MCP Papers in Press. Published on October 2, 2006 as Manuscript M600320-MCP200

# Improved Validation of Peptide MS/MS Assignments Using Spectral Intensity Prediction

Shaojun Sun[1], Karen Meyer-Arendt[2], Brian Eichelberger[2], Robert Brown[3], Chia-Yu Yen[1], William M. Old[2], Kevin Pierce[2], Krzysztof J.Cios[1,4,5], Natalie G. Ahn[2,6], Katheryn A. Resing[2,†]

[1]Dept. of Computer Science and Engineering, Univ. of Colorado at Denver and Health Sciences Center, CO (UCDHSC), [2]Dept. of Chemistry and Biochemistry, Univ. of Colorado, Boulder, CO (UCB), [3]Yale University, New Haven, CN, [4]Dept. of Computer Science, UCB, and [5]Dept. of Preventive Medicine and Biometrics, UCDHSC, and [6]Howard Hughes Medical Institute, UCB.

[†] Datasets and software programs are available upon request from the corresponding author.
Corresponding author:
Katheryn A. Resing
Department of Chemistry and Biochemistry
University of Colorado
Boulder, CO 80309-0215
Phone: 303-735-4019
FAX: 303-492-2439
Katheryn.Resing@Colorado.edu

**Running Title:** Automating Manual Analysis of MS/MS Spectra

**Keywords:** Shotgun proteomics, MS/MS ion current, similarity score, theoretical MS/MS spectra, manual analysis.

**Abbreviations:**

ΔCN: difference between 1st and 2nd ranked assignments by XCorr

DTA: text file summary of MS/MS information

IntFrag score: Proportion of ion current assigned to internal fragment ions

MAE: Manual Analysis Emulator program

MS: Mass Spectrometer

MS/MS: Fragmentation spectra

PIC: Proportion of Ion Current

Sim: Similarity score

XCorr: Cross-correlation score by Sequest

## ABSTRACT

A major limitation in identifying peptides from complex mixtures by shotgun proteomics is the ability of search programs to accurately assign peptide sequences using mass spectrometric fragmentation spectra (MS/MS). Manual analysis is used to assess borderline identifications; however, it is error-prone and time consuming, and criteria for acceptance or rejection are not well defined. Here we report a Manual Analysis Emulator (MAE) program which evaluates results from search programs by implementing two commonly used criteria: (1) consistency of fragment ion intensities with predicted gas phase chemistry, and (2) whether a high proportion of the ion intensity (PIC = Proportion of Ion Current) in the MS/MS can be derived from the peptide sequence. To evaluate chemical plausibility, MAE utilizes similarity (Sim) scoring against theoretical spectra simulated by MassAnalyzer software (Zhang, Z., *Anal. Chem*. 2004, *76:*1002-8) using known gas phase chemical mechanisms. The results show that Sim scores provide significantly greater discrimination between correct and incorrect search results than achieved by Sequest XCorr scoring or Mascot Mowse scoring, allowing reliable automated validation of borderline cases. To evaluate PIC, MAE simplifies the DTA text files summarizing the MS/MS spectra and applies heuristic rules to classify the fragment ions. MAE output also provides data mining functions, which are illustrated by using PIC to identify spectral chimeras, where two or more peptide ions were sequenced together, as well as cases where fragmentation chemistry is not well predicted.

## INTRODUCTION

Recent advances in genome sequencing, mass spectrometry (MS) instrumentation, and chromatographic methods allow high throughput identification of peptide fragmentation spectra (MS/MS) from samples as complex as whole cell extracts (1,2). For very complex samples, the best results are obtained from ion trap mass spectrometers, due to their fast scanning rate and ability to rapidly shift between MS and MS/MS modes during data collection. However, when operated for high data collection rate, mass accuracy and resolution are compromised. It was recognized early on that scores from search programs showed poor discrimination between correct and incorrect sequence assignments when using ion trap MS/MS data to search large protein databases (3). Methods have been developed to specify thresholds for acceptance, either by searching datasets against an inverted sequence database of similar size to identify false positive scores (4), or by statistical analysis of multiple scores and results from normal searches (4,5). Using methods such as these, limits on search program scores or combinations of scores can be set to yield low number of false positives (6), but also will produce large false negative rates (4,7). This problem is more acute when larger databases are used.

To minimize false negatives, investigators often reduce the acceptance threshold in order to capture more information. Several methods have been developed to filter the resulting false positives, based on agreement between sequence composition of the peptides and their behavior on ion exchange or reversed phase chromatography (7,8), probability of missed cleavages (9), exact mass measurements (8), or differences in scores between the top ranking peptides and lower ranked candidates (10,11). Methods have also utilized intensity information in statistical or machine learning approaches for validation (12-15). Nevertheless, manual analysis of MS/MS spectra by an experienced annotator, who examines each spectrum for chemical plausibility, is regarded by many as the best method for validating borderline cases (16-18). In particular,

manual analysis considers other fragment ion types not evaluated by the search program, evaluates fragment ion intensities for chemical plausibility, and whether a peptide assignment is based primarily on noise peaks. However, manual analysis lacks uniform criteria, it can be error prone (17), and it is impractical with large datasets.

Accurate methods of predicting fragment ion intensity would allow the evaluation of chemical plausibility to be automated. Recently, the known gas phase chemistry mechanisms of peptides (12,19,20) were incorporated into a kinetic model for peptide fragmentation by Z. Zhang and implemented in the program MassAnalyzer, which simulates MS/MS spectra including relative fragment ion intensities (21,22). Similarity scoring between observed MS/MS spectra and the theoretical spectra generated by MassAnalyzer showed excellent discrimination between correct *vs.* incorrect assignments for LCQ and LTQ MS/MS of standard peptides and digests of purified proteins. This suggests that simulated spectra can be used to automate the evaluation of chemical plausibility for validating search results in complex samples. Such an approach complements methods that match MS/MS spectra to libraries of previously observed spectra (23,24), because it can assess any predicted peptide sequence and can be rapidly adapted to different mass spectrometers or sample preparation methods. Therefore, one of our goals was to test the performance of MassAnalyzer-generated spectra on multidimensional LC/MS/MS ("MuDPIT" (25)) datasets collected on tryptic digests.

Here we report a "ManualAnalysisEmulator" ("MAE") program, developed to automate key aspects of manual analysis, minimize subjective decisions, and enable high-throughput processing. We evaluated MAE performance with datasets of varying complexity (Table 1) and found substantial discriminating power of a Similarity (Sim) score using the theoretical spectra to evaluate the chemical plausibility of Sequest and Mascot search results. In addition, we developed a score to measure the proportion of the MS/MS ion current accounted for by the

peptide sequence (Proportion of Ion Current, or PIC). A commonly used test for manual analysis is to account for most of the ion intensity in the MS/MS spectrum (16-18). However, MAE analyses revealed that MS/MS spectra with borderline PIC scores were often correctly assigned, but included more than one peptide ion in the MS isolation window (chimera spectra). MAE also provides functions for MS/MS data mining, which were effective for identifying unexpected fragmentation chemistries. Thus, MAE provides a useful platform for assessing the validity of search program results and for mining information about gas phase chemistry from large proteomics datasets.

## METHODS

**Data collection, processing of raw files, database searching and analysis of peptide and protein assignments.** The datasets utilized in this study are summarized in Table 1, and details for analysis of each sample are given in Table 1 footnotes. (These datasets have been deposited with PeptideAtlas and all Raw files, DTA files, and program reports of analyses not in Supplementary Data can be obtained from the authors upon request.) Briefly, the complex MuDPIT samples were derived from the soluble proteins of human erythroleukemia K562 cells, either unfractionated or fractionated by gel filtration chromatography, as previously described (7). K562 proteins or standard proteins were alkylated on Cys with iodoacetamide, and proteolyzed with unmodified trypsin. The resulting peptide digests were analyzed by reverse phase LC-MS/MS performed on LCQ Classic, Deca XP, or LTQ-Orbitrap mass spectrometers (ThermoElectron) or fractionated by off-line strong cation exchange (SCX) chromatography before LC-MS/MS.

*[insert Table 1]*

Raw data was centroided during data collection, using vendor default parameters. DTA file summaries for the MS/MS spectra were generated from the Raw files using TurboSequest 3.0 (Extract_MSN module), with intensity threshold = 10,000, peptide mass tolerance = 2.5 Da (average mass), grouping of 1-5 scans and minimum ion count = 35. An in-house script concatenated DTA files into a Mascot Generic File for Mascot searches. Mascot (version 1.9) and TurboSequest searches were carried out against the IPI protein database (version 3.1) or a database where each protein sequence in the same IPI database was inverted to read from C- to N-terminus. Search parameters allowed 2.5 Da (avg) peptide mass tolerance and 1.0 Da (avg) fragment ion mass tolerance, as previously optimized to maximize the number of peptides identified (7), allowing only fully tryptic products, with up to two missed cleavages. In-house parsers were used to extract search results into an Oracle 9i database. In addition to MAE, we utilized in-house MSPlus and IsoformResolver programs (7) to analyze the search results. The MSPlus program evaluates consensus between the first Sequest and the top two Mascot search results, filtering out cases with unlikely physicochemical properties (in this study, validation required that the number of basic residues was consistent with observed charge on the parent ion and the SCX elution properties (7), and excluded unlikely trypsin missed cleavage products (9)). IsoformResolver generates protein profiles (either based on score thresholds or MSPlus evaluation), by resolving ambiguous protein assignments due to peptide isoforms that the MS/MS cannot differentiate, then reports the minimum number of proteins that account for the identified peptides.

