment of guidelines for disclosure of statistical analyses used to establish the accuracy of assignments.[13] In part, instrumentation improvements lead to the intransigence of the false identification problem as lower-signal species move into

Correspondence to: David E. Clemmer.

identification range with increased analytical performance capabilities. Typically such species produce lower-quality spectra leading to suspect assignments. There is a constant need to develop methods to improve the confidence of peptide ion assignments.

To dramatically improve the accuracy of assignments in proteomics studies, the measurement of new characteristics attributable to dataset features is required. As an example consider the enabling effect of MS/MS experiments. Whereas, the precursor ion mass is insufficient to allow identification of peptide ions in complex proteomics samples, the addition of MS/MS information allows accurate assignments in many cases. A question that arises is how will the new distinguishing characteristics be produced? Some advocate chemometric approaches to elucidate distinguishing characteristics buried in proteomics datasets. For example, ongoing work consists of efforts to predict ion fragmentation distributions (including ion intensities)[14–25] as well as LC retention[26–31] in order to provide increased identification accuracy. Finally, improved separations of dataset components can be used to enhance peptide ion assignments. Examples include the use of increased mass accuracy permitting more stringent mass matching thresholds for protein database searches[32–34] as well as precursor and fragment ion intensity matching that includes the use of LC retention time profiles[35,36].

The work presented here describes the use of an additional precursor ion trait –ion mobility– to evaluate peptide ion assignments. Specifically, the use of mobility data obtained from LC-MS/MS analyses of the yeast proteome is evaluated as a means for improving peptide ion identification. Briefly, similar to ion mobility spectrometry (IMS) experiments performed previously,[37–40] peptide ion composition is related to measured ion mobilities in order to determine the general effect that the presence of specific amino acid residues have on the overall mobilities of database ions. Upon establishing this relationship for groups of peptide ions, the ability to match drift times (t_D) with peptide ions based solely on amino acid composition has been evaluated. A simple peptide ion identification scoring scheme for data that can be produced on current commercial instrumentation (Synapt HDMS, Waters) is discussed. Finally, it is noted that this work is related to that attempting to predict t_Ds of peptide ions using artificial neural networks (ANNs).[41]

Experimental

General

Data from the analysis of a yeast proteome was provided by Waters Corporation. IMS techniques,[42–46] instrumentation,[47–54] and theory,[55–59] as well as the combination of LC with IMS-MS instrumentation[60–65] have been discussed elsewhere. Here only a brief description of methods related to the collection of the tryptic digest data is presented.

800 ng of a tryptic digest of S. cerevisae was injected onto a Trapping and Nanoscale column configuration using a nanoACQUITY (Waters) UPLC system. Peptides were separated on the UPLC prior to being electrosprayed into the entrance orifice of the Synapt HDMS (Waters) instrument. Peptide ions were stored in the Trap Travelling Wave (T-Wave) located at the front of the IMS (Ion Mobility Separation) T-Wave device. Periodically, ion packets from the Trap T-Wave were pulsed into the IMS T-Wave cell where ions were separated due to their mobilities through a buffer gas (N_2 for these experiments) under the influence of a drift voltage that is rapidly transmitted along adjacent electrostatic lenses in the IMS T-Wave cell. The repetition of this voltage transmission (wave) provides periodic separation of ions according to their mobilities. Most ions have mobilities that are lower than the transmission rate of the T-Wave voltage causing them to “roll” back and be separated in subsequent waves. The residence times in the T-Wave cell can be calibrated to ion mobilities and thus to collision cross sections. After exiting the IMS

T-Wave cell, ions are transmitted through a Transfer T-Wave collision cell into a time-of-flight (TOF) MS device for mass analysis. The collision energy of the Transfer T-Wave is increased on an alternate scan basis producing approximately 10 low energy and 10 high energy spectra across each chromatographic peak.

Yeast Digest Samples

Yeast strain W303 (MATa ura3-52 leu2-3 leu2-112 trp1-1 ade2-1 his3-11 can1-100) was grown at 30 °C to exponential phase (A600 = 0.8) in rich YEPD medium (2% w/v glucose, 2% w/v bactopeptone, 1% w/v yeast extract). Cells were harvested by centrifugation and washed with water to remove any traces of growth medium. Cells were resuspended in ice-cold water and broken with glass beads using a Minibead beater (Biospec Products, Bartlesville, OK) for 40 s at 4 °C. Cell debris was pelleted in a microcentrifuge for 15 min (13,000 rpm; 4 °C) and supernatants collected for further analysis.

400 μg of protein was suspended in 44 μL of 50 mM ammonium bicarbonate solution containing 0.1% Rapigest (Waters Corporation) and heated at 80 °C for 15 minutes. Dissulfide bonds were reduced by addition of DTT (5 mM) and incubation at 60 °C for ½ an hour. Protein samples were then alkylated with addition of iodoacetamide (10 mM) and incubation at 23 °C for 1 hour in the dark. Trypsin (1:50 trypsin:protein) was added to the protein solution and the sample was incubated for 16 hours at 37 °C. Rapigest was then removed by adding TFA to a final concentration of 0.5%, incubating at 37 °C for 45 minutes and spinning down at 13000 rpm for 20 minutes.

UPLC settings

800ng of the tryptic sample was loaded onto a 180 μm × 20 mm Trapping column and washed with 30 column volumes of solvent A (99.9% H2O, 0.1% Formic acid). Peptides are separated on this column and a 75 μm × 200 mm using 1.3, 0.7 and 0.44% per minute gradient increases in solvent B (99.9% ACN, 0.1% formic acid) starting from an initial mixture of 99:1 solvent A:solvent B. A total separation time of 60, 90 and 120 minutes was used for the LC separation resulting in a total experimental time of 180, 270 and 360 minutes for the replicate runs. A flow rate of 300 $nL{\cdot}min^{-1}$ is used to perform the LC separation and the eluent is directed into a capillary ESI tip for direct electrospray into the mass spectrometer.

Mass Spectrometer Settings

To perform the mobility separation, the IMS T-Wave height is set to 40 V during transmission. The wave velocity was set at 600 m/s. These settings resulted in a total separation time of 13.7 ms. Nitrogen gas pressure in the IMS T-Wave was maintained at 3.27 mBar. The TOF mass spectrometer was operated in “V” mode with a resolving power of $>2\times10^4$ FWHM and a mass accuracy of 3 ppm RMS. MS/MS experiments were performed using the IdentityE mode.[66] Here conditions in the Transfer T-Wave located behind the IMS T-Wave cell are alternated between those that favor transmission of precursor ions (Collision Energy 0 V) and those that induce precursor ion dissociation (Collision Energy ramped from 19 to 45 V). Product ions produced under these conditions have the same chromatographic retention time and the same ion mobility as their precursor. Precursor and product ion mass spectra were acquired over the mass range 50 to 2000 amu with an acquisition rate of 0.9 s per spectrum. A total of 10,000 MS/MS spectra were generated and subjected to protein database searches using the Waters ProteinLynx Global Server (PLGS) and IdentityE software suite.

Peptide Ion Identification

Ion mobility enhanced MSE spectra were submitted to the PLGS software suite for protein database searches. Mass tolerances used for database searches were 5 ppm and 10 ppm for precursor and product ions, respectively. At least two unique peptides of greater than a 95% probability were required for a protein to be reported. A forward/reverse protein database search strategy was implemented to limit the number of proteins reported. For these datasets utilized in this study the protein false discovery rate was set to 1%.

