MULTIFACTORIAL MODELING OF ION ABUNDANCE IN TANDEM MASS SPECTROMETRY EXPERIMENTS

BY

ZEESHAN FAZAL

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Animal Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:
Professor Sandra Luisa Rodriguez-Zas

ABSTRACT

Tandem mass spectrometry (MS/MS) encompasses the enzymatic digestion of proteins, usually with trypsin, followed by additional fragmentation of the resulting peptides into ions. The mass spectrum that relates the peak intensity (or abundance) to the mass-to-charge (m/z) of the ions is then used to deduce the sequence of amino acids in the ion and corresponding peptide. Attempts have been made to identify factors that influence the ion peak intensity which have been challenged by high dimensional and multi-factorial nature of the MS/MS data. The objective of this study was to identify and characterize the variables associated with ion intensity and validate the findings on separate data sets which were accomplished by implementing ten-fold cross-validation. Ion intensity measurements from 6,548,340 ion fragments formed from 61,543 peptides corresponding to 7,761 proteins obtained from the National institute of Standards and Technology were analyzed. The ion data set was divided into 10 data sub-sets and a 10-fold cross-validation analysis was undertaken. The identification and characterization of the explanatory variables in each of the 10 training data sets was accomplished by applying linear fixed-effect model framework. A stepwise variable selection approach was used to identify the explanatory variables that were associated with ion intensity. Results from the stepwise analysis were used in a final mixed effects model including protein as a random effect to allow consideration of the covariation between ion fragments from the same protein. Several factors had a significant (p-value < 0.00005) association with ion intensity across all 10 data sets. Charge state of the precursor and resulting fragment ion was associated with ion intensity. The highest intensities were observed both in low charged peptides that produce low charged ions. The numbers of basic amino acids in the peptide and resulting ion were associated with ion intensity. Peptides with no basic amino acids were associated with highest intensities; however

ions with lower number of basic amino acids had lower ion intensity relative to ions with one or more basic amino acids. The numbers of Proline (P) on the peptide and resulting ion were also associated with ion intensity. Peptides and ions with lower number of P had been associated with highest ion intensities. Several residues and group of amino acids were consistently associated with ion intensity. The property that was highly significantly associated with intensity most frequently at various locations relative to the N or C termini was residue charge. Residues that have neutral charge proximal to the N terminus were associated with higher intensities. The results from this study expand the understanding on peptide fragmentation patterns and could be used to improve algorithms for peptide identification and simulation in MS/MS experiments.

ACKNOWLEDGMENTS

First of all, I would like to thank ALLAH for providing me the ability to comprehend and analyze the task before me, the facility to keep our noses to the grindstone throughout, the sustenance to keep going at times of extreme stress and tension, and the capability to finally complete the task. I would like to show my appreciation to my adviser Dr. Sandra Rodriguez-Zas for all our fruitful discussions, without which this project would probably never have been completed. She also provided me ample opportunities to improve myself in many aspects of my fledgling scientific career, including scientific reasoning, presentation and writing skills. Additionally, I would also like to thank her for being very flexible by allowing me to pursue my interests. I would also like to thank my committee members, Dr. Jonathan Sweedler and Dr. Maria Villamil, who offered guidance and support. And last but not least, I would like to express thanks to my wife Sana Zeeshan, my lab members and my parents for their unconditional support and for putting up with me in the most trying of times.

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW[1]

Overview of Protein Identification Methods

Many years ago, the Edman degradation was used to identify protein sequences.[1] This process involves the reaction of N terminus amino acid of a peptide with a chemical reagent that cleaves the amino acid from the peptide. The resultant compound is then used to identify the amino acids present in that peptide. This process involves many chemical reactions that can take several days and the sequenced peptide cannot be longer than 50-60 residues in length. In the early 90s, the identification of proteins and peptide was revolutionized by the use of mass spectrometry (MS).[2] The identification of proteins and peptides using MS is much more sensitive and faster because this technique uses the mass of amino acids residues and can fragment peptides in seconds rather than in days. Proteins are often digested with trypsin enzyme and the resulting peptides are further subject to MS analysis. The amino acids sequence can be determined with tandem mass spectrometry (MS/MS) which uses Collision Induced Dissociation (CID) to further fragment peptides.[3]

Bottom-Up Approach

The identification of protein sequence and Post-Translational Modifications (PTMs) is widely done by a bottom-up approach. In the bottom-up approach, target proteins are digested with trypsin and the resulting peptides are further analyzed by Electrospray Ionization (ESI) or Matrix Assisted Laser Desorption/Ionization (MALDI). In these MS techniques, the protein and

[1] The part of chapter has been published as poster paper in a conference organized by IEEE. The copyright owner, IEEE has given permission to reuse text in thesis/dissertation. © 2011 IEEE. Reprinted, with permission, from Z. Fazal, B. R. Southey, A. Sadeque, J. V. Sweedler, S. L. Rodriguez-Zas. Model of ion intensity from tandem mass spectra for improved peptide identification and simulation. Nov. 2011.(59)

peptide molecular ions are put into the gas phase without fragmentation. Analysis in ESI or MALDI is done in two steps. In the first step, the intact peptide masses are determined and these peptide ions are further fragmented in gas phase to provide information on their sequence and modifications. The bottom-up approach is very useful for protein identification because tryptic peptides are easily solubilized and separated, tasks that are more difficult for the parent proteins.[4] The bottom-up approach has limitations including that only a few peptides can be detected and even fewer may provide a useful fragmentation pattern. This approach is useful for identification of PTMs and splice variants.[4] The bottom-up approach is also used in peptide mass fingerprinting and MS/MS.

In peptide mass fingerprinting, the mass of peptide acquired from an MS scan are compared to the known mass of a peptide. This mass is obtained by "in silico" cleavage of protein sequences in a database using the same specificity as the enzyme that was used in the experiment.[5] One of the disadvantages is that peptide mass fingerprinting requires pure protein or simple mixtures of proteins. The purification of proteins limits the throughput of the peptide mass fingerprinting approach. Another disadvantage is the need of several peptides to identify proteins uniquely.[5]

In MS/MS, peptide ions are isolated in the mass analyzer and subjected to further dissociation to produce product ion fragments. Masses of the fragment ions are used to deduce the amino acid sequence of the original precursor ion which forms the basis for de novo sequencing by MS/MS.[5] Mass spectrometry has enhanced its capabilities of mass determination and mass resolving by using MS/MS with chromatographic separation techniques. A common combination is liquid chromatography-MS/MS (LC-MS/MS) which separates compounds chromatographically before they are introduced to the ion source and mass spectrometer.[6]

Top-Down Approach

The top-down approach is a MS technology which preserves the post-translationally modified forms of proteins by analyzing them intact, rather than analyzing the peptides produced from the digestion of proteins by trypsin enzyme. In this approach, intact protein ions produced by ESI are put into the gas phases which are further fragmented in the mass spectrometer providing the molecular masses of both proteins and fragment ions. This analysis can provide information on sequence of protein and PTMS if sufficient numbers of fragment ions are observed.[4] The main advantage of top-down approach is doing the MS/MS experiment on an intact protein ion. This feature ensures the availability of complete sequence for analysis and helps in better characterization of protein and any PTM. In addition, the time-consuming protein digestion required for bottom-up methods is eliminated.[7]

Top-down proteomics is a relatively young field compared to bottom-up proteomics, and has some limitations. The major disadvantage of the top down approach is the complex spectra acquired from multiply charged proteins which limits the approach to isolated proteins or simple protein mixtures. Also, the top down approach does not work best with intact proteins larger than 50KDa. A greater understanding of fragmentation patterns of multiply charged ions is required, including the effect of precursor ion charge state, the role of protein primary, secondary and tertiary structure, and the contribution of PTMs to widely adopt top down approach. Finally, bioinformatics tools for top-down proteomics are in early stages of their development compared to those for bottom-up proteomics.[5]

Shotgun Proteomics

Shotgun proteomics supports further progress in the identification of proteins individually or in a complex mixture. Shotgun proteomics is named after shotgun DNA sequencing, in which

small sequencing reads are used to reconstruct the long DNA sequence.[8] Shotgun proteomics identifies the proteins from the mass spectra of their digested peptides. Complex protein mixtures are digested into peptides by sequence-specific proteolysis and resulting peptides are further analyzed by mass spectrometry. Each peptide is isolated in the mass spectrometer and characterized in MS/MS, which involves the fragmentation of peptides into many smaller fragments and measuring the mass spectrum. These MS/MS spectra are used to identify the components of peptide and therefore the parent proteins.[8]

Collision Induced Dissociation

Collision Induced Dissociation is extensively used in MS experiments for peptide and protein identification. In CID, ions collide with neutral atoms and some of an ion's kinetic energy is converted into internal energy and the ion will decompose if there is enough internal energy to break the chemical bond.[9] This method fragments peptides at their amide bond and this leads to the formation of *b*-ions if the charge is retained at the N terminus of the peptide or *y*-ions if the charge is retained on the C terminus. This technology is also known as Collisionally Activated Dissociation (CAD).

High-energy Collisions

When a precursor ion is activated to kinetic energy of one kilovolt or higher, this will excite the electron states of the precursor ion and will produce a broad internal disturbance. Theoretically, all possible fragmentations occur with some probability. High-energy CID spectra are usually the results of single collisions between precursor ions and atoms of the inert gas and results in a wide range of fragmentation pathways. Helium is mostly used as an inert gas for high-energy CID because it is not expensive and has high ionization potential but does not cause large scattering of the precursor ions.[9]

Low-energy Collisions

In low-energy CID, precursor ions have kinetic energies from a few electron volts to a few hundred electron volts. Low-energy CID activates vibrational states of electrons which produce narrow internal disturbances. The type of products ions generated from low-energy CID strongly depends on the internal energy disturbance. The type of ions observed from low-energy CID can be changed by increasing the collision energy. The resulting masses of product ions do have a strong influence on the MS/MS spectrum for low-energy CID. Low-energy CID spectra are a results of multiple collisions and reflect multiple cleavage reactions. Mostly Xenon and Argon gases are used as inert gases for low-energy CID.[9]

Type of Ions formed in Tandem Mass Spectrometry (MS/MS)

Different type of ions can be formed in low energy collisions. The most commonly observed ions in low energy collisions are *a*, *b* and *y*. The fragment ions that have a charge of 1 or more will be detected. Fragment ions that are charged at the N terminus are classified as either *a*, *b* or *c*. The fragment ions that are charged at the C terminus are classified as either *x*, *y* or *z*. The counts of residues in a fragment ion are represented by a subscript. The fragment formed by combination of *y* and *a* type ions that have only one side chain is known as immonium ion. These ions are represented with the one letter code for the respective amino acid. The ions described above are all from singly charged positive ions. There are some negative ions as well. Negative ions are the same as positive ions but with one proton per charge removed rather than added. Peptide fragmentation is not done sequentially, i.e. the fragmentation event does not proceed from N terminus one amino acid at a time down to C terminus. Fragmentation occurs randomly rather than sequentially.

Overview of Protein and Peptide Identification Methods

As of April 2012, the majority of proteomic data is generated by MS/MS. These techniques have generated hundreds to tens of thousands of fragment ion spectra per hour of data. The main computational and statistical challenges in proteomics are the assignment of fragment ion spectra to peptide sequences, the identification of proteins from known peptide sequence, and the determination of their intensities. Several algorithms have been developed to assign peptide sequences to ion spectra. These algorithms can be classified into three major categories: 1) Peptide Identification by Sequence Search Algorithms, 2) Peptide Identification by Spectral Matching, 3) Peptide Identification by *de novo* Sequencing.

Peptide Identification by Sequence Search Algorithms

A large numbers of MS/MS database search algorithms have been developed for peptide identification and mostly used commercial software are SEQUEST[10], MASCOT[11], X!TANDEM[12], and OMSSA[13]. All these commercial software works in a similar fashion. These algorithms take experimental MS/MS spectra as an input and compare it against the theoretical spectra which are constructed for peptides from the searched database to find a match.[14] Some established peptide fragmentation rules or parameters are used to calculate the theoretical fragmentation patterns. Users can select different parameters such as types of fragment ions, monoisotopic versus average mass, peptide ion charge state, parent ion mass tolerance, and enzymatic digestion constrain to restrict the database search space.[14] The score is computed according to functions that measure the degree of similarity between experimental and theoretical peptide.

