

Multifactorial Understanding of Ion Abundance in Tandem Mass Spectrometry Experiments

Zeeshan Fazal¹, Bruce R Southey¹, Jonathan V. Sweedler^{2,3} and Sandra L. Rodriguez-Zas^{1,3,4*}

¹Department of Animal Sciences, University of Illinois Urbana-Champaign, IL, USA

²Department of Chemistry, University of Illinois Urbana-Champaign, IL, USA

³Department of Statistics, University of Illinois Urbana-Champaign, IL, USA

⁴Institute of Genomic Biology, University of Illinois Urbana-Champaign, IL, USA

Abstract

In a bottom-up shotgun approach, the proteins of a mixture are enzymatically digested, separated, and analyzed via tandem mass spectrometry. The mass spectra relating fragment ion intensities (abundance) to the mass-to-charge are used to deduce the amino acid sequence and identify the peptides and proteins. The variables that influence intensity were characterized using a multi-factorial mixed-effects model, a ten-fold cross-validation, and stepwise feature selection on 6,352,528 fragment ions from 61,543 peptide ions. Intensity was higher in fragment ions that did not have neutral mass loss relative to any mass loss or that had a +1 charge state. Peptide ions classified for proton mobility as non-mobile had lowest intensity of all mobility levels. Higher basic residue (arginine, lysine or histidine) counts in the peptide ion and low counts in the fragment ion were associated with lower fragment ion intensities. Higher counts of proline in peptide and fragment ions were associated with lower intensities. These results are consistent with the mobile proton theory. Opposite trends between peptide and fragment ion counts and intensity may be due to the different impact of factor under consideration at different stages of the MS/MS experiment or to the different distribution of observations across peptide and fragment ion levels. Presence of basic residues at all three positions next to the fragmentation site was associated with lower fragment ion intensity. The presence of proline proximal to the fragmentation site enhanced fragmentation and had the opposite trend when located distant from the site. A positive association between fragment ion intensity and presence of sulfur residues (cysteine and methionine) on the vicinity of the fragmentation site was identified. These results highlight the multi-factorial nature of fragment ion intensity and could improve the algorithms for peptide identification and the simulation in tandem mass spectrometry experiments.

Keyword: Fragment ion; Peptide ion; MS/MS; Proton mobility; Stepwise selection; Neutral mass loss; Proline

Introduction

Identification of proteins by bottom-up shotgun proteomics begins with the enzymatic digestion of protein mixtures that are separated using techniques like liquid chromatography. The resulting separation eluent is gasified and ionized in the ion source of a mass spectrometer and undergoes a first round of Mass Spectrometry (MS) analysis. This MS analysis can be done in a high resolution Fourier Transform Mass Spectrometers (FTMS) such as the LTQ-Orbitrap mass spectrometer. The resulting peptide ions are further separated into fragment ions in a second round of MS analysis using Collision Induced Dissociation (CID) or other methods such as Electron Transfer Dissociation (ETD), Pulsed-Q-Dissociation (PQD), and high energy C-trap dissociation and Higher Energy Collision Dissociation (HCD) or a combination of these methods. The second MS analyses result in spectrum relating the intensity of the fragment ions (indicating the fragment ion abundance) to the mass/charge (m/z) ratio. Data dependent tandem Mass Spectrometry (MS/MS) provides information from the first and second MS scans that is used in database search or spectrum-to-spectrum algorithms to identify the peptides and associated proteins [1].

The accuracy of peptide identification depends on the detection of all, most, or informative fragment ions generated during the MS/MS experiment. In order for a fragment ion to be detected and used to infer a residue sequence, the fragment ion abundance must surpass the minimum intensity threshold for peak detection. Prior studies have identified several instrumental and chemical factors influence the abundance of ion formation and associated intensity [2-7]. However, limitations of the studies have impacted the replicability of the findings. Small data sets (typically less than 5000 peptides), consideration of

single factors or dimensions (i.e., analysis within fragment ion type), and simple models and mining approaches (i.e., one-way ANOVA, decision trees) that do not account for dependencies between fragment ion observations (i.e., fragment ions within a peptide, peptide ions within a protein) challenge the accuracy and precision of the results. Use of larger data sets, simultaneous consideration of multiple ion intensity factors, and multi-dimensional methodology enhances the accuracy, precision and complement the previous characterization of the associations between factors and ion intensities. An evaluation of the simultaneous impact of 230 factors on the intensity of more than 6.3 million fragment ions while accounting for the co-variation between fragment ions within a peptide and peptides within a protein was undertaken.