Data Analysis

To provide the best estimation of intrinsic amino acid size parameters it was necessary to filter the datasets to group ions into those that may contain structural similarities. For the work performed here, the first filter requirement was that peptide ions be doubly charged. The second filter criterion removes all peptides with missed cleavages to allow use only of peptide ions where the location of the protons is known. Next peptide ions were divided into those containing a c-terminal arginine or lysine residue. Finally, within these two subgroups, the peptides were further divided by length (number of amino acids). Size paramterization was performed as described below for each of these groups of peptide ions. Matrix manipulation was achieved using the MATLAB software suite.[67]

Results and Discussion

Derivation of Size Parameters

To determine the contribution of each amino acid residue to the overall size of the peptide ions, those sequences estimated to exhibit similar gas-phase structures are selected (see selection criteria above and discussion below). As described previously, from ion mobility measurements for the peptide ions within a parameterization set,[37–40] it is possible to establish a system of equations relating size (ion mobility) to the amino acid composition using equation 1,

$$\sum_{j=1}^{n} X_{ij} p_j = y_i. \tag{1}$$

In equation 1, *i* and *j* represent a given peptide ion in the parameterization set (*i* = 1 to *m*, where *m* is the total number of peptides in the set) and the given amino acid residue (*j* = 1 to *n* where *n* is the number of separate amino acids), respectively. *X* represents the frequency of occurrence of the *j*th amino acid in the *i*th peptide of the parameterization set. The variable *y* is related to the ion mobility (represented here by a calibrated t_D) of the *i*th peptide ion. For these experiments *y* is calibrated to obtain a reduced t_D. Because peptide ion size is correlated to mass, it is necessary to calibrate the system such that differences in *y* within a subset of peptide ions are associated with peptide composition and sequence rather than differences in mass alone. That is, dividing the t_D of a peptide ion by that of a “model” peptide ion of the same mass (obtained from a second-order polynomial fit to the t_D versus molecular weight data) captures the variability in *y* at given masses. This variability is presumably determined largely by differences in peptide amino acid composition and sequence. Finally, because the ratio of t_D values is the same as the ratio that would be obtained for collision cross sections, values of *p* are referred to as intrinsic “size” parameters. In equation 1, *p* represents the intrinsic size parameter of the *j*th amino acid.

Because equation 1 represents a linear system of *m* equations with *n* coefficients, it can be written in matrix form as[40,68,69]

$$Xp=y, \tag{2}$$

where X is a $m \times n$ matrix, p is a vector of n components, and y is a vector of m components. It is straightforward to solve for individual intrinsic size parameters using,[68,69]

$$p=(X^T X)^{-1} X^T y. \tag{3}$$

The $m/n = 1$ diagonal of the variance-covariance matrix of the size parameters (M^p) provides the variance for the size parameter p_n where[69]

$$M^p=\frac{s}{m-n}(X^T X)^{-\mathbf{1}}, \tag{4}$$

and[69]

$$S=y^T \hat{\mathrm{r}} . \tag{5}$$

In equation 5, $\hat{\mathrm{r}}$ corresponds to the residuals ($\hat{\mathrm{r}} = y - Xp$) of the individual equations.[70] Errors representing one standard deviation can be obtained as the square root of the variance for each intrinsic size parameter. For the study presented here, the size parameters have been determined for groups of peptide ions having the same length within the lysine- and arginine-terminated subgroups (see above). The size parameters for the c-terminal residues have been maintained at the previously reported values of 1.230 and 1.150 for lysine and arginine, respectively.[37] This has been performed in order to remove any effect that might treat these parameters as “compensating” residues due to their single occurrence in every peptide ion sequence.

Figure 1A shows the values of the intrinsic amino acid size parameters obtained for doubly-charged, arginine-terminated peptide ions containing 12 amino acid residues. Several trends are worth noting. First, nonpolar aliphatic residues generally have larger intrinsic size parameters (i.e., they have a greater contribution to peptide ion size) than polar aliphatic residues. This is very similar to the trend observed for singly-charged, lysine-terminated peptides and it has been suggested that stronger interactions between the charge site and polar residues may account for the difference in size.[37] Another similarity is that the size parameters for the aromatic residues are intermediate in value to those of the nonpolar aliphatic and the polar aliphatic residues. Additionally, the size parameters for proline and glycine are relatively small. When compared to the previous work,[37] the size parameter for valine obtained from this peptide ion group is relatively large. The intrinsic size parameters for histidine and cysteine are the smallest determined for this parameterization set. Finally, it should be noted that the size parameter errors for cysteine, histidine, methionine, and tryptophan are relatively larger than those of other residues. This can be attributed to the relatively low level of occurrence of these amino acids in the peptide ion group used to obtain parameters. For example, the numbers of occurrence of these respective peptides in the 102 peptides in this group are 7, 3, 12, and 14, respectively. In comparison, alanine occurs 118 times within the same parameterization set.

Previously it has been demonstrated that intrinsic size parameters can be used to predict peptide ion collision cross sections.[37–40] The study showed that predictions improved upon restricting the sizes and types of peptide ions used to obtain the parameters. The reasoning for the improvement is that ions exhibiting similarities in length, composition (i.e., no missed cleavages), charge, and C-terminal residue (R or K) are more likely to adopt related gas-phase conformations; these similarities would be reflected in the intrinsic amino acid size parameters and thus lead to greater prediction accuracy for peptides within a subset. Indeed, in a previous study collision cross section prediction accuracy decreased by as much as a factor of two when size parameters from one parameterization set were used in cross section calculations for another set.[39] For the present study, seventeen peptide subgroups have been extracted from the annotated proteome dataset. Figure 1B shows the average size parameters obtained from each of the parameterization sets (peptides of different length) for arginine- and lysine-terminated peptides. For the former, average values were obtained from intrinsic size parameters determined for peptides having residue lengths of 7, 8, 9, 10, 11, 12, 13, and 14 to 15. The last grouping is required because of an insufficient number of peptide ions containing either 14 or 15 amino acid residues. For lysine-terminated peptides, size parameters from peptide groups with lengths of 7, 8, 9, 10, 11, 12, 13, 14, and 15 residues were obtained. Figure 1B shows that similar trends in size parameters are obtained for the different peptide ion subgroups.

Predicting Peptide Ion Drift Times

Size parameters can be used with amino acid composition to predict reduced t_Ds using equation 1. Because peptide ion t_D values are calculated for the ions used to obtain parameters, the calculations can be termed retrodictions. Previously we have shown that retrodictions are very similar in accuracy to bona fide predictions and therefore we shall use the term predictions throughout this work.[39] The predicted t_Ds can be compared with experimental values to assess the quality of the intrinsic size parameter determination for each dataset. As an example consider the peptide ion $[\mathrm{NTTIPTK+2H}]^{2+}$ from the heat shock protein SSC1. This seven-residue peptide ion has a t_D peak centered at 36.31 bins. From a polynomial fit to the t_D versus molecular weight data, it is observed that a “model” peptide of the same *m/z* (774.4 Da) would have a peak centered at a t_D of 36.65 bins. Thus the reduced t_D for $[\mathrm{NTTIPTK+2H}]^{2+}$ would be 0.991 (36.31/36.65). The predicted reduced t_D would be calculated according to equation 1 as $X_N p_N + X_T p_T + X_I p_I + X_P p_P + X_K p_K$ ($0.143 \times 0.883 + 0.429 \times 0.967 + 0.143 \times 1.003 + 0.143 \times 0.936 + 0.143 \times 1.23$). The calculated reduced t_D for this peptide is 0.993 corresponding to a drift bin value of 36.40. This is within 0.25% of the 36.31 value associated with the peak. This is significantly more accurate than the 36.65 value (0.94%) obtained from the polynomial fit to the t_D versus molecular weight data. Supplementary Table 1 shows a comparison of experimental and theoretical $\mathrm{t_D}$ values for all peptides used in this study. On average, experimental and theoretical $\mathrm{t_D}$s agree to within ±1.8%.

