*i*PE-MMR: An integrated approach to accurately assign monoisotopic precursor masses to tandem mass spectrometric data

Hee-Jung Jung[1], **Samuel O. Purvine**[2], **Hokeun Kim**[1], **Vladislav A. Petyuk**[2], **Seok-Won Hyung**[1], **Matthew E. Monroe**[2], **Dong-Gi Mun**[1], **Kyong-Chul Kim**[1], **Jong-Moon Park**[1], **Su-Jin Kim**[1], **Nikola Tolic**[2], **Gordon W. Slysz**[2], **Ronald J. Moore**[2], **Rui Zhao**[2], **Joshua N. Adkins**[2], **Gordon A. Anderson**[2], **Hookeun Lee**[3], **David G. Camp II**[2], **Myeong-Hee Yu**[4], **Richard D. Smith**[2,*], and **Sang-Won Lee**[1,*]

[1]Department of Chemistry, Korea University, 1, 5-ka, Anam-dong, Seongbuk-gu, Seoul 136-701, Republic of Korea

[2]Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352, USA

[3]Lee Gil Ya Cancer and Diabetes Institute, Gachon University of Medicine and Science, 7-45-ka, Songdo-dong, Yeonsu-ku, Incheon 406-840, Republic of Korea

[4]Functional Proteomics Center, Life Sciences Division, Korea Institute of Science and Technology, 136-791, Hawalgok-dong, Seongbuk-gu, Seoul 130-650, Republic of Korea

Abstract

Accurate assignment of monoisotopic precursor masses to tandem mass spectrometric (MS/MS) data is a fundamental and critically important step for successful peptide identifications in mass spectrometry based proteomics. Here we describe an integrated approach that combines three previously reported methods of treating MS/MS data for precursor mass refinement. This combined method, "*i*ntegrated Post-Experiment Monoisotopic Mass Refinement" (*i*PE-MMR), integrates steps: 1) generation of refined MS/MS data by DeconMSn; 2) additional refinement of the resultant MS/MS data by a modified version of PE-MMR; 3) elimination of systematic errors of precursor masses using DtaRefinery. *i*PE-MMR is the first method that utilizes all MS information from multiple MS scans of a precursor ion including multiple charge states, in an MS scan, to determine precursor mass. By combining these methods, *i*PE-MMR increases sensitivity in peptide identification and provides increased accuracy when applied to complex high-throughput proteomics data.

Introduction

With the development of high-resolution hybrid mass spectrometry platforms (e.g. Quadrupole Ion Trap/Fourier Transform-Ion Cyclotron Resonance and Quadrupole Ion Trap/Orbitrap), MS-based proteomics measurements can yield thousands of high-resolution mass spectra along with tens of thousands of MS/MS spectra from a single LC-MS/MS analysis. These hybrid instruments allow acquiring MS/MS data correlated with accurate

CORRESPONDING AUTHOR FOOTNOTE, Sang-Won Lee: Department of Chemistry, Korea University, 1, 5-ka, Anam-dong, Seongbuk-gu, Seoul 136-701, Korea; Tel: +82-2-3290-3137; Fax: +82-2-3290-3121; sw_lee@korea.ac.kr, Richard D. Smith: Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352, USA; Tel: +1-509-371-6576; Fax: +2-509-376-7722; rds@pnl.gov.

precursor masses from high-resolution MS spectra (i.e. MS and MS/MS data produced simultaneously). This simultaneous approach of hybrid mass spectrometers greatly improves overall throughput by obtaining "free" MS/MS data simultaneous with MS data acquisition (i.e. no additional time to acquire MS/MS data).

A common experimental strategy using hybrid LC-MS/MS instruments requires the acquisition of a low resolution MS spectrum, obtained "on-the-fly" from a short time domain signal (ca. 180 ms, instead of utilizing full time domain signal of ca. 720 ms for the final spectrum), next this spectrum is analyzed for selection of precursor ions for MS/MS fragmentation in the ion trap. After the MS/MS analysis, *m/z* information for the selected precursor ions is recorded in an exclusion list for a fixed time duration to prevent reacquisition of MS/MS spectra for the same precursor ion. The next MS/MS spectrum is then obtained for a different peptide, having lower intensities. While this data-dependent LC-MS/MS experimental design allows for many peptide ions to be analyzed and increases the chance of identifying low abundant peptides, it frequently selects peptide ions for fragmentation that are relatively "weak" (i.e. at an early point in LC elution time). Even though the corresponding MS/MS spectra of the weak precursor ions may be of sufficient quality due to the higher sensitivity of the MS/MS event over the MS measurement, as a result of the selective accumulation of ions over a narrow mass range in the ion trap, the lower MS peak intensity makes it likely that the determined monoisotopic mass of the precursor ion will be less accurate than otherwise achievable; i.e. the number of peptide ions in the MS spectrum is low, resulting in deviations from expected isotopic distributions. The resultant non-statistical distributions with potentially missing peaks can lead to errors in monoisotopic mass determination (especially for larger peptides). This behavior is referred to as the "1–Da problem." Previously, we estimated that ~ 40% of precursor masses of conventionally processed tandem mass spectrometric data contained wrong monoisotopic mass information, which is attributed to this 1-Da problem.[1]