The first objective of this study was to gain a comprehensive and accurate characterization of the factors associated with fragment ion intensity fluctuation in MS/MS experiments. A stepwise feature selection strategy that considered each potential factor in the context of the other factors, and a hierarchical model that accommodates the dependencies between ion measurements were used to identify

*Corresponding author: Sandra L. Rodriguez-Zas, Institute of Genomic Biology, University of Illinois Urbana-Champaign, IL, USA, E-mail: rodrgzszs@illinois.edu

Received December 18, 2012; Accepted January 25, 2013; Published January 29, 2013

Citation: Fazal Z, Southey BR, Sweedler JV, Rodriguez-Zas SL (2013) Multifactorial Understanding of Ion Abundance in Tandem Mass Spectrometry Experiments. J Proteomics Bioinform 6: 023-029. doi:10.4172/jpb.1000256

Copyright: © 2013 Fazal Z, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

important factors influencing fragment ion abundance. The second objective was to confirm the factors associated with ion intensity across independent validation data sets. A ten-fold cross-validation approach was applied to compare the estimates and significance levels of the explanatory variables associated with fragment ion intensity across 10 data sets. The performance of the estimates from each data set to predict fragment ion intensity in the remaining 9 data sets was also evaluated.

Materials and Methods

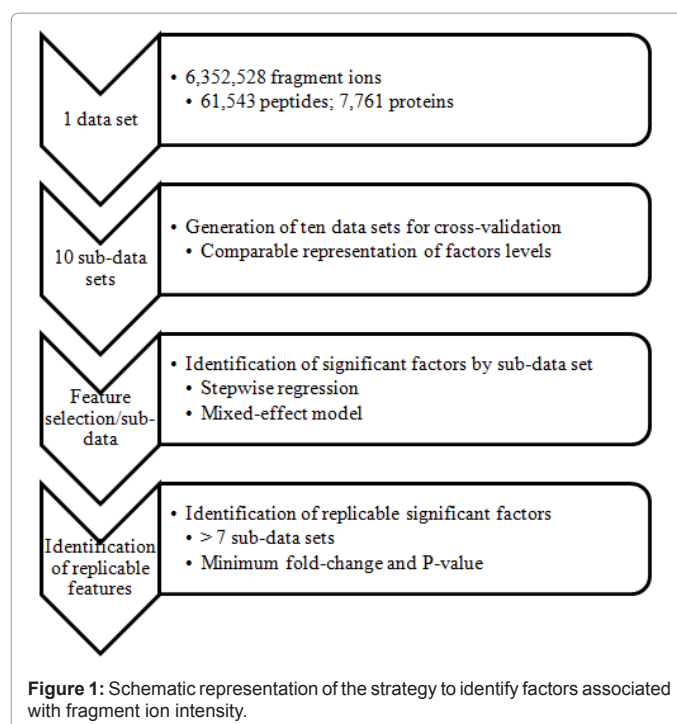
Data set

The fragment ion intensity data generated from a CID ion trap mass spectrometer was obtained from the National Institute of Standards and Technology (NIST) database (Mouse build, May 24, 2011, <http://peptide.nist.gov/>). The NIST-provided assignments of the spectra to peptides were done using the SEQUEST [8], MASCOT [9], X!TANDEM [10], and OMSSA [11] programs. The scores reported by each search engine were normalized using results of the search against a combined forward (correct) and reversed (incorrect) sequence library. Among the search engines, the best normalized score (or expectation value) was used for the annotation of the fragment ion to a peptide identifier. Identification threshold included up to two missing cleavages and one non-tryptic terminus in the peptide. Precursor and product ion tolerances of ± 2 and ± 0.8 m/z, respectively were used.

Accurate fragment ion annotation is critical to study the association between residues and fragment ion intensity. Thus, several annotation-related criteria were implemented. First, only annotated b- and y-fragment ions in the NIST repository were considered. Second, fragment ions shorter than three residues were not considered because short sequences cannot be accurately annotated and prevent the study of residue effects within three positions of the fragmentation site. Third, the NIST preferred fragment ion sequence was exclusively considered when more than one residue sequence was available for an ion fragment. Fourth, peptides assigned Post-Translational Modifications (PTMs) were excluded because these PTMs influence fragmentation and this study aims at identifying general trends. The intensity measurements from 6,352,528 fragment ions formed from 61,543 peptides corresponding to 7,761 proteins were analyzed. A \log_{10} transformation of the fragment ion intensity values resulted in a Normal distribution of the response variable. The strategy to identify replicable factors and covariates associated with fragment ion intensity encompassed four steps: 1) general data set, 2) partition of the data set into 10 sub-data sets for cross-validation, 3) modeling, analysis and feature selection of each sub-data set, and 4) identification of factors and covariates consistently associated with intensity across sub-data set. Figure 1 offers a schematic representation of the strategy. A detailed description of each step follows.

Cross-validation

A ten-fold cross-validation strategy was used to reliably identify and characterize the factors associated with ion intensity on independent data sets. At the first stage, the preprocessed NIST data was divided into 10 data sets. Any dependencies between fragment ions obtained from the same peptide ion were minimized by assigning all the fragment ions from the same peptide and protein into the same data set. All 10 data sets have comparable number of fragment ions and distribution of ions across the levels of the factors evaluated. Consideration of the representation of proteins and factors within data set prevented the exact same number of fragment ions across data sets. At the second stage, each of the 10 data sets was analyzed separately. The resulting



models including the factors significantly associated with fragment ion intensity were validated on the remaining 9 data sets. The adequacy of each of the 10 models to describe ion intensity on the other 9 data sets was evaluated by comparing the Mean Square Error (MSE).