To better understand the efficacy of a size parameter prediction of the data, it is instructive to make comparisons to the polynomial fit for a group of peptide ions. Figure 2 shows the ratios of predicted and experimental t_Ds obtained for both the size parameter fit and the polynomial fit. These have been performed for arginine-terminated peptide ions of 12 amino acid residues in length using the size parameter values depicted in Figure 1A. In comparison, all predicted t_D values are within 8% of experimental values using the polynomial fit. All predicted t_Ds are within 5% of experimental values using the size parameters. Additionally, the data for the size parameter fit is more compressed around the unity line indicating a higher level of accuracy. This increased density of data points in this region is an indication of the t_D prediction improvement obtained when using size parameters. Another way to visualize this improvement is to compare the number of ions in the parameterization group that are accurately predicted to within ±1%. The 1% accuracy threshold has been selected as being representative of the typical experimental accuracy of ion mobility measurements.[43] Use of size parameters results in ~40% of all predictions meeting this accuracy threshold; the use of a polynomial fit to t_D versus molecular weight data results in ~18% of all predictions reaching this same level of accuracy. Thus there is more than a 2-fold improvement in predictive capabilities using the size parameters compared to the polynomial fit. This advantage exists for higher accuracy thresholds as well. For example, an improvement of a factor of ~1.7 is observed for predictions that are within 2% of experimental values. Here we note that size parameters obtained from peptide ions of this size provide the most accurate predictions. That said, the average improvement for arginine-terminated peptides of all sizes using the 1% accuracy threshold is ~50%. For all comparisons reported here, a second-order polynomial fit has been used because it has been shown to provide the greater prediction accuracy compared to higher-order polynomials and a linear least squares fit.

Although the discussion has focused on the superiority of the size parameters in predicting t_Ds to within 1% and 2% of experimental values, it is worthwhile considering the range of accuracy over which this advantage holds. Consider Figure 3 which shows the average fraction of the peptides correctly predicted as a function of accuracy threshold. Again a comparison is drawn between the prediction capabilities of the size parameter fit and those of the polynomial fit to t_D versus molecular weight data. The data shown in Figure 3 suggests that a significant advantage in predictive capabilities is attainable using intrinsic size parameters over an accuracy threshold range of ±0.5% to ±6%. At higher accuracy threshold values, both models do nearly as well in predicting t_D values.

Peptide Ion Assignments

To determine how intrinsic size parameters would aid peptide identification efforts, it is useful to consider two factors influencing the quality of the fit. This is accomplished by comparing the predictions obtained for specific peptide ions with those that would be obtained for nearly all peptide ion sequences at the same *m*/*z* values. Consider the peptide ion $[\mathrm{QAYAVSEK+2H}]^{2+}$ from the 60S ribosomal protein L4 A. Using the polynomial fit to the t_D versus molecular weight data for the eight-residue peptides, a reduced t_D for the peptide ion $[\mathrm{QAYAVSEK+2H}]^{2+}$ is determined to be 1.037. The predicted reduced t_D obtained using the appropriate intrinsic size parameters is 0.995. Thus, the prediction accuracy is ~0.041 or ~4.1%. A sampling of the complete list of lysine-terminated peptide ions ranging in length from 7 to 10 amino acids and within 0.01 Da of the precursor ion mass (894.45 Da) yields ~7.13×10^5 separate sequences. Predicted drift t_Ds for all possible peptide sequences have been computed using the intrinsic size parameters obtained from the 7-, 8-, 9-, and 10-residue, lysine-terminated peptide ion groups. It is observed that ~4% of all isobaric sequences have predicted t_Ds that are within the prediction accuracy (±4.1%) of the experimental sequence. In a sense, this prediction accuracy for incorrect peptide ion assignments can be considered a false discovery rate and will be useful in formulating a peptide ion identification scoring scheme outlined below. Thus, for this peptide ion, the predicted reduced t_D outperforms those obtained for ~96.0% of nearly all possible sequences at the same *m*/*z*.

From such an analysis of interfering sequences, one can determine the degree of overlap at different prediction accuracy thresholds. This is shown in Figure 4A. Here consider only the trace with the solid square symbols as this represents data for peptide sequences matching the mass (894.45 Da) of the peptide ion $[\mathrm{QAYAVSEK+2H}]^{2+}$. As the prediction accuracy threshold increases from 0.005 to 0.030 the fraction of total peptide ion sequences within the required threshold value for a match with the experimental value increases slowly from ~0.00 to ~0.02. Going from a prediction accuracy threshold of 0.030 to 0.040, the fraction of total sequences predicted accurately doubles to ~0.04. Above this value, the fraction of predicted sequences increases dramatically to 0.21, 0.63, and 0.88 at accuracy thresholds of 0.050, 0.060, and 0.070, respectively. Above an accuracy threshold of 0.070, the fraction of predicted sequences begins to level off approaching a value of 1 resembling a sigmoidal dependence. The data can be fitted with an expression for the sigmoidal curve intensity (I) according to,[71]

$$I=A+\frac{B-A}{1+e^{-\frac{(x-x_0)}{w}}}, \qquad 6)$$

where the variables A and B represent the minimum and maximum values of the sigmoidal curve (0 and 1 in this case), respectively. The variables x_0 and w represent the prediction accuracy threshold value associated with the inflection point of and a width factor of the sigmoidal curve, respectively. Using a prediction accuracy threshold value of ~0.060 to represent x_0 and a value of ~0.006 for w, the data for competitive assignments to the peptide ion $[QAYAVSEK+2H]^{2+}$ can be fit as shown in Figure 4A.

The comparison of overlapping competitive peptide ion assignments can be performed for other assigned peptide ions from the proteome database. For example, Figure 4A also shows data for accurately predicted interfering sequences having the same masses as the peptide ions $[EAYVPATK+2H]^{2+}$ and $[LNLFLSTK+2H]^{2+}$ from the proteins suppressor protein STM1 and isocitrate dehydrogenase, respectively. The data for competitive assignments of the former peptide ion also reveals a sigmoidal dependence albeit x_0 and w values are shifted to higher values (~0.120 and ~0.009, respectively). The curve obtained for the latter peptide ion reveals a pseudo-sigmoidal dependence where the x_0 and w values are shifted to lower values (~0.007 and ~0.003, respectively). The reduced t_Ds for the peptide ions $[EAYVPATK+2H]^{2+}$ and $[LNLFLSTK+2H]^{2+}$ are 1.104 and 1.023. Thus it is observed that as the reduced t_D increases, values for x_0 and w providing the best fit to the data increase as well. This observation is somewhat intuitive as a histogram of reduced t_Ds at a given m/z value reveals a Gaussian distribution centered about 1.000. That is, the majority of the reduced t_Ds are close to unity. Therefore, higher prediction accuracy thresholds would be required to obtain matches between competitive ion assignments and experimental features exhibiting reduced t_Ds that are significantly removed from 1.000.

To obtain a mathematical expression for a simple scoring scheme it is possible to use the data presented in Figure 4A. Examination of this data shows the dependence of a false discovery rate on two factors. The first factor is the overall prediction accuracy and the second factor is the magnitude of the reduced t_D of the experimental peak. As described above and demonstrated in Figure 4A, these two factors are correlated. One way to estimate potential false discovery rates for dataset features is to reconstruct sigmoidal curves (Figure 4A) for given reduced t_D values. As a first approximation this can be accomplished by examining the dependence of w and x_0 on reduced t_D. In Figure 4B and 4C this dependence is depicted for w and x_0, respectively. Here the dependence is derived as a function of the deviation of the reduced t_D from unity (d). The deviation is the fraction difference of the reduced t_D from the “model” peptide ion obtained from the polynomial fit to t_D versus molecular weight data. For the peptide ions $[QAYAVSEK+2H]^{2+}$, $[EAYVPATK+2H]^{2+}$, and $[LNLFLSTK+2H]^{2+}$ having reduced t_Ds of 1.037, 1.104, and 1.023 the deviation values are 0.037, 0.104, and 0.023, respectively. A linear least squares fit of the data in Figures 4B and 4C provides the dependence of the sigmoidal curve variables on d. For the w and x_0 variables this dependence is $0.0803\times d+0.0013$ and $1.1489\times d-0.0022$, respectively.