There have been several approaches for overcoming this issue.[2,3,4,5,6,7,8,9] Gygi and coworkers used a relatively large mass tolerance (50 ppm) for the first database search and subsequently subjected the search results to more stringent filtering (8 ppm).[4] This approach significantly reduced the false positive (FP) rate with only a minimal loss of peptide information. They also showed that accurate determination of precursor mass increased the number of phosphopeptide identification[5] and improved the accuracy in stable-isotope-based quantitation.[6] Similar results were obtained using a low resolution ion trap, but utilizing zoom scans in the data acquisition method.[7] Mann and co-worker described MaxQuant, which collectively considers peptide features as three-dimensional objects in *m/z*, elution time, and peak intensity, and determines accurate monoisotopic masses for the peptide features by calculating intensity-weighted averages of all MS peak centroids in the three-dimensional objects.[8] By applying this method to high resolution Orbitrap data, mass accuracy in the sub-ppm range was achieved, greatly reducing the 1-Da problem. DeconMSn performs THRASH deisotoping to provide better precursor masses, and was shown to effectively reduce incorrect monoisotopic peak assignments. Additionally, DeconMSn performs a summation of MS spectra over the retention time period to provide isotopic distributions with better signal-to-noise ratios.[9] It was also demonstrated that the use of summed spectra resulted in improved mass accuracies, along with an increased number of identified peptides over conventional Extract_MSn generation of DTAs. An independent approach, encoded in the software DtaRefinery[10,11], was demonstrated to eliminate systematic errors associated with monoisotopic masses of peptide identifications by using partial knowledge of peptide identifications and multiple regressions for mass accuracy dependencies on different experimental parameters.

PE-MMR was developed to provide accurate assignment of precursor masses to tandem mass spectrometry data.1 It first deisotopes every MS feature in the MS dataset using the THRASH algorithm12 and clusters the MS features of similar monoisotopic mass (within 10 ppm), but different LC elution times (defined by chromatographic peak width) into an unique mass class (UMC). 13 PE-MMR then uses Extract_MSn to generate MS/MS data (DTA files). It subsequently calculates neutral precursor masses from the MS/MS data (i.e. from $[M+H]^+$ mass of DTA files) and performs subtraction of multiples of 1.00235 Da, up to three times.1 PE-MMR uses the resultant four possible precursor masses to find matches with the UMC mass, as calculated by intensity-weighted averaging of monoisotopic masses of component MS features in each UMC. When matches are found, PE-MMR replaces the precursor masses of the MS/MS data with UMC masses. By considering possibilities of selecting an incorrect isotopic mass (i.e. by considering original mass and −1/−2/−3 Da subtracted masses), and using experimentally observed mass information from a UMC, instead of using mass information from a single spectrum to represent the mass of a peptide, it demonstrated an effective solution to the 1-Da problem in many cases, as well as improvements in mass measurement accuracy for identified peptides.

Here, we report substantially improved results by combining DeconMSn, PE-MMR and DtaRefinery into an integrated data analysis pipeline for tandem mass spectrometric data. This combined method, coined "*i*ntegrated post-experiment monoisotopic mass refinement" (*i*PE-MMR)14, uses DeconMSn to generate MS/MS data (DTA files) resulting in an initial correction of the 1-Da problem. It subsequently uses the DTA files to further correct and refine the precursor masses for the MS/MS data using a modified version of PE-MMR. The resultant refined DTA files are finally subjected to systematic error correction using DtaRefinery. Compared to the conventional method (Extract_MSn) or individual methods (DeconMSn or PE-MMR) of precursor mass correction, *i*PE-MMR resulted in a 20% increase in the number of identified peptides, with a decrease in the false positive rate, and an increase in mass measurement accuracy.

Experimental

Sample preparation

Preparation of the yeast proteome sample has been previously described in detail.1 Briefly, *S. cerevisiae* haploid strains Y2805 (*MATα pep::his3 prb1-Δ1.6R can1 his1-200 ura3-52)* and AF-2 (*HMLα or HMRa ho ade2-1 trp1-1 can1-100 leu2-3,112 his3-11,15 ura3-1 ssd1*) were used and the soluble fraction of cell lysate was used for subsequent tryptic digestion.