Peptide identification by sequence search algorithms is schematically illustrated in Figure 1. For each experimental spectrum, SEQUEST[10] calculates the cross correlation score

(*Xcorr*) for all the candidate peptides retrieved from the database. To compute *Xcorr* score, the intensity of all the peaks in the experimental spectrum are normalized, peaks with low intensity are removed, and all m/z values in the spectrum are rounded off to the next integer to generate a new experimental spectrum (X). In the next step, the theoretical spectrum (Y) is generated for each candidate peptide in the database using fragmentation rules. Finally, a correlation function *Corr (t)* is used to give the *Xcorr* score. This *Xcorr* score is then fitted into a Weibull distribution and the resulting distribution is then used to compute a *p*-value.[15] The algorithm results in a list of ion spectra matched to peptide sequences which are always ranked according to the specific score. Only the peptides with the highest score are considered for further analysis. However, the best scoring peptides are not always correct. The reasons why SEQUEST may fail to assign correct peptides include deficiencies of the scoring scheme, low MS/MS spectrum quality, fragmentation of multiple peptide ions, presence of homologous peptides, incorrect determination of charge state and peptide mass, restricted search space of database, and presence of novel peptides.[14]

Peptide Identification by Spectral Matching

A major problem in database search algorithms is the repeated identification of the same peptides which is time consuming. In the spectra matching approach, the experimental mass spectra of correctly identified peptides are used to generate the spectral library. Any unknown mass spectrum can be identified by comparing it to all spectra in the spectral library to determine the best spectral similarity or match.[16] Recently, a number of tools such as SpectraST[17] and X!P3[18] have been developed to support peptide identification by spectral matching. This approach outperforms the database search algorithms in speed, sensitivity and error rate.[17] A major limitation of the spectral matching approach is that peptides that were absent in respective

spectral libraries can not be identified. At this time when no proteome map has been completed, spectral library matching approaches can be used more effectively as a first pass in an incremental search strategy.[19]

Peptide Identification by *de novo* Sequencing

In *de novo* sequencing, peptide sequence is read from the fragmented ion spectrum explicitly. In the initial applications of this technique, the amino acid sequences were read manually. Recently, a series of tools such as PEAKS[20] and PepNovo[21] have been developed to automatize this task. The main advantage of this approach over MS/MS database search algorithms is in the way that the spectra are identified when no exact peptide sequence is present in the databases. This technique is suitable for protein identification in species that have limited genome sequence information, or for modified proteins that include polymorphisms and PTMs.[19] The *de novo* sequencing method is computationally intense and requires high quality ion spectrum. Protein identification requires the matching of peptides extracted from MS/MS spectra using de novo algorithms, e.g. using BLAST, against the sequence of known proteins in the databases. This strategy is tedious in high throughput proteomics environment. A more accurate approach is to start with database search algorithms and then apply the *de novo* sequencing technique to the remaining unidentified ion spectra.[22]

Influence of Post-Translational Modifications on Protein and Peptide Identification

Proteins are generated through a well-defined biological mechanism called protein biosynthesis. The process begins with transcription of genes into messenger RNA (mRNA) which is later translated into a protein.[23] The activity of the protein is generally regulated by chemical modifications called PTMs. Post-translational modifications refer to changes in the

polypeptide chain that result from the enzymatic addition of a chemical group or a large molecule onto one or more amino acid side chains.[24] A protein which went through a PTM process may either exhibit a mass increase or decrease relative to its molecular weight calculated from its amino acid sequence.[25] For example, phosphorylation of a Tyrosine (Y) residue which leads to a mass increment of 80Da; an accurate spectra is required for its detection. If the mass of a modified peptide is not enough to determine the nature of PTM and its location then that particular peptide is further analyzed by MS/MS. The complexity of post-translationally modified protein sample and characteristics of the modified peptides further increase the problem. The ionization and detection efficiency of the peptide is influenced by their size and physiochemical properties, which eventually make them difficult to identify in a high background of other noise. Moreover, identification of post-translationally modified amino acid is sometimes difficult due to lack of complete ion series in the MS/MS experiment.[25]

There are several computational methods for automated annotation of PTMs in peptides.[24] These methods analyse MS and MS/MS data while taking into account the change in mass values after PTM and sometimes neutral mass losses as well as diagnostic ions for the PTM of interest. In the initial step, an initial database search is performed by these algorithms with only a few sets of modifications to minimize the search space and to avoid possible false positive PTM assignment. In the next step, the molecular weights of the unmodified proteins and resulting peptides are calculated from the protein sequences obtained by the initial search. Thus, it becomes easy to identify whether an unassigned MS/MS spectra can be matched to modified peptides while taking into account putative modifications and their change in mass values and the list of predicted and observed peptide masses.[24]

Computational tools (e.g. PTMClust) for the assignment of PTMs to peptide sequences are very useful. However, these sequence annotation tools should be used with much care because they may cause false positive results when mass accuracy or signal-to-background level of MS and MS/MS data are not enough to make unambiguous results. This is very important in modification-specific proteomics where protein identification is often done by the detection and sequencing of only few PTM peptides per protein. Peptides with multiple modifications or different types of modifications may also complicate MS and MS/MS data interpretation.[24] Mass spectrometry allows the identification and detailed characterization of post-translationally modified proteins but it relies on the accurate enrichment or separation of post-translationally modified protein and moreover, isolation of the modified peptides before carrying out the MS analysis. It is very clear that new techniques and sensitive analytical tools have to be developed to fully understand the modification specific proteomics.[26]

Databases of Tandem Mass Spectrometry (MS/MS)

NIST Library of Peptide Ion Fragmentation Spectra

The National Institute of Standards and Technology (NIST, http://peptide.nist.gov/) has developed a peptide mass spectral library which is an extension of the NIST/EPA/NIH Mass Spectral Library.[27] The NIST library makes available to the public peptide reference data generated by mass spectrometry. The NIST library contains MS/MS of peptide ions produced by digestion of proteins by trypsin enzymes. Initial assignments of spectra to peptides were done using SEQUEST[10], MASCOT[11], X!TANDEM[12], and OMSSA[13]. The Reported scores for the different search engine were normalized using results of searching against a combined forward (correct) and reversed (incorrect) sequence library. The best normalized score (or expectation

value) among the search engines were used for further processing. In general, peptides were permitted to have to up to two missed cleavages and one non-tryptic terminus (semitryptic). Parent and fragment ion tolerances of 2 and 0.8 m/z, respectively, were generally used. The spectra information in the NIST library includes mass to charge ratio (m/z) values, intensities of ion products from protonated peptide ions and the peptide to which the MS corresponds. These spectra can be used to identify or validate peptide by matching experimental peptides to spectrum in the library. Electrospray ionization in LC-MS/MS experiments was used to generate all spectra from the peptide ions in the library. The majority of spectra were generated with ion trap mass spectrometers while some spectra were generated from low energy collision cell instruments. As of March 2012, NIST includes MS from several libraries including mouse, rat, human, E. coli, and yeast.

PRIDE (PRoteomics IDEntifications database)

PRIDE is a centralized and public repository for proteomic data (http://www.ebi.ac.uk/pride/).[28] PRIDE includes mass spectrometric evidence for peptide and protein including PTMs. As of March 2012, PRIDE database currently contains information from 21,087 experiments, 8,131,516 proteins, 47,680,048 peptides, 4,498,365 unique peptides, and 283,187,118 spectra from 60 species including several model organisms. In total, 17 animal species are represented in the PRIDE database covering 84.4% of all proteins and 74.3% of all peptides. The largest number of protein and peptide correspond to human.[28]

PeptideAtlas

PeptideAtlas (http://www.peptideatlas.org/) is a multi-organism publicly available database of peptides identified in MS/MS experiments.[29] The initial build of PeptideAtlas began as a collection of peptides from a group of human and drosophila shotgun MS/MS data sets. The

database also includes description of the sample from where the peptides and proteins were obtained, how frequently the peptides were observed and how these peptides mapped to the genome.[29] Raw MS files are collected from the scientific community and pipeline tools are used to process these files. These tools perform database searching automated validation of the results using the Trans Proteomic pipeline. Sequence searching is done by SEQUEST[10] or other sequence search algorithms followed by validation of best hits with PeptideProphet, a program that models the true and false spectrum-peptide match populations, and assigns the probability of being true to each match. Results from PeptideProphet are combined using ProteinProphet to derive protein-level probabilities. SpectraSt[17], a spectral library building tool is used to make a consensus spectrum library containing all the observed peptide ions and all information is then loaded to PeptideAtlas database for use.

Prediction of Spectra

More accurate algorithms for high throughput protein identification require a better understanding of peptide fragmentation mechanisms is important. Most peptide identification methods predominantly use the m/z values of fragment ions and information given by the intensity of the MS/MS spectra is less used.

Statistical Prediction of spectra

A proteomics workflow consists of protein separation followed by identification of individual proteins by MS.[30] Algorithms such as SEQUEST[10] and MASCOT[11] have been developed for identification of peptide sequence. Both algorithms assume that peptide fragmentation occur uniformly giving equal proportion of each possible ion. However, this assumption may not apply. Several studies have investigated the factors that may influence the intensity of ions formed.

Tabb et al.[31] investigated the effects of residues on either side of the fragmentation on spectra from double charge peptides and using an ion trap mass spectrometer. The difference in intensity between the C and N terminus fragment peaks that produce *b*- and *y*-ions was taken and normalized to a range between -1 and +1 (termed the N bias of the residue) was studied. A +1 N bias indicates that the observed fragment ions were formed from the N terminus fragment and -1 N bias indicates that all the observed fragment ions were formed from the C terminus fragment. A N bias of 0 indicates that ions are formed from each terminus in equal amount. Results from Tabb et al.[31] showed that Isoleucine (I), Leucine (L) and Valine (V) have strong bias for C terminus cleavage for *y*-ions and Proline (P), Glycine (G) and Serine (S) has increased bias for N terminus cleavage.

Breci et al.[32] investigated the effect of cleavage using ion trap MS and double charged tryptic peptides ion. Breci et al.[32] analyzed cleavage N terminus to P and calculated the ratio of the intensity of ions produced at X-P (X can be any of 20 amino acids) cleavage to the intensity of all ions formed for that peptide. This ratio is called the bond cleavage ratio. Authors considered *a*-, *b*-, and *y*-ions formed for each residue by replacing X N terminus to P. Breci et al.[32] also investigated cleavage C terminus to P and observed that these cleavages are not frequently observed. Most frequent X-P cleavages were observed when X is V, Histidine (H), Aspartic Acid (D), I, and L. The least abundant cleavages occurred when X is G or P. Breci et al.[32] concluded that the cleavage at X-P is predictable and there are combinations of factors that determine how and when fragmentation occur N terminus to P.

Huang et al.[33] used CID ion trap MS/MS and divided a peptide data set into 9 strata based on charge state, basic residues and number of P residues in the peptide. A heat map was used to depict the intensity of specific type of ion formed for each pair of amino acids flanking the

cleavage site. Huang et al.[33] reported strong bias towards C terminus cleavage after acidic residues for *y*-ions and that cleavage was higher on the C terminus side of the D and Glutamic acid (E) when Arginine (R) is present at C terminus. Also I, L, and V increase the *b*-ion intensity when present C terminus to the fragmentation site when taking into account of relative proton mobility. Huang et al.[33] concluded that the amino acids I, L, and V were associated with higher *y*-ion intensity in doubly charged ions.

Khatun et al.[34] analyzed fragmentation using MALDI TOF/TOF mass spectrometry, collecting statistics using curated set of 2,459 MS/MS spectra. Khatun et al. [34] observed a strong bias for N terminus cleavage associated with P and G and for C terminus cleavage associated with V, Glutamine (Q) and H. Authors found the association of Cysteine (C), D, E, R, and Lysine (K) with C terminus cleavage. Another finding was the N terminus cleavage bias for Tryptophan (W). These results indicated a preference for C terminus cleavage following all charged amino acids which includes acidic and basic amino acids. In addition, a C terminus cleavage bias was also observed for the polar residues C and Tyrosine (Y) along with the non-polar residues V and Phenylalanine (F).

Kapp et al.[35] investigated the association between intensity and multiple factors simultaneously. A database of 5,500 unique peptide MS/MS spectra acquired in an ion trap mass spectrometer digested with trypsin enzyme was considered. Peptides were classified into mobile, non-mobile, and partial-mobile. A peptide is non-mobile if the charge on peptide is greater than or equal to number of R residues. A peptide is mobile if the charge on peptide is greater than the total number of basic residues (R, K, and H). Peptides that do not classify into mobile or non-mobile groups are called partial-mobile. The dataset was divided in to 9 strata based on the charge state and relative proton mobility. Authors reported results from singly, doubly and triply

charged peptides. It was observed that charge state on peptide and residue composition of peptide can have dramatic effect on the cleavage of peptide. Kapp et al.[35] also analyzed the cleavage effects of residue pairs. It was indicated that D has strong effect on the C terminus cleavage for singly charged peptides that had the proton localized to an R residue. Peptide ions with mobile proton, the cleavage bias shifted to the N terminus side of P.

Barton et al.[30] developed a database of mass spectra from 5,448 peptides digested with trypsin enzyme acquired in time-of-flight mass spectrometer. The *y*- and *b*-ions were analyzed separately. The mass of the ion, amino acid composition of peptide, and amino acid residues on either side of the fragmentation site were associated with ion intensity. The ion mass is directly proportional to the relative ion intensity for *b*-ions while the relationship is inverse relation for *y*-ions. Intensity was lower when the peptide was mobile compared to those peptides that have a non- or partially-mobile proton. Presence of P in the peptide decreased ion intensity for *y*-ions and does not have significant effect for the *b*-ions. C terminus H at the fragmentation site can be seen to the lower ion intensity more than any other amino acid. Proline C terminus to fragmentation site tends to increase ion intensity and K decreased ion intensity more than any other amino acid at C terminus. For N terminus I, L, and V increased ion intensity for *y*-ions for doubly charged peptides not containing H or P.

Zhou et al.[36] developed a Bayesian neural network to identify different features that have a significant role in ion intensity prediction. This model was used to identify the features that have significant association with the ion intensity of fragmentation spectra. A library of features that were supposed to have significant influence on peptide fragmentation was used. Results from the model showed that cleavage is not necessary to occur at the two ends of peptide and it can occur in the middle of the peptide because of the specificity of trypsin enzyme. Zhou et al.[36]

evaluated the factors on three classes of peptides: mobile, non-mobile and partial mobile. Results showed that P enhances cleavage at its N terminus bond in mobile peptides. On the other hand, D and E appeared to have inhibitory effect on cleavage at N terminus bond. Isoleucine and V were found to enhance cleavage at C terminus whereas G and Asparagine (N) residues have the inhibitory effect on the cleavage at the C terminus. For non-mobile peptides, results showed that P still enhanced the cleavage at N terminus but to a lesser extent than in the case of mobile peptides. Aspartic acid was the most influential residue to enhance the cleavage at C terminus. Arginine has a greater influence at cleavage at C terminus and the other two basic residues K and H also have the same effect but to a lesser extent than R. The benefit of this approach included the various factors that can be numerically analyzed so that large number of features can be directly comparable at one time.[36]

Sun et al.[37] presented a Bayesian Peptide Detection Algorithm (BPDA) for identification of peptide from MS instruments. All possible combinations of all possible peptide candidates originated from well-defined peaks of the raw spectra are evaluated by BPDA in order to minimize the mean squared error of the interpreted spectrum to the observed spectrum. Enumerated advantages of BPDA include: the algorithm looks for the optimal among all possible interpretations of the MS spectra, isotopic peaks and charge states are considered, peptide probabilities are estimated and many parameters possess physical meaning because they come directly from the observation of the mass spectra. BPDA is a global-based approach which looks for the optimal solution through Gibbs sampling rather than working on a local region at a time which typically requires more computation.