The XCorr, Mowse, and MAE scores for the DTA files for Sample 2 are given in Supplementary Table 1 (most of the analyses in this study were done on this dataset) and the protein profile is given in Supplementary Table 2. When evaluating search results where the number of identifiable MS/MS is not known, we report false discovery rates [FDR= FP/(TP +

FP); FP = false positives, TP = true positives], either estimating the number of FPs using the inverted database search, as recommended by Weatherly et al. (6) or by directly determining the FPs by manual analysis. When evaluating results with the sequence inverted protein database, we report false positive rates [FPR = FP/(TN+FP); TN=true negatives], because in this analysis, FDR is meaningless, as it equals one under every condition (TP = 0). In some analyses, a high confidence subset of a dataset was used; this subset required that both Sequest and Mascot agreed on a sequence, that either XCorr or Mowse was above score thresholds yielding 0 FPs from an inverted sequence database search, that Sequest RSP score = 1, and that the sequence satisfied charge state, SCX, and missed cleavage rules of MSPlus.

**MAE program.** MAE was built using C++ to carry out three major steps: (1) simplify each low resolution MS/MS spectrum by eliminating noise and combining isotope clusters into single peaks, (2) calculate a Similarity (Sim) score which evaluates chemical plausibility based on relative fragment ion intensities when comparing an observed MS/MS spectrum to a theoretical spectrum, and (3) identify all possible fragment ions, including those not considered by conventional search programs, based on heuristic rules used in manual analysis, and calculate a Proportion of the total Ion Current (PIC) score for each MS/MS spectrum which accounts for fragment ion assignments. MAE inputs and outputs are described in Fig. 1. Theoretical spectra were generated using MassAnalyzer 1.03 (21,22), except that average parent and fragment ion masses were reported. In order to achieve rapid access to these spectra, we constructed a database of the +1 to +4 charge forms for all tryptic peptides allowed by the search strategy. The theoretical MS/MS database consisted of 8,332,846 spectra, formatted in a three level tree by the first three amino acids in the sequence. MAE required 77.74 msec on a Pentium 4 computer (CPU 2.6G) to evaluate a candidate sequence, when compiled with $C^{++}$ 6, student version.

*[insert Fig. 1]*

The output files of MAE include (1) simplified DTA (sDTA) files, created after noise processing and removal of isotope peaks, (2) a spreadsheet summary, reporting scores from search programs and MAE, as well as MSPlus evaluation of consensus between search programs and consistency between observed and predicted physicochemical properties, and (3) a concatenated summary report for each DTA file that focuses on information used in our manual analyses and data mining studies (an example is shown in Supplementary Fig. 1). The concatenated summary report includes the candidate peptide sequence, observed and calculated parent ion masses, major fragment ions masses and intensities sorted by ion type and charge, the primary and alternative annotations of type of the major fragment ions and the mass difference from predicted values, the results of heuristic (true/false) tests, information about observed neutral losses from the parent ion, and a summary of search program scores and MAE-generated scores which evaluate the chemistry, such as Sim (Eq. 3), PIC (Eq. 4), Intfrag (Eq. 5), and the ratios between observed intensities of each ion and those predicted by the theoretical spectra.

**Simplifying the information in DTA files.** To generate the sDTA files, noise peaks from each MS/MS spectrum are evaluated and removed. (Here, we describe sDTA processing for the LCQ data. Similar issues exist for the LTQ-Orbitrap, although analysis of the LTQ-Orbitrap dataset, Sample 6, utilized simpler methods for processing as described in Table 1, footnote 1). The first step is to create an "ion list" of the most intense DTA ions (a DTA ion is one line in the DTA file) equal to seven or fourteen times the number of amino acids in the candidate peptide for singly or multiply charged cases, respectively. This division rule encompasses 98% of the obvious ions in a survey of the high intensity, richly fragmenting MS/MS spectra in Sample 2 and 5. The remaining DTA ions are categorized as “bulk noise ions”.

In order to simplify the spectrum, the lines remaining in the ion list after removing the bulk noise are grouped into ”clusters" of DTA ions and combined, as described in the legend for Fig.

2. The resulting single peak is referred to as an "Average Mass Ion", with intensity equal to the sum of all the included DTA ions. The m/z and intensity of the Average Mass Ions are calculated using Eq. 1 and Eq. 2, showing an example of a cluster with three fragment ions. To distinguish the processed Average Mass Ions from the original data in the DTA file, we refer to the weighted average m/z of each reprocessed peak as "M/z", where $I_k$ and $M_k$ are intensity and m/z of DTA ions within each cluster which are combined during processing of the DTA file.

$$I = \sum I_k = 2656 + 19046 + 5015 = 26717 \quad (1)$$

$$M/z = \frac{\sum (I_k M_k)}{\sum I_k} = \frac{1103.3 \times 2656 + 1103.5 \times 19046 + 1104.5 \times 5015}{2656 + 19046 + 5015} = 1103.67 \quad (2)$$

The resulting *I* and *M/z* values for each Average Mass Ion are written to the sDTA. An error correction of 1.0002 × M/z was applied to each M/z value, after analyzing the high quality assignments in the list (e.g., Supplementary Fig. 2). A second noise threshold was then determined, where Average Mass Ions with *I* greater than the mean + 2.8 × standard deviation (s.d.) of the bulk noise ions were classified as "major ions" for the next step of processing. This threshold was chosen to eliminate >95% of noise ions in spectra where MS/MS had apparently been carried out on an MS noise peak. Such cases were identified when spectra were very weak, no MS/MS was observed for ions with similar m/z and reverse phase retention time but with higher signal, and no peptide tag sequence could be identified. A third noise threshold was used in calculating PIC score as described below, excluding any Average Mass Ion below 3000 counts. Supplementary Fig. 3 summarizes the method used to set this threshold. The effect of simplifying clusters by combining ions with similar m/z values is evaluated in Supplementary Fig. 4.

**Similarity scoring.** The Similarity (Sim) score is the ratio between the sum of the geometric mean and the sum of the arithmetic mean of ions in each spectrum, normalizing the two spectra to the total intensity in the spectrum, as described by Zhang (19):

$$Similarity = \frac{\sum \sqrt{kI_A I_B}}{\sum \frac{kI_A + I_B}{2}} = \frac{\sum \sqrt{I_A I_B}}{\sqrt{(\sum I_A)(\sum I_B)}}, \quad where \quad k = \frac{\sum I_B}{\sum I_A} \qquad (3)$$

where $I_A$ is the summed intensity of the observed ions that make up each Average Mass Ion and $I_B$ is the intensity of each ion in the theoretical spectrum. To reduce the impact of the remaining noise ions, any ion in the sDTA that was not in the theoretical DTA and that had intensity less than 0.5% of the total observed ion intensity, was not considered in the Sim scoring. This filter was the only noise processing used for the preliminary analysis of the LTQ/Orbitrap dataset. When comparing theoretical and experimental spectra for Sim scoring for the LCQ mass spectrometers, a mass tolerance of ±(0.45 Da + 0.00085 × M/z) was allowed for the 3D traps, determined from error analysis of the M/z values for fragment ions (Supplementary Fig. 2), and +/- 0.2 Da for the LTQ/Orbitrap dataset.

**Annotating the fragment ions.** In the next step, the Average Mass Ions are annotated by comparison with average m/z values of possible theoretical fragment ions, including sequence-specific *a*, *b*, and *y* ions, multiply dehydrated *b* ions (up to four), internal fragmentation products, and $b_{n-1}$+18 Da ions produced by C-terminal rearrangements (20). Multiple charges are considered for all except internal fragment ions (produced from cleavage of two peptide bonds). MAE assigns fragment ions in the sDTA in two stages, based on how closely the observed M/z values match theoretical average masses. First, ions with narrow mass tolerance ±(0.25 Da + 0.00045 × M/z) are assigned in the order listed below; this mass tolerance was chosen to include >95% of the fragment ion assignments in MS/MS spectra of standard peptides

where parent ion counts were ≤ 80% of the target values, in order to minimize inaccuracy due to space charging. Second, ions with greater mass tolerance ± (0.45 Da + 0.00085 × M/z) are assigned to accommodate cases with higher mass error. This larger window was determined from error measurements for replicate MS/MS of the same peptide observed in several datasets, and captures cases where there are D/N or Q/E isoforms, where the 2$^{nd}$ or 3$^{rd}$ isotopic peak was sequenced, or where the isotopic envelope of an intense fragment ion is split into two Average Mass Ions by MAE processing.

Only one fragment ion classification is made for each Average Mass Ion, even when several are possible (for example, analysis of the 2$^{nd}$, 4$^{th}$, and 5$^{th}$ peptide bond in DTA files from Sample 5 indicates that 8% of doubly charged ions are isomeric with a singly charged ion). Rules commonly utilized in manual analysis are employed to assign the most likely annotation, but all other possibilities are listed in the summary output. Fragment ions are classified in the following order: (1) parent ion and ions generated by neutral losses of water/ammonia, the quanidinium side-chain (only allowed when Arg is present), or $H_2CO_3$, (2) singly charged "canonical" $b_n$ and $y_n$ ions produced by cleavage at one peptide bond, (3) singly charged canonical $a_n$ ions, (4) singly charged dehydrated/deammoniated canonical $a_n$, $b_n$, and $y_n$ ions, and C-terminal rearrangements ($b_{n-1}$+18), (5) singly charged multiple dehydrated/ deammoniated canonical $b_n$ ions, (6) doubly charged canonical $b_n$ and $y_n$ ions, (7) doubly charged canonical $a_n$ ions, (8) doubly charged dehydrated/deammoniated canonical $a_n$, $b_n$, and $y_n$ ions, and C-terminal rearrangements ($b_{n-1}$+18), (9) doubly charged multiple dehydrated/deammoniated canonical $b_n$ ions, (10) triply charged canonical $b_n$ and $y_n$ ions, (11) triply charged canonical $a_n$ ions, (12) triply charged dehydrated/deammoniated canonical $a_n$, $b_n$, and $y_n$ ions, and C-terminal rearrangements ($b_{n-1}$+18), (13) triply charged multiple dehydrated/deammoniated canonical $b_n$

ions, (14) *b* ions generated by internal fragmentation, (15) *a* ions from internal fragmentation, (16) dehydrated/deammoniated *a* and *b* ions from internal fragmentation.