With the prediction accuracy dependencies on reduced t_D deviation established, it is possible to construct estimated false discovery rate curves that are specific for dataset features of given reduced t_Ds. This is accomplished by substituting the w and x_0 dependencies as well as values for A and B into equation 6 yielding,

$$I=\frac{1}{1+e^{-\frac{[x-(1.1489d-0.0022)]}{0.0803d+0.0013}}}. \tag{7)}$$

It is instructive to consider the false discovery rate at the limits of high- and low-confidence matches to experimental reduced t_Ds. A low-confidence assignment would consist of a small reduced t_D deviation and a large prediction accuracy threshold. Using values of d = 0 and x = 0.15 (a worst case scenario based on examination of database values), the exponential expression in equation 7 would approach zero and the fraction of competitive peptides predicted accurately becomes 1. A high-confidence assignment where d = 0.15 and x = 0, would result in prediction accuracy values approaching 0 as the exponential expression approaches 3.44×10^5.

A simple scoring scheme for aiding peptide ion identification can be devised based on equation 7. Because the power in the exponential expression in equation 7 essentially determines the false discovery rate, this expression can be used to provide a score for potential sequence matches. For example, the power expression ranges from −117.07 to 12.74 for low- and high-confidence matches, respectively. A scoring scheme can be set up of the form

$$S=\left[k-\frac{[x-(1.1489d-0.0022)]}{0.0803d+0.0013}\right]L. \tag{8)}$$

Here, k is an arbitrary variable used to shift the scoring range onto a positive scale. L is an arbitrary variable used to scale the output score. Values of 117.08 and 0.7703 for k and L, respectively, provide output scores that range from 0 to 100 for nearly all peptide sequences.

To evaluate the new scoring approach, consider the peptide ion $[\mathrm{VSGVSLLALWK+2H}]^{2+}$ from the 40S ribosomal protein S23 which has a reduced t_D of 1.047. The predicted reduced t_D for this peptide ion is 1.036. Using x = 0.011 and d = 0.047, S is determined to be 96.38. In the yeast proteome database used to derive the intrinsic size parameters (both arginine- and lysine-terminated peptide ions), there are 7 different peptide ions that are within ~1 Da of the molecular weight (1171.702 Da) of the peptide ion $[\mathrm{VSGVSLLALWK+2H}]^{2+}$. None have higher scores than the correct peptide; scores for these sequences range from 90.15 to 95.80. Here we note that caution should be used with such a scoring scheme especially when comparing values for species for which reduced t_D deviations are significantly different. That said, the results shown above for a peptide ion exhibiting moderate prediction accuracy and reduced t_D deviation are encouraging and suggest that in the future, a similar approach may be useful in helping to weed out false positive identifications by indicating more probable matches to experimental data.

Additional comparisons of peptide ion scores are presented in Table 1. Here, the scores for 10 peptide sequences (selected at random) are listed. For half of the comparisons, the assigned peptide sequence yields the highest score when compared to other database peptide sequences within ~1 Da in mass. In two other instances the assigned peptide sequence yields the second highest score. In the remaining three instances, the assigned peptide score is the median score or higher. It is instructive to consider the cases where the assigned peptide ion didn't score as highly as other sequences. For example, the peptide ion [IGTDIQDNK+2H]$^{2+}$ yields the relatively high score of 96.31. However, it is the fourth highest score within a group containing nine total peptide sequences. Scores of the three other peptide sequences range from 96.51 to 97.07. These values are very similar to that obtained for the assigned peptide ion. In this situation, several peptides that are within ~1 Da of the assigned peptide ion in mass have predicted t_Ds that are similar to the experimental t_D. As such, the clustering of such high scores does not warrant discarding the assigned peptide ion sequence. Rather, additional evidence would be required to confirm the peptide ion assignment.

It is instructive to consider the peptide ion sequences that have a higher score rank than the sequence associated with the correct identification (Table 1). Three of the assigned peptide ions have scores yielding a rank of 3 (or lower). Two of these peptide ions have the highest mass fraction of polar residues compared with all other sequences in Table 1. The third peptide ion is one of the top five ions with respect to mass fraction of polar residues. Overall these three peptide ions have a higher average mass fraction of polar residues (52.5±15.2%) compared to the other sequences (30.9±12.7%) in Table 1. Currently, no sequence correlation can be drawn between incorrect peptide ion assignments and the identified ions presumably because of the limited number of comparisons available. Additionally no correlation can be made to exact peptide composition. However, it is noted that the incorrect sequences of higher rank for all three peptides also contain a higher mass fraction of polar residues than those sequences of lower rank. Consider the peptide ion [IIENAEGSR+2H]$^{2+}$ having a mass fraction of polar residues of 46.4%. The two database peptides ions with scores of higher rank are [AQELAEATR+2H]$^{2+}$ and [VLQDSGLEK+2H]$^{2+}$. These peptide ions have mass fractions of polar residues of 49.3% and 59.4%, respectively. The average mass fraction of polar residues for the other scored peptide ions is 34.4±14.8%. This weak correlation suggests that the fraction of polar residues in peptide ion sequences can influence the scoring capability of the approach. That said, because of the limited amount of data, only a note of caution can be suggested in the scoring of such peptides. A greater elucidation of the effect of peptide ion sequence and composition on overall ion scores (and size parameters) requires the development of much larger proteome databases.

Improving Peptide Identification Capabilities Using Ion Mobilities

Several factors need to be addressed in order to improve the ability to aid peptide ion identification with ion mobility data. These include improvements in ion mobility instrumentation as well as to the method employed to determine instrinsic size parameters for different amino acids. As mentioned above, the development of instrumentation of higher resolving power would provide greater accuracy in the determination of ion mobilities and by association increased accuracy of intrinsic size parameters for different amino acids. In a related manner, higher resolving power may also allow the removal of interfering species affecting the mobility determination of peaks in proteomics mixtures. It is noted that a newer version of the Synapt HDMS system has recently been commercialized affording ~3 times greater resolving power. Additionally, careful studies of T-Wave separation parameters should be explored. It may be possible that many high-mobility species are travelling at the velocity of the voltage wave and are thus not separated as efficiently as other species.

Improvements in the determination of intrinsic size parameters may be enhanced by instrumentation developments in a different manner. For example, higher resolving power may allow the resolution of peptide ion conformer types (e.g., helices, partial helices, globules, and elongated structures). The resolution of structural types should allow increased parameterization of peptide ion subgroups. This would require the determination of correlations between peptide ion composition and (or) sequence to conformer types. It may also be necessary to employ other methods such as molecular dynamics simulations to assign structural types. Another factor that should aid the determination of conformer types is the construction of much larger databases. Many more sequence measurements would be required. As a note of caution, with larger databases comes the problem of increasing numbers of false positives. This is particularly problematic for data to be used in the determination of intrinsic size parameters. It is noted that a weighting factor can be incorporated into equation 3.[69] Such a weighting factor may include the probability score obtained from protein database searches.

It is instructive to consider the relevance of using intrinsic amino acid size parameters to validate peptide ion assignments. In a recent publication, Zubarev and Gorshkov and their coworkers described how they addressed a basic tenet of both analytical and engineering sciences, the tenet being a requirement for "…the use of a technique for a model validation materially different (complementary) from the one employed in the model creation".[72] For peptide ion identification in proteomics analyses, a verification model that is not based on a fragment ion interpretation is required. Employing retention-time modeling algorithms, the authors found many peptide sequences, even those with high scores, illustrating significant deviations from the theoretical retention times. The observation was largely attributed to the effects of chimeric spectra as opposed to bias in the independent retention-time models. Similarly the present work illustrates how predicting the mobility and the use of a statistical strategy can provide increased specificity of database search results. Therefore, the use of accurate mass, retention-time, and ion mobility all as independent metrics of peptide validation should significantly reduce the false positives in complex mixture analysis.