Linear Model and Model Selection Techniques

Linear Model

Linear models are often used to answer one of the important questions. How can an observed quantity y be explained by a number of other quantities x_1, x_2.....,x_n? A simple model to answer this question is the linear model.

$$\mathbf{y_i = \beta_0 + \beta_1 x1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon_n}$$

where: y is the response or dependent variable, β_0, β_1, β_2....., β_n are the coefficients or parameters of the model, x_1, x_2..... x_n are the explanatory variables or independent variables, and ε_n is the error term that accounts for uncertainties. Explanatory variables can be continuous (regressors) or discrete (factors). In the study of ion intensity, y_i is ion intensity. x_1, x_2.....,xn can include ion type, proton mobility, and relative ion mass. Linear models are better understood and easier to interpret than most of other models and the methods of analysis and inference are better developed because of their simplicity. Models including a random effect other than the residual are known as mixed effects models. Model assumptions include: 1) the residuals are independent, 2) the residuals are normally distributed, 3) the residuals have a mean of 0, and 4) the residuals have constant variance.[38] These assumptions can be relaxed. Test statistics (e.g. *F*-test statistic) are indicators of the association between the variables and are calculated from the sample. The test statistics value computed from the sample is compared to the theoretical distribution of values of the test statistics to assess the probability that the observed value occurred by chance. This probability is known as *p*-value. In other words, the *p*-value reflects the probability that the calculated test statistics lies within the theoretical distribution. Calculation of *p*-value depends upon the test statistics used. In a linear fixed effects model, the *F*-statistic is calculated as a ratio between Mean Sum of Square (MSS) and Mean

Square Error (MSE). Then the *p*-value is calculated as the area under the appropriate null sampling distribution of *F* that is bigger than the observed *F*-statistics. If the observed test statistics value is very distant from zero, then the estimate is unlikely to have happen by chance and the null hypothesis of no association is rejected. If the observed test statistics value is very close to zero, then the estimate is likely to have arisen by chance and the null hypothesis of no association is kept.

Model Selection Approaches

The identification of explanatory variables that contribute the most to explain the variation of the response variables can be accomplished through one of many model selection approaches. In forward selection, an empty model with no explanatory variables is considered initially.[39] For each independent variable, an *F* statistics and associated *p*-value that quantify the contribution of that independent variable to explain the response variable are computed. The most significant *p*-value is compared to an "entry" significance level threshold and the variable is entered into the model if the *p*-value is more significant than the threshold. Then, the *p*-value of the remaining variables not yet included is computed, conditional on the previously entered variable(s) and the next most significant variable that has a *p*-value lower than the threshold is entered into the model. This evaluation process is then repeated until no additional variable has a *p*-value smaller than the threshold. In forward selection, a variable that is selected remains in the model.[39]

In backward selection, the initial model includes all possible explanatory variables.[40] The *p*-value of each variable is calculated and the variable that has the least significant *p*-value that is less significant than the "stay" *p*-value is removed from the model. In other words, the variable with the largest *p*-value exceeding the specified cut-off value (e.g. 0.15) is then removed from

the model. Once a variable is removed, the *p*-values of all the remaining variables are recomputed and the variable that has the least significant *p*-value that is less significant than the "stay" *p*-value is removed from the model. Backward selection deletes the variable one by one from the model until the remaining explanatory variables have all *p*-values more significant than the "stay" threshold. At each step, the variable showing the smallest contribution to the model is deleted. This process continues until no remaining variables have *F* statistic *p*-values above the specified threshold. Once removed from the model, a variable cannot be added to the model again.[40]

Stepwise selection is the modified form of forward selection. Like in forward, the initial model is empty.[39, 40] Once an explanatory variable enters into the model, the *p*-values of the other variables in the model are recomputed. If any of the variables now has a *p*-value less significant than the "stay" *p*-value threshold, then the variable is removed from the model. The variable(s) that are removed from the model are subsequently evaluated for potential re-entry in the model in the same way that all other variables that have not entered into the model. Stepwise like forward stops when there is no variable left outside of model that has a *p*-value more significant than the "entry" threshold and all the variables in the model have *p*-values more significant than the "stay" threshold and every variable included in the model is significant at the specified threshold (e.g. 0.15).[39, 40]

In the Least Angle Regression (LAR) method, all coefficients (b_i) are set to zero initially.[41] From the set of explanatory variables, the LAR algorithm finds one predictor (x_i) that has the highest correlation with the response variable y. A further step is taken to increase the coefficient (b_i) in the direction of the sign of the correlation with y and then the algorithm computes the residual (e = y - yhat) along the way. The LAR algorithm stops when another

predictor (x_j) has the same correlation with the residual as x_i. By moving this way, the algorithm ensures that both predictors have a common correlation with the current residual. In next step, algorithm increase the coefficient (b_i,b_j) in their joint least square direction until a third predictor (x_k) has the same correlation with the residual as previous two predictors had. In this way, the LAR algorithm will continue until all the predictors are in the model.[41]

Tibshirani[42] proposed the Least Absolute Shrinkage and Selection Operator (LASSO) technique that estimates linear regression coefficients by a L1-constrained least squares approach. The LASSO is usually used to estimate the regression parameters $\mathbf{b} = (b_1....,b_p)$ in the model

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xb} + \boldsymbol{\varepsilon}$$

where **y** is response variable, $\boldsymbol{\mu}$ is the overall mean, **X** is the design matrix, and $\boldsymbol{\varepsilon}$ is the independent and identically distributed normal errors with mean 0 and unknown variance σ. The LASSO technique derived from constrained form of ordinary least squares methods. In constrained ordinary least squares, the sum of the regression coefficients is restricted to be smaller than a selected parameter threshold. The LASSO parameter estimates minimize the residual sum of squares subject to the constraint that the sum of the absolute values of the coefficients is less than the selected threshold. Due to nature of this type of constraint, LASSO tends to produce some coefficients that are exactly zero and gives interpretable models.[43]

Model assessment is critical in the process of identifying the explanatory variables associated with a response variable. There are various indicators of the model adequacy that are a function of the variation explained by the model and that not explained by the model. The component not explained by the model is characterized by the residuals or difference between the observed and predicted values. The mean square error is a commonly used indicator of the

variation not explained by the model. Mean square error is the approximate average of the squared deviations of observed values from the fitted or predicted. Another indicator of model adequacy is the R-square statistics that quantifies the total variation explained by the model. High R-square is usually considered to be good for selecting the best models. The model assumptions include that the residuals are normally distributed, with zero mean. The simplest assumptions are that the residuals are independent and have a common variance.[38]

Regardless of the approach used to build a model, the adequacy of a model to describe a response variable should always be checked. The adequacy of a model can be evaluated in the same data set used to estimate the model parameters. A potential disadvantage is that the estimates may be optimized for the particular data set analyzed and be less suitable for other data sets. Alternatively, the suitability of the explanatory variables can be evaluated on an independent data set or by dividing the data set into three different subsets called training, testing and validation data. The training data is used to fit the model and the validation data is used to obtain the prediction error. This prediction error is then used to decide the explanatory variables that need to be included in the final model. K-fold and leave-one-out are two approaches to generate training and validating data sets.

In K-fold cross-validation, the dataset is divided into K equal subsets. Each subset should have similar representation of the levels of the explanatory variables considered. Then, the model is fitted on K-1 of the subsets to get the estimates evaluated on the remaining subset using MSE measure of prediction accuracy. This process is repeated for all K subsets.[40]

Leave-one-out cross-validation is a particular case of K-fold cross-validation. In this technique, all minus one observation are used for training, and testing is done on the single observation which was left out. An accuracy estimate resulting from the leave-one-out cross-

validation is nearly unbiased although the estimate has high variance leading to unreliable estimates.[44]

Motivation for research

Proteomic techniques are important to identify and characterize proteins in samples. The proteomics workflow consists of the separation of proteins which is followed by the identification of the individual proteins using MS. Proteins are often digested using the trypsin enzyme and the resultant peptides are further subject to further MS analysis.[45] Tandem mass spectrometry (MS/MS) encompasses the enzymatic digestion of proteins followed by additional fragmentation of the resulting peptides into ions. The mass spectrum that relates the peak intensity (or abundance) to the mass-to-charge (m/z) of the ions is then used to deduce the sequence of amino acids in the ion and corresponding peptide. The accuracy of the algorithms that identify the peptides based on matches of the spectra observed to that expected based on known peptide sequences peaks is directly dependent on the precision of the spectra peaks. The peptide fragmentation pattern in MS/MS experiments is a complex process that is influenced by number of factors.[30] Peak intensity varies across ions within a peptide and the ions that do not surpass a minimum peak threshold may not be detected. Ions also vary on the type (e.g. *b*, *y*, *a*), charge (e.g. 1 to 6), length and amino acid sequence and content of specific amino acids such as basic ones. Furthermore, peptides can have different charges, PTMs (e.g. amidation, glycosylation), proton mobility (mobile, non-mobile or partial- mobile), neutral mass loss, length and amino acid sequence that can all influence peak intensity. Attempts have been made to identify the factors that can potentially influence the ion peak intensity. However, these attempts have been challenged by the high dimensional and multi-factorial nature of the MS/MS data.

The goal of this study is to develop a comprehensive model that describes MS/MS spectra intensity while accounting for multiple factors, and gain insights into the factors that influence ion intensity.

CHAPTER 2: Multifactorial modeling of Ion Abundance in Tandem Mass Spectrometry Experiments

INTRODUCTION

Mass spectrometry (MS) has become a standard technique to identify proteins and peptides in samples. In the first stage of tandem mass spectrometry (MS/MS) experiments, the proteins are first extracted from the sample and then are digested with an enzyme, typically trypsin.[45] In the second stage, the peptides (also known as parental ions or precursor ions) obtained from the enzymatic digestion are separated based on mass on the first MS stage and further fragmented into ions (also known as fragment ions). A MS graph relating the intensity of the fragment ions to the mass/charge (m/z) ratio is generated. The combination of m/z associated with the multiple intensity peaks is used to deduce the amino acids in the ion. The amino acid sequences that are deduced from the masses of the fragment ions are in turn used to identify the peptides in the sample.

Two main issues influence the accuracy of peptide identification in MS/MS studies: 1) the ability to obtain all possible ion fragments, and 2) the quantity in which each ion fragment is generated, also known as ion intensity. Both issues can be combined into the ability of the MS/MS technique to detect the fragment ions. The ions necessary to identify a peptide varies. Not all ion fragments are equally informative; however a fraction of all possible ions is sufficient to identify a peptide. In order for an ion to be detected, the ion has to be present in an amount that surpasses the minimum intensity threshold for peak detection. Many factors influence the intensity of ion formation.[30, 31]

Several studies have investigated the factors that influence ion intensity. Some of these studies and factors are reviewed by Barton and Whittaker.[46] The accuracy and precision of the findings in these previous studies are limited by the information or methodology used. With respect to data limitations, a relatively small set of ions from selected peptides and proteins not representative of the whole proteome spectrum was used in many studies.[33, 35] With respect to methodology, different limitations can be identified. First, few factors were considered simultaneously. In these cases, the failure to account for multiple sources of variation simultaneously could result in inaccurate conclusions. For example, if two factors that could influence intensity are highly correlated and only the most significant is considered, the second may be dismissed. The most important limitation of previous studies that have tried to identify the factors that influence ion intensity is that the information from each ion was incorrectly assumed to be independent. Data from ions formed from the same peptide are correlated and failure to account for these covariances will result in biased conclusions. Another limitation in most studies of ion intensity determinants is that the results from one data set were not confirmed on comparable and independent data sets. This situation has resulted in a failure to confirm most factors identified on one study.[46]

A more complete understanding of the factors influencing ion intensity is necessary to improve the accuracy of the peptide identification approaches. Both, the identification of factors associated with ion intensity and characterization of the association are necessary. This study has two main objectives. The first objective was to identify variables that impact ion intensity in MS/MS while addressing the limitations of prior studies. This objective was accomplished by using a very large data set, a stepwise feature selection strategy that considered each potential explanatory variable in the context of the other variables, and a hierarchical model that

accommodates numerous potential explanatory variables and the correlation between ion measurements. Second objective was to validate the identified variables and the characterized trends across data sets. This goal was accomplished by applying a 10-fold cross-validation approach and through two complementary strategies. The estimates and significance levels of the explanatory variables associated with intensity were compared across 10 data sets. In addition, the estimates from each data set were used to predict the ion intensity on the remaining 9 data sets and to evaluate the accuracy of each set of predictions.

MATERIALS AND METHODS

Data set

Ion intensity data generated from collision induced dissociation (CID) ion trap mass spectrometer was obtained from the NIST database (Mouse build, May 24, 2011, http://peptide.nist.gov/). Initial assignments of spectra to peptides were done using SEQUEST[10], MASCOT[11], X!TANDEM[12], and OMSSA[13]. Reported scores for the different search engine were normalized using results of searching against a combined forward (correct) and reversed (incorrect) sequence library. The best normalized score (or expectation value) among the search engines were used for further processing. In general, peptides were permitted to have to up to two missed cleavages and one non-tryptic terminus (semitryptic). Parent and fragment ion tolerances of 2 and 0.8 m/z, respectively, were generally used. Records from categories with limited representation were removed to minimize the risk of identifying false associations due to insufficient representation. In this study, *b*- and *y*-ions were used because of their annotation was available in NIST database. Observations from other ions such as *a*-,*c*-,*x*-, and *z*-ions were removed because of their minor representation (~5%) in dataset. Observations from peptide with

PTMs were also removed. The preferred ion amino acid sequence was selected when more than one amino acid sequence was available for an ion fragment. Ion intensity measurements from 6,548,340 ion fragments formed from 61,543 peptides corresponding to 7,761 proteins were analyzed. A log base 10 transformation of the ion intensity values resulted in a closer to Normal distribution of the response variable.