**Additional scores generated by MAE.** Several additional scores are generated by MAE; only two of these are used in this study. The Proportion of Ion Current (PIC) score is applied to each set of major ions (excluding those with $I_{M/z}$ <3000 counts), and is defined by Eq. 4:

$$PIC = \frac{\sum I_{matched}}{\sum I} \quad (4)$$

Where $I_{matched}$ is the intensity of each assigned Average Mass Ion that matches to a theoretical fragment ion, and $I_A$ is the intensity of each observed Average Mass Ion in the sDTA file.

Internal fragment ions have high probability of generating combinatorial redundancies; therefore, a score to assess internal fragments was added to the MAE output. The internal fragmentation ion score ("Intfrag") evaluates the percentage of observed fragments accounted for by internal cleavages, and is defined by Eq. 5:

$$Intfrag = \frac{\sum I_i}{\sum I} \quad (5)$$

where $I_{int}$ is the intensity of each observed ion identified as an internal fragmentation product.

**Rules for fragment ions generated by secondary cleavages.** In order to decide between alternative ion types and to minimize chance assignments due to the large number of combinatorial possibilities, heuristic rules for fragment ion annotations based on simple chemical rules for multi-step cleavages (indicated by → ) or parallel reactions that are expected to be independent of each other (indicated by || ) are applied as follows:

1. Do not assign a dehydrated (Δ) or deammoniated ion (∇), if the corresponding unmodified form is absent (using the rule that "unmodified → Δ/∇"). We use the standard Δ symbol for dehydration, but in some cases, the ion shows mass closer to a deammoniated ion.

Stochastic variation in intensity of different isotope peaks produces ambiguity between dehydrated and deammoniated ions (see expanded view panels in Fig. 2 that show individual ions, keeping in mind that dehydration and deammoniated ions are 1 Da apart and have overlapping isotope peaks). Therefore, we utilize a ∇ symbol for potentially deammoniated ions, instead of the more common $-NH_3$ nomenclature, in order to emphasize that the evidence for the observed deammoniated ions is inadequate to distinguish from dehydration.

2. Do not assign an *a* ion if the corresponding *b* ion is absent, except for *a* ions generated by internal fragmentation. This rule can be expressed as: "$b \rightarrow a$".

3. Do not assign a dehydrated/deammoniated (Δ/∇) *a* ion if the ratio of intensities between the Δ/∇ *a* ion and its corresponding *a* ion ($R_a$) is significantly different from the ratio of intensities between the Δ/∇ *b* ion and its corresponding *b* ion ($R_b$) generated by cleavage of the same peptide bond. This assumes that the two reactions are independent: "(a $\rightarrow$ Δ/∇ *a*) || (*b* $\rightarrow$ Δ/∇ *b*)". This rule is utilized when $|R_a-R_b| < 0.15 (R_a+R_b)$ to avoid misclassifying cases where small changes could be due to stochastic differences in ion counting.

4. Do not assign a product of C-terminal cleavage to Pro, unless no other alternative is possible. This rule is based on known chemistry of Pro (12,19), and forces the program to change the assignment order for fragment ions when this situation occurs.

5. Do not assign internal fragment ions unless there is moderately high likelihood for chemical cleavage at both ends. This rule assumes that the two cleavages are independent: "(cleavage at site 1) || (cleavage at site 2)", where site 1 and site 2 represent the peptide bonds defining the ends of the internal fragment ion. We assume that a fragmentation event has similar chemistry whether it represents the first or second cleavage. Thus, we estimate the likelihood of cleavage at each peptide bond from the intensities of the canonical *a, b* or *y* ion generated by

fragmentation at each site, which are directly related to the reactivity of that site. The probability of obtaining an internal fragmentation between the $s^{th}$ and $t^{th}$ peptide bonds is estimated by Eq. 6:

$$Probability_{st} = \frac{\sum I_s}{\sum I} \times \frac{\sum I_t}{\sum I} \quad (6)$$

where $I_s$ and $I_t$ are the intensities of each observed ion identified as sequence-specific *a, b,* or *y* ions representing cleavage at the *s* and *t* peptide bonds, and *I* is the intensity of each observed ion of any type except the internal fragment ions.

6. For internal fragment ions, the order of assignment prioritizes *b* over *a* fragment ions and unmodified fragment ions over Δ/∇-modified *a* or *b* ions (for example $\Delta a_n$). This rule assumes that the loss of $H_2O/NH_3$ and CO are independent: "($b \rightarrow a$) || (unmodified $b \rightarrow \Delta/\nabla b$)".

7. When two or three Δ/∇ derivatives of a *b* ion are present (e.g., $b_n$, $\Delta b_n$, $\Delta\Delta b_n$, $\Delta\Delta\Delta b_n$), the related ions must follow intensity patterns that assume sequential reactions: "unmodified → Δ → ΔΔ → ΔΔΔ" For example, a set of three related ions should show intensity patterns $b_n \geq \Delta b_n \geq \Delta\Delta b_n$, $b_n \leq \Delta b_n \leq \Delta\Delta b_n$, or $b_n \leq \Delta b_n \geq \Delta\Delta b_n$, and exclude the pattern $b_n \geq \Delta b_n \leq \Delta\Delta b_n$. If the set fails this test, the ΔΔ form should not be assigned. Similar patterns should be assessed for sets of four ions; if the set fails the test, the ΔΔΔ ion should not be assigned and the first three ions should be reconsidered. This annotation test considers alternative assignments and low intensity ions in the sDTA files when testing for the unmodified element in this series, and requires that the peptide sequence has sufficient Ser/Thr to account for the multiple dehydration events (# Ser/Thr ≥ # of Δs allowed).

## RESULTS

**Processing of vendor software-generated DTA files to simplify MS/MS spectra.** Our first goal was to test whether MassAnalyzer theoretical spectra could be used in place of manual analysis to evaluate chemical plausibility. To do this, we must score for similarity between the observed and theoretical spectra. For optimum alignment of the two spectra, we must take into account the processing that occurs when raw data files are created during the typical data collection mode used for high-throughput proteomics profiling, as well as the way that the information is extracted from those files by the vendor software in creating the text DTA files that summarize the MS/MS information. In the following discussion, the m/z entries in the original DTA files are referred to as "DTA ions" and a group of DTA ions corresponding to one fragment ion is referred to as a "cluster". Although the clusters of DTA ions resemble isotope peaks, particularly for the more intense singly charged ions, they are in reality created by centroid processing (e.g., expanded views in Fig. 2A,B). For multiply charged or weak singly charged ions, the resulting spectra can be difficult to interpret, because isotopic peaks cannot be reconstructed reliably and the charge information is obscured (e.g., Fig. 2, expanded panel for 1082-1084 Da).

*[insert Fig. 2]*

A solution to these problems is to simplify the clusters into one peak. Ideally, we would like to use the monoisotopic peak, but computationally it is easiest to identify a cluster by locating its most intense DTA ion. The peak corresponding to the monoisotopic mass may not be the most intense DTA ion, or it may be absent in low intensity, high m/z ions or cases where the centroid processing distorts the distribution of the ion intensity. Thus, the average m/z of the original fragment ions is more reliably reconstructed from the information in the DTA files, than is the monoisotopic m/z. MAE processing of the DTA information to remove noise and simplify

ion clusters into single peaks is described in the legend to Fig. 2. Briefly, MAE separates DTA ions into two classes: "major ions" and "bulk noise ions". Clusters in the major ion list are then combined to produce "Average Mass Ions", by calculating the weighted average mass of the DTA ions within a -2.0 to +2.5 Da window of the most intense DTA ion in the cluster (Eq. 2), but terminating at the point that DTA ions are >1.0 Da apart. We utilize the unit designation "M/z" for these averaged ions to distinguish them from the unprocessed data. An example of the resulting simplified DTA (sDTA) file is illustrated in Fig. 2C.

**Characterizing the sDTA information.** Accuracy of the DTA signal processing by MAE was assessed by examining mass errors of fragment ion assignments (theoretical m/z – observed M/z) of the cases where we were confident of the peptide identification. The mass error distribution of all *b* and *y* fragment ions showed an offset of the mean from zero (mean = 0.1535, s.d. = 0.492), indicating a systematic error in mass determination (Supplementary Fig. 2B), which varied with fragment ion M/z (Supplementary Fig. 2A). This was not an artifact due to combining cluster lines, because the same systematic error was observed in unprocessed clusters where we could identify the monoisotopic peaks. It appears to be an aspect of the mass inaccuracy of ThermoElectron 3D ion traps and has been observed in other labs (18,26,27), but was not observed with the LTQ/Orbitrap when MS/MS were collected in the LTQ. After applying a correction factor for the LCQ datasets, the resulting error distribution after the mass correction is shown in Supplementary Fig. 2C,D (mean = –0.035, s.d. = 0.430). Applying this correction factor to a larger dataset collected a year later (Supplementary Fig. 2E,F) showed similar results (mean = +0.088, s.d. = 0.347). The range of observed errors is well within the range expected for 3D ion traps (26,27), and indicates that ions within 1.5 Da of each other cannot be accurately distinguished.