LC-MS/MS experiments

The tryptic yeast peptides were separated using a modified version of our ultra high-pressure dual on-line solid phase extraction/capillary reverse-phase liquid (DO-SPE/cRPLC) system, having a maximum operating pressure of 10,000 psi, which is described in detail elsewhere. 15 Briefly, the first dimensional separation is performed on a SCX column (150-µm inner diameter × 360-µm outer diameter × 30-cm length, 5-µm, 200-Å pore size, PolyLC), operated by an Agilent 1100 nano-LC system. The system was equipped with two RPLC columns (75-µm inner diameter × 360-µm outer diameter × 80-cm length, 3-µm, 300-Å pore size, Jupiter, Phenomenex, Torrance, CA, USA) for the second dimensional separation, which was operated using two ISCO pumps (100 DM). The sample was fractionated on-line into 14 salt gradients by SCX separation (0, 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 70, 100% salt fractions and 5% acetonitrile elution fraction) and then each of the fractions were alternatively directed to the two RPLC columns. In total, 14 sets of RPLC/MS/MS data with a gradient time of 180 min were used for this study.

A 7-tesla Fourier transform ion cyclotron resonance mass spectrometer (FTICR, LTQ-FT, Thermo Electron, San Jose, CA, USA) was used to collect the mass spectra. MS precursor ion scans (m/z 500 – 2,000) were acquired in the profile mode with an AGC target value of 1.0×10^6, a mass resolution of 1.0×10^5, and a maximum ion accumulation time of 1,000 ms. The mass spectrometer was operated in the data-dependent tandem MS mode; the three most abundant ions detected in the precursor MS scan were dynamically selected for MS/MS experiments simultaneously incorporating a dynamic exclusion option (exclusion mass width low: 1.10 Th; exclusion mass width high: 2.10 Th; exclusion list size: 120; exclusion duration: 30 s) to prevent reacquisition of the MS/MS spectra of the same peptides. Collision-induced dissociations of the precursor ions were performed in an ion trap (LTQ) with the collisional energy and isolation width set to 35% and 3 Th, respectively.

Database search

The MS/MS data were searched against a composite target-decoy database, containing the SGD database (orf_trans.fasta, ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/) and its reversed complements, using SEQUEST (Version 27, Revision 12) on Linux cluster systems having 32 – 35 nodes. The database was an alternate composite target-decoy database[16] that uses protein reversal as a decoy sequence. The tolerance was set to 10 ppm (see supplementary figure 1) or 0.1 Da or 3 Da for precursor ions and 1 Da for fragment ions. The oxidation of methionine (15.994920 Da) was used as a variable modification. The hydrolysis of asparagine (0.98402 Da) was used in analysis of deamidation. The search results were filtered using an estimated FP rate.[16] The FP rate of the peptide assignment was estimated through a composite target/decoy database search. The values of Xcorr and the ΔCn threshold for the 1% FP rate were used to obtain peptide IDs.

Generation of unique mass classes (UMCs) and UMC mass

As displayed in Figure 1, *i*PE-MMR combines several analysis steps of MS data de-isotoping and unique mass class clustering and MS/MS data refinements by DeconMSn, PE-MMR, DtaRefinery. As described in previous detail, the MS data sets arising from the LC-MS/MS experiments were submitted directly as a batch to ICR-2LS,[14] and the isotopic distributions and charge states of the peptide ions in the mass spectra were deconvoluted using the THRASH algorithm developed by Horn et al.[12] In an LC-MS/MS experiment, a peptide MS peak is detected over a period of time during LC elution. *i*PE-MMR groups such mass spectral peaks that have similar monoisotopic masses (within a mass tolerance of 10 ppm), but different LC elution times into a unique mass class (UMC). All of the UMCs observed in an LC-MS/MS experiment were recorded in an XML-formatted file (denoted herein a "UMC list"). We used a simple approach to generate UMCs by using mass tolerance and the requirement of detection in sequential mass spectra. Briefly, when a monoisotopic mass was first calculated, the mass was searched in subsequent MS scans with a mass tolerance of ±10 ppm. The sequential mass spectral peaks having masses within this tolerance were grouped into a UMC until there were no corresponding peaks for more than five consecutive mass spectra. Mass spectral peaks had to be observed in at least two scans to compose a UMC.

We note that a UMC uses neutral monoisotopic masses (not *m/z* values) to cluster MS features. A UMC contains monoisotopic mass information from every MS feature of different charge state within 10 ppm mass tolerance and elution requirements. After generation of UMCs, *i*PE-MMR calculates an intensity-weighted average monoisotopic mass value of up to 20 top intensity component masses of a UMC to represent each UMC, which we termed the "UMC mass".