The objectives of this study were two-fold: 1) to identify and characterize the variables associated with ion intensity, and 2) to validate these findings on separate data sets. These objectives were accomplished using a ten-fold cross-validation strategy. First, the preprocessed NIST data was divided into 10 data sets. Dependencies between data sets were minimized by assigning all the fragment ions within a protein to the same data set. The suitability of each K data set to cross-validate the other data sets was optimized by ensuring that all the data sets have comparable representation of all the explanatory variables evaluated. All 10 data sets have comparable number of observations. Consideration of the representation of proteins and factors within data set prevented the exact same number of observations across data sets. Second, each of the 10 data sets was considered a training data set and analyzed separately. The resulting models including the variables significantly associated with 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

The identification and characterization of the explanatory variables associated with ion intensity in each of the 10 training data sets was accomplished following a two-step strategy. First, within a linear fixed-effect model framework, a stepwise variable selection approach was used to identify the explanatory variables that were associated with ion intensity in each K data

set and Mallows' Cp is used to assess the fit of a regression model that has been estimated using ordinary least squares. Second, the final set of explanatory variables was evaluated in a hierarchical mixed-effects model including the random effect of protein.

Multi-level Classification Variables

The discrete multi-level or classification explanatory variables considered simultaneously in the first stage fixed effects and second stage mixed effects analysis of each of the 10 datasets were: combination of peptide and ion charge (levels 11, 21, 22, 31, 32, 33: where the first and second number of a level indicate that peptide and ion charge state, respectively, and the third denotes that charge is equal or greater than 3), proton mobility (levels mobile, non-mobile and partial mobile), neutral mass loss (levels -17,-18 -3,-4,None: where -17 corresponds to water loss, -18 corresponds to ammonia loss, -3 corresponds to H_2S+H (-34), Cl (-35), or HCl (-36) loss , -4 corresponds to either C_2H_5O+H; (-46), C_2H_5O (-45), or CO_2O, $CONH_2$ (-44) loss and None correspond to no neutral mass los), ion type (*y* or *b*-ion series), combination of number of Arginine (R) residues in the peptide and ion (levels 00, 10, 11: where the first and second number correspond to the number of residues on the peptide and ion, respectively and 1 denote number of R is equal or greater than 1), combination of number of Lysine (K) residues in the peptide and ion (levels 00, 10, 11: where the first and second number correspond to the number of residues on the peptide and ion, respectively and 1 denote number of K is equal or greater than 1), combination of number of Histidine (H) residues in the peptide and ion (levels 00, 10, 11: where the first and second number correspond to the number of residues on the peptide and ion, respectively and 1 denote number of H is equal or greater than 1), combination of number of Proline residues (P) in the peptide and ion (levels 00, 10, 11, 20, 21, 22: where the first and second number correspond to the number of residues on the peptide and ion, respectively and 2

denote number of P is equal or greater than 2), and combination of number of basic residues (R, K, H) in the peptide and ion (levels 00, 10, 11, 20, 21, 22: where the first and second number correspond to the number of residues on the peptide and ion, respectively and 2 denote number of basic residues is equal or greater than 2), and protein (7,761). Explanatory variables that combine the number (count) of specific residues or groups of residues (basic) in the peptide and ion fragment were evaluated instead of fitting peptide and ion counts separately because the value at the ion level is dependent on the value at the peptide level. For example, a peptide that has 1 basic amino acid can only form ions with 0 or 1 basic amino acids.

Binary Classification Variables

The binary (presence or absence = 1 or 0, respectively) explanatory variables considered in the first stage fixed effects and second stage mixed effects analysis of each of the 10 datasets were: each of the 20 amino acids at position 1 to 9 from either the N or C termini of the ion (e.g. AN1 denote presence or absence of Alanine at position 1 from the N terminus and RC9 denote presence or absence of R at position 9 from the C terminus), basic or positively charged residues (R, K or H) at position 1 to 9 from either the N or C termini, aliphatic residues (Valine -V, Leucine -L, Isoleucine -I, and Methionine –M) at position 1 to 9 from either the N or C termini, aromatic residues (Phenylalanine -F, Tyrosine -Y, Tryptophan –W, H) at position 1 to 9 from either the N or C termini, tiny residues (Glycine –G, Alanine –A, Serine -S) at positions 1 to 9 from either the N or C termini, small residues (Threonine -T, Proline or iminoacid -P, Aspartate –D, Asparagine –N, Cysteine -C) at positions 1 to 9 from either the N or C termini, large residues (V, L, I, M, P, F, W) at position 1 to 9 from either the N or C termini, polar residues(Aspartate –D, Glutamate –E, Asparagine –N, Glutamine –Q) at position 1 to 9 from either the N or C termini, hydrophobic residues (Cysteine –C, A, G, V, I, L, M, F, Y, W, K, P) at

position 1 to 9 from either the N or C termini, charged neutral residues (A, N, C, Q, G, I, L, M, F, P, S, T, W, V) at position 1 to 9 from either the N or C termini, charged negative or acidic residues (D, E) at position 1 to 9 from either the N or C termini, hydroxyl residues (S, T, Y) at position 1 to 9 from either the N or C termini, sulfuric residues (C, M) at position 1 to 9 from either the N or C termini, amide residues (N, Q) at position 1 to 9 from either the N or C termini, and the following subgroups of amino acids that do not overlap completely with the previous tested properties: tiny and very small amino acids (S, T, P, A, G) at position 1 to 9 from either the N or C termini, true aromatic excluding H (F, Y, W) at position 1 to 9 from either the N or C termini, acidic and amide amino acids (D, E, Q, N) at position 1 to 9 from either the N or C termini.

Continuous Explanatory Variables

The continuous explanatory variables evaluated in the first and second stage models were: relative ion size (ranging from 0.047 to 0.98) defined as ion number by length of peptide, relative ion weight (ranges from 0.05 to 0.98).

Variable Selection

Linear Model

$$\mathbf{y_i = \beta_0 + \beta_1 x1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon_n}$$

where: y_i is the response or dependent variable, β_0, β_1, β_2....., β_n are the coefficients or parameters of the model, x_1, x_2..... x_n are the explanatory variables or independent variables, and ε_n is the error term that accounts for uncertainties. Explanatory variables can be continuous (regressors) or discrete (factors). In the study of ion intensity, y_i is ion intensity. x_1, x_2.....,xn can include ion type, proton mobility, and relative ion mass.

The stepwise selection used in the first step of the approach is the modified form of forward selection. Like in forward selection, the initial model is empty. Once an explanatory variable enters into the model the *p*-values of the other variables in the model are recomputed. If any of the variables now has a *p*-value less significant than the "stay" *p*-value threshold, then the variable is removed from the model. The variable(s) that are removed from the model are subsequently evaluated for potential re-entry in the model in the same way that all other variables that have not entered into the model. The stepwise approach, like forward, stops when there is no variable left outside of model that has a *p*-value more significant than the "entry" threshold and all the variables in the model have *p* -values more significant than the "stay" threshold.[39, 40] The threshold criteria for variable entrance and permanence in the final model were set to *p*-value < 0.00005. These stringent criteria were aimed at minimizing the identification of false positive associations in consideration of the 10 data sets analyzed and the large number of observations within data set.

Analysis

The feature selection approach in the first stage was implemented in a least-squares framework using PROC GLMSELECT.[47] Once the final set of significant explanatory variables was identified in each of the 10 data sets, these variables were considered in a mixed effects model that included protein as a random effect in the second stage. This block effect allowed the consideration of the covariation between ion fragments from the same protein. The linear mixed effects model was fitted in a restricted maximum-likelihood framework using PROC MIXED.[48]

RESULTS AND DISCUSSION

Table 1 summarizes the distribution of observations across the levels of discrete explanatory variables and covariates in each of the 10 data sets. Table 2 summarizes the results (corresponding to the discrete or class explanatory variables significantly (p-value < 0.00005) associated with 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 explanatory variable in the final model. Table 3 summarizes the results corresponding to the binary (no or yes = 0 or 1) and continuous explanatory variables significantly associated (p-value < 0.00005) with ion intensity in 7 or more data sets. Supplementary Table 1 lists the results corresponding to the binary or continuous variables significantly associated with ion intensity in at least one data set.

The performance of the estimates from the 10 training data sets to predict ion intensities on the remaining validating data sets was comparable to the performance on the original training data sets. The average difference between the training and validation mean square errors among the 10 data sets expressed in absolute (0.0942) and relative (0.0709) terms was 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 ion intensity data set. This conclusion is also confirmed by the narrow range of estimates of the association between factors and ion intensity observed in the 10 data sets and summarized in Tables 2 and 3 and in Supplementary Table 1.

Several studies concur that the type of ion formed is associated with the ion intensity.[46] We found that the intensity of y-series ions has a higher intensity than b-series ions (Table 2) and this trend is in agreement with reports.[30, 31, 33, 35, 49]. Tabb et al.[31], Kapp et al.[35], and Huang et al.[33]

used CID ion trap MS/MS while Barton et al.[30] used time-of-flight mass spectrometry. Our results are also consistent with the rules explained in *de novo* sequencing tutorial which indicated that the *y*-ion series intensities will often be the most prominent peaks in the spectrum.[50] The association between neutral mass loss and ion intensity detected in this study is supported by previous reports. In agreement with the trend observed in this study, Tabb et al.[49] reported that mass spectra were negatively affected by neutral losses of molecules.

In this study the association between ion intensity and the combined charge state at the peptide and ion level was considered instead of evaluating peptide and the fragment ion charges separately. The same model specification was used to evaluate the association between ion intensity and combined number of specific amino acids at the peptide and ion level. The rational 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. In this study, the combined peptide-ion charge state was associated with ion intensity. This finding is consistent with previous reports.[33-35, 45, 46] Khatun et al.[34] used MALDI TOF/TOF mass spectrometry for fragmentation analysis. The higher intensities were observed in peptides and ions that had the lowest charge state. Barinaga et al.[51] reported a decrease in intensity of the daughter ions with increase in the parent charge-state. The maximum difference in intensity was found between peptide-ion combinations that have charge of one and peptide-ion combinations that have a charge of 3 or more. Most differences in intensities between levels of peptide-ion charge combinations were significant in all 10 data sets.

Our results on the association between charge state and ion intensity are also consistent with previous reports on the association between the relative proton mobility of the peptide and ion intensity. The relative proton mobility of a peptide has been associated with ion intensity.[30,

[33, 35, 45, 46] The mobile proton theory indicates that sequence features that hinder proton mobility (such as basic residues) also hinder ion intensity because more energy is required to induce fragmentation.[46, 52] The explanation is that migration of a charge is required for cleavage initiation and to induce fragmentation. Barton et al.[46] concluded that the charge state contributes to the proton mobility and that hindered proton mobility is associated with limited fragmentation and spectra dominated by few large peaks. This conclusion is consistent with our results showing a statistically significant association between ion intensity and both proton mobility and charge state in all 10 data sets. Peptides classified as mobile are associated with higher intensity relative to peptides classified as non-mobile.

In the present study, the relationship between the number of basic residues in the peptide and ion and intensity was characterized (Table 2). Previous studies assessed the association of a simpler indicator, presence of basic residues.[46] Presence of basic residues has been linked to ion intensity.[30, 31, 33, 34, 46, 53] In this study, the combined number of basic amino acids on the peptide and ion was associated with intensity. The highest intensities were observed in peptides with no basic amino acid. However within peptide, ions with lower number of basic amino acids had lower intensity relative to ions with one or more basic amino acids. Thus, the maximum fold difference in intensity was observed between peptides (and consequently ions) with no basic amino acids relative to ions with no basic amino acids originated from peptides with 2 or more basic amino acids. This statistically significant difference was observed on all 10 data sets. The trend observed at the ion level is consistent with prior suggestions that basic residues retain a proton and thus more energy is required to transfer the proton from the basic side chain to the peptide backbone.[46, 52] Our findings at ion level are also consistent with Tabb et al.[31] reported that fragment ions lacking basic amino acids are associated with lower ion intensity and

fragments ions with basic amino acids are associated with higher ion intensity.

Consistent with the association between basic residues and ion intensity reported, in this study the number of K, R, and H on the peptide and resulting ion were positively and significantly associated with intensity (Table 2). The highest intensities were observed in peptides and associated ions with high number (1 or more) of K relative to ions with no K regardless of number of K in peptides they originated. This trend is consistent with that observed for the number of R in the peptide and intensity. Likewise, within number of K in the peptide, ions with higher number of K had higher intensity relative to ions with low number of K. Thus, the maximum fold difference in intensity was observed between peptides and resulting ions that have one or more K relative to peptides and resulting ions with no K. This statistically significant difference was observed on all 10 data sets. Although statistically significant, the difference in ion intensity between peptide and ions that have different number of K was lower than for R. Peptides and associated ions carrying one or more R had higher intensities than ions with no R regardless of the number of R in the peptide they originated. Peptides and associated ions with no H had higher intensities than peptides and ions carrying one or more H regardless of number of H in peptide. Elias et al.[53] found that the fraction of basic residues such as R, K, and H in the fragment ion is important in determining the intensity of fragment ions for *b*-series ions but not for *y*-series ions.