In combining DTA ions in a cluster, two difficult cases were seen, when very intense singly charged fragment ions produced clusters up to 5 Da wide or when two different fragment ions appeared with monoisotopic m/z values within 3 Da of each other. For the first case, clustering of the high intensity DTA ions often produced two adjacent Average Mass Ions differing by 1.5-2.0 Da, one with high signal intensity and one with low intensity, where the most intense Average Mass Ion was within 0.5 Da of the expected mass in 100% of 23 randomly chosen cases. Thus, the Average Mass Ion provided a reasonable representation of the major fragment ion that was well within the error distribution for fragment ions in high confidence cases. The second, weaker peaks were as much as 2 Da larger than the main peak; they accounted for most of the outlier values in the error analysis in Supplementary Fig. 2.

For the second case, we evaluated spectra where it was likely that corresponding *b* and *y* ions were present and within 4.5 Da of each other (Supplementary Fig. 4). We surveyed 48 randomly chosen spectra of $MH^+$ and $MH_2^{+2}$ parent ions containing this type of potential ambiguity. Thus, the method of combining clusters into Average Mass Ions combined only those ions separated by less than ~2 Da, rather than the 4.5 Da which might be expected from the window size, most likely due to termination of ion summation when gaps >1.0 Da were encountered. This resolution was similar to that expected, given the distribution of errors for all Average Mass Ions (Supplementary Fig. 2 shows 3 × s.d. = 1.25 Da).

After the signal processing, the resulting ion list of Average Mass Ions is written to a simplified DTA (sDTA) that is then used as input to other MAE functions (Fig. 1). To test whether processing into sDTA files removed any critical information, sDTA files in a small MuDPIT dataset of K562 cell proteins (Sample 2, Table 1) were tested in searches with Mascot and Sequest, and the results were compared against those obtained when searching using the unprocessed DTA files. Overall, Mascot Mowse scores for correct assignments remained the

same or increased by 1-5% due to better matching of the high mass ions, and the scores for incorrect hits decreased slightly, most likely due to the removal of isotopic peaks. The net effect was a small increase in discrimination. On the other hand, >90% of Sequest XCorr scores decreased with the sDTA files, sometimes by as much as 25%, which we attributed to the reduction in noise. Sequest SP and ion scores increased, because these scores are sensitive to the more accurate matching between observed M/z and predicted m/z values. (SP evaluates the presence of continuous "runs" of *b* and *y* ions, while ion score evaluates the percentage of predicted *b* and *y* ions that are observed.) Importantly, no correct Sequest or Mascot assignments made with DTA files were lost by searching using the sDTA files. These results are consistent with other studies reporting similar improvements upon deisotoping and removing noise from DTA files (28). Because the purpose of our study is to evaluate search results on DTA files, all other searches in this study were done using DTA files. However, this test demonstrated that the sDTA files lose no significant information after processing by MAE, and in fact the processing increases the likelihood that predicted and observed fragment ions can be matched.

**Similarity (Sim) scoring against theoretical MS/MS spectra generated by MassAnalyzer.** We next tested the performance of DTA files *vs.* sDTA files in Sim scoring of the Sample 2 dataset. To do this, we generated two sets of theoretical spectra, (1) those generated by a version of the MassAnalyzer program that calculates isotope peaks of the fragment ions, which were compared to DTA ion m/z values, or (2) those generated by a modified version of MassAnalyzer that calculates average m/z values of the fragment ions, which were compared to sDTA Average Mass Ion M/z values. We then evaluated Sim scores comparing theoretical spectra to either sDTA or DTA files. Analysis of the Sample 2 dataset showed a bimodal distribution for the Sim scores with sDTA files (Fig. 3A, open symbols). The higher scoring class in the complex test sample (sDTA, Fig. 3A) aligned well with the score

distribution of 175 validated MS/MS from a dataset collected on standard proteins (Fig. 3C, Sample 1 in Table 1), while the lower scoring class aligned with false positives generated by searching against inverted protein sequences (Fig. 3B). In addition, the ratio of areas of the two classes in the complex test sample was similar to our previous estimate of the ratio between normal tryptic MS/MS and other MS/MS in this dataset (approx. 1:3, when $MH_2^{+2}/MH_3^{+3}$ duplicated DTA files are included). Thus, the two peaks roughly corresponded to the expected ratio of correct *vs.* incorrect assignments.

*[insert Fig. 3]*

In contrast to the results with the sDTA files, Sim scores against unprocessed DTA files revealed less separation between the two classes as well as a shift in the high scoring peak to lower values (Fig. 3A). On the other hand, false positives generated by searching against an inverted sequence database revealed no significant differences between DTA and sDTA files (Fig. 3B). These results showed that the processing of spectra into sDTA files led to increased Sim scores for correct assignments, but little change for incorrect assignments, resulting in improved discrimination.

Several tests were carried out to assess whether the high *vs.* low scoring peaks in the biphasic distribution corresponded to correct *vs* incorrect assignments. (1) In earlier studies, 70% of the 1,838 MS/MS in Sample 2 (after removing spectra with low signal intensity) were evaluated manually. Manual analysis was carried out on all MS/MS where the Sequest ΔCN score was ≥ 0.08, or where Sequest and Mascot agreed on the peptide assignment (7,9). In addition, any of the top five Mascot assignments were examined whenever Sim was ≥ 0.45, which validated lower ranked sequences. Together, these confirmed 925 MS/MS assignments as correct, including those where the correct assignment was a lower ranked "hit" or where only Sequest or Mascot alone could correctly identify the spectrum. In all cases, those MS/MS that

were manually validated as correct showed Sim >0.47 (Fig. 3D). (2) In contrast, the remaining 30% of cases that were not evaluated manually showed very low Mowse or XCorr scores, and were most likely incorrect. All distributed to the low range of Sim score, less than 0.47. A random sampling of 45 cases were then further evaluated; manual analysis showed that all of their top-ranked Sequest and top two ranked Mascot assignments were incorrect. (3) Finally, the two alternative charge forms of each multiply charged DTA, generated by the LCQ, were examined for incorrect charge forms (referred to as "decoys" in Supplementary Table 1, column 5) in those cases where the correct form could be identified. All of these incorrect assignments showed Sim score less than 0.5. Taken together, this extensive analysis of the Sample 2 dataset showed that the range of Sim scores for the true positive class was nearly identical to that observed for the standard peptides (Fig. 3D *vs* Fig. 3C). This indicates that the two peaks in the bimodal distribution of Sim scores effectively distinguishes true positive from false positive assignments.

**ROC analyses.** To further evaluate the discriminating power of Sim scoring with complex samples, we carried out Receiver Operating Characteristic (ROC) analyses; ROC curves compare sensitivity (true positives identified) *vs.* specificity (1-false positive rate (FPR)), where the area under the curve correlates positively with discriminating power, i.e., the ability to achieve high sensitivity with low numbers of false positives. The 925 manually validated MS/MS of Sample 2 provided an ideal dataset for this analysis, because the spectra adequately sampled the full range of scores, not just the scores that were above an acceptance threshold. Thus, the manually validated cases showed a wide range of Mowse and XCorr ranging between 17-196 and 1.14-7.85, respectively (Supplementary Table 1, considering only cases where Mascot or Sequest correctly identified the spectrum). This was similar to the range seen with standard peptides (7). Furthermore, the validated cases accounted for most of the high scoring

peak in the bimodal distribution of Sim scores (Fig. 3D), and was similar to our previous estimate of 920-930 MS/MS that would be identifiable if all correct sequences (fully tryptic, allowing up to two missed cleavages) with Mowse or XCorr below acceptance thresholds could be captured (7). Thus, the manually validated dataset from Sample 2 satisfied our criteria for ROC analyses of XCorr, Mowse, and Sim scoring.

*[insert Fig. 4]*

Using this dataset, ROC analyses were carried out for the $MH^+$, $MH_2^{+2}$, and $MH_3^{+3}$ charge states, comparing Sim, Mowse, and XCorr scores of the manually validated subset (Fig. 4). Sensitivity was plotted as the number of DTA files with score greater than an acceptance threshold, expressed as the true positive rate (TP/(TP +FN), where FN includes both low scoring cases and those where the search program did not identify the correct assignment as the top ranked "hit". Specificity was evaluated from the number of FP in an inverted database search that passed the acceptance thresholds, expressed as the false positive rate (FPR = FP/(FP + TN). From the ROC curves, it was clear that Sim scoring showed significant improvement in discrimination compared to Mowse and XCorr when applied to $MH^+$ and $MH_3^{+3}$ ions, as well as modest improvement with $MH_2^{+2}$ ions. For many cases with low XCorr or Mowse scores, Sim performed dramatically better than XCorr or Mowse at capturing correct assignments; this could be seen by plotting Sim against either XCorr or Mowse (Fig. 3E,F). Similar results were seen when specificity was normalized to only those peptides top-ranked by Mascot (see Supplementary Fig. 5, Supplementary Table 3). Thus, the ROC analysis revealed significant improvement in discriminating power by using relative fragment ion intensity information and Sim rescoring to evaluate search results.

To further evaluate the effect of this improvement, we examined the number of MS/MS validated by Sim when holding FDR constant. Sim was used to rescore Mascot results, because

more of the correct assignments were top-ranked by Mascot than by Sequest. Using score thresholds that allowed 1.5% FPR in Sample 2 (Supplementary Table 3), Sim rescoring of top-ranked Mascot assignments accepted 829 peptides, while Mowse and XCorr accepted 714 and 616 peptides, respectively (Table 1). The superiority of Mowse over XCorr is consistent with other studies (e.g., Kapp et al. (29)). We also evaluated other methods used to validate lower scoring MS/MS. The Sequest ΔCN parameter is often used to improve sensitivity and specificity by allowing use of lower XCorr thresholds without an increase in FPR (10). We found in a previous study that commonly accepted thresholds for XCorr + ΔCN validated 720 assignments in Sample 2 (9). In addition, we validated 833 cases with in-house MSPlus software, which evaluates consensus between Mascot and Sequest, then filters out FPs based on physicochemical properties (7), as described in Methods. Because MSPlus validated cases that included most of the peptides identified by XCorr + ΔCN, our further analyses compared results between Mowse, MSPlus, and Sim.