Generation and refinement of MS/MS data

One of the major issues when using tandem mass spectra for accurate peptide identification is inaccuracy in determining the monoisotopic masses of the precursor peptide ions selected for fragmentation. As described in detail previously,[1] the conventional method of determining monoisotopic masses of precursor ions suffers from errors in deisotoping due to the non-statistical behavior of experimental isotopic distribution.

In this study, DeconMSn was used to process MS/MS data and generate the peak lists required by peptide identification tools (DTA files in the case of SEQUEST) (Figure 1a). Unlike conventional methods, DeconMSn performs THRASH deisotoping of the precursor ions (including spectral summation across multiple MS scans when signal levels are low) and determines monoisotopic masses for the precursors. For DeconMSn analysis, MS/MS fragmentation scans containing less than five ion peaks were not considered.

We modified the original algorithm of PE-MMR to maximize the effect of utilizing DeconMSn-generated DTA files (Figure 1b). The improved PE-MMR algorithm now calculates neutral precursor masses from a DTA and performs subtraction of multiples of 1.00235 Da (the mass difference between the adjacent isotopic peaks) from the neutral precursor mass, up to three times, and the addition of 1.00235 Da (Figure 1b, *mass corrections*). It subsequently compares the five masses (original DTA neutral mass, three masses from subtractions, and one mass from addition of 1.00235 Da) with the UMC masses within ± 10 scans (*range scanning*). This step of range scanning is now performed for each of these five mass candidates, even if there is direct matching of the original DTA neutral mass to the UMC mass of the scan number. The original PE-MMR did not perform range scanning when there is direct matching between the original DTA neural mass to the UMC mass of the scan number. This change was made in order to account for the cases in which the UMC of the scan number is comprised of masses from isotopically substituted peaks only, and thereby the UMC mass is inaccurate. For example, when a peptide was measured in three different charge states (e.g. +1, +2 and +3), only two charge states (i.e. +2, +3) may be abundant enough to produce correct isotopic distributions to give a correct monoisotopic mass. If the third charge state (+1) is not strong enough and the monoisotopic masses of this charge state is consistently differed by one Da from the true value, these masses can be grouped into a different UMC having a mass difference of one Da from the true value. Previously, when a DTA mass matched to the UMC of a wrong mass (UMC comprised of +1 charge state masses), it used the UMC mass without searching for better matches. However, with the improved PE-MMR algorithm, it is possible to find the correct mass (UMC masses of +2 and +3 charge state), even for the DTA of the +1 charge state as PE-MMR still looks for the matches of other possible masses within ± 10 scans (in this case, 1 Da smaller masses that match to the UMC mass of the +2 and +3 charge state).

If a match was found within a tolerance of ± 25 ppm, PE-MMR replaces the precursor mass of the MS/MS file with the UMC mass. If no match was found, it either uses the DTA files with no mass refinement or filters them out. We note that a single MS/MS scan can generate multiple DTA files with different precursor masses if multiple matches were found. On average, ca. 15% of final DTA files after *i*PE-MMR were resultant of the multiple matches. After applying target-decoy cutoff criteria of 1% false positive rate, ca. 11% of final peptide identifications were corresponding to multiple matches. Finally, *i*PE-MMR uses the one peptide identification that has the highest Xcorr score.

The resultant MS/MS files after PE-MMR processing were subjected to DtaRefinery (Figure 1a) with default settings. DtaRefinery performs a preliminary database search using X! Tandem[17] and these prior knowledge on peptide identifications is utilized in multivariate calibration to reduce systematic instrumental mass errors. Next *i*PE-MMR generates MS/MS

files that are refined and calibrated, which are subsequently subjected to a SEQUEST database search. After the SEQUEST search, the target-decoy method16 was used to filter the search results using an estimated false positive rate of 1% as described above. The processing times of individual analysis steps (i.e. DeconMSn, PE-MMR, DtaRefinery) of *i*PE-MMR are provided in supplementary table 1. All of the PE-MMR processes of *i*PE-MMR were written in Java (Java Runtime Environment package version 1.5, ORACLE), and every component programs of *i*PE-MMR are publically available14.