The number of P on the peptide and resulting ion were associated with intensity (Table 2). The highest intensities were observed in peptides with higher number of P and also in ions with higher number of P within peptide. The maximum difference in intensity was found between peptides (and consequently ions) with one P relative to peptides and ions with two or more P. Association of P in the peptide and ion and intensity is may be due to the number of rings that

may interfere with fragmentation. Most differences in intensities between levels of peptide-ion charge combinations were significant in all 10 data sets.

We studied the link between amino acid and groups of residues that share physicochemical properties located at or within 9 positions from the N or C termini of the ion (Table 3). The study of individual and groups of amino acids 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 amino acids. The important role of groups of amino acids sharing similar properties relative to individual amino acids in ion intensity suggests that well-known properties are key on fragmentation and the resulting ion intensity. The consistent association between amino acids within a group and intensity augments the statistical precision to identify significant associations. Several amino acids and groups of amino acids were consistently associated with ion intensity across data sets.

The property that was significantly associated with intensity at various locations relative to the N or C terminus was residue charge. Amino acids that have negative charge at various positions from the N terminus (position 2, and 4 to 8) were associated with higher intensities. Likewise, amino acids with neutral charge at various positions from the N terminus (positions 2 and 4 to 8) are associated with higher intensities. This result is consistent with findings that D augmented the likelihood of cleavage in its proximity.[34] The trend between positively charged amino acids and intensity was less consistent or significant. This may be due to the significant associations between a particular basic amino acid and intensity identified simultaneously in this study.

Several properties not commonly associated with ion intensity were detected in this study. We identified an association between aliphatic amino acids and ion intensity (Table 3). The

aliphatic nature of amino acid was associated with intensity in this study. Residues with aliphatic (I, V, L) or quasi-aliphatic properties (M) located in the C terminus were associated with lower intensity levels. Although the same group of amino acids has a positive association with fragment ion intensity when located at the N terminus. Our results are also consistent with the work of Breci et al.[32] concluding that cleavage N terminus to P (X-P) was favored when the X residue was V, H, D, I, and L. Tabb et al.[54] also concluded that intense fragments ions are formed to the N terminus side of a P residue. Also, branched aliphatic residues located N terminus to the fragmentation site increase the intensity of *y*-series ions.[30]

Another property associated with ion intensity was iminoacids (Table 3). Negative associations between intensity and iminoacids acids (P, D,N,C) at various positions from the C and N terminus (position 3 to 9) were identified. In addition, consistent positive associations between intensity and sulfur containing amino acids (C and M) at various positions to the C and N terminus were identified.

Our research confirms that the location of basic residues impact ion intensity.[30] The most frequent residues associated with ion intensity at various locations on the ion fragment were R, H and P. The role of basic amino acids is not unexpected considering that trypsin usually cleaves a protein after R or K but not H. Previous studies have also reported association between the position of basic residues in the ion and ion intensity.[30, 31, 46, 53]

In this study, R proximal to the C terminus were associated with higher intensities meanwhile R located in high proximity to the N terminus were associated with lower intensities. The association between R and ion intensity found in this study is consistent with multiple reports.[30-33, 35, 45, 46, 49, 53] Khatun et al.[34] concluded that for singly charged peptides, basic amino acids had a strong effect on ion intensities when not adjacent to the fragmentation site. In

addition, Khatun et al.[34] concluded that R augmented the likelihood of cleavage in its proximity. Also, H distant from the C terminus (position 9) had a positive association with intensity in all 10 datasets. Our finding is also consistent with a report by Tsaprailis et al.[55] indicating that the side chain of H can attack its own C terminus bond, thus enhancing cleavage at the C terminus. Wysocki et al.[56] also indicated that cleavage is promoted near H in many peptides. The association between K and intensity was less wide-spread. Lysine located distant from C terminus (position 5) was associated with lower intensity. Consistent with our findings, Zhou et al.[36] concluded that R has a strong positive association with cleavage toward the C terminus and a similar yet weaker trend is also observed for K and H. Huang et al.[33] reported that cleavage C terminus to acidic residues (D and E) is favored in peptides that have double charge state provided there is a basic residue (H) in the peptide. Our results are also consistent with the findings of Tabb et al.[54] that the side chain of D has the ability to cleave the peptide bond on its C terminus. This cleavage enhancement corroborates a mechanism for D dependent cleavage reported by Gu et al.[57] Both D and E inhibit and promote cleavage to the N and C terminus, respectively.[36]

In this study, P at various locations in the ion was associated with ion intensity. Presence of P in high proximity to the C terminus was associated with higher ion intensity. However, P distant from the C or N terminus was associated with lower intensity. Other studies have also reported an association between the presence of P and intensity.[30, 33, 46, 53] Kapp et al.[35] noted that P had a significant association with ion intensity when located N terminus from the cleavage site. Consistent with our findings, Barton et al.[30] reported that P increased the ion intensity when located at the C terminus from the cleavage site. Khatun et al.[34]concluded that P augmented the likelihood of cleavage in its proximity. Zhou et al.[36] also observed that P enhanced cleavage

when located at the N terminus while having a negative effect on cleavage when located at the C terminus. The results in this study are also supported by reports that peptide cleavage, and thus ion intensity, was favored N terminus to P.[32]

The present study uncovered novel or less recognized associations between individual amino acids and ion intensity. Tryptophan located at various positions from the C terminus had a positive association with intensity. Likewise, Khatun et al.[34] concluded that W augmented the likelihood of cleavage in its proximity. Arnold et al.[58] also indicated that cleavage is favored on the C terminus of W. A positive association between N proximal to the C terminus (position C2) and intensity was identified on the 9 data sets analyzed. A negative association between N at the most terminus N position was also identified in 8 data sets. Positive associations between G at two most proximal positions from the C terminus (position 2 and 3) and intensity were identified. Negative associations between G proximal to N terminus (position 1) in 8 datasets were analyzed.

In this study, relative ion size (or length) was negatively associated with intensity while relative ion weight was positively associated with intensity (Table 3). In addition, lengthier ions had higher intensity. Thus, for a given ion length, ions from shorter peptides exhibited lower intensities. Multiple studies have reported that the relative weight of the ion was associated with fragment ion intensity.[30, 46, 53] Two studies reported a non-linear association between relative ion mass and intensity. Results from the evaluation of non-linear trends for relative ion weight and size in this study suggested that quadratic trends could be confounded with other model terms. Thus, the previous reports of non-linear trends could be the artifact of other factors not considered in the corresponding analysis.

This study used well characterized spectra from accurately identified peptides. A valuable consideration could involve the analysis of unidentified fragments. This approach was not used in this study because our model included the consideration of protein which allows taking in account the covariation between ion fragments from the same protein.

CONCLUSIONS

Results from the present 10-fold analysis show that ion intensities are associated with a number of factors. The type of ions produced (*b*- and *y*- series ions) was strongly associated with ion intensity. The highest intensities were observed both in low charge state peptides that produce low charge state ions. Neutral mass loss and proton mobility were also associated with the intensity of the ion in the spectra.

Peptides with no basic residues had high intensities meanwhile ions with lower number of basic residues within a peptide had low intensities. Consistent with the association between basic residues and ion intensity, that the number of K, R, and H in the peptide and ion were positively associated with ion intensity. Peptides and ions that have higher number of P showed higher intensities relative to peptides and ions with two or more P. In addition to presence, the location of specific residues or residues sharing physiochemical properties relative to the fragmentation site were associated with ion intensity. For example, aliphatic amino acids had a positive association with ion intensity when located proximal to the N terminus. The use of multi-factorial and hierarchical models allowed the identification and characterization of factors associated with MS/MS ion intensity. Understanding the factors associated with ion intensity patterns can help in the improvement of database search and spectrum-to-spectrum algorithms for identification of peptides and proteins in MS experiments.

REFERENCES

(1) Edman, P.; Högfeldt, E.; Sillén, L. G.; Kinell, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **1950**, *4*, 283.

(2) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **1996**, *379*, 466-469.

(3) Polce, M. J.; Ren, D.; Wesdemiotis, C. Dissociation of the peptide bond in protonated peptides. *J. Mass Spectrom.* **2000**, *35*, 1391-1398.

(4) Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **2006**, *314*, 65-66.

(5) Wehr, T. Top-Down versus Bottom-Up Approaches in Proteomics. *LCGC NORTH AMERICA* **2006**, *24*, Feb 27,2012.

(6) Kenneth, B. T.; Moseley, M. A.; Leesa, J. D.; Carol, E. P. Capillary liquid chromatography/mass spectrometry. *Mass spectrometry reviews* **1994**, *13*, 431-457.

(7) Reid, G. E.; McLuckey, S. A. 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* **2002**, *37*, 663-675.

(8) Marcotte, E. M. How do shotgun proteomics algorithms identify proteins? *Nat. Biotechnol.* **2007**, *25*, 755-757.

(9) Papayannopoulos, I.A. The interpretation of collision-induced dissaciation tandem mass spectra of peptides. *Mass spectrometry reviews* **1995**, *14*, 14.

(10) Eng, J.; McCormack, A.; Yates, J. 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* **1994**, *5*, 976.

(11) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551-3567.

(12) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466-1467.

(13) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958-964.

(14) Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol. Biol.* **2007**, *367*, 87-119.

(15) Klammer, A. A.; Park, C. Y.; Noble, W. S. Statistical calibration of the SEQUEST XCorr function. *J. Proteome Res.* **2009**, *8*, 2106-2113.

(16) Stein, S. E.; Scot, D. R. Optimization and testing of mass-spectral library search algorithms for compound identification. *Nat. Methods* **2007**, *4*, 10.

(17) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655-667.

(18) Craig, R.; Cortens, J. P.; Beavis, R. C. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1844-1850.

(19) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787-797.

(20) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337-2342.

(21) Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77*, 964-973.

(22) Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **2006**, *5*, 652-670.

(23) Chung, C.; Liu, J.; Emili, A.; Frey, B. J. Computational refinement of post-translational modifications predicted from tandem mass spectrometry. *Bioinformatics* **2011**, *27*, 797-806.

(24) Larsen, M. R.; Trelle, M. B.; Thingholm, T. E.; Jensen, O. N. Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *BioTechniques* **2006**, *40*, 790-798.

(25) Guerrera, I. C.; Kleiner, O. Application of mass spectrometry in proteomics. *Biosci. Rep.* **2005**, *25*, 71-93.

(26) Jensen, O. N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **2004**, *8*, 33-41.

(27) Stein, S. E. Peptide Mass Spectral Libraries. http://www.nist.gov/mml/chemical_properties/data/peplibdatareport.cfm (accessed 04/05, 2012).

(28) Vizcaino, J. A.; Cote, R.; Reisinger, F.; Foster, J. M.; Mueller, M.; Rameseder, J.; Hermjakob, H.; Martens, L. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **2009**, 9, 4276-4283.

(29) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *3*4, D655-8.

(30) Barton, S. J.; Richardson, S.; Perkins, D. N.; Bellahn, I.; Bryant, T. N.; Whittaker, J. C. Using statistical models to identify factors that have a role in defining the abundance of ions produced by tandem MS. *Anal. Chem.* **2007**, *7*9, 5601-5607.

(31) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R.,3rd. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *7*6, 1243-1248.

(32) Breci, L. A.; Tabb, D. L.; Yates, J. R.,3rd; Wysocki, V. H. Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **2003**, *7*5, 1963-1971.

(33) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem.* **2005**, *7*7, 5800-5813.

(34) Khatun, J.; Ramkissoon, K.; Giddings, M. C. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem.* **2007**, *7*9, 3032-3040.

(35) Kapp, E. A.; Schutz, F.; Reid, G. E.; Eddes, J. S.; Moritz, R. L.; O'Hair, R. A.; Speed, T. P.; Simpson, R. J. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **2003**, *7*5, 6251-6264.

(36) Zhou, C.; Bowler, L. D.; Feng, J. A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* **2008**, 9, 325.

(37) Sun, Y.; Zhang, J.; Braga-Neto, U.; Dougherty, E. R. BPDA - a Bayesian peptide detection algorithm for mass spectrometry. *BMC Bioinformatics* **2010**, *1*1, 490.

(38) Verran, J. A.; Ferketich, S. L. Testing linear model assumptions: Residual analysis. *Nursing Research* **1987**, *3*2, 127-130.

(39) Rencher, A. C. In *Methods of Multivariate Analysis;* John Wiley & Sons Inc.,New York: New York, 1995.

(40) Neter, J. In *Applied linear statistical models;* Irwin: Chicago, 1996; , pp 720.

(41) Efron, B.; Hastie, T.; Jhonstone , I.; Tibshirani, R. Least Angle Regression. *The Annals of Statistics* **2004**, *3*2, 407-499.

(42) Tibshirani, R. Regression Shrinkage and Selection via Lasso. *J. R. Statistic. Soc* **1996**, *5*8, 267-288.

(43) Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **1997**, *1*6, 385-395.

(44) Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association* **1983**, *7*8, pp. 316-331.

(45) Schutz, F.; Kapp, E. A.; Simpson, R. J.; Speed, T. P. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans.* **2003**, *3*1, 1479-1483.

(46) Barton, S. J.; Whittaker, J. C. 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.* **2009**, *2*8, 177-187.

(47) SAS Institute. Cary, N. The GLMSELECT procedure (Experimental). http://support.sas.com/rnd/app/papers/glmselect.pdf (accessed 04/05, 2012).

(48) SAS Institute. Cary, N. The MIXED Procedure. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mixed_toc.htm (accessed 04/05, 2012).

(49) Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R.,3rd. Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. *Anal. Chem.* **2003**, *7*5, 2470-2477.

(50) Guzzetta, A. De Novo Peptide Sequencing Tutorial. http://ionsource.com/tutorial/DeNovo/introduction.htm (accessed 04/05, 2012).

(51) Barinaga, C. J.; Edmonds, C. G.; Udseth, H. R.; Smith, R. D. Sequence determination of multiply charged peptide molecular ions by electrospray? Ionization tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **1989**, 3, 160-164.