Conclusions

Intrinsic size parameters for amino acid residues have been determined for a variety of peptide groups obtained from a yeast proteome database. In general the size parameters are very similar to those obtained from singly-charged, lysine-terminated peptide ions indicating a degree of similarity between the types of structures (or elements of structure) formed by singly- and doubly-charged peptide ions. Additionally, the size parameters are very similar for peptides of very different lengths (from 7 to 15 residues). These size parameters have been used to predict ion mobilities (t_Ds). Predictions of t_Ds using intrinsic size parameters are more accurate than predictions obtained from polynomial fits to t_D versus molecular weight data. This ability is proposed as a means to aid peptide ion identification and a simple scoring scheme has been introduced.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge support for the development of new instrumentation by a grant from the National Institutes of Health (1RC1GM090797-01). We are also grateful to Professor Rob Beynon from the University of Liverpool and Professor Chris Grant from the University of Manchester for providing the yeast samples. Finally, we are grateful to Tim Riley of the Waters Corporation for coordinating the sharing of data and helpful discussions.

References

1. Scoble HA, Martin SA, Biemann K. Peptide sequencing by magnetic deflection tandem mass spectrometry. Biochem J. 1987; 245:621–622. [PubMed: 3663180]

2. Johnson RS, Martin SA, Biemann K. Collision induced fragmentation of (M+H)+ ions of peptides. Side chain specific sequence. Int J Mass Spectrom Ion Processes. 1998; 86:137–154.

3. Hunt DF, Yates JR III, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. Proc Natl Acad Sci USA. 1986; 83:6233–6237. [PubMed: 3462691]

4. Wolters DA, Washburn MP, Yates JR. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Anal Chem. 2001; 73:5683–5690. [PubMed: 11774908]

5. Washburn MP, Wolters D, Yates JR. Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. Nat Biotechnol. 2001; 19:242–247. [PubMed: 11231557]

6. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. J Proteome Res. 2003; 2:43–50. [PubMed: 12643542]

7. Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD, Springer DL, Pounds JG. Toward a Human Blood Serum Proteome: Analysis by Multidimensional Separation Coupled with Mass Spectrometry. Mol Cell Proteom. 2002; 1:947–955.

8. Wu SL, Choudhary G, Ramstrom M, Bergquist J, Hancock WS. Evaluation of Shotgun Sequencing for Proteomic Analysis of Human Plasma Using HPLC Coupled with either Ion Trap or Fourier Transform Mass Spectrometry. J Proteome Res. 2003; 2:383–393. [PubMed: 12938928]

9. Zhou M, Lucas DA, Chan KC, Issaq HJ, Petricoin EF, Liotta LA, Veenstra TD, Conrads TR. An Investigation into the Human Serum “Interactome”. Electrophoresis. 2004; 25:1289–1298. [PubMed: 15174051]

10. Shen Y, Jacobs JM, Camp DG II, Fang R, Moore RJ, Smith RD, Xiao W, Davis RW, Tompkins RG. Ultra-High-Efficiency Strong Cation Exchange LC/RPLC/MS/MS for High Dynamic Range Characterization of the Human Plasma Proteome. Anal Chem. 2004; 76:1134–1144. [PubMed: 14961748]

11. Rose K, Bougueleret L, Baussant T, Böhm G, Botti P, Colinge J, Cusin I, Gaertner H, Gleizes A, Heller M, Jimenez S, Johnson A, Kussmann M, Menin L, Menzel C, Ranno F, Rodriguez-Tomé P, Rogers J, Saudrais C, Villain M, Wetmore D, Bairoch A, Hochstrasser D. Industrial-Scale Proteomics: From Liters of Plasma to Chemically Synthesized Proteins. Proteomics. 2004; 4:2125–2150. [PubMed: 15221774]

12. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H, Paik Y, Yoo J, Ping P, Pounds J, Adkins J, Qian X, Wang R, Wasinger V, Wu CY, Zhao X, Zeng R, Archakov A, Tsugita A, Beer I, Pandey A, Pisano M, Andrews P, Tammen H, Speicher DW, Hanash SM. Overview of the HUPO Plasma Proteome Project: Results from the Pilot Phase with 35 Collaborating Laboratories and Multiple Analytical Groups, Generating a Core Dataset of 3020 Proteins and a Publicly-Available Database. Proteomics. 2005; 5:3226–3245. [PubMed: 16104056]

13. Tabb DL. What’s driving the false discovery rate? J Proteome Res. 2008; 7:45–46. [PubMed: 18081243]

14. Barton S, Richardson S, Perkins D, Bellahn I, Bryant T, Whittaker J. Using Statistical Models To Identify Factors That Have a Role in Defining the Abundance of Ions Produced by Tandem MS. Anal Chem. 2007; 79:5601–5607. [PubMed: 17579495]

15. Schutz F, Kapp E, Simpson R, Speed T. Deriving statistical models for predicting peptide tandem MS product ion intensities. Biochem Soc Trans. 2003; 31:1479–1483. [PubMed: 14641094]

16. Elias J, Gibbons F, King O, Roth F, Gygi S. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotech. 2004; 22:214–219.

17. Frank A, Pevzner P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Anal Chem. 2005; 77:964–973. [PubMed: 15858974]

18. Tanner S, Shu H, Frank A, Mumby M, Pevzner P, Bafna V. InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. Anal Chem. 2005; 77:4626–4639. [PubMed: 16013882]

19. Wan Y, Chen T. PepHMM: A hidden Markov model based scoring function for tandem mass spectrometry. Anal Chem. 2006; 78:432–7. [PubMed: 16408924]

20. Bern M, Cai Y, Goldberg D. Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. Anal Chem. 2007; 79:1393–1400. [PubMed: 17243770]

21. Colinge J. Peptide Fragment Intensity Statistical Modeling. Anal Chem. 2007; 79:7286–7290. [PubMed: 17713966]

22. Klammer A, Reynolds S, Bilmes J, MacCoss M, Noble W. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics. 2008; 24:i348–356. [PubMed: 18586734]

23. Zhang Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. Anal Chem. 2004; 76:3908–3922. [PubMed: 15253624]

24. Zhang Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides with Three or More Charges. Anal Chem. 2005; 77:6364–6373. [PubMed: 16194101]

25. Frank AM. Predicting Intensity Ranks of Peptide Fragment Ions. J Proteome Res. 2009; 8:2226–2240. [PubMed: 19256476]

26. Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Pasa-Tolic L, Lipton MS, Auberry KJ, Strittmatter E, Shen Y, Zhao R, Smith RD. Anal Chem. 2003; 75:1039–1048. [PubMed: 12641221]

27. Strittmatter EF, Kangas LJ, Petritis K, Mottaz HM, Anderson GA, Shen Y, Jacobs JM, Camp DG II, Smith RD. Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. J Proteome Res. 2004; 3(4):760–769. [PubMed: 15359729]

28. Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA. An Improved Model for Prediction of Retention Times of Tryptic Peptides in Ion Pair Reversed-phase HPLC Its Application to Protein Peptide Mapping by Off-Line HPLC-MALDI MS. Mol Cell Proteomics. 2004; 3.9:908–919. [PubMed: 15238601]

29. Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Pasa-Tolic L, Lipton MS, Auberry KJ, Strittmatter EF, Shen Y, Zhao R, Smith RD. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. Anal Chem. 2003; 75:1039–1048. [PubMed: 12641221]

30. Palmblad M, Ramstrom M, Markides KE, Hakansson P, Bergquist J. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. Anal Chem. 2002; 74:5826–5830. [PubMed: 12463368]

31. Oh C, Zak SH, Mirzaei H, Buck C, Regnier FE, Zhang X. Neural network prediction of peptide separation in strong anion exchange chromatography. Bioinformatics. 2007; 23(1):114–118. [PubMed: 17092987]

32. Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. J Proteome Res. 2010; 9(2):1138–1143. [PubMed: 19938873]

33. Xu H, Freitas MA. A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data. BMC Bioinformatics. 2007; 8:133. [PubMed: 17448237]

34. Mann M, Kelleher NL. Precision proteomics: The case for high resolution andhigh mass accuracy. Proc Natl Acad Sci USA. 2008; 105:18132–18138. [PubMed: 18818311]

35. Levin Y, Jaros JA, Schwarz E, Bahn S. Multidimensional protein fractionation of blood proteins coupled to data-independent nanoLC–MS/MS analysis. J Protoemics. 2010; 73(3):689–695.