Fig. 5A summarizes the overlap between Mowse, MSPlus, and Sim validated peptide assignments, along with the FPs identified in the manual analysis. There were 670 MS/MS validated by all three methods, with no FPs detected. MAE Sim rescoring of Mascot results identified 159 more cases, of which 11 were FPs (7 could be filtered out by applying physicochemical tests). Of the 159, 80 were not validated by MSPlus, because they were low scoring or there was no Sequest/Mascot consensus. On the other hand, Mascot or MSPlus validated 44 or 84 additional cases (with 2 or 10 FPs); most of these had Sim scores just below the acceptance threshold, as discussed later.

*[insert Fig. 5]*

**Protein profiling results.** Confidence in the new identifications could be assessed by the manner in which the new cases validated by MAE-Sim affected the protein profile compared to

Mowse alone. We utilized the IsoformResolver protein profiling program to carry out this comparison (Supplementary Table 2), and evaluate the degree to which new peptide assignments made by MAE-Sim captured information about new peptide sequences and new proteins. If MAE validation of peptide assignments captured a substantial number of false positives, many new protein identifications supported by only one peptide should have been added. Instead, we found that the new peptides captured by MAE-Sim had their largest effect on increasing sequence coverage of proteins already identified, with proportionally few new proteins, suggesting low false positives in the new peptide identifications.

We first compared the results of Sim rescoring of Mascot results with those from Mowse (summarized in Fig. 5B and Supplementary Table 4). Using Mowse acceptance thresholds, 228 proteins were identified based on 395 unique peptide sequences. MAE-Sim increased support for 44 of the proteins identified by Mowse (19%) including 32 proteins previously supported by only one or two peptides, and added 33 new protein identifications (14%). Among the 13 cases that were unique to Mascot (all supported by only one peptide), two were found to be false positive assignments. MAE-Sim added 9 manually confirmed FPs (all supported by only one peptide), but 7 of these were eliminated upon application of physicochemical filters, so no net effect on FPs was shown. After removing the 9 FPs from the MAE-Sim protein list, only 24 new proteins were identified, compared to 44 proteins identified by Mascot that enjoyed additional support by MAE-Sim. These results are consistent with the population sampling nature of MuDPIT, where additional sampling will more likely identify peptides from proteins that were previously identified.

**MAE data mining functions provide detailed analyses of MS/MS anomalies.** We observed two anomalies in comparing standard protein digests to complex MuDPIT samples. First, the percentage of the expected tryptic peptides identified in the Sample 2 dataset (88%)

was lower than that in the standard peptide dataset (97%); these represented cases with moderate Sim scores that were uniquely caught by Mascot and MSPlus. Second, the Sim score distribution of manually validated cases in the Sample 2 dataset extended farther into the lower range for the complex MuDPIT dataset than the standard protein dataset (Fig. 3C *vs.* 3D), with proportionately more validated MS/MS between 0.35-0.60. These effects revealed that a larger proportion of correct assignments with borderline Sim scores appear in datasets of complex protein mixtures, compared to simple protein samples. In order to further characterize these borderline cases, we turned to other scoring functions of MAE, the Proportion of Ion Current (PIC) and Intfrag scores. PIC is used in manual analysis to assess how well a peptide sequence accounts for the total fragment ion current in the MS/MS; because it accounts for more ion types than considered by search engines, it captures orthogonal information about how well the assigned sequence fits the spectrum. The IntFrag score assesses the amount of the ion current present as internal fragment ions generated by cleavage of two peptide bonds. A high IntFrag score occurs when many fragment ions are matched by random chance to internal fragment ions, and can be used together with PIC to evaluate plausibility.

The PIC and IntFrag scores require the major Average Mass Ions to be classified by fragment ion type, including canonical *a*, *b*, or *y* ions generated by peptide bond cleavage, as well as noncanonical ions from other cleavages, as described in Methods. Such annotations revealed fragmentations that were not modeled well by MassAnalyzer. Of interest were the problems encountered with: (1) internal fragment ions generated by cleavage of two peptide bonds or (2) fragment ions generated by multiple dehydration of *b* ions. Fig. 6A shows an example of internal fragmentation, where one cleavage is generated N-terminal of Pro and the other cleavage is generated at hydrophobic amino acids which also yield intense *y* ions (…FIQNV…). Most internal fragment ions were predicted by the theoretical spectra, but often

with lower intensities than observed. The multiply dehydrated *b* ions were observed in peptides with multiple Ser/Thr residues, often containing acidic residues and often in singly charged peptides. Up to four dehydration events could be observed in $b_n$ ions, depending on the number of Ser and Thr residues. In some cases, the multiply dehydrated fragment ions showed greater intensity than their corresponding undehydrated forms (Fig. 6B, e.g., $b_{11}$, $b_{12}$, $b_{13}$), particularly when the parent ions were singly charged. Such fragment ion types were poorly predicted by the theoretical spectra, and may account for part of the lower performance of Sim scoring with $MH^+$ parent ions, compared to $MH_2^{+2}$ or $MH_3^{+3}$ parents (Fig. 4).

*[insert Fig. 6]*

Annotation of internal fragment ions and multiply dehydrated *b* ions required new methods to avoid a "combinatorial explosion" of redundant possibilities which would complicate the automated assignment process. The high number of possible fragment ions increases the probability of incorrectly assigning an observed ion, especially with larger parent ions. Therefore, rules were implemented to filter false positives by prioritizing fragment ion assignments, as described in Methods.

The accuracy of the fragment ion assignment process carried out by MAE was assessed using the high confidence MS/MS subset of Sample 2 dataset (Table 1), where the automated annotations for 52 randomly chosen spectra were compared to annotations assigned manually by an experienced analyzer. We evaluated 1,357 fragment ions (not including internal fragment ions) assigned by manual analysis *vs.* MAE, and found only 53 MAE annotations that differed from manual analysis. Of these, six represented cases where multiple ions were possible and different alternatives were chosen by MAE *vs.* an experienced annotator. Twenty-eight were plausible fragment ions that were not annotated during manual analysis of the original DTA file, but were annotated by MAE after sDTA processing led to intensity enhancement of a weak ion

cluster above the noise threshold. Only 19 MAE annotations (1.4%) were considered wrong or unlikely by the experienced annotator. This rate was comparable to the frequency (1.7%) at which two experienced annotators differed in their ion assignments after examining the same spectra. Taken together, the analysis shows that MAE accurately recapitulated the annotations of manual analysis.

**Impact of spectral chimera in the borderline scoring cases.** Using the PIC and IntFrag scores, we explored MS/MS with borderline Sim scores that were more common in complex samples like the Sample 2 dataset (Fig. 3D), compared to simpler samples like the standard protein dataset (Fig. 3C). We first focused on cases where more internal fragments were annotated than were plausible (IntFrag score >0.20) and the PIC score was moderate (0.45-0.60). In many cases, the internal fragment ions and unidentified fragment ions were absent in other DTA files representing the same peptide. This suggested that such ions represented chemical contamination due to commingling of fragment ions from two parent ions with similar m/z that eluted simultaneously on reverse phase HPLC ("spectral chimera"). The analysis of a spectral chimera is illustrated in Fig. 7, where two peptides with the same m/z were partially resolved on SCX and observed in fractions 6 and 8, but coeluted in SCX fraction 7. The MS/MS in fractions 6 and 8 reported two different peptides with very high Mowse, XCorr, Sim, and PIC scores. The chimera MS/MS in fraction 7 contained all major ions seen in both peptides individually. This example is an ideal situation for analyzing a spectral chimera, because the peptides in this case had the same charge, and the first and second highest Mascot “hits” for the chimera MS/MS represented the two peptides identified in SCX fractions 6 and 8. Furthermore, the two peptides contributed approximately equal intensity to the chimera, which also made it easier to deconvolute the two peptide sequences.

*[insert Fig. 7]*

When PIC was plotted against Sim, it was clear that a borderline Sim scoring case was more likely to show a moderate PIC score (Fig. 3G,H and Supplementary Fig. 6), indicating that many of these were chimera spectra. Manual analysis strongly supported the identifications in many of the borderline cases (e.g., adjacent scans or SCX fractions had another charge form or more intense MS/MS of the same sequence, and there was a clear agreement with expected fragmentation behavior at Gly, Asp, Glu, and Pro). Among these cases, we found 32 spectral chimera candidates, where a tag sequence (30) or continuous set of adjacent amino acids was identified among the unannotated ions and unlikely internal fragment ions, supporting the presence of a second peptide. Furthermore, the presence of chemical contamination was very common in complex samples compared to simple samples; 60% of manually validated peptides in the Sample 2 dataset had PIC <1.0, whereas only 5% (9 of 175) of the validated peptides in the standard protein dataset had PIC <1.0. This is similar to estimates made by Zhang et al. (31). We found that the presence of chemical contamination reduces the discrimination in scoring search results. This was apparent in ROC plots of subsets of the Sample 2 dataset that were filtered to remove cases with PIC < 0.95; these showed higher discrimination as the PIC score stringency increased and chemical contamination decreased (Supplementary Fig. 5D-F). Thus, the data suggests the borderline Sim scoring cases are spectral chimeras that can be flagged using the Intfrag and PIC scores.