(52) Wysocki, V. H.; Resing, K. A.; Zhang, Q.; Cheng, G. Mass spectrometry of peptides and proteins. *Methods* **2005**, *3*5, 211-222.

(53) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *2*2, 214-219.

(54) Tabb, D. L.; Friedman, D. B.; Ham, A. J. Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat. Protoc.* **2006**, 1, 2213-2222.

(55) Tsaprailis, G.; Nair, H.; Zhong, W.; Kuppannan, K.; Futrell, J. H.; Wysocki, V. H. A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Anal. Chem.* **2004**, *7*6, 2083-2094.

(56) Wysocki V. H.; Tsaprailis G.; Smith L. L; Breci L. A. Mobile and localizedprotons: A framework for understandingpeptidedissociation. *J Mass Spectrom* **2000**, *3*5, 1399-1406.

(57) Gu, C.; Tsaprailis, G.; Breci, L.; Wysocki, V. H. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal. Chem.* **2000**, *7*2, 5804-5813.

(58) Arnold, R. J.; Jayasankar, N.; Aggarwal, D.; Tang, H.; Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.* **2006**, 219-230.

(59) Fazal, Z.; Southey, B. R.; Sadeque, A.; Sweedler, J. V.; Rodriguez-Zas, S. L.; In Model of ion intensity from tandem mass spectra for improved peptide identification and simulation. 2011 *IEEE International Conference on Bioinformatics and Biomedicine Workshops* **2011**, 994-996.

Figures:

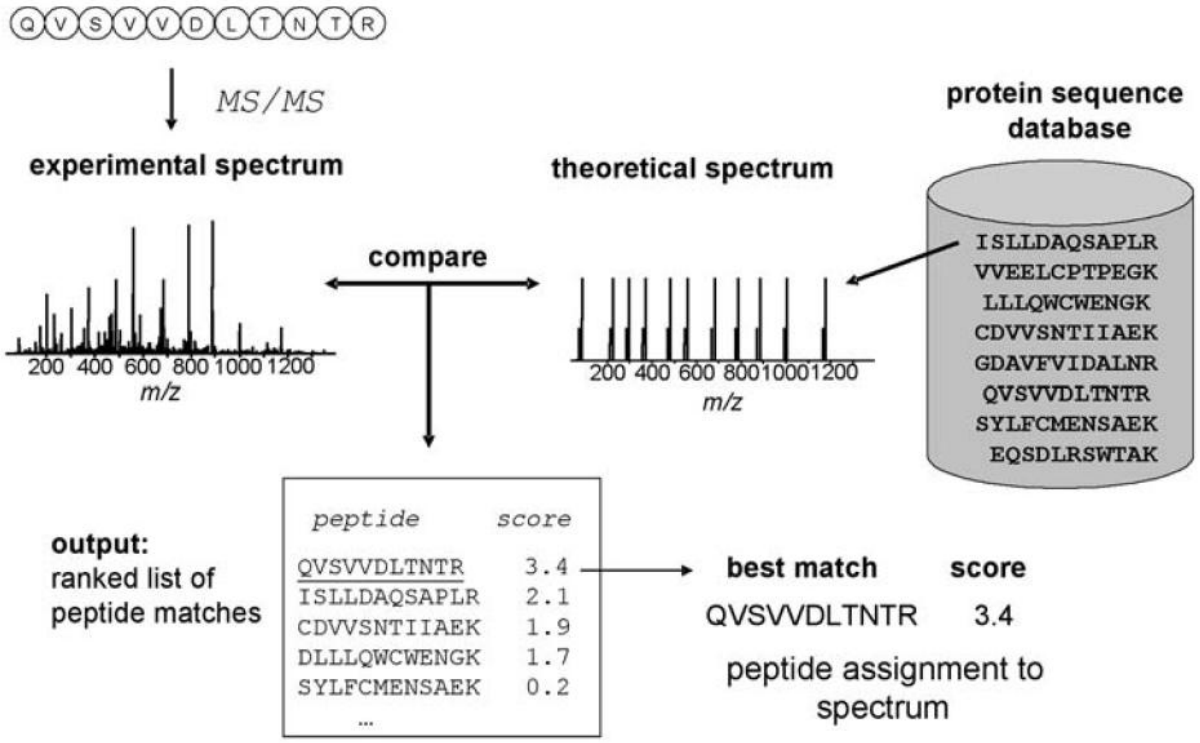

Figure1: Tandem mass spectrometry (MS/MS) database searching. Acquired MS/MS spectra are correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the spectra. Candidate peptides are ranked according to the computed score, and the highest scoring peptide sequence (best match) is selected for further analysis [2].

[2] Springer and the Methods in Molecular Biology, 367, 2007, 87-119, Protein identification by tandem mass spectrometry and sequence database searching, Nesvizhskii AI, 3; with kind permission from Springer Science and Business Media.

Table 1. Percentage of Ion Fragment Observations by Level of Potential Explanatory Variables of Intensity, Range of Covariates, and Descriptive Statistics for Each of the Ten Training Data Sets Analyzed.

		Datasets									
		1	2	3	4	5	6	7	8	9	10
Ion Type Series	b	49.97	49.90	49.95	49.95	50.06	49.96	50.03	49.97	50.01	49.98
	y	50.03	50.10	50.05	50.05	49.94	50.04	49.97	50.03	49.99	50.02
Neutral Mass Loss	-18	16.48	16.51	16.59	16.61	16.55	16.54	16.52	16.53	16.53	16.51
	-17	17.41	17.42	17.44	17.37	17.51	17.50	17.48	17.44	17.51	17.39
	-3x	9.83	9.81	9.79	9.81	9.77	9.84	9.80	9.81	9.81	9.86
	-4x	1.22	1.25	1.23	1.25	1.28	1.24	1.25	1.21	1.25	1.25
	None	55.06	55.02	54.95	54.96	54.90	54.88	54.97	55.01	54.90	55.00
Proton Mobility	Mobile	53.76	53.76	53.76	53.76	53.77	53.76	53.76	53.76	53.77	53.77
	Non	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09	3.09
	Partial	43.14	43.14	43.14	43.14	43.14	43.14	43.14	43.14	43.14	43.14
Peptide Ion Charge	1 1	4.86	4.86	4.86	4.86	4.86	4.86	4.86	4.86	4.86	4.86
	2 1	45.71	45.74	45.75	45.81	45.78	45.76	45.81	45.75	45.79	45.77
	2 2	10.89	10.85	10.85	10.79	10.81	10.84	10.78	10.84	10.80	10.82

Table 1(Contd.)

		Datasets									
		1	2	3	4	5	6	7	8	9	10
Peptide Ion Charge	3 1	14.58	14.52	14.51	14.56	14.56	14.57	14.60	14.58	14.53	14.59
	3 2	18.18	18.22	18.24	18.13	18.21	18.16	18.15	18.14	18.17	18.19
	3 3	5.79	5.80	5.80	5.85	5.77	5.81	5.80	5.83	5.85	5.77
Peptide Ion Basic	0 0	0.56	0.55	0.56	0.55	0.54	0.56	0.56	0.57	0.55	0.58
	1 0	20.00	19.95	19.98	19.97	20.04	19.99	20.05	19.96	19.97	19.98
	1 1	19.64	19.69	19.70	19.62	19.60	19.66	19.59	19.71	19.66	19.62
	2 0	6.10	6.05	6.07	6.08	6.07	6.08	6.11	6.08	6.09	6.08
	2 1	28.72	28.74	28.69	28.85	28.84	28.72	28.72	28.66	28.76	28.78
	2 2	24.97	25.02	24.99	24.92	24.92	24.99	24.97	25.02	24.96	24.96
Peptide Ion R	0 0	48.16	48.12	48.29	48.24	48.17	48.15	48.09	48.17	48.18	48.15
	1 0	21.76	21.71	21.72	21.71	21.79	21.77	21.77	21.66	21.76	21.71
	1 1	30.08	30.17	29.99	30.04	30.04	30.08	30.14	30.17	30.06	30.14
Peptide Ion K	0 0	34.73	34.78	34.72	34.64	34.76	34.72	34.85	34.72	34.80	34.81
	1 0	26.24	26.13	26.18	26.23	26.20	26.18	26.27	26.23	26.21	26.20

Table 1(Contd.)

		Datasets									
		1	2	3	4	5	6	7	8	9	10
Peptide Ion K	1 1	39.03	39.09	39.10	39.14	39.04	39.10	38.88	39.04	38.98	38.99
Peptide Ion P	00	43.94	44.05	44.02	44.02	44.09	43.97	44.07	44.07	44.04	43.98
	1 0	12.60	12.61	12.65	12.74	12.64	12.66	12.68	12.67	12.73	12.74
	1 1	18.51	18.52	18.47	18.45	18.48	18.48	18.43	18.45	18.49	18.41
	2 0	4.26	4.18	4.25	4.19	4.16	4.27	4.23	4.24	4.16	4.18
	2 1	8.15	8.03	8.01	8.07	8.07	8.08	8.07	8.01	8.05	8.10
	2 2	12.54	12.62	12.59	12.54	12.56	12.55	12.52	12.56	12.53	12.59
Peptide Ion H	00	64.44	64.37	64.35	64.44	64.42	64.44	64.38	64.48	64.35	64.36
	10	11.69	11.71	11.73	11.75	11.73	11.72	11.74	11.69	11.73	11.78
	11	23.87	23.92	23.92	23.81	23.85	23.84	23.88	23.83	23.92	23.86
Number of Proteins		7736	7741	7739	7740	7734	7740	7744	7750	7746	7743
Number of Peptides		61348	61353	61342	61351	61339	61339	61306	61319	61347	61342
Ion Length Range		1-59	1-65	1-59	1-64	1-64	1-62	1-65	1-61	1-61	1-64

Table 1(Contd.)

	Datasets									
	1	2	3	4	5	6	7	8	9	10
Rel Ion Weight	0.07-0.98	0.05-0.98	0.07-0.98	0.06-0.98	0.06-0.98	0.06-0.98	0.07-0.98	0.07-0.98	0.05-0.98	0.06-0.98
N.observations	654838	654184	654919	654658	654699	655011	655011	654356	655333	655331

Neutral Mass Loss: -1x = water (-17) or ammonia (-18), -3x = H_2S+H (-34), Cl (-35), or HCl (-36), and -4x = C_2H_5O+H; (-46), C_2H_5O (-45), or CO_2O, $CONH_2$ (-44); Peptide Ion: combination of charge state or residue counts in the peptide and resulting fragment ion; Ion Length Range: range of ion length or size; Rel Ion Weight range: range of relative ion mass values; N. observations: number of ion intensity observations in the data set.

Table 2. Average Estimates and Standard Errors of the Discrete Multi-level Explanatory Variables Significantly Associated with Fragment Ion Intensity and Number of Final Models Including the Variables Across the Ten Training Data Sets.

Effect[a]	level[b]	N Est[c]	Mean Est[d]	Range Est[e]	Mean SE[f]
Ion Type or Series	b	10	2.43902	0.022079	0.003931
	y	10	2.48188	0.012819	0.003889
Neutral Mass Loss	-17	10	2.40819	0.005631	0.003567
	-18	10	2.47720	0.005638	0.003555
	-3	10	2.34425	0.018845	0.006341
	-4	10	2.36713	0.006271	0.003803
	None	10	2.70547	0.004205	0.003359
Proton Mobility	Mobile	10	2.43249	0.009937	0.003391
	Non-Mobile	10	2.37827	0.012061	0.005146
	Partial-Mobile	10	2.57059	0.006004	0.003367
Peptide Ion Charge	11	10	2.54415	0.012570	0.004421
	21	10	2.53822	0.007113	0.003533
	22	10	2.39153	0.010151	0.003902
	31	10	2.45893	0.010439	0.004051
	32	10	2.49686	0.008988	0.003958
	33	10	2.33300	0.008896	0.004484
Peptide Ion Basic	00	10	2.44681	0.033485	0.009803
	10	10	2.44073	0.009946	0.004540
	11	10	2.54670	0.008479	0.003740
	20	10	2.26515	0.010319	0.004836

Table 2(Contd.)

Effect[a]	level[b]	N Est[c]	Mean Est[d]	Range Est[e]	Mean SE[f]
Peptide Ion Basic	21	10	2.48135	0.005018	0.002980
	22	10	2.58196	0.010615	0.004020
Peptide Ion H	00	10	2.48055	0.009345	0.003329
	10	10	2.42398	0.010281	0.003985
	11	10	2.47683	0.007227	0.004271
Peptide Ion K	00	10	2.44890	0.003875	0.003365
	10	10	2.44568	0.008848	0.003924
	11	10	2.48676	0.013049	0.004361
Peptide Ion P	00	10	2.51051	0.008746	0.003681
	10	10	2.40631	0.008925	0.003961
	11	10	2.33326	0.011327	0.004691
	20	10	2.50710	0.008018	0.003831
	21	10	2.43541	0.007525	0.004191
	22	10	2.57011	0.009685	0.004890
Peptide Ion R	00	10	2.45980	0.005589	0.003435
	10	10	2.41091	0.009105	0.003979
	11	10	2.51064	0.022590	0.004509

[a] Neutral Mass Loss: water (-17) or ammonia (-18), -3x = H_2S+H (-34), Cl (-35), or HCl (-36), and -4x = C_2H_5O+H; (-46), C_2H_5O (-45), or CO_2O, $CONH_2$ (-44); Peptide Ion: combination of charge state or residue counts in the peptide and resulting fragment ion;

[b] Levels of each explanatory variable ;

[c] number of training data sets significant at *p*-value < 0.00005;

[d] Average estimate across all 10 data sets;

[e] Range of estimates across all 10 data sets;

[f] Average standard error across all 10 data sets;

Table 3. Average Estimates and Standard Errors of the Binary Explanatory Variables Significantly Associated with Fragment Ion Intensity and Number of Final Models Including the Variables in at Least Seven of the Ten Training Data Sets.