36. Blackburn K, Mbeunkui F, Mitra SK, Mentzel T, Goshe MB. Improving Protein and Proteome Coverage through Data-Independent Multiplexed Peptide Fragmentation. J Proteome Res. 2010; 9(7):3621–3637. [PubMed: 20450226]

37. Valentine SJ, Counterman AE, Hoaglund-Hyzer CS, Clemmer DE. Intrinsic Amino Acid Size Parameters from a Series of 113 Lysine-Terminated Tryptic Digest Peptide Ions. J Phys Chem B. 1999; 103:1203–1207.

38. Henderson SC, Li J, Counterman AE, Clemmer DE. Intrinsic Size Parameters for Val, Ile, Leu, Gln, Thr, Phe, and Trp Residues from Ion Mobility Measurements of Polyamino Acid Ions. J Phys Chem B. 1999; 103:8780–8785.

39. Valentine SJ, Counterman AE, Clemmer DE. A Database of 660 Peptide Ion Cross Sections: Use of Intrinsic Size Parameters for Bona Fide Predictions of Cross Sections. J Am Soc Mass Spectrom. 1999; 10:1188–1211. [PubMed: 10536822]

40. Counterman AE, Clemmer DE. Volumes of Individual Amino Acid Residues in Gas-Phase Peptide Ions. J Am Chem Soc. 1999; 121:4031–4039.

41. Wang B, Valentine S, Plasencia M, Raghuraman S, Zhang X. Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry. BMC Bioinformatics. 2010; 11:182. [PubMed: 20380738]

42. For a review of IMS techniques see (and references therein): St Louis RH, Hill HH. Ion Mobility Spectrometry in Analytical Chemistry. Crit Rev Anal Chem. 1990; 21:321–355.

43. For a review of IMS techniques see (and references therein): Clemmer DE, Jarrold MF. Ion Mobility Measurements and their Applications to Clusters and Biomolecules. J Mass Spectrom. 1997; 32:577–592.

44. Liu Y, Clemmer DE. Characterizing Oligosaccharides Using Injected-Ion Mobility/Mass Spectrometry. Anal Chem. 1997; 69:2504–2509. [PubMed: 21639386]

45. Liu Y, Valentine SJ, Counterman AE, Hoaglund CS, Clemmer DE. Injected-ion Mobility Analysis of Biomolecules. Anal Chem. 1997; 69:728A.

46. For a review of IMS techniques see (and references therein): Bohrer BC, Merenbloom SI, Koeniger SL, Hilderbrand AE, Clemmer DE. Biomolecule Analysis by Ion Mobility Spectrometry. Annu Rev Anal Chem. 2008; 1(10):1–10.

47. Wittmer D, Luckenbill BK, Hill HH, Chen YH. Electrospray Ionization Ion Mobility Spectrometry. Anal Chem. 1994; 66:2348–2355.

48. Hoaglund CS, Valentine SJ, Sporleder CR, Reilly JP, Clemmer DE. Three-Dimensional Ion Mobility/TOFMS Analysis of Electrosprayed Biomolecules. Anal Chem. 1998; 70:2236–2242. [PubMed: 9624897]

49. Gillig KJ, Ruotolo B, Stone EG, Russell DH, Fuhrer K, Gonin M, Schultz AJ. Coupling High-Pressure MALDI with Ion Mobility/Orthogonal Time-of Flight Mass Spectrometry. Anal Chem. 2000; 72:3965–3971. [PubMed: 10994952]

50. Hoaglund-Hyzer CS, Li J, Clemmer DE. Mobility Labeling for Parallel CID of Ion Mixtures. Anal Chem. 2000; 72:2737–2740. [PubMed: 10905301]

51. Hoaglund-Hyzer CS, Clemmer DE. Ion Trap/Ion Mobility/Quadrupole/Time-of-Flight Mass Spectrometry for Peptide Mixture Analysis. Anal Chem. 2001; 73:177–184. [PubMed: 11199963]

52. Hoaglund-Hyzer CS, Lee YJ, Counterman AE, Clemmer DE. Coupling Ion Mobility Separations, Collisional Activation Techniques, and Multiple Stages of MS for Analysis of Complex Peptide Mixtures. Anal Chem. 2002; 74:992–1006. [PubMed: 11925002]

53. Tang K, Shvartsburg AA, Lee HN, Prior DC, Buschbach MA, Li FM, Tolmachev AV, Anderson GA, Smith RD. High-Sensitivity Ion Mobility Spectrometry/Mass Spectrometry Using Electrodynamic Ion Funnel Interfaces. Anal Chem. 2005; 77:3330–3339. [PubMed: 15889926]

54. Koeniger SL, Merenbloom SI, Valentine SJ, Jarrold MF, Udseth HR, Smith RD, Clemmer DE. An IMS-IMS Analogue of MS-MS. Anal Chem. 2006; 78:4161. [PubMed: 16771547]

55. Revercomb HE, Mason EA. Theory of Plasma Chromatography Gaseous Electrophoresis - Review. Anal Chem. 1975; 47:970–983.

56. Mason, EA.; McDaniel, EW. Transport Properties of Ions in Gases. Wiley; New York: 1988.

57. Shvartsburg AA, Jarrold MF. An exact hard-spheres scattering model for the mobilities of polyatomic ions. Chem Phys Lett. 1996; 261:86–91.

58. Mesleh MF, Hunter JM, Shvartsburg AA, Schatz GC, Jarrold MF. Structural information from ion mobility measurements: effects of the long-range potential. J Phys Chem. 1996; 100:16082–86.

59. Wyttenbach T, von Helden G, Batka JJ, Carlat D, Bowers MT. Effect of the long-range potential on ion mobility measurements. J Am Chem Soc. 1997; 8:275–82.

60. Valentine SJ, Kulchania M, Srebalus Barnes CA, Clemmer DE. Multidimensional separations of complex peptide mixtures: a combined high-performance liquid chromatography/ion mobility/time-of-flight mass spectrometry approach. Int J Mass Spectrom. 2001; 212:97–109.