Model

A linear mixed effects model was used to describe the fragment ion intensity. There were 14 discrete multi-level or classification explanatory variables considered: Combination of peptide and ion charge (levels: 11, 21, 22, 31, 32, 33 where the first and second number of a level indicate the peptide and fragment ion charge states, respectively and 3 denotes charge $\geq +3$); peptide ion proton mobility [3] (levels: mobile, non-mobile and partial mobile); fragment ion neutral mass loss (levels: -18 for H_2O , -17 for NH_3 , -3 for H_2S+H , -4 for C_2H_5O+H , C_2H_5O , CO_2O , or $CONH_2$, and None for no loss); fragment ion type (y- or b-ion series); combination of the number of arginine (R), histidine (H) or lysine (K) residues in the peptide and fragment ions (levels: 00, 10, 11-where the first and second number correspond to the number of residues on the peptide and fragment ions, respectively, and 1 denotes numbers ≥ 1); combination of number of proline residues (P) or number of basic (R, K, or H) residues in the peptide and fragment ions (levels: 00, 10, 11, 20, 21, 22- where the first and second number correspond to the number of residues on the peptide and fragment ions, respectively and 2 denotes numbers ≥ 2); and unique protein identifier (7,761 levels). The factors that combine the number (count) of specific residues or groups of residues (basic) in the peptide and fragment ion were evaluated instead of fitting peptide ion and fragment ion counts separately because the value at the fragment ion level is dependent on the value at the peptide level. For example, a peptide ion that has one basic residue can only form fragment ions with zero or one basic residue. The multi-level variables focused on basic (positively charged) residues and P because associations between these residues and ion intensity have been previously reported [3,5].

The 210 binary (presence or absence=1 or 0, respectively) factors

considered that represented individual residues, groups of residues that share physicochemical properties or groups of residues based on the reduced alphabet at position 1 (proximal) to 3 (most distal) from the product ion fragmentation site. The physicochemical properties were: basic or charged positive (R, K or H residues); aliphatic (A, glycine-G, valine-V, leucine-L, or isoleucine-I residues); aromatic (phenylalanine-F, tyrosine-Y, tryptophan-W residues); tiny (A, G, P, serine-S residues); small (aspartate-D, asparagine-N, threonine-T residues); large (cysteine-C, glutamate-E, F, H, I, K, L, M, P, glutamine-Q, R, V, W, Y) residues; hydrophobic (A, F, V, W residues); neutral (A, C, F, G, I, L, M, N, Q, P, S, T, V, W, Y residues); acidic or charged negative (D, E residues); hydroxyl (S, T residues); sulfur (C, M residues); amide (D, N, Q residues). The reduced residue alphabet that do not overlap completely with the previous tested properties: STPAG (S, T, P, A, and G residues-neutral, tiny with the exception of T and hydrophilic with the exception of A; and MILV (M, I, L, and V residues - neutral, large, hydrophilic with the exception of V); DENQ (D, E, Q, and N residues - acidic and polar or neutral). These binary factors are denoted X # where X represents either the single-letter amino acid code denoting the specific residue or residue group and # represents denotes the position (1, 2 or 3) relative to fragmentation. For example, A1 is the indicator of the presence or absence alanine (A) at position 1 from (i.e., next to) fragmentation and DENQ3 is the indicator of the presence or absence of an D, E, Q, or N residue located three residues from the site of fragmentation. The presence or absence residues tested was characterized relative to the fragmentation site irrespectively of being N or C termini of the fragment ion. In contrast, other studies (e.g. [Barton SJ et al. [5], Zhang Q et al. [12]) only considered whether the fragmentation is on the N or C termini of the fragment or residue. The later variable specification does not permit to study the impact of non-termini residues and confounds the effect of terminal residue with fragment ion type. The simultaneous consideration of fragment ion type and all residue positions up to three residues from the fragmentation site in this study permitted to separate specific residual effects from fragment ion type effects and evaluates the association between specific residues two or three positions from the fragmentation site (e.g. R2 or R3) or at the fragmentation site (R1).

The continuous explanatory variables considered were relative

fragment size, defined as the ratio between the number of residues in fragment ion and peptide ion, and the relative fragment mass, defined as the ratio between the mass of the fragment ion and peptide ion. These variables were evaluated because previous studies have suggested that the effect of the size or mass of the fragment ion relative on intensity is dependent on the size or mass of the fragment ion [6].