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
ChargeN N2	10	0.31818	0.04401	0.010540
ChargeN N5	10	0.10284	0.01732	0.002889
ChargeN N6	10	0.06366	0.03205	0.003479
ChargeN N7	10	0.03888	0.01992	0.003526
ChargeO N2	10	0.34340	0.05107	0.010423
ChargeO N5	10	0.11028	0.01404	0.002480
ChargeO N6	10	0.07456	0.02385	0.003052
ChargeO N7	10	0.04862	0.02008	0.003083
H C9	10	0.05821	0.02771	0.005889
MILV C2	10	-0.02530	0.01598	0.001613
MILV N1	10	0.08147	0.10592	0.003027
R C2	10	0.06211	0.01549	0.005345
RelativeIon_Size	10	-1.00294	0.10487	0.024507
RelativeIonWeight	10	0.53649	0.12411	0.025384
R N2	10	-0.05285	0.02065	0.007295
P C2	10	-0.02114	0.01644	0.003083
Sulfur N3	10	0.03210	0.02657	0.004633
Aliph C4	9	-0.01354	0.00603	0.001488
ChargeN N4	9	0.09203	0.04960	0.004077

Table 3(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
ChargeO N4	9	0.09535	0.04671	0.003711
D N3	9	-0.01884	0.07454	0.003336
G C1	9	0.03054	0.00705	0.002815
P C1	9	0.10493	0.05017	0.003622
Philic N9	9	0.12122	0.02747	0.002878
Phobic N9	9	0.12125	0.02832	0.003388
R N1	9	-0.04436	0.01530	0.003716
N C2	9	0.03276	0.02313	0.003197
Sulfur C2	9	0.03533	0.02319	0.004812
ChargeN N8	8	0.05096	0.02275	0.003418
ChargeO N8	8	0.05448	0.02047	0.002941
G N1	8	-0.10934	0.05036	0.004317
HRK N2	8	0.20615	0.05149	0.010769
N N1	8	-0.07242	0.04438	0.004782
P N1	8	-0.20144	0.15790	0.005742
N C4	8	0.02059	0.00907	0.003282
G C2	8	0.01479	0.01109	0.002515
C C7	8	0.12851	0.04157	0.027714
D N4	8	-0.01578	0.00545	0.002969
ChargeO C1	7	0.03554	0.02032	0.002057
ChargeP N7	7	-0.04212	0.01544	0.004365

Table 3(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Hydroxyl N1	7	-0.03788	0.15034	0.003815
Imino C4	7	-0.08843	0.00938	0.003157
IminoN3	7	-0.07365	0.01217	0.003193
Imino N4	7	-0.06860	0.00912	0.003046
P C6	7	-0.08871	0.00673	0.003294
R C3	7	0.06390	0.02335	0.005421
K C5	7	-0.03585	0.01598	0.003848
M N1	7	-0.04678	0.03400	0.006577
L N2	7	0.01342	0.00692	0.002012
Sulfur C3	7	0.06527	0.13188	0.011102
W C1	7	0.03539	0.01340	0.005992
Hydroxyl N1	7	-0.03788	0.15034	0.003815

[a] Binary explanatory variables: XCn: X is any amino acid, C denotes C terminus and n is Position of the residue relative to the C terminus (e.g. C1 denotes at C terminus, C2 denotes one position from the C terminus); XNn: X is any amino acid, N denotes N terminus and n is Position of the residue relative to the N terminus (e.g. N1 denotes at N terminus, N2 denotes one position from the N terminus); Charge O, N, P: residues that have neutral, negative or positive charge respectively; Hydroxyl: Group of hydroxyl residues; Large: Group of large amino acids; Aliph: Group of Aliphatic amino acids; Sulfur: Group of sulfuric residues; DENQ: Group of specific amino acids with similar physicochemical properties.;

[b] number of training data sets significant at p-value < 0.00005;

[c] Average estimate across all 10 data sets;

[d] Range of estimates across all 10 data sets;

[e] Average standard error across all 10 data sets;

APPENDIX

Supplementary Table 1. Average Estimates and Standard Errors of the Binary Explanatory Variables Significantly Associated with Fragment Ion Intensity and Number of Final Models Including the Variables in at Least One of the Ten Training Data Sets.

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
ChargeN N2	10	0.31818	0.04401	0.010540
ChargeN N5	10	0.10284	0.01732	0.002889
ChargeN N6	10	0.06366	0.03205	0.003479
ChargeN N7	10	0.03888	0.01992	0.003526
ChargeO N2	10	0.34340	0.05107	0.010423
ChargeO N5	10	0.11028	0.01404	0.002480
ChargeO N6	10	0.07456	0.02385	0.003052
ChargeO N7	10	0.04862	0.02008	0.003083
H C9	10	0.05821	0.02771	0.005889
MILV C2	10	-0.02530	0.01598	0.001613
MILV N1	10	0.08147	0.10592	0.003027
R C2	10	0.06211	0.01549	0.005345
RelativeIonSize	10	-1.00294	0.10487	0.024507
RelativeIonWeight	10	0.53649	0.12411	0.025384
R N2	10	-0.05285	0.02065	0.007295
P C2	10	-0.02114	0.01644	0.003083
Sulfur N3	10	0.03210	0.02657	0.004633
Aliph C4	9	-0.01354	0.00603	0.001488
ChargeN N4	9	0.09203	0.04960	0.004077

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
ChargeO N4	9	0.09535	0.04671	0.003711
D N3	9	-0.01884	0.07454	0.003336
G C1	9	0.03054	0.00705	0.002815
P C1	9	0.10493	0.05017	0.003622
Philic N9	9	0.12122	0.02747	0.002878
Phobic N9	9	0.12125	0.02832	0.003388
R N1	9	-0.04436	0.01530	0.003716
N C2	9	0.03276	0.02313	0.003197
Sulfur C2	9	0.03533	0.02319	0.004812
ChargeN N8	8	0.05096	0.02275	0.003418
ChargeO N8	8	0.05448	0.02047	0.002941
G N1	8	-0.10934	0.05036	0.004317
HRK N2	8	0.20615	0.05149	0.010769
N N1	8	-0.07242	0.04438	0.004782
P N1	8	-0.20144	0.15790	0.005742
N C4	8	0.02059	0.00907	0.003282
G C2	8	0.01479	0.01109	0.002515
C C7	8	0.12851	0.04157	0.027714
D N4	8	-0.01578	0.00545	0.002969
ChargeO C1	7	0.03554	0.02032	0.002057
ChargeP N7	7	-0.04212	0.01544	0.004365

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Hydroxyl N1	7	-0.03788	0.15034	0.003815
Imino C4	7	-0.08843	0.00938	0.003157
Imino N3	7	-0.07365	0.01217	0.003193
Imino N4	7	-0.06860	0.00912	0.003046
P C6	7	-0.08871	0.00673	0.003294
R C3	7	0.06390	0.02335	0.005421
K C5	7	-0.03585	0.01598	0.003848
M N1	7	-0.04678	0.03400	0.006577
L N2	7	0.01342	0.00692	0.002012
Sulfur C3	7	0.06527	0.13188	0.011102
W C1	7	0.03539	0.01340	0.005992
ChargeP C2	6	-0.07979	0.01375	0.003155
ChargeP C6	6	-0.02321	0.01070	0.002768
ChargeP N3	6	-0.08843	0.08585	0.004560
Imino C5	6	-0.08377	0.00908	0.003189
Imino C7	6	-0.08297	0.01282	0.003439
Imino C9	6	-0.05137	0.00996	0.003913
Imino N2	6	-0.08282	0.01149	0.003227
Imino N6	6	-0.08500	0.00651	0.003188
Imino N7	6	-0.08376	0.00599	0.003324
Imino N8	6	-0.08392	0.00806	0.003437

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
MILV C3	6	-0.01421	0.00584	0.001616
P C3	6	-0.11615	0.00962	0.003163
P C8	6	-0.07137	0.00519	0.003689
P N5	6	-0.08463	0.00378	0.003118
P N9	6	-0.07872	0.01058	0.003702
R C4	6	0.05341	0.01454	0.005072
Tiny C1	6	0.03109	0.02750	0.001920
ChargeP N8	6	-0.04709	0.05704	0.004071
K C1	6	-0.02626	0.01759	0.003945
Hydroxyl N3	6	-0.01109	0.00451	0.001932
C C3	6	0.11642	0.04428	0.025457
Aliph C1	5	-0.03501	0.02058	0.001852
ChargeO N3	5	0.07178	0.01219	0.005190
ChargeP C9	5	-0.07449	0.00443	0.002886
ChargeP N9	5	-0.06890	0.01632	0.003836
DENQ C3	5	0.04946	0.04254	0.002687
HRK C9	5	-0.06969	0.00601	0.002871
HRK N9	5	-0.06734	0.01789	0.003911
L N1	5	-0.02772	0.00664	0.003170
N C3	5	0.03490	0.01088	0.003136
Phobic N1	5	0.06446	0.10119	0.003141

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Q C3	5	0.02284	0.01164	0.002767
ChargeP N6	5	-0.03176	0.00816	0.004454
I N1	5	0.03263	0.01424	0.003911
Aliph C5	5	-0.00941	0.00491	0.001463
H C5	5	-0.02895	0.01028	0.004214
R C1	5	0.03439	0.01736	0.005338
Amide N3	5	-0.00836	0.07278	0.003094
S N2	5	-0.01446	0.00821	0.002452
Aliph C6	5	-0.00785	0.00252	0.001515
H C7	5	-0.02247	0.01101	0.004621
Hydroxyl N4	5	-0.00950	0.00127	0.001946
Q C8	5	0.00745	0.05428	0.003970
Sulfur N4	5	0.02589	0.01019	0.004965
ChargeN N3	4	0.05959	0.03700	0.005748
ChargeO C3	4	0.06472	0.01572	0.003512
H C4	4	-0.03743	0.01667	0.004087
Imino C3	4	-0.11603	0.00770	0.003248
Imino C8	4	-0.07328	0.00560	0.003628
Imino N5	4	-0.08313	0.00608	0.003129
Imino N9	4	-0.07726	0.00798	0.003681
K C4	4	-0.03621	0.01217	0.003884

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Large N1	4	-0.02865	0.00267	0.002851
P C5	4	-0.08180	0.00263	0.003177
P C7	4	-0.08295	0.00648	0.003453
P C9	4	-0.05044	0.00790	0.003907
P N2	4	-0.08615	0.01919	0.003205
P N6	4	-0.08427	0.00720	0.003169
P N7	4	-0.08768	0.01022	0.003320
P N8	4	-0.08985	0.00830	0.003429
Small N1	4	-0.00522	0.06497	0.003296
S C1	4	0.01923	0.00639	0.002651
Sulfur C1	4	0.02987	0.01553	0.004473
D C2	4	-0.02003	0.02224	0.002851
HRK N4	4	-0.05425	0.09847	0.004596
Sulfur C5	4	0.02775	0.01625	0.005195
Sulfur C9	4	0.03744	0.01579	0.006517
V N2	4	-0.01280	0.00673	0.002448
ChargeO N9	4	0.01277	0.00147	0.002447
D N2	4	-0.01784	0.00388	0.003250
N C5	4	0.01588	0.00421	0.003278
Sulfur N5	4	0.02531	0.00574	0.005002
W C8	4	0.03642	0.00489	0.007986

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
ChargeN C3	3	0.02107	0.06684	0.004081
ChargeP C4	3	-0.03308	0.00634	0.003014
ChargeP N1	3	-0.09572	0.09392	0.004517
DENQ C1	3	0.01765	0.00890	0.001990
HRK C2	3	-0.07162	0.01731	0.002929
HRK C3	3	-0.05176	0.00513	0.003052
HRK N1	3	-0.13363	0.02714	0.004166
HRK N3	3	-0.10329	0.07238	0.004066
HRK N6	3	-0.03413	0.00151	0.004470
H C3	3	-0.05874	0.01215	0.004042
Imino C6	3	-0.08838	0.00186	0.003298
K C3	3	-0.04728	0.01556	0.004168
P C4	3	-0.09251	0.01120	0.003119
P N3	3	-0.07023	0.00102	0.003164
P N4	3	-0.06713	0.00624	0.003048
S N1	3	-0.05737	0.02403	0.004183
Y N1	3	0.09097	0.10690	0.005382
ChargeP C5	3	-0.02687	0.02194	0.003010
ChargeP N4	3	-0.03712	0.03201	0.005115
DENQ N3	3	0.03736	0.05770	0.003826
HRK C8	3	0.00217	0.06060	0.002933

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Hydroxyl C1	3	-0.01864	0.00864	0.002543
N C1	3	0.03570	0.03120	0.003080
A N2	3	0.01420	0.00499	0.002477
F C2	3	-0.01937	0.00676	0.003105
Hydroxyl C2	3	0.01120	0.00506	0.002072
Philic C2	3	0.01002	0.00820	0.001608
R C5	3	0.03190	0.00346	0.004994
MILV C5	3	-0.00943	0.00221	0.001628
N N3	3	0.01583	0.00599	0.003024
Sulfur C4	3	0.02627	0.00634	0.005168
Sulfur N6	3	0.02556	0.00295	0.005169
Sulfur N7	3	0.02640	0.00855	0.005543
T N1	3	0.02748	0.00347	0.004877
W N1	3	0.04396	0.00227	0.007922
Aliph C3	2	-0.01164	0.00012	0.001553
Amide C3	2	-0.04546	0.00035	0.004016
ChargeN N1	2	0.17044	0.09188	0.005762
ChargeO C4	2	0.05757	0.07200	0.003150
ChargeO N1	2	0.13939	0.00504	0.004617
ChargeP N2	2	0.20526	0.01341	0.011023
DENQ N1	2	0.10296	0.03343	0.003897