Results and Discussions

Increased accuracy and sensitivity of iPE-MMR for peptide identification

Figure 2 compares the number of peptide identifications (Figure 2a), mass accuracy distributions (Figure 2b), and SEQUEST Xcorr distributions (Figure 2c) obtained by the conventional method (Extract_MSn), PE-MMR, DeconMSn, and *i*PE-MMR. As no attempt was made to calibrate the precursor masses of Extract_MSn generated DTA data, a somewhat wider tolerance of 0.1 Da was used for the SEQUEST search for these data. For other data, however, precursor mass tolerance of 10 ppm was used. As previously reported, PE-MMR resulted in a 12% increase in peptide identifications compared to the conventional method; DeconMSn also produced a similar increase. By combining the two methods, *i*PE-MMR further increased the number of peptide identifications from 7,286 to 8,710 (ca. 20% increase), within the same false positive rate of 1%. We note that the number of explainable MS/MS spectra within a FP rate of 1% was significantly increased by applying the *i*PE-MMR algorithm (Figure 2c). While 29,870 DTA files contained forward database hits among the total of 251,357 in the conventional method, 40,523 DTA files resulted in forward database hits after the *i*PE-MMR process. The result is an additional 10,653 MS/MS spectra (an increase of approximately 35%) assigned to forward peptide hits at an FDR of 1% after *i*PE-MMR process.

With the use of intensity-weighted average monoisotopic masses and elimination of systematic mass errors, *i*PE-MMR improved mass measurement accuracy (MMA) of peptide identification significantly: from 4.48 ± 2.08 ppm with Extract_MSn, 0.44 ± 1.53 with PE-MMR, and -0.03 ± 1.01 ppm with DeconMSn to 0.05 ± 0.94 ppm with *i*PE-MMR analysis (Figure 2b).

iPE-MMR is a monoisotopic mass determination method utilizing multiple scans and multiple charge states of a peptide

One of the mysteries in MS-based proteomics is that less than 20% of MS/MS data from routine LC-MS/MS experiments can be assigned to peptide sequences. This lack of coverage has been explained, in part, by a number of causes: presence of covalent modifications, alternative splicing, co-fragmentation, low MS/MS sensitivities, non-peptide components, and other contaminants. Along with others, we observed that inaccurate assignment of monoisotopic precursor masses to MS/MS data was one of major explanations. There have been many methods for determining monoisotopic precursor masses associated with MS/MS data, and these approaches can be classified into three categories: single scan/single charge state (SSSC), multiple scan/single charge state (MSSC), and multiple scan/multiple charge state (MSMC) methods. The single scan/single charge state (SSSC) method includes Extract_MSn and others. The SSSC method relies on a single isotopic distribution measured in a single MS spectrum. While most simple, these are subjected to inaccurate assignment of precursor monoisotopic mass. For example, using Extract_MSn, one of the SSSC methods, we estimated that 40% of the final peptide identifications were corrected after PE-MMR by 1, 2, or 3 Da, relative to the Extract_MSn generated DTAs. Figures 3a and 3b compare fragment annotations obtained using Extract_MSn generated DTA mass and *i*PE-MMR

corrected mass. The annotated MS/MS spectrum of the highest ranking peptide, using original DTA mass, was not only poorly correlated with the theoretical fragmentation pattern, but also with a peptide match to the reverse sequence database (i.e. a false peptide identification). However, the *i*PE-MMR algorithm corrected the precursor mass by 1.00235 Da and the resultant MS/MS spectrum was annotated with high confidence and matched the forward sequence database. The SSSC methods utilize only one MS measurement, even though there may be many more MS measurements of the same precursor ions at higher intensities and alternate charge states found in the LC-MS/MS experiment. The SSSC method often fails to identify the correct peptide whenever the referenced MS spectrum exhibited a non-statistical isotopic distribution. Conversely, *i*PE-MMR utilizes MS spectra of higher intensities to calculate precursor mass resulting in fewer errors.