Analysis

The identification and characterization of the variables (factors and covariates) associated with fluctuations in the fragment ion intensity in each of the 10 training data sets was accomplished following a two-step strategy. In the first step, a stepwise variable selection approach was used to identify the variables that were associated with ion intensity in the linear model of each data set. In this approach, the initial model was empty and the association between each explanatory variable and the response variable (intensity) was assessed in terms of the *p*-value. The variable with the lowest (most significant) *p*-value that is also lower than an entry threshold was incorporated into the model. The association of the remaining variables with intensity, adjusted by the previously entered variable, was recomputed and again the variable with the lowest *p*-value was entered into the model. Once a variable enters into the model, the *p*-values of the previously included variables are recomputed and variables with a *p*-value higher than a stay threshold are removed from the model. The Mallows' Cp statistic [13], a measure of model fit, was computed at the end of each step. The process is repeated until no new variable can be included or removed from the model [14] or if the current model failed to improve the Cp. The entry and stay threshold were set to *p*-value<0.00005. This stringent criterion corresponds to a Bonferroni-adjusted *p*-value<0.05 that accounts for multiple testing across the multiple explanatory variables. These steps minimized over-fitting, identification of false positive associations in consideration of the 10 data sets analyzed and the influence of the large number of observations within data set. The stepwise selection was implemented in each of the 10 data sets using the SAS 9.2 GLMSELECT procedure [15].

In the second step of our analytical strategy, the final set of explanatory variables identified by the stepwise selection was evaluated in a mixed-effects model including all selected variables as fixed

Variable ^a	Level ^b	1	2	3	4	5	6	7	8	9	10
Ion Type	b	50.4	50.3	49.9	50.2	50.3	50.2	50.3	50.6	50.3	50.4
Mass loss	None	55.0	54.8	55.0	54.9	54.8	54.4	54.6	54.0	54.5	55.0
Proton	Mobile	53.5	53.1	54.2	52.7	55.7	51.3	53.2	53.0	55.3	53.5
Protein	N	777	776	776	776	776	776	776	776	776	777
Peptide	N	6087	5840	5248	5783	6380	6852	5881	7431	6056	6087
Fragment	N	634K	609K	514K	592K	656K	723K	601K	782K	621K	634K
Charge	2 1	45.0	45.0	46.5	44.6	45.4	44.2	45.5	45.3	45.1	45.0
Basic	2 1	28.4	28.2	29.0	28.8	27.7	28.3	28.2	29.4	28.8	28.4
Proline	0 0	41.2	43.2	45.3	42.4	42.2	45.1	42.0	46.1	43.8	41.2
Arginine	0 0	48.2	46.5	47.8	50.0	49.0	49.1	48.8	48.2	46.8	48.2
Lysine	1 1	38.1	37.5	38.7	40.2	39.4	41.4	39.6	39.9	37.7	38.1
Histidine	0 0	62.0	63.2	62.9	64.2	67.3	63.9	64.2	64.4	66.5	62.0

^aFragment ion Type: b-series; Neutral Mass Loss: none; Protein, Peptide and Fragment Ion N: number of proteins, peptide ions and fragment ions, respectively; Charge 2 1: charge state in the peptide ion of +2 and resulting in fragment ion of +1; Basic 2 1: two or more basic residues (H, R, and K) in the peptide ion and one basic residue in the fragment ion proline 0 0: no proline in the peptide ion and no proline in the fragment ion; arginine 0 0: no arginine in the peptide ion and no arginine in the fragment ion; Lysine 1 1: one lysine in the peptide ion and one lysine in the fragment ion; histidine 0 0: no histidine in the peptide ion and no histidine in the fragment ion.

^bLevels of the factor. First and second positions in the level denote the level of the indicator in the peptide ion and in the fragment ion, respectively.

Table 1: Percentage of fragment ions for the most representative level of selected factors that influence fragment ion abundance across the ten data sets analyzed (complete description is in Supplementary Tables 1 and 2).

Variable ^a	Level ^b	N Est ^c	Average Est ^d	Range Est ^e	Average SE ^f
Neutral Mass Loss	-17	10	2.4780	0.0760	0.0085
	-18	10	2.5416	0.0739	0.0085
	-3	10	2.4066	0.0691	0.0101
	-4	10	2.4343	0.0720	0.0086
	None	10	2.7675	0.0743	0.0084
Proton Mobility	Mobile	10	2.4823	0.0731	0.0084
	Non-Mobile	10	2.4558	0.1019	0.0093
	Partial-Mobile	10	2.6388	0.0650	0.0084
Peptide-	11	10	2.5186	0.0928	0.0089
Fragment Ion	21	10	2.5966	0.0618	0.0085
Charge	22	10	2.4287	0.0702	0.0086
	31	10	2.5719	0.0766	0.0087
	32	10	2.6178	0.0690	0.0087
	33	10	2.4201	0.0723	0.0089
Peptide-	00	10	2.6012	0.1593	0.0128
Fragment Ion	10	10	2.5509	0.0822	0.0088
Basic	11	10	2.5863	0.0738	0.0086
	20	10	2.3724	0.0771	0.0090
	21	10	2.5066	0.0538	0.0083
	22	10	2.5363	0.0688	0.0086

Variable ^a	Level ^b	N Est ^c	Average Est ^d	Range Est ^e	Average SE ^f
Peptide-	00	9	2.5076	0.0935	0.0083
Fragment Ion	10	9	2.5276	0.0660	0.0086
K	11	9	2.5534	0.0621	0.0088
Peptide-	00	10	2.5336	0.0897	0.0084
Fragment Ion	10	10	2.4919	0.0731	0.0086
R	11	10	2.5513	0.0837	0.0088
Peptide-	00	10	2.5934	0.0554	0.0085
Fragment Ion	10	10	2.5247	0.0736	0.0086
P	11	10	2.5270	0.0677	0.0085
	20	10	2.4853	0.0964	0.0090
	21	10	2.5051	0.0848	0.0087
	22	10	2.5183	0.0874	0.0086