61. Srebalus Barnes CA, Hilderbrand AE, Valentine SJ, Clemmer DE. Resolving Isomeric Peptide Mixtures: A Combined HPLC/Ion Mobility-TOFMS Analysis of a 4000-Component Combinatorial Library. Anal Chem. 2002; 74:26–36. [PubMed: 11795805]

62. Moon MH, Myung S, Plasencia M, Hilderbrand AE, Clemmer DE. Nanoflow LC/Ion Mobility/CID/TOF for Proteomics: Analysis of a Human Urinary Proteome. J Proteome Research. 2003; 2:589–597. [PubMed: 14692452]

63. Taraszka JA, Kurulugama R, Sowell R, Valentine SJ, Koeniger SL, Arnold RJ, Miller DF, Kaufman TC, Clemmer DE. Mapping the Proteome of Drosophila melanogaster: Analysis of Embryos and Adult Heads by LC-IMS-MS Methods. J Proteome Res. 2005; 4:1223–1237. [PubMed: 16083272]

64. Valentine SJ, Plasencia MD, Liu X, Krishnan M, Naylor S, Udseth HR, Smith RD, Clemmer DE. Toward Plasma Proteome Profiling with Ion Mobility-Mass Spectrometry. J Proteome Res. 2006; 5:2977–2984. [PubMed: 17081049]

65. Liu X, Valentine SJ, Plasencia MD, Trimpin S, Naylor S, Clemmer DE. Mapping the Human Plasma Proteome by SCX-LC-IMS-MS. J Am Soc Mass Spectrom. 2007; 18:1249–1264. [PubMed: 17553692]

66. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S. Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem. 2005; 77(7):2187–2200. [PubMed: 15801753]

67. The Mathworks, Inc. http://www.mathworks.com/products/matlab/

68. Leon, SJ. Linear Algebra with Applications. 3. Macmillan; New York: 1990. p. 208

69. http://en.wikipedia.org/wiki/Numerical_methods_for_linear_least_squares.

70. Bethea, RM.; Duran, BS.; Boullion, TL. Statistical Methods for Engineers and Scientists. 2. Marcel Dekker; New York: 1985.

71. Ledvij, M. Curve fitting made easy. http://physik.uibk.ac.at/hephy/muon/origin_curve_fitting_primer.pdf

72. Goloborodko AA, Mayerhofer C, Zubarev AR, Tarasova IA, Gorshkov AV, Zubarev RA, Gorshkov MG. Empirical approach to false discovery rate estimation in shotgun proteomics. Rapid Commun Mass Spectrom. 2010; 24:454–462. [PubMed: 20069687]

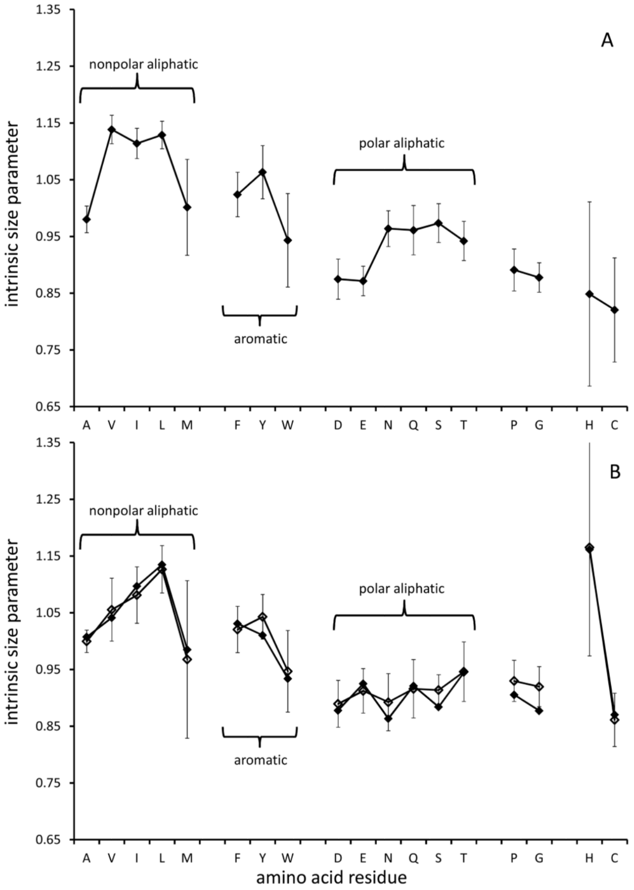

Figure 1.

Plot A shows the intrinsic amino acid size parameters derived from a group of doubly-charged peptide ions that are 12 residues in length and each contain a single, c-terminal arginine residue. Intrinsic size parameters have been grouped by types of amino acids. Size parameters for proline and glycine are presented separately in light of their propensity to disrupt α-helical structure in solution. Histidine and Cysteine size parameters are also shown separately because of their relative infrequent occurrence in the proteome database. Error bars represent one standard deviation about the mean. Plot B shows the average size parameters obtained from doubly-charged peptide ions of different residue lengths. Solid and open diamonds represent average values for lysine- and arginine-terminated peptide ions, respectively. Error bars represent one standard deviation about the mean.

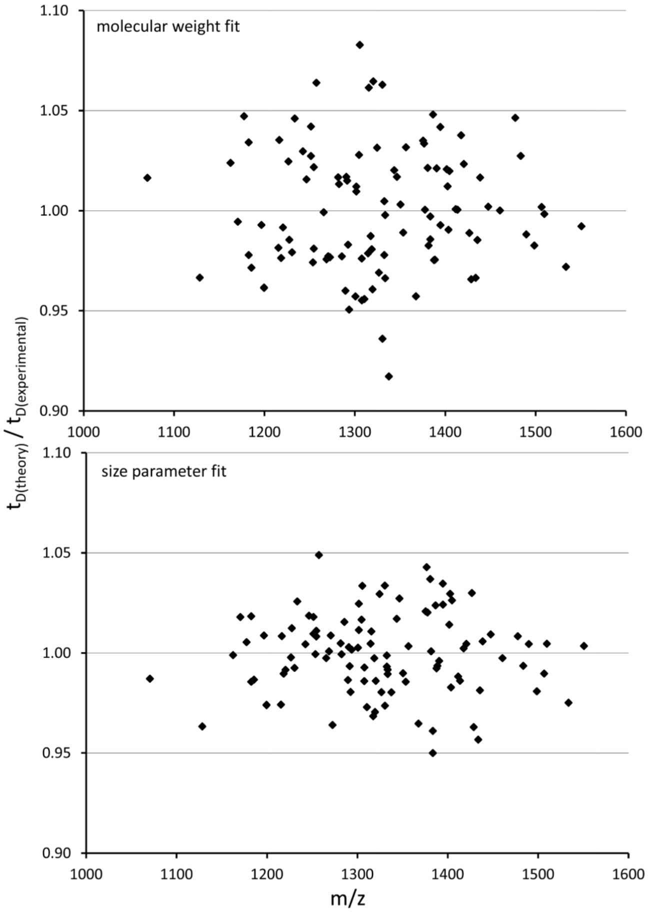

Figure 2.

Dot plots showing prediction accuracies of t_Ds for individual peptide ions as a function of molecular weight. Prediction accuracy is depicted as the ratio of the predicted t_D to the experimental t_D. The top plot shows the prediction accuracy obtained by employing a polynomial fit to t_D versus molecular weight data. The bottom plot shows the prediction accuracy obtained by using intrinsic size parameters. Data in these plots are obtained from peptide sequences containing 12 residues and a single c-terminal arginine residue.

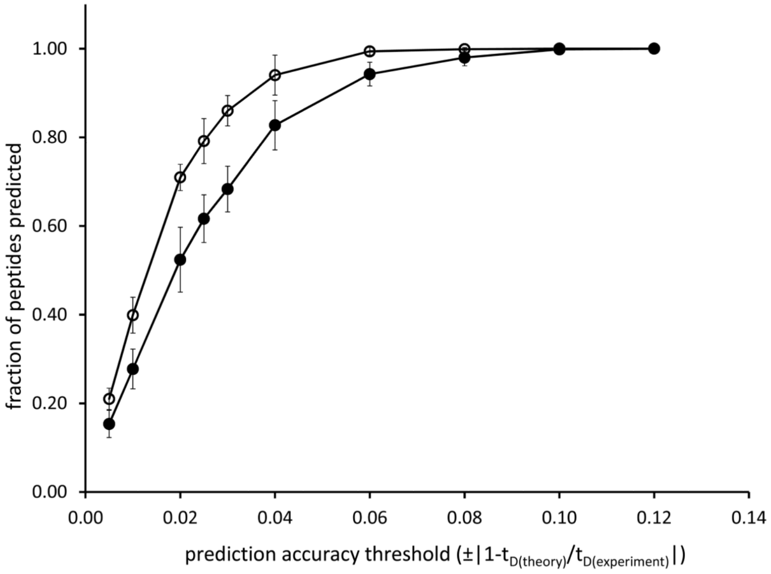

Figure 3.