Multiple scan/single charge state (MSSC) methods take multiple MS spectra for an *m/z* precursor ion into account when determining monoisotopic masses, but do not consider other charge states measured simultaneously. DeconMSn and MaxQuant can be classified as a MSSC method. For example, DeconMSn can perform spectral summation over a period of LC elution times and calculates monoisotopic masses by performing THRASH on this improved MS spectrum. By utilizing *m/z* precursor spectra from multiple MS spectra, these methods effectively correct inaccurate determination of isotopically substituted peaks for the monoisotopic peak. These methods, however, still suffer inaccurate assignments especially when the MS spectra of the particular charge state are all weak and spectral summation does not significantly improve precursor mass determination. An example of this case is shown in Figure 3c, where the DeconMSn generated DTA resulted in a non-tryptic peptide (V.SM*FEEVRVDPVLEYLDNRILDLI.M, * = oxidation, NTT = 0) with fragments poorly annotated. Upon *i*PE-MMR analysis, however, the same MS/MS spectrum was matched to a tryptic peptide (K.VFQVGNIVHPSVVVSNDEENNELVR.T, NTT=2) which differed in mass by 1 Da (Figure 3d). Close inspection of the precursor ions of charge state +2 revealed that they were weakly observed at all times during their LC elution, and indicated a [M+H]+ of 2794.4337 Da even after spectral summation (Figure 4a–c). On the other hand, the +3 charge state of the corresponding peptide was measured at higher intensities (ca. 400 times higher) and deisotoping the isotopic distribution of the +3 charge state resulted in a [M+H]$^+$ of 2793.4171 Da (Figure 4e–g). During the *i*PE-MMR process, +2 and +3 ions of this peptide were grouped into two different UMCs as they are differed by 1 Da; two DTAs for this MS/MS spectrum were generated that contained two different UMC masses, as two of the masses among five candidate masses were matched to the UMC masses of these two UMCs. While both corresponded to forward hits within the FP rate of 1%, the peptide identification from the DTA that contained the UMC mass of the +3 charge state (−1 Da corrected) has a higher Xcorr value than that of the +2 charge state (5.83 versus 2.58), with fragments better annotated. A DTA mass of the +2 precursor ion was replaced by that of the +3 charge state (Figure 4d), and the peptide sequence from the original DTA of the +2 charge state was discarded as one MS/MS spectrum of a peptide would result in one peptide identification barring co-fragmentation of peptides. While this particular case may not increase the number of correct peptide identifications in the final peptide list as the DTA of +3 charge state resulted in the same peptide identification, *i*PE-MMR reduces false positives from inaccurately assigned monoisotopic masses. *i*PE-MMR searches for a mass match using all five different masses (from subtractions and additions of 1.00235 Da from the experimental value) to a UMC that contains every monoisotopic mass from all measured charge states in multiple MS scans and considers all the matches as candidates before they are discarded after the database search except the peptide sequence with the highest peptide identification score, making *i*PE-MMR a true multiple scan/multiple charge state (MSMC) based method. It is worth noting that the PE-MMR component in *i*PE-MMR is improved with the assistance of DeconMSn as it performs summation of MS spectra of weak intensities and provides better precision in determining peak position. As described above,

the spectral summation may still suffer inaccurate determination of monoisotopic peaks, but improves the precision of each peak in a given isotopic distribution. This corrects the case where the original DTA data contained precursor masses outside of the mass tolerance (25 ppm) of the PE-MMR component, even after mass corrections of multiples of 1.00235 Da, due to severely non-statistical distributions and inaccurate peak positions. In the preceding case, PE-MMR processing fails to provide accurate monoisotopic mass information as it fails to find matches to the UMC masses. DeconMSn, however, improves the S/N ratio of the MS spectra and provides more precise peak positions from the summed spectra.

iPE-MMR increases accuracy in identifying peptide modifications: deamidation

Figure 5 illustrates the effect of accurate determination of monoisotopic masses on the accuracy of peptide identifications. MS/MS data from the 50% salt fraction was used for this figure. The difference between experimental and calculated peptide masses (delta mass, DelM), as a function of Xcorr, are plotted with the results from Extract_MSn generated DTAs and *i*PE-MMR generated DTAs. A set of separate SEQUEST searches, with a precursor tolerance of 3 Da, was performed for these data; static values for Xcorr and ΔCn were used to filter the results (ΔCn ≥ 0.1, Xcorr ≥ 1.9 for +1 peptides, Xcorr ≥ 2.2 at +2 peptides, Xcorr ≥ 3.2 for ≥ +3 peptides). Compared to Extract_MSn, *i*PE-MMR improves the 1-Da problem as the identifications near delta masses of 1 or 2 are significantly reduced in the case of *i*PE-MMR (Figure 5d and 5e), as well as improving the mass measurement accuracy distribution of a delta mass of 0.

However, despite combining two independent methods for accurate monoisotopic mass assignment, 223 (ca. 7%) of the final DTA files after *i*PE-MMR are still observed to contain monoisotopic masses of 1-Da higher than the predicted masses of peptide identification. A manual inspection of the MS spectra of these peptide ions revealed that the MS spectra of these ions were actually corresponding to masses of 1 Da higher, such as in Figure 6b. Despite the high intensity and well-defined isotopic distribution with great S/N ratio, the isotopic distributions of the peptide peaks of all the different charge states resulted in precise monoisotopic masses of 1 Da higher than the predicted monoisotopic mass of the peptide identification. That is, deisotoping empirically calculates the correct monoisotopic mass in this case. This observation led to consideration of deamidation modification in the SEQUEST searches for these DTAs. Deamidation is a common modification that transforms asparagine into aspartic acid and increases peptide mass by 0.98402 Da for every occurrence.[18] In fact, a recent study showed that 70 – 80% of -Asn-Gly- sequence were deamidated during sample digestion.[19] When the same *i*PE-MMR generated DTAs were subjected to a SEQUEST search with precursor mass tolerance of 10 ppm and deamidation option as dynamic modification, 163 (ca. 73%) of 223 DTA files resulted in the same stripped peptide sequences but with deamidation at asparagine. Figure 5f was obtained after identifications near DelM = 1 and 2 were corrected by SEQUEST search results with deamidation on asparagines. We note that there are still 60 (1.8%) and 30 (0.9%) forward hits at, respectively, DelM = 1 and 2 even after considering deamidation (Figure 5f). These may be due to a misidentification of peptide sequences by SEQUEST searches or remaining errors in determining the true monoisotopic mass.