^aVariables significant at $p < 0.00005$ in seven or more data sets and minimum difference between most extreme levels (average estimates) $\geq 0.05 \log_{10}$ (ion intensity); Neutral Mass Loss: ammonia (-17 Da) or water (-18 Da), -3 = H₂S+H (-34 Da), Cl (-35 Da), or HCl (-36 Da), and -4 = C₂H₅O+H (-46 Da), C₂H₅O (-45 Da), CO₂, or CONH₂ (-44 Da); Peptide-Fragment Ion Charge: combination of charge state or residue counts in the peptide ion and fragment ion; Peptide ion basic: combination of the number of residues in the peptide ion and fragment ion that are basic (H, R, and K); Peptide-Fragment Ion K: combination of the number of residues in the peptide ion and fragment ion that are K; Peptide-Fragment Ion P: combination of the number of residues in the peptide ion and fragment ion that are P; Peptide-Fragment Ion R: combination of the number of residues in the peptide ion and fragment ion that are R.

^bLevels of each variable. Peptide-Fragment Ion: first and second position in the level denotes the level of the indicator in the peptide ion and in the fragment ion, respectively.

^cNumber of training data sets significant at $p < 0.00005$.

^dAverage of the \log_{10} (intensity) estimates, adjusted for all other variables in the model, across all 10 data sets.

^eRange of the \log_{10} (intensity) estimates across all 10 data sets.

^fAverage standard error of the \log_{10} (intensity) estimates across all 10 data sets.

Table 2: Average estimates and standard errors of the discrete multi-level factors significantly associated with fragment ion intensity and number of data sets (out of 10) that detected the factor.

effects and protein as a random effect. The consideration of a random protein effect is critical for two reasons; to account for the covariation between ions from the same peptide and protein and, use of the correct experimental unit and degrees of freedom for variable testing. This is a novel model to study the association between variables and ion intensity. The final linear mixed effects model was fitted for each data set in a restricted maximum-likelihood framework using the SAS 9.2 MIXED procedure [16].

Results and Discussion

All 10 data sets from the NIST repository had comparable number of fragment ions and distribution of fragments across levels of the factors evaluated. Table 1 summarizes the representation of the most frequent level for selected factors across all 10 data sets. Supplementary tables 1-3 list all the descriptive statistics for each of the 10 data sets analyzed. Noteworthy distributional characteristics include that on average, across the 10 data sets, b-fragment ions comprised 50.2% of all fragment ions, 54.6% of all fragment ions with no neutral mass loss, followed by approximately 17% of fragment ions losing H₂O or NH₃. The distribution of peptide ion charge +2, +3 and +1 was 56%, 39%,

Variable ^a	N Est ^b	Average Est ^c	Range Est ^d	Average SE ^e
Binary:				
HRK3	10	-0.1264	0.0644	0.0034
HRK2	9	-0.0789	0.0284	0.0036
HRK1	8	-0.1023	0.0422	0.0042
P1	10	0.0783	0.0300	0.0030
P2	10	-0.0719	0.0281	0.0030
P3	10	-0.0798	0.0255	0.0030
H3	8	0.0492	0.0445	0.0049
Sulfur3	9	0.0825	0.1389	0.0112
Sulfur2	8	0.0350 ^f	0.0316	0.0049
Continuous:				
Relative Fragment Size	10	-0.6955	0.1748	0.0241
Relative Fragment Mass	10	0.8009	0.1693	0.0250

^aVariables significant at p -value < 0.00005 in seven or more data sets and minimum average estimates $\geq 0.05 \log_{10}$ (ion intensity); Binary explanatory variable nomenclature: X#: X denotes the residue (individual letter) or residue group (word or group of letters corresponding to the residues), and # denotes the position of the residue or group relative to the production ion fragmentation site (e.g. C2 denotes C located at the second to last position from the fragmentation terminus); HRK: group of residues that have positive charge and are large and hydrophilic (H, R, and K); P: denotes residue P that has neutral charge, is hydrophilic, imino and tiny; N: denotes residue N that has neutral charge, is hydrophilic, amide and small; H: denotes residue H that has a positive charge, is hydrophilic and large; Sulfur: group of residues that have sulfur (C, and M); HRK: group of residues that have positive charge, and are large and hydrophilic; Relative Fragment Size: ratio between the fragment ion and peptide ion length; Relative Fragment Mass: ratio between the fragment ion and peptide ion mass.

^bNumber of training data sets significant at p -value < 0.00005 .

^cAverage of the \log_{10} (intensity) estimates, adjusted for all other variables in the model, across the number of data sets in which the variable appeared. For the binary variables estimate is the difference in \log_{10} (ion intensity) between presence and absence of the residue or residue group at a specific location. For the continuous variables estimate is the change in ion intensity per unity change in the continuous variable.