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
D C3	2	0.03062	0.00197	0.004185
D N1	2	0.05925	0.05376	0.005071
E N1	2	-0.05808	0.06445	0.004856
FYW C1	2	0.03142	0.00670	0.002437
HRK N7	2	-0.03610	0.00419	0.004301
HRK N8	2	-0.03393	0.00705	0.004417
Imino N1	2	-0.23278	0.03076	0.005739
MILV C1	2	-0.02702	0.00889	0.001796
Q C1	2	0.03625	0.00686	0.003012
STPAG C1	2	0.02602	0.00388	0.001905
STPAG N2	2	-0.01857	0.00147	0.001785
V N1	2	-0.08426	0.01386	0.005300
A C2	2	-0.01495	0.00364	0.002406
ChargeP C7	2	-0.01689	0.00495	0.002790
ChargeP C8	2	-0.01712	0.00753	0.002943
DENQ C4	2	0.01996	0.00905	0.002176
F C1	2	-0.01864	0.00767	0.003390
F N1	2	-0.02734	0.00475	0.004601
G N2	2	-0.01511	0.00611	0.002511
HRK C6	2	-0.02116	0.01384	0.002661
K C6	2	-0.02837	0.01635	0.003742

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
STPAG C3	2	0.01061	0.00709	0.001668
Amide N1	2	-0.01514	0.00008	0.003616
A N5	2	0.01088	0.00027	0.002462
C C5	2	0.11384	0.01512	0.022935
DENQ C8	2	0.01112	0.00396	0.002351
DENQ N9	2	-0.01060	0.00373	0.002157
I C5	2	-0.01444	0.00369	0.003003
I C8	2	-0.01490	0.00059	0.003636
M C3	2	-0.12188	0.01192	0.026525
Q C4	2	0.01391	0.00002	0.002920
Small N5	2	-0.00942	0.00132	0.001833
Sulfur N9	2	0.03132	0.00937	0.006355
V C6	2	-0.01190	0.00050	0.002692
W N2	2	0.02565	0.00328	0.005736
A C1	1	-0.01853	0.002647	0.002647
Acidic C3	1	0.01666	0.002301	0.002301
Aliph N1	1	0.02925	0.003131	0.003131
Amide C1	1	0.04020	0.002288	0.002288
Amide C2	1	0.02198	0.002348	0.002348
Amide C8	1	0.05867	0.004406	0.004406
A N3	1	0.01845	0.002543	0.002543

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Aroma C1	1	0.02603	0.002460	0.002460
Basic C2	1	-0.06981	0.003016	0.003016
Basic C4	1	0.05782	0.004674	0.004674
Basic C8	1	-0.01938	0.002736	0.002736
Basic N1	1	-0.15267	0.004024	0.004024
Basic N3	1	-0.06219	0.005889	0.005889
Basic N4	1	-0.03592	0.005277	0.005277
Basic N7	1	-0.04208	0.004188	0.004188
ChargeN C2	1	-0.01672	0.002054	0.002054
ChargeN C4	1	0.07736	0.005012	0.005012
ChargeN C8	1	0.04094	0.004611	0.004611
ChargeO C8	1	0.04953	0.003237	0.003237
D C1	1	-0.03759	0.003143	0.003143
HRK C4	1	-0.04422	0.003024	0.003024
HRK C5	1	-0.02871	0.002889	0.002889
H C6	1	-0.02931	0.004449	0.004449
Imino C1	1	0.09725	0.003557	0.003557
K N1	1	0.04461	0.003428	0.003428
Large C8	1	0.04861	0.003416	0.003416
MILV C4	1	-0.01157	0.001695	0.001695
Philic C3	1	0.05893	0.005187	0.005187

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Philic C9	1	0.12727	0.002785	0.002785
Phobic C3	1	0.05416	0.005329	0.005329
Phobic C9	1	0.12367	0.003217	0.003217
STPAG C8	1	0.04652	0.003703	0.003703
T C1	1	0.02324	0.002670	0.002670
A C4	1	0.00987	0.002557	0.002557
A C5	1	0.01395	0.002634	0.002634
Aliph N6	1	-0.00757	0.001645	0.001645
Amide C5	1	0.01097	0.001805	0.001805
Amide N2	1	-0.00842	0.001859	0.001859
A N4	1	0.01109	0.002338	0.002338
Aroma C3	1	0.01002	0.002535	0.002535
Aroma N1	1	0.01901	0.003504	0.003504
C C4	1	0.13739	0.028307	0.028307
C C6	1	0.11379	0.027274	0.027274
C C8	1	0.12228	0.029711	0.029711
C C9	1	0.12924	0.032929	0.032929
ChargeN N9	1	-0.01501	0.002465	0.002465
ChargeO C7	1	0.00804	0.001787	0.001787
ChargeP C3	1	0.03086	0.005645	0.005645
DENQ N6	1	0.01030	0.002399	0.002399

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
D N5	1	-0.01245	0.003149	0.003149
D N7	1	-0.02026	0.003629	0.003629
FYW C6	1	0.01328	0.002606	0.002606
FYW C9	1	0.01335	0.003263	0.003263
FYW N1	1	0.01561	0.003506	0.003506
G C5	1	-0.01281	0.002671	0.002671
G N3	1	-0.01186	0.002577	0.002577
H C2	1	0.02239	0.005298	0.005298
H C8	1	-0.02641	0.005044	0.005044
I C1	1	-0.01369	0.002947	0.002947
I C4	1	-0.01521	0.003048	0.003048
I N5	1	-0.01504	0.003023	0.003023
Large N2	1	0.00723	0.001504	0.001504
M C8	1	0.02859	0.006330	0.006330
M N4	1	0.02146	0.004990	0.004990
M N7	1	0.02423	0.005604	0.005604
N C8	1	0.01747	0.003965	0.003965
N N2	1	0.01555	0.003077	0.003077
Philic C8	1	-0.01077	0.002054	0.002054
Q C2	1	0.01251	0.002875	0.002875
Q N1	1	-0.01950	0.003920	0.003920

Supplementary Table 1(Contd.)

Effect[a]	N Est[b]	Mean Est[c]	Range Est[d]	Mean SE[e]
Q N2	1	-0.01338	0.002776	0.002776
R C6	1	0.02040	0.004935	0.004935
R N4	1	-0.03435	0.008093	0.008093
R N5	1	-0.03488	0.007946	0.007946
R N9	1	-0.03675	0.008892	0.008892
STPAG N3	1	-0.00816	0.001709	0.001709
S N3	1	-0.01481	0.002465	0.002465
S N4	1	-0.01127	0.002487	0.002487
Sulfur C6	1	0.02179	0.005313	0.005313
Sulfur C8	1	0.03011	0.006101	0.006101
Sulfur N8	1	0.02742	0.005975	0.005975
Tiny N2	1	-0.00804	0.001646	0.001646
V N5	1	-0.01279	0.002599	0.002599
W C3	1	0.03534	0.006198	0.006198
W C6	1	0.03667	0.007101	0.007101
W N7	1	0.03270	0.007484	0.007484
Y C1	1	-0.01786	0.003792	0.003792
Y C2	1	-0.01528	0.003654	0.003654

[a] Cx: Position of the residue relative to the C terminus (e.g. C1 denotes at C terminus, C2 denotes one position from the C terminus); Ny: Position of the residue relative to the N terminus (e.g. N1 denotes at N terminus, N2 denotes one position from the N terminus); Charge O, N, P: residues that have neutral, negative or positive charge respectively; Phobic: Group of amino acids with hydrophobic properties; Philic: Group of amino acids with hydrophilic properties; Aliph: Group of amino acids with aliphatic properties; Hydroxyl: Group of hydroxyl residues; Amide: Group of amide residues; Large: Group of large amino acids; Small: Group of small amino acids ; Tiny: Group of Tiny amino acids; Sulfur: Group of sulfuric residues; STPAG: Group of specific amino acids sharing same physicochemical properties; FYW: Group of specific amino acids sharing same physicochemical properties; MILV: Group of specific amino acids sharing same physicochemical properties; DENQ: Group of specific amino acids sharing same physicochemical properties;

[b] number of training data sets significant at *p*-value < 0.00005;
[c]Average estimate across all 10 data sets;
[d]Range Est: Range of estimates across all 10 data sets;
[e]Average standard error across all 10 data sets;

SAS CODE EXAMPLE

```
proc glmselect data=zeeshan.firstout;

class protein_id charge_ion_charge ProtonMobility NeutralMaasLoss ion_type pep_ion_R pep_ion_K pep_ion_H

pep_ion_P pep_ion_basic;

model l10inten= protein_id rel_ion_size NeutralMaasLoss protonmobility charge_ion_charge pep_ion_R pep_ion_K

pep_ion_H pep_ion_P Pep_ion_basic An1-An9 Ac1-Ac9 Cn1-Cn9 Cc1-Cc9 Dn1-Dn9 Dc1-Dc9 En1-En9 Ec1-Ec9

Fn1-Fn9 Fc1-Fc9 Gn1-Gn9 Gc1-Gc9 Hn1-Hn9 Hc1-Hc9 In1-In9 Ic1-Ic9 Kn1-Kn9 Kc1-Kc9 Ln1-Ln9 Lc1-Lc9

Mn1-Mn9 Mc1-Mc9 Nn1-Nn9 Nc1-Nc9 Pn1-Pn9 Pc1-Pc9 Qn1-Qn9 Qc1-Qc9 Rn1-Rn9 Rc1-Rc9 Sn1-Sn9 Sc1-Sc9

Tn1-Tn9 Tc1-Tc9 Vn1-Vn9 Vc1-Vc9 Wn1-Wn9 Wc1-Wc9 Yn1-Yn9 Yc1-Yc9 HRKn1-HRKn9 HRKc1-HRKc9

STPAGn1-STPAGn9 STPAGc1-STPAGc9 DENQn1-DENQn9 DENQc1-DENQc9 MILVn1-MILVn9 MILVc1-

MILVc9 FYWn1-FYWn9 FYWc1-FYWc9 Basicn1-Basicn9 Basicc1-Basicc9 Phobicn1-Phobicn9 Phobicc1-

Phobicc9 Philicn1-Philicn9 Philicc1-Philicc9 ChargeOn1-ChargeOn9 ChargeOc1-ChargeOc9 ChargeNn1-

ChargeNn9 ChargeNc1-ChargeNc9 ChargePn1-ChargePn9 ChargePc1-ChargePc9 Acidicn1-Acidicn9 Acidicc1-

Acidicc9 Hydroxyln1-Hydroxyln9 Hydroxylc1-Hydroxylc9 Aliphn1-Aliphn9 Aliphc1-Aliphc9 Sulfurn1-Sulfurn9

Sulfurc1-Sulfurc9 Amiden1-Amiden9 Amidec1-Amidec9 Aroman1-Aroman9 Aromac1-Aromac9 Iminon1-Iminon9

Iminoc1-Iminoc9 Largen1-Largen9 Largec1-Largec9 Smalln1-Smalln9 Smallc1-Smallc9 Tinyn1-Tinyn9 Tinyc1-

Tinyc9 rel_ion_size  ion_number  ion_type promw / include=1 selection =stepwise (select=SL SLE=.00005

SLS=.00005 choose=cp);

score data=all out=zeeshan.out1 p=pred r=resid;

run;


 proc mixed data=zeeshan.firstout covtest;

class protein_id charge_ion_charge NeutralMaasLoss ProtonMobility pep_ion_P pep_ion_R pep_ion_K

pep_ion_basic Ion_Type;
```

```
model l10inten=rel_ion_size NeutralMaasLoss ProtonMobility charge_ion_charge pep_ion_R pep_ion_K
pep_ion_P pep_ion_basic An4 Ac1 Ac2 Ac5 Cc7 Dn1 Dn3 Dn5 Dc2 Dc4 Dc5 Ec8 Fc2 Gn1 Gn3 Gn6 Hn1 Hn3 Hn4
Hc1 Hc2 Hc3 Hc4 Hc5 Hc6 Hc7 Hc8 Hc9 In1 Kn1 Kn2 Kc2 Lc4 Mc3 Nn1 Nn3 Nc2 Pn1 Pn5 Pn6 Pn7 Pn8 Pn9 Pc1
Pc2 Pc3 Pc4 Pc6 Pc7 Rn1 Rn2 Rc1 Rc3 Rc4 Rc5 Rc6 Sn1 Vn2 Vn3 Wn1 Wn2 Wn3 Wn4 Wc1 Yn1 HRKn2
HRKn3 HRKn5 HRKn8 HRKc2 STPAGn3 STPAGn4 STPAGc3 DENQc3 MILVn1 MILVn2 MILVc1 MILVc2
MILVc5 MILVc6 MILVc8 FYWn5 FYWn9 FYWc3 PhobicN1 PhobicN2 PhilicN9 ChargeON2 ChargeON4
ChargeON5 ChargeON6 ChargeON7 ChargeON9 ChargeOC9 ChargeNN2 ChargeNN4 ChargeNN5 ChargeNN6
ChargeNN7 ChargeNN9 ChargeNC1 ChargeNC3 ChargeNC7 AliphC4 AliphC7 SulfurN1 SulfurN3 SulfurN4
SulfurN5 SulfurN6 SulfurN7 SulfurC1 SulfurC2 SulfurC3 SulfurC5 AmideC4 LargeN3 LargeN7 LargeN8 LargeC6
LargeC7 LargeC8 LargeC9 SmallN1 TinyN2 TinyC1 TinyC5 Ion_Number  Ion_Type promw / solution;
random protein_id;
lsmean charge_ion_charge NeutralMaasLoss ProtonMobility pep_ion_P pep_ion_R pep_ion_K pep_ion_basic
Ion_Type / diff;
ODS OUTPUT SolutionF=zeeshan.sol1;
ODS OUTPUT LSMeans=zeeshan.lsm1;
ODS OUTPUT Diffs=zeeshan.difflsm1;
ODS OUTPUT CovParms=zeeshan.covarparms1;
run;
```