Plots showing the fraction of database peptides predicted accurately for given prediction threshold values. Open- and solid-circles represent data obtained from predictions with intrinsic size parameters and predictions obtained from a polynomial fit to t_D versus molecular weight data, respectively. Data points represent average values obtained from peptide sequences of ranging in length from 7 to 15 amino acid residues. Error bars represent one standard deviation about the mean. Prediction accuracy threshold (*x*-axis) represents the deviation from the experimental values expressed as a fraction; a prediction accuracy threshold of *x*=0.03 would provide the fraction of all database peptides predicted to within ±3% of the experimental values.

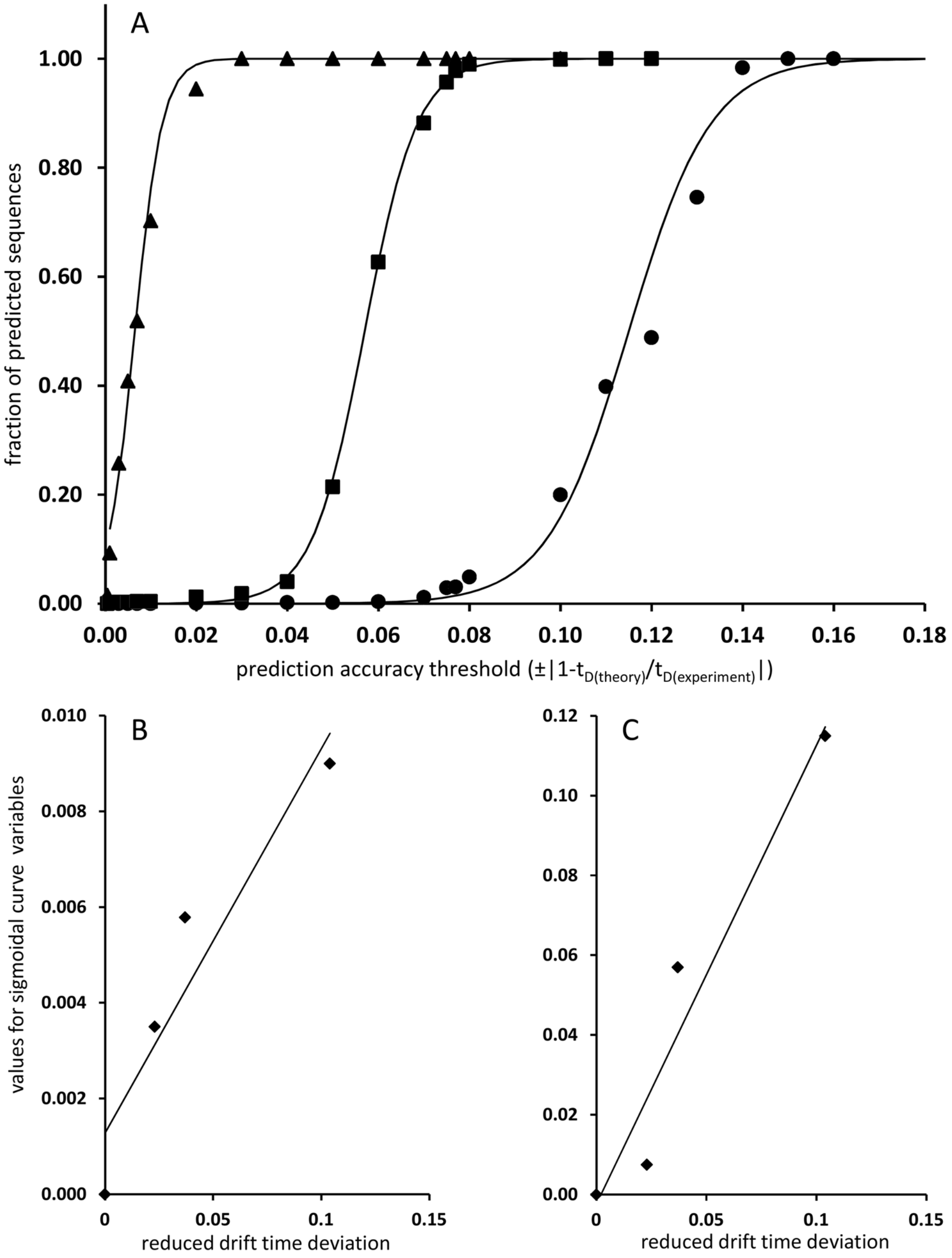

Figure 4.

Plot A shows the percentage of all possible 7-,8-,9-, and 10-residue sequences meeting or exceeding the prediction thresholds for t_Ds of the select peptide ions [EAYVPATK +2H]$^{2+}$(solid circles), [QAYAVSEK+2H]$^{2+}$(solid squares), [LNLFLSTK+2H]$^{2+}$(solid triangles). Because these values show the overlap between all possible sequences of the same *m*/*z* and the correct prediction, the value can be thought of as a false discovery rate (see text for details). Unique peptide ions from the complete list of all possible peptide ions that are within 0.01 Da of the select peptides are used in this analysis. The three different datasets have been fitted with sigmoidal curves (solid traces) using equation 6. Plot B shows the *w* values (equation 6) for the three different curves as a function of reduced t_D deviation, *d* (see text for details) as well as a zero value. Also shown is a linear least-squares fit of the data (solid line). Plot C shows the x_0 values (equation 6) for the three different curves as a function of *d* (see text for details) as well as a zero value. Also shown is a linear least-squares fit of the data (solid line).

Table 1

Scores for assigned peptide ions: comparison to sequences of similar mass.

Protein Accession[a]	Protein[b]	Peptide[c]	S[d]	Rank[e]	Number of Sequences[f]
P39741	60S ribosomal protein L35 OS Saccharomyc	QIAFPQR	92.65	1	6
P16862	6 phosphofructokinase subunit beta OS Sa	AVAEAIQAK	87.57	1	7
P53622	Coatomer subunit alpha OS Saccharomyces	IWDISGLR	95.57	1	6
P12398	Heat shock protein SSC1 mitochondrial OS	IIENAEGSR	93.23	3	7
P08524	Farnesyl pyrophosphate synthase OS Sacch	IGTDIQDNK	96.31	4	9
P38972	Phosphoribosylformylglycinamidine syntha	VLNLPSVGSK	92.63	2	5
P38910	10 kDa heat shock protein mitochondrial	TASGLYLPEK	90.00	3	5
P10614	Lanosterol 14 alpha demethylase OS Sacch	GVIYDCPNSR	90.91	1	7
Q03161	Glucose 6 phosphate 1 epimerase OS Sacch	GGIPLVFPVFGK	84.70	2	6
P48570	Homocitrate synthase cytosolic isozyme O	SDLVDLLNIYK	96.73	1	7

[a] Protein accession number obtained from the Protein Knowledgebase (UniProt KB at http://www.uniprot.org/uniprot/)

[b] First 40 characters of the protein name in the Protein Knowledgebase

[c] Tryptic peptide sequence obtained from the LC-MS/MS analysis

[d] Peptide sequence score using predicted t_D values and equation 8 (see text for details).

[e] Peptide sequence score rank compared with all other parameterization peptide sequences within ~1 Da of the precursor ion mass

[f] Total number of sequences within ~1 Da of the precursor ion mass