^dRange of the \log_{10} (intensity) estimates across the number of data sets in which the variable appeared.

^eAverage standard error of the \log_{10} (intensity) estimates across the number of data sets in which the variable appeared.

^fBorderline chemically significant.

Table 3: Average estimates and standard errors of the binary and continuous explanatory variables significantly associated with fragment ion intensity and number of data sets (out of 10 sets) that detected the variable.

and 5%, respectively. The majority of the peptide ions were classified as mobile proton ions (53.5% of all ions) followed by partial mobile ions (43.3%). The majority of the fragment ions are from peptide ions that have two or more basic residues (25.6%) and no P (43.4%). The fragment ion length ranged from 3 to 62 residues and the relative fragment ion mass ranged from 8% to 98% of the peptide ion mass. Each data set included on average 635,252 fragment ions, from approximately 6,155 peptide ions corresponding to 776 proteins.

Table 2 summarizes the results corresponding to the multi-level discrete factors associated with fragment ion intensity. Factors were considered both statistically and chemically significantly associated with fragment ion intensity if their p -value < 0.00005 in at least 8 data sets and the minimum difference between most extreme levels was at least 0.05. This threshold corresponds to a difference between most extreme levels of 2% relative to the average fragment ion intensity. The results include the average and range of estimate and average standard error of the estimate across the 10 data sets and number of data sets including the significant factor in the final model. Table 3 summarizes the results corresponding to the binary (presence or absence) and continuous explanatory variables significantly associated with ion intensity if the p -value < 0.00005 across at least 8 data sets and the minimum average difference between presence or absence or per unit of the continuous variable was at least 0.05 \log_{10} (intensity). Supplementary table 4 lists the results corresponding to all the binary and continuous variables significantly associated with ion intensity in at least one data set at p -value < 0.00005 with no required minimum difference between levels.

Validation

The performance of the estimates from any one of the 10 (training) data sets to predict fragment ion intensities on the remaining 9 (validating) data sets was comparable to the performance in the original data set. The average difference between the training and validation MSE among the 10 data sets expressed in absolute (0.05) and relative (0.81) terms were very low. The consistency of results across data sets indicates that the training data sets (and thus estimates resulting from the analysis) were a good representation of the general fragment ion intensity data set. This conclusion is also confirmed by the narrow range of estimates of the association between factors and fragment ion intensity observed in the 10 data sets and summarized in tables 2 and 3 and in supplementary table 4.

Tables 2 and 3 highlight the complexity and multi-factorial nature of fragment ion abundance in MS/MS experiments. The association between neutral mass loss and fragment ion intensity detected in this study is supported by previous research reporting that mass spectra were negatively affected by neutral losses of molecules [17]. The present study, offered additional understanding of the impact of neutral mass losses. The fragment ion intensity was higher in the absence of neutral mass loss, distantly followed by NH_3 (-17 Da) and H_2O (-18 Da) loss, and by higher than 30 Da losses corresponding to $\text{H}_2\text{S}+\text{H}$, $\text{C}_2\text{H}_5\text{O}+\text{H}$, $\text{C}_2\text{H}_5\text{O}$, CO_2O , or CONH_2 .

In this study, the association between fragment ion intensity and the combined charge state at the peptide and fragment ion levels was considered instead of evaluating peptide and the fragment ion charges separately. The same model specification was used to evaluate the association between fragment ion intensity and combined number of specific amino acids at the peptide and ion level. The rationale for this model specification stems from the fact that the maximum charge state (or count of any amino acid) of the fragment ion is limited by the charge state of the peptide ion. A significant association between the combined

peptide-fragment ion charge state and intensity was detected on all 10 data sets. This finding is consistent with previous reports [2,3,6,18-20]. The higher intensities were observed in fragment ions that had the lowest charge state (+1 charge) relative to higher ion states ($\geq +3$ charge). A similar negative trend has been reported [21]. Additional insights gained from this study are that the average fragment ion intensity did not differ substantially across peptide ion charge states and the trend was for peptide ions that have $\geq +3$ charges to exhibit the highest intensities (Table 2).

Peptide ions classified for proton mobility as non-mobile had lowest intensity of all mobility levels (Table 2). Our results on the association between proton mobility and fragment ion intensity are consistent with the trends that we reported for peptide-fragment ion charge and with previous reports [2,3,5,6,19,22]. Migration of a charge is required for cleavage initiation and to induce fragmentation. The mobile proton theory indicates that sequence features that hinder proton mobility (such as basic residues) also hinder the fragment ion intensity because more energy is required to induce fragmentation [6,23]. The charge state contributes to the proton mobility and hindered proton mobility is associated with limited fragmentation and spectra dominated by few large peaks [6].