The presence of spectral chimera may explain why MSPlus was able to identify many cases that were not validated by MAE Sim scores, but were identified by Mowse or MSPlus (Fig. 5A). Of the 84 cases identified by Mascot and/or MSPlus, but rejected by MAE, all showed Sim scores in the borderline scoring region, just below acceptance thresholds. All showed moderate PIC and high IntFrag scores, strongly suggesting they were spectral chimeras, and included the 32 spectral chimera candidates identified from tag sequences, described above. On the other

hand, MAE uniquely identified 80 peptides which were rejected by MSPlus because of low combined Mowse/XCorr scores, lack of consensus between Sequest and Mascot, or Sequest score RSP ≠ 1. Together MAE and MSPlus rescoring of Mascot results validated the search results for 95% (881 of the 920-930 expected) of the DTA files that we estimated could be identified as tryptic peptides. In a preliminary test to optimize data capture, we lowered Sim score to 0.45 and used PIC scores along with physiochemical filters to remove FPs. The results showed 94% of validated search results could be captured by this simple search strategy.

**Analysis of other datasets.** Finally, we tested the performance of MAE-Sim with additional MuDPIT datasets (Samples 3 and 4, Table 1, respectively collected on LCQ Classic and LCQ DecaXP instruments), where methods were utilized to achieve higher sampling of low abundance ions. A larger proportion of the DTA files were in the low scoring class (Sim 0.0-0.5), than seen in Sample 2, because higher sampling depth resulted in more sequencing of source-generated fragment ions. The result was a very large number of high quality MS/MS that were not identifiable as fully tryptic peptides; this type of MS/MS often show FP assignments in Sequest and Mascot searches. Nevertheless, the Sim scores for these datasets yielded bimodal distributions (Supplementary Fig. 7) similar to that seen with Sample 2 (Fig. 3D), and the number of DTA files identified by Sim scoring *vs.* Mowse in Samples 3 and 4 increased by 28% and 29%, respectively (Table 1), with FDR=1.3-1.6% confirmed by two independent measurements. Analysis of MS/MS from simple peptide mixtures collected on a LTQ-Orbitrap MS instrument also showed improvement in data capture by Sim compared to Mowse, XCorr or XCorr+ΔCN, shown by greater separation of the high scoring class to yield a bimodal distribution, (Supplementary Fig. 8). Thus, MAE showed improved performance in capturing new information from MuDPIT datasets taken under different conditions and different mass spectrometers.

## CONCLUSIONS

This study describes the development and applications of a novel program which automates principles utilized in manual analysis of MS/MS spectra. MAE provides several scores and data mining tools, but here we focused on two major metrics: (1) Similarity scoring against theoretical spectra based on a chemical model of peptide gas phase chemistry (Sim score), and (2) Proportion of Ion Current which accounts for number of fragments consistent with the peptide assignment (PIC score). To compare results from simple *vs.* complex samples, we used a dataset of standard peptides derived from purified proteins and several MuDPIT datasets of complex samples of human proteins, one of which had been extensively annotated by manual analysis. Reprocessing to simplify the information in the DTA files and using average masses enhanced the accuracy of fragment ion assignments, particularly for weak spectra. Excellent discrimination compared to Sequest or Mascot scoring was demonstrated by ROC analysis of the search results for the manually validated MuDPIT cases (Fig. 4), and Sim score distributions showed bimodal distributions with greater separations between correct *vs* incorrect assignments (Fig. 3). Preliminary results with LTQ MS/MS data show similar improvement in peptide identification, consistent with the previous report by Z. Zhang showing comparable performance of MassAnalyzer with LCQ vs LTQ instruments (21,22). We conclude that MAE Sim scores are well suited to provide an objective validation of chemical plausibility in search results, replacing manual analysis for unmodified peptides.

The most dramatic improvement in discrimination by Sim compared to Mowse or XCorr scoring was achieved with triply charged peptide ions, with good improvement also for singly charged ions and moderate improvement for doubly charged ions (Fig. 4, Supplementary Fig. 5). The latter results are due to the fact that doubly charged ions usually produce a large number of fragment ions (10). Their spectra produce high scores in Mascot or Sequest because of the large

number of ions and high quality sequence tags available for matching. It is likely that the higher complexity of chemistry for singly charged peptides limited the discrimination achieved with those peptides, but our results indicate several ways that the MassAnalyzer simulations can be modified to improve those theoretical spectra. These interesting leads for exploring new gas phase chemical mechanisms illustrate the usefulness of MAE in data mining studies.

The excellent improvement for triply charged forms is due to the fact that cleavages typically cover only a part of the peptide sequence, and different fragment ion charge forms produce partially overlapping coverage of the sequence. However, search programs such as Sequest and Mascot consider a large number of possible fragment ions; consequently, there is a much larger probability that a match will be made by random chance for the larger, more highly charged parent ions. Because MassAnalyzer generates theoretical spectra with a smaller number of fragment ions and Sim scoring gives higher weight to the high to moderate intensity ions, the likelihood of random matches is reduced. The result is a better fit to the observed spectrum, which both lowers the overall range of scores for incorrect assignments and raises the scores for correct assignments, leading to improved discrimination. The excellent performance of theoretical spectra for highly charged peptides will be important in studies which analyze large processing products, such as in identifying discriminators from serum profiling (32).

An important outcome of this study was the number of spectral chimeras that were revealed by the PIC scoring, allowing us to evaluate the impact of chimeras in these experiments. Our use of the manually validated dataset was critical to reveal this, because chimeras were far less common in the simple peptide mixtures from standard proteins. Chimeras resulting from limited peak capacity of the chromatographic methods typically used in analysis of complex samples present a critical problem for studies that quantify protein abundances using stable isotope labeling or peak ion intensity measurements, when using low resolution mass spectrometers.

The presence of spectral chimeras also may explain why the new faster scanning MS instruments do not achieve the expected improvements in data capture, because the higher sensitivity of these instruments would increase the likelihood that two different peptide ions would be sequenced together. Our preliminary data show in fact that the intermediate scoring class, corresponding to chimeras, is increased in LTQ-Orbitrap datasets (Supplementary Fig. 7). An increased number of chimera spectra may explain why the ΔCN method is less effective at recovering low XCorr hits in the LTQ dataset, as also reported recently (33). The use of MAE PIC and Intfrag scoring will aid in determining the presence of chimera spectra, in order to evaluate the prevalence of this problem.

Overall, the performance of the scoring methods described in this study indicate that nearly complete capture of the expected MS/MS for tryptic peptides in complex MuDPIT datasets can now be achieved (and preliminary studies indicate they improve analysis of nontryptic peptides, as well). Combined with its data mining functions, MAE provides a novel strategy to significantly improve the amount of information that can be gained about the composition of proteins in complex samples.

**ACKNOWLEDGEMENTS**

We are greatly indebted to Dr. Zhong-qi Zhang (Amgen) for providing theoretical spectra and for advice in reproducing the program in our laboratory and Dr. Karen Jonscher for assistance with manual analysis and data collection. We also thank Karen Kafadar (UC Denver Health Science Center) and Dr. Robin Knight (UC Boulder) for advice on statistical analysis and Alex Mendoza for assistance with data management in early phases of this project. This work was supported by a Jane and Charlie Butcher Seed Grant Award in Genomics and Biotechnology (KAR and KC), NIH R01 CA87648 (KAR), and a predoctoral fellowship in Molecular Biophysics to KMA (T32 GM065103).

## REFERENCES


1. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data. *Mol. Cell. Proteomics* **4,** 1419-1440

2. Russell, S. A., Old, W., Resing, K.A., and Hunter, L. (2006) Proteomic informatics. *Int Rev Neurobiol.* **61,** 127-57

3. MacCoss, M. J., Wu, C. C., and Yates, J. R.,III (2002) Probability-based validation of protein identifications using a modified Sequest algorithm. *Anal. Chem.* ***74,*** 5593-5599

4. Moore, R. E., Young, M. K., and Lee, T. D. (2002) QScore: an algorithm for evaluating Sequest database search results. *J. Am. Soc. Mass Spectrom*. **13,** 378-386

5. Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem.* **74,** 5383-5392

6. Weatherly, D. B., Atwood, J. A., III, Minning, T. A., Cavola, T. A., Tarleton R. L., and Orlando, R. (2005) A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics*. **4,** 762-772

7. Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, J. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76,** 3556-3568

8. Norbeck, A. D.; Monroe, M. E.; Adkins, J. N.; Anderson, K. K.; Daly, D. S., and Smith, R. D. (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *J. Am. Soc. Mass. Spectrom.* **16,** 1239-1249

9. Yen, C.-Y., Russell, S., Mendoza, A., Meyer-Arendt, K., Sun, S., Cios, K., Ahn, N. G., and Resing, K. A. (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.* **78,** 1071-1084

10. Sadygov, R. G., Cociorva, D., and Yates, J. R.,III (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1,** 195-202

11. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768-774

12. Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35,** 1399-1406

13. Elias, J. E., Gibbons, F. D., King, O.D., Roth, F. P., Gygi, S.P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. **22,** 214-219

14. Narasimhan, C., Tabb, D. L., VerBerkmoes, N. C., Thompson, M. R., Hettich, R. L., and Uberbacher, E. C. (2005) MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem*. **77,** 7581-7593

15. Havilio, M., Haddad, Y., and Smilansky, Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem*. **75,** 435-444

16. Eddes, J. S., Kapp, E. A., Frecklington, D. F., Connolly, L. M., Layton, M. J., Moritz, R. L., and Simpson, R. J. (2002) CHOMPER: a bioinformatics tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. J. *Proteomics.* **2,** 1097-1103

17. Johnson R. S., Davis, M.T., Taylor, J. A., and Patterson, S.D. (2005) Informatics for protein identification by mass spectrometry. *Methods* **35,** 223-236

18. Chen, Y., Kwon, S. W., Kim, S. C., and Zhao, Y. (2005) Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res.* **4,** 998-1005

19. Wysocki, V.H., Resing, K.A., Zhang, Q., and Cheng, G. (2004) Mass spectrometry of peptides and proteins. *Methods* **35,** 211-222

20. Paizs, B. and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24,** 508-548

21. Zhang, Z. (2004) Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal Chem.* **76,** 3908-3922

22. Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem.* **77,** 6364-6373

23. Craig, R., Cortens, J.C., Fenyo, D., Beavis, R.C. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **5,** 1843-1849

24. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S, and MacCoss, M. J. (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries *Anal Chem.* in press.

25. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J.R. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676-682

26. Cox, K. A., Cleven, C. D., and Cooks, R. G. (1995) Mass shifts and local space-charge effects observed in the quadrupole ion-trap at higher resolution. *Int. J. Mass Spectrom. Ion Process.* **144,** 144, 47-65

27. Jonscher K. R., and Yates J. R., III (1997) The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal Biochem*. **244,** 1-15

28. Wehofsky, M., and Hoffmann, R. (2002) Automated deconvolution and deisotoping of electrospray mass spectrometry. *J Mass Spectrom.* **37,** 223-229

29. Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., and Simpson, R. J. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* **5,** 3475-3490

30. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399

31. Zhang, N., Li, Z.-J., L., Ye, M., Pan, S., Schwikowski, B., and Aebersold, R. (2005) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics.* **5,** 4096-4106

32. Bons, J. A., Wodzig, W. K., and van Dieijen-Visser, M. P. (2005) Protein profiling as a diagnostic tool in clinical chemistry: a review. *Clin Chem Lab Med.* **43,** 1281-1290

33. Xie, H., and Griffin, T. J. (2006) Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics. *J. Proteome Res.* **5,** 1003-1009

## FIGURE LEGENDS

**Figure 1: High-level view of the MAE algorithm.** Elements within the dotted line represent those in MAE. Elements above this box represent the search components (MSPlus, Mascot, Sequest) and the MassAnalyzer, which is used to calculate theoretical spectra for candidate peptides in the MSPlus output file. The three inputs to MAE include (i) the DTA file, listing the $MH^+$ for the parent peptide based on the observed ion m/z and the assumed charge, along with m/z and intensities of fragment ions after centroiding, (ii) the candidate peptide sequence(s) from the search result (obtained from the MSPlus.csv common data interface generated by in-house software), and (iii) the theoretical MS/MS spectrum generated by MassAnalyzer, listing m/z and intensity of predicted fragment ions for candidate peptide sequence(s). Three MAE outputs include (i) the simplified sDTA file, (ii) the summary file listing annotations of the major ions in the sDTA file, and (iii) the MAE scores (e.g., Sim, PIC, y-b/y+b and Intfrag), which are written to the MSPlus file. Finally, MAE evaluates whether the scores are above user input thresholds for acceptance and writes the outcome to the MSPlus file. The list of validated peptides is input into an in-house IsoformResolver program to produce a protein profile.

**Figure 2: Processing of fragment ion information in DTA files to remove isotope ions and noise.** To illustrate the handling of the information in the DTA file, an example was chosen where three MS/MS spectra (panel A) were summed to produce a DTA file (panel B). This DTA gave a high confidence assignment to the peptide sequence SVEMHHEALSEALPGDNVGFNVK (XCorr 4.92, Mowse 94). A triply charged form of the same peptide coeluted with the doubly charged form, increasing the confidence in the assignment. Panel C shows the MAE-reprocessed spectrum (sDTA). In each panel, expanded

views of three mass regions, containing a weak doubly charged ion and two stronger singly charged ions, are shown to illustrate how DTA and sDTA files are processed (Left panel: the full MS/MS spectrum; Left expanded panel: 1082-1084 Da; Middle expanded panel: 1321-1324 Da; Right expanded panel: 1433.5-1437 Da). (A) Each of the three MS/MS spectra that were combined to make the final DTA file are displayed. The mass accuracy is reported to three decimals by the vendor software. (B) Visualization of the final DTA file generated by the vendor Extract_MSN software. The file contains 1,313 m/z entries, including noise ions, most of which are too low to see in the spectrum. The mass accuracy is given to one decimal place by the software. For the weak, doubly charged ion (1082-1084 Da), note that the process of centroiding and summation used to generate the DTA file produces several apparent fragment ions with no obvious isotope envelope; thus, deconvoluting charge information and identifying the monoisotopic ion is difficult. (C) Visualization of the final simplified DTA (sDTA) file after MAE processing. Many of the noise ions were removed and overall signal of most remaining ions increased. The three expanded panels show resulting fragment ions of 1083.73 Da ($y_{20}^{+2}$, pred m/z = 1083.72), 1322.03 Da ($b_{12}$, pred m/z = 1322.45), and 1435.11 ($b_{13}$, pred m/z = 1435.61), where the clustering process produces combined ions with greater intensity. Clusters are identified by choosing the most intense DTA ion as the starting point and assigning ions within its cluster in three steps: (1) check for ions within ±1 Da of the starting point; (2) check for additional ions within ±1 Da of those identified in the first step; (3) check for additional ions within ±1 Da of those identified in the second step. All ions in the cluster must be within -2.0 to +2.5 Da from the starting point; we find in practice that the most intense DTA ion is almost always within +1.5 Da of the monoisotopic mass. The process is repeated with the next most intense ion in the ion list, until all ions have been assigned to a cluster. For example, to produce the ion at 1083.73 Da, MAE first selects the most intense ion in the region (1084.2 Da) and looks

for adjacent ions between 1082.2-1086.7 Da, extending in one Da "steps". The extension of the cluster stops if ions fail to appear within 1 Da, or if ion(s) within 1 Da have intensity below 3000 counts (e.g., 1082.1 Da). The 1.0 Da window corresponds to the maximum difference expected from the default centroid processing parameters. Thus, MAE first clusters DTA ions within ± 1 Da of the 1084.2 peak (1083.5 and 1084.3 Da), then clusters adjacent DTA ions extended by 1 Da (1082.8 Da), followed by DTA ions further extended by 1 Da from 1082.8 (none observed, therefore search is stopped for this cluster). All DTA ions within this cluster are then combined to produce an ion with weighted average mass 1083.73 Da and intensity equal to the sum of each DTA ion. Note that several small DTA ions are within these limits, but are not included in the calculation because they are classified as bulk noise (2.8 × s.d., see Methods).

**Figure 3. Sim and PIC scoring results for standard peptides and the test shotgun proteomic dataset.** A small MuDPIT dataset of K562 proteins, the manually validated subset of this dataset, and a manually validated dataset of peptides from standard proteins (Samples 1 and 2 in Table 1) were analyzed by Sim, XCorr, Mowse, and PIC score, as described in Methods. Two types of searches were carried out: a "normal search" with the IPI database, allowing up to two missed tryptic cleavages and missed cleavage at KP or RP, or an "inverted search" with same parameters but using a database where each protein entry had its sequence inverted from C- to N-terminus. Only peptides with observed mass >950 Da and with standard deviation of DTA ion intensities >1,000 counts were included. (A) Sim score distribution for Sample 2, using sDTA files (◊) or unprocessed DTA files (♦). A clear bimodal distribution is achieved using sDTA files compared to unprocessed files, implying higher discrimination between correct and incorrect assignments. The ratio of areas under the two peaks in this distribution is ~1:3, consistent with the expected number of tryptic peptide MS/MS in this dataset (ref. 7). (B) Sim

score distribution for the Sample 2 dataset, searched against an inverted protein sequence database where all sequences are false positives. Little difference can be seen between sDTA files (△) and unprocessed DTA files (▲). (C) Sim score distribution for the standard protein dataset, processed to sDTA files and searched against a normal database (□) or an inverted database (△). (D) Sim score distribution for the Sample 2 dataset, showing all assignments (◊), correct identifications validated manually (□), and identifications from an inverted database search (△). (E) Comparison of XCorr *vs.* Sim and (F) Mowse *vs.* Sim for the Sample 2 dataset. The data form two clusters, where all cases with Sim score > 0.53 have been validated by manual analysis (see Supplementary Fig. 6). Many validated assignments with high Sim could not be captured by Mowse or XCorr, whose values are below high confidence thresholds. (G) Sim *vs* Proportion of Ion Current (PIC) for the Sample 2 dataset. Cases with moderate PIC scores appear to have two parent peptide ions cosequenced during MS/MS. (H) Sim *vs* PIC for false positives generated by searching the Sample 2 dataset against the inverted sequence database. This control shows that the distribution of scores for incorrect assignments closely resembles the low scoring peak between Sim = 0 to 0.5 for normal assignments (panel D). Few incorrect assignments occur within the range for manually validated assignments (Sim > 0.5), reflecting good discrimination by Sim.

**Figure 4. Receiver operating characteristic (ROC) analyses comparing MAE and Mascot.** True positive rates are plotted against false positive rates at varying Sim, Mowse, or XCorr values, for singly (A), doubly (B), and triply (C) charged states of the parent ions. The manually validated (correctly assigned) subset of DTA files from the Sample 2 dataset was used in these analyses. True positive rates (sensitivity) are calculated from the number of correctly

identified cases determined by manual validation; false positive rates (1-specificity) are measured by searching the full dataset against the inverted sequence protein database. The area under the curve reports discriminatory power of each score, and increases with Sim compared to Mowse or XCorr scoring.