Higher counts of basic residues (R, K or H) in the peptide ion were associated with lower fragment ion intensities (Table 2). This result is consistent with the proton mobility theory. Low basic residue counts in the fragment ion were associated with low intensities (Table 2). Presence of basic residues has been linked to fragment ion intensity [2,4-6,18,24]. The fraction of basic residues in the fragment ion is important in determining the intensity of product ions for b-series fragment ions [24]. Our findings at fragment ion level are consistent with reports that fragment ions lacking basic amino acids were associated with lower fragment ion intensity than fragment ions that contain basic amino acids [4]. When the number of each basic residue was considered separately, the peptide-fragment ion combination having one or more K (or R) had the highest intensities and fragment ions that have one or more K (or R) have higher intensities than those with no K (or R). This trend is similar to the trend between number of basic residues in the fragment ion and intensity. At the peptide ion level, there is no clear relationship between the number of K or R and ion intensity (Table 2).

Peptide and fragment ions with no P had the highest fragment ion intensity of all peptide-fragment ion combination levels. Higher counts of P on the peptide ion were associated with lower intensities (Table 2). The association between the P count and fragment ion intensity detected in this study is consistent with previous reports [25]. Higher intensities were associated higher P counts in the fragment ion. This result confirms work that associated the presence of P at the product ion fragmentation site and intensity [5].

The model parameterization used uncovered the distinct association between fragment ion intensity and the count of basic residues or count of P. This unexpected result may be due to the different impact of these residues on fragmentation that correspond to the first and second MS stages of the MS/MS experiment. Another possible reason for the apparent opposite trend at the peptide and fragment ion levels is the different distribution of ion fragments with 0, 1, and ≥ 2 residues between levels (Table 2).

A previously unreported systematic evaluation of the association between fragment ion intensity and residues or residue groups at distances up to three positions from the fragmentation site was undertaken. Consistent negative associations between the presence of

basic or positively charged residues (R, K, or H) at all three positions next to the fragmentation site (HRK1, HRK2, HRK3) and intensity were detected (Table 3). The position of the residues is relative to the terminal residue of the fragment ion, irrespectively of being N or C termini of the fragment ion. This result is consistent with the negative association between number of basic residues in the fragment ion and intensity previously described (Table 2). Several studies have reported associations between basic residues and fragment ion intensity [2,4-6,17,19,20,22,24-26]. For singly charged peptide ions, basic amino acids have a strong effect on product ion intensities even when not adjacent to the fragmentation site [18].

The presence of P next to the fragmentation site (P1) had a positive association with fragment ion intensity, while the presence of P farther from the fragmentation site had a negative association with fragment ion intensity (Table 3). Several studies have reported an association between the presence of P and intensity [2,3,5,6,22,24]. The positive association between P and intensity is consistent with previous reports that P enhanced fragmentation when located at the N terminus of fragmentation [25-27]. The change in association with distance from the fragmentation site is consistent with reports that P augmented the likelihood of fragmentation in its proximity [18].

At position 3 from the fragmentation site, H (H3) had a positive association with fragment ion intensity (Table 3). This relationship is further supported by a positive association between H next to the fragmentation site (H1) that was statistically significant (p -value < 0.0005) and borderline chemically significant (Supplementary table 4). Our finding is also consistent with reports that the side chain of H can attack its own C terminus bond thus enhancing fragmentation at the C terminus and that fragmentation is promoted near H in many peptides [20,28,29].

Novel factors influencing fragment ion intensities were uncovered and previously reported factors were confirmed based on the simultaneous study of residues and residue groups that share properties (Table 3). The association between specific groups of residues and fragment ion intensity suggests that the shared properties are important on fragmentation patterns and ion intensity. The consistent association between residues within a group and intensity augments the statistical precision to identify significant associations. In addition, the study of individual and groups of residues allowed the disentanglement between the effects of an amino acid is due to a particular feature or due to a common physical or chemical property shared with other residues.

Several physicochemical properties not commonly associated with fragment ion intensity were detected in this study. A consistent positive and statistically significant association between product ion intensity and sulfur residues (C and M) on the vicinity of the fragmentation site was identified (Table 2). The difference in ion intensity between presence and absence of sulfur residues reached chemical significance when located three residues from fragmentation and was border line when located two residues from fragmentation (Supplementary table 4). A positive and statistically significant (p -value < 0.00005) association between N next to fragmentation (i.e., N1) and ion intensity was detected in all 10 data sets (Table 3). However, the difference in fragment ion intensity between presence and absence of E did not reach chemical significance (Supplementary table 4). This relationship is consistent with reports that E inhibits fragmentation [25]. A relationship between acidic residues and ion intensity has also been reported [2]. In this study, relative fragment ion size (or length) was negatively associated with intensity while relative fragment ion mass was positively associated with ion intensity (Table 3). The relationship between relative fragment

mass and fragment ion intensity detected in this study is consistent with previous studies [5].

Conclusions and Future Studies

Results from the present ten-fold cross-validation analysis of intensities from 6,352,528 fragment ions formed from 61,543 peptide ions corresponding to 7,761 proteins confirmed the multi-factorial nature of fragment ion abundance. Fragment ion type, peptide ion proton mobility, fragment ion neutral mass loss, and peptide and fragment ion characteristics (e.g. charge, number of basic residues, and number of P) had a significant association with fragment ion intensity. Residue groups were also associated with fragment ion intensity.

Novel findings include the strong relationship between fragment ion intensity and sulfur residues, the opposite association between P and fragment ion intensity across distance from the fragmentation site, and the opposite association between fragment ion intensity and charge (or number of basic residues) at the peptide and fragment ion levels. These findings can be used to further refine the sensitivity and specificity of the protein identification algorithms. The detection of simultaneous and significant associations between basic residue count and location relative to the fragmentation site suggest the impact of residues per se and in the context of sequence composition.

The results from this study characterize the fundamental ion intensity trends from an extensive yet particular database. Additional insights from the database considered can be gained from the analysis of unidentified ions, and from the consideration of PTMs. The former study was not undertaken because our model included the important effect of protein and unidentified ions cannot be assigned to a protein. The later study was not undertaken because the influence of PTMs in fragmentation could add a level of complexity and could have bias the fundamental characterization undertaken in this study. Based on our results, PTM effects must be evaluated conditional on ion-type, proton mobility, neutral mass loss, charge and content of basic residues and P. Analysis of intensities obtained from techniques other than CID and ion trap could shed light on trends that are technique-dependent. Likewise, the majority of the cleavages in the database analyzed are tryptic (73%) and the analysis of data from other cleavage systems could offer additional insights. Understanding the factors associated with fragment ion intensity patterns will enhance the sensitivity and specificity of database search and spectrum-to-spectrum algorithms for identification of peptides and proteins in MS experiments.

Acknowledgements

The support of NIH/NIDA (Grant Numbers: R21DA027548 and P30DA018310) and COMSATS Institute of Information Technology are greatly appreciated.

References

1. Kim MS, Kandasamy K, Chaerkady R, Pandey A (2010) Assessment of resolution parameters for CID-based shotgun proteomic experiments on the LTQ-Orbitrap mass spectrometer. *J Am Soc Mass Spectrom* 21: 1606-1611.
2. Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, et al. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal Chem* 77: 5800-5813.
3. Kapp EA, Schütz F, Reid GE, Edes JS, Moritz RL, et al. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* 75: 6251-6264.
4. Tabb DL, Huang Y, Wysocki VH, Yates JR 3rd (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem* 76: 1243-1248.
5. Barton SJ, Richardson S, Perkins DN, Bellahn I, Bryant TN, et al. (2007) Using

- statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. *Anal Chem* 79: 5601-5607.
6. Barton SJ, Whittaker JC (2009) Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom Rev* 28: 177-187.
 7. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA (1999) De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 6: 327-342.
 8. Eng J, McCormack AL, Yates J (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5: 976-989.
 9. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20: 3551-3567.
 10. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20: 1466-1467.
 11. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3: 958-964.
 12. Zhang Q, Perkins B, Tan G, Wysocki VH (2011) The role of proton bridges in selective cleavage of ser-, thr-, cys-, met-, asp-, and asn-containing peptides. *International Journal of Mass Spectrometry* 30: 108-117.
 13. Mallows CL (1973) Some comments on C_p . *Technometrics* 15: 661-675.
 14. Neter J, Kutner M, Nachtsheim C, Wasserman W (1996) *Applied Linear Statistical Models*. McGraw-Hill Companies Inc. New York.
 15. SAS Institute. Cary N The GLMSELECT procedure (experimental)2012.
 16. Pfeifer LF, Adams GP, Pearson RA, Singh J (2012) 174 ultrasound biomicroscopic image attributes of ovarian follicles in heifers. *Reprod Fertil Dev* 25: 236
 17. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 75: 2470-2477.
 18. Khatun J, Ramkissoon K, Giddings MC (2007) Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal Chem* 79: 3032-3040.
 19. Schütz F, Kapp EA, Simpson RJ, Speed TP (2003) Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem Soc Trans* 31: 1479-1483.
 20. Wysocki VH, Tsaprailis G, Smith LL, Brezi LA (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* 35: 1399-1406.
 21. Barinaga CJ, Edmonds CG, Udseth HR, Smith RD (1989) Sequence determination of multiply charged peptide molecular ions by electrospray-ionization tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* 3: 160-164.
 22. Huang Y, Tseng GC, Yuan S, Pasa-Tolic L, Lipton MS, et al. (2008) A data-mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. *J Proteome Res* 7: 70-79.
 23. Craig R, Cortens JP, Beavis RC (2005) The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom* 19: 1844-1850.
 24. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol* 22: 214-219.
 25. Zhou C, Bowler LD, Feng J (2008) A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* 9: 325.
 26. Brezi LA, Tabb DL, Yates JR 3rd, Wysocki VH (2003) Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal Chem* 75: 1963-1971.
 27. Tabb DL, Friedman DB, Ham AJ (2006) Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat Protoc* 1: 2213-2222.
 28. Tsaprailis G, Nair H, Zhong W, Kuppannan K, Futrell JH, et al. (2004) A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Anal Chem* 76: 2083-2094.
 29. Willard BB, Kinter M (2001) Effects of the position of internal histidine residues on the collision-induced fragmentation of triply protonated tryptic peptides. *J Am Soc Mass Spectrom* 12: 1262-1271.