**Figure 5. Peptides and proteins identified from Mowse, MSPlus, and MAE scoring.** (A) Peptide identifications: DTA files from the Sample 2 MuDPIT dataset were scored by Mascot using Mowse thresholds of 38, 40, and 43, and rescored by MAE using Sim thresholds of 0.59, 0.52, and 0.49, respectively for $MH^+$, $MH_2^{+2}$, and $MH_3^{+3}$ ions (FPR = 1.5%, based on Supplementary Table 3). Observed peptides count each spectrum separately. False positives indicated in parentheses were manually validated, and all corresponded to proteins supported by only one peptide. For Sim, 12 FPs were estimated from the ROC analyses, in agreement with the 11 FP identified by manual analysis. Similar results were obtained for Mowse and MSPlus analyses, and overall the observed FDR=2.3% (21/913). The three methods identified 670 peptides in common, 159 peptides unique to MAE compared with Mowse, and 84 peptides unique to MSPlus compared with MAE. (B) Protein identifications were summarized using IsoformResolver, an in-house protein profiling program which groups proteins by isoforms and presents the minimum number of proteins that account for the unique peptide sequences (7). The three methods identified 215 proteins in common, while 33 were observed by MAE but not Mowse and 23 were observed by MSPlus but not MAE. Mowse results were completely overlapping with MSPlus, because MSPlus accepts all peptides above Mowse thresholds. Of the 228 proteins validated by Mowse thresholds, 87 were altered by MAE rescoring, and 154 were unaffected. MAE identified 64 new unique peptide sequences, which increased support for 44 proteins previously identified by Mowse (111 new peptides) and identified 33 new proteins (48

new peptides). MAE also rejected 44 unique peptides previously accepted by Mowse, which decreased support for 27 proteins, including 13 proteins previously supported by only one peptide. In 3 proteins, MAE removed the same number of peptides as it added, yielding no net change in support. Manual analysis showed that all but one of the peptides rejected by MAE were false positives.

**Figure 6. Examples of MS/MS where the gas phase chemistry is incompletely predicted by theoretical spectra.** Summary of two spectra, showing observed Average Mass Ions in sDTA files (top panels) and theoretical spectra generated by MassAnalyzer (bottom panels). (A) Example of a peptide where multiple internal fragment ions are observed. Most of the internal fragments were generated by cleavage first between $Ile_7$-$Pro_8$, and second within an "active region" at QNVP. As a result, $b_7$ shows significantly lower intensity than predicted, due to its depletion following internal fragmentation. (B) Example of a peptide showing unusual dehydration. This is a singly charged peptide with two Ser, one Thr, and two acidic groups. Multiple dehydrations from $b_{11}$ and $b_{12}$ representing neutral loss of one, two, or three water molecules are observed, consistent with the number of Ser/Thr residues on these fragment ions. Note that the larger $b_{13}$ fragment ion shows less dehydration. Multiple dehydration is also observed from the parent ion. Annotations of major ions are indicated below each spectrum. Canonical *b* and *y* ions are respectively shown by ┐ and └ symbols above and below the sequence. Dehydrated ions are represented as triangles, and *a* ions are indicated. Internal fragment ions are shown by bars below the sequence. Multiply charged canonical fragment ions are shown above or below the singly charged ions.

**Figure 7. An example of a spectral chimera and the MS/MS of peptides that contribute to the spectrum.** The three panels show sDTA files for MS/MS spectra taken at the same RP elution time and m/z value for peptide ions detected in three adjacent SCX fractions (SCX 6, 7, 8). The top and bottom panels show two peptides where the sequence assignment was unambiguous, with PIC score >90% and Sim score > 0.75. (Top panel, XCorr = 5.5, Mowse = 111, Sim = 0.87, PIC = 1.0; Bottom panel, XCorr = 5.0, Mowse = 71, Sim = 0.75, PIC = 1.0) The middle panel shows all major ions seen in each of the two peptides, where MSPlus identified LLQA… as the top "hit" (XCorr = 4.5, Mowse = 76, Sim = 0.64, PIC = 0.63), and identified VTIA… as the second hit (Mowse = 39, Sim = 0.60, PIC = 0.81). Fragment ions corresponding to the second sequence are underlined. Together, the two peptide sequences accounted for 98% of the ion current in the chimera MS/MS.

**Table 1. LC/MS/MS datasets used in this study.** [1]

| Samples[2] | Total DTA files[3] | XCorr | Mowse | | MAE-Sim rescoring of Mascot | |
|---|---|---|---|---|---|---|
| | | Peptides ID'd[4] | Peptides ID'd[4] | Proteins ID'd | Peptides ID'd[4] | Proteins ID'd |
| 1. Standard proteins dataset[5] | (175) | 127 | 136 | 6 | 172 | 6 |
| 2. Small test dataset[6] | 2,943 | 616 | 714 | 228 | 829 | 248 |
| High confidence subset[7] | (545) | 511 | 533 | 184 | 531 | 182 |
| Manually validated subset[8] | (925) | 576 | 666 | 213 | 785 | 233 |
| 3. Four repeats; shallow RP gradient[9] | 108,458 | not determined | 5,793 | 621 | 7,427 | 910 |
| 4. Two repeats + gas phase fractionation[10] | 85,742 | not determined | 7,509 | 780 | 10,702 | 1,271 |
| 5. Gel filtration subset[11] | (1,448) | 1,380 | 1,425 | 756 | 1,425 | 756 |
| 6. LTQ-Orbitrap dataset[12] | 32,502 | 873 | 1,047 | N.D. | 1,282 | N.D. |

[1] Score distributions for Sample 1,2 are shown in Fig. 3, Samples 3-5 shown in Suppl. Fig. 6, Sample 6 shown in Suppl. Fig. 7.

[2] Samples 1,2,4,5 were collected on an LCQ Classic; Sample 3 was collected on an LCQ XP; Sample 6 was collected on an LTQ/Orbitrap. Details on LTQ/Orbitrap data collection are given in the figure legend to Suppl. Fig. 7. Target values for the LCQ ion traps were $4 \times 10^8$ ions for Samples 2 and 3 and $5 \times 10^8$ ions for Samples 1, 4, and 5 in full MS scan mode, and $2 \times 10^7$ ions for Samples 1-6 in MS/MS mode; MS/MS spectra were acquired utilizing an exclusion window of 2 Da and normalized collision energy of 34 units. Criteria for inclusion for Samples 1-5 were peptide MW > 950 Da and std. dev. of intensities for unprocessed DTA ions > 1,000 counts.

[3] For some datasets or subsets of datasets, only the number of DTA files used in this study are shown.

[4] False positive thresholds for XCorr, Mowse, and Sim were determined by searching the Sample 2 dataset against the inverted sequence database. FPR=FP/total DTA files = 1.5%, in all cases.

[5] The six standard proteins were CYC: BOVIN CYTOCHROME C, PHS2: RABIT Glycogen phosphorylase, CASB: BOVIN Beta-casein, ALBU: BOVIN Serum albumin, OVAL: CHICK Ovalbumin, TRFE: BOVIN Serotransferrin, G3P: RABIT Glyceraldehyde 3-phosphate dehydrogenase, CATA: BOVIN Catalase, PPB: ECOLI Alkaline phosphatase. Each protein digest was run once individually by RP-LC/MS/MS, with gradient 0-70% ACN/45 min, m/z =350-1500 Da (ref. 7).

[6] Small MuDPIT dataset of K562 soluble proteins. This is the primary dataset analyzed throughout this study, containing 1,838 MS/MS and 2,943 DTA files, the latter which includes computer-generated $MH_2^{+2}/MH_3^{+3}$ alternatives for each multiply charged case. Detailed results are in Suppl. Tables 1 and 2. K562 soluble proteins were digested with trypsin and separated by SCX-HPLC. Each SCX fraction was run once by RP-LC/MS/MS, with RP gradient 0-21% ACN/30 min, 21-35% ACN/10 min, and 35-70% ACN/10 min, scanning m/z =350-1500 Da (Method 1). (Details in ref. 7).

[7] High confidence subset requires MSPlus SumScore > 6. SumScore requires consensus between Mascot and Sequest search results and sums Mowse and XCorr, after normalizing Mowse to the XCorr scale (7).

[8] Manual validation confirmed assignments for 925 MS/MS identified by Mascot and Sequest, out of 920-930 MS/MS that we estimated were identifiable (7).

[9] Large MuDPIT dataset of K562 soluble proteins. Each SCX fraction was analyzed in four repeat RP-LC/MS/MS runs, with RP gradient 0-12.6% ACN/27 min, 12.6-18.9% ACN/45 min, 18.9-35% ACN/22 min, and 35-70% ACN/7 min, scanning m/z =350-1500 Da (Method 2).

[10] Large MuDPIT dataset of K562 soluble proteins. Each SCX fraction was analyzed once by Method 1, once by Method 2, and once by Method 2 with gas phase fractionation over three overlapping m/z mass ranges (300-798, 790-1038, 1030-1750 Da) (7).

[11] Derived from a large MuDPIT dataset of K562 soluble proteins (~600,000 DTA files) separated by gel filtration and proteolyzed, then separated by SCX-HPLC and analyzed by Method 2 RP gradient with gas phase fractionation over 10 overlapping m/z mass ranges as previously described (7). A subset of 1,448 very high confidence assignments from this dataset were selected for error analysis and behavior of fragment ion Average Ion Mass, shown in Suppl. Fig. 2E,F and Suppl. Fig. 4.

[12] LC/MS/MS analysis of a digest of 50 standard proteins (ABRF standard mix) analyzed twice by RP-LC/MS/MS as described in Suppl. Fig. 7 legend.

**Sun et al., Figure 1**