Figure 6 shows an example of a corrected identification by considering deamidation. As shown in comparison of the two annotated spectra (Figure 6c and 6d), the SEQUEST search, a "pattern-match" type peptide identification, failed to distinguish the two since both annotated spectra were virtually identical when a wide precursor mass tolerance of 3 Da was used. They could only be distinguished after accurate assignment of monoisotopic mass was given. Previously, we demonstrated that the PE-MMR process significantly improved the accuracy of glycopeptides identifications, which are identified by the same modification (i.e. Asn to Asp), through the correct assignment of monoisotopic masses. Likewise, herein, prior

assignment of accurate monoisotopic masses to SEQUEST searches prevented incorrect identification of deamidated peptides. While it is likely that high-resolution ECD MS/MS data20 and carefully controlled proteome sample preparation21 would be required for unambiguous determination of deamidation, it remains critical to accurately assign the true monoisotopic masses for all members in the typically massive MS/MS datasets encountered in a standard proteomics experiment in order to correctly assign peptide deamidation and other peptide modifications of 1 Da or less.

Conclusions

We have developed an integrated approach for more accurate assignment of precursor masses for MS/MS data. Unlike previous methods, *i*PE-MMR uses every monoisotopic mass of a peptide from every measured charge state for all relevant MS data in an LC/MS/MS analysis to assign accurate precursor masses. With the increased precision of precursor masses provided by DeconMSn and the effective removal of systematic errors associated with UMC masses obtained using DtaRefinery, *i*PE-MMR enhances not only mass accuracy, but also increases the number of peptide identifications. Accurate assignment of monoisotopic precursor masses enabled limiting mass tolerances for database search to 10 ppm and effectively reduced search space. This allowed the identification of peptide modifications such as deamidation with significantly less ambiguity.

While we combined DeconMSn, PE-MMR and DtaRefinery into an integrated data analysis pipeline for tandem mass spectrometric data, *i*PE-MMR is not restricted to this application. *i*PE-MMR can seamlessly incorporate other precursor mass analysis tools that are compatible in a data format (assuming none will be perfect), such as the RAPID algorithm22 for deisotoping, and should provide significant additional benefits. As exemplified in this study, "open proteomics software forum" for not only sharing software and their source codes but also collaborating for various integrations hold great promises for advancing proteomics informatics and we suggest the initiation of such efforts.

Supplementary Material

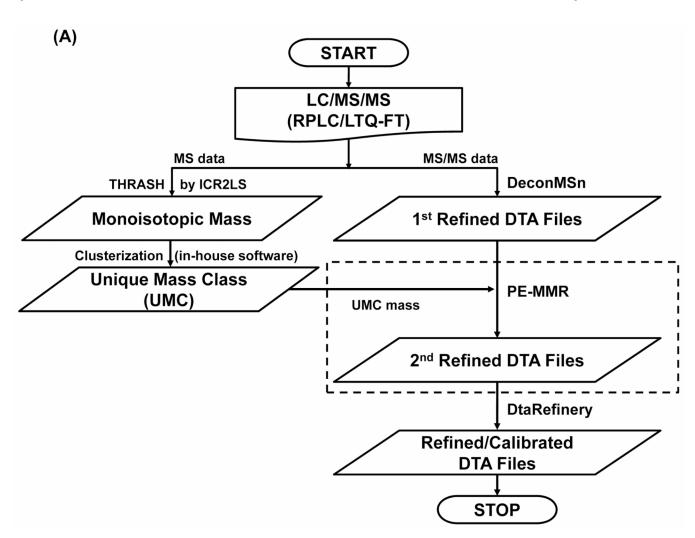Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Shin B, Jung H-J, Hyung S-W, Kim H, Lee D, Lee C, Yu M-H, Lee S-W. Mol. Cell. Proteomics. 2008; 7:1124–1134. [PubMed: 18303012]

2. Luethy R, Kessner DE, Katz JE, McLean B, Grothe R, Kani K, Faca V, Pitteri S, Hanash S, Agus DB, Mallick P. J. Proteome Res. 2008; 7:4031–4039. [PubMed: 18707148]

3. Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. J. Proteome Res. 2010; 9:1138–1143. [PubMed: 19938873]

4. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. Nat. Biotechnol. 2006; 24:1285–1292. [PubMed: 16964243]

5. Bakalarski CE, Haas W, Dephoure NE, Gygi SP. Anal. Bioanal. Chem. 2007; 389:1409–1419. [PubMed: 17874083]

6. Bakalarski CE, Elias JE, Villen J, Haas W, Gerber SA, Everley PA, Gygi SP. J. Proteome Res. 2008; 7:4756–4765. [PubMed: 18798661]

7. Victor KG, Murgai M, Lyons CE, Templeton TAB, Moshnikov SA, Templeton DJ. J. Am. Soc. Mass Spectrom. 2010; 21:80–87. [PubMed: 19850495]

8. Cox J, Mann M. Nat. Biotechnol. 2008; 26:1367–1372. [PubMed: 19029910]

9. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD. Bioinformatics. 2008; 24:1021–1023. [PubMed: 18304935]

10. Petyuk VA, Mayampurath AM, Monroe ME, Polpitiya AD, Purvine SO, Anderson GA, Camp DG 2nd, Smith RD. Mol. Cell. Proteomics. 2010; 9:486–496. [PubMed: 20019053]

11. Petyuk VA, Jaitly N, Moore RJ, Ding. J, Metz TO, Tang K, Monroe ME, Tolmachev AV, Adkins JN, Belov ME, Dabney AR, Qian WJ, Camp DG 2nd, Smith RD. Anal. Chem. 2008; 80:693–706. [PubMed: 18163597]

12. Horn DM, Zubarev RA, McLafferty FW. J. Am. Soc. Mass Spectrom. 2000; 11:320–332. [PubMed: 10757168]

13. Zimmer JS, Monroe ME, Qian WJ, Smith RD. Mass Spectrom. Rev. 2006; 25:450–482. [PubMed: 16429408]

14. *i*PE-MMR and the component softwares such as DeconMSn, PE-MMR, DtaRefinery, and ICR2LS are publically available from http://omics.pnl.gov/software/.

15. Min H-K, Hyung S-W, Shin J-W, Nam HS, Ahn SH, Jung HJ, Lee S-W. Electrophoresis. 2007; 28:1012–1021. [PubMed: 17309056]

16. Elias JE, Gygi SP. Nat. Methods. 2007; 4:207–214. [PubMed: 17327847]

17. Fenyö D, Beavis RC. Anal. Chem. 2003; 75:768–774. [PubMed: 12622365]

18. Aswad, DW., et al. Deamidation and Isoaspartate Formation in Peptides and Protein. Aswad, DW., editor. Boca Raton: CRC series in analytical biotechnology; 1995. p. 1-259.

19. Krokhin OV, Antonovici M, Ens W, Wilkins JA, Standing KG. Anal. Chem. 2006; 78:6645–6650. [PubMed: 16970346]

20. Yang H, Fung EYM, Zubarev AR, Zubarev RA. J. Proteome Res. 2009; 8:4615–4621. [PubMed: 19663459]

21. Ren D, Pipes GD, Liu DJ, Shih LY, Nichols AC, Treuheit MJ, Brems DN, Bondarenko PV. Anal. Biochem. 2009; 392:12–21. [PubMed: 19457431]

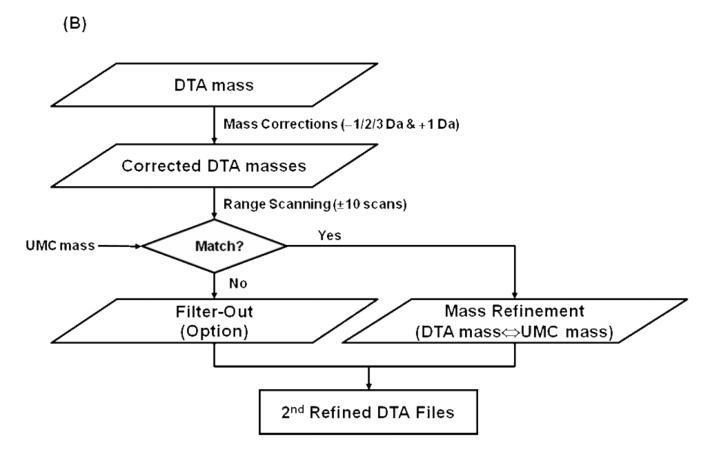22. Park K, Yoon JY, Lee S, Paek E, Park H, Jung H-J, Lee SW. Anal. Chem. 2008; 80:7294–7303. [PubMed: 18754627]

(A)

Figure 1.
(A) Overall data flow chart of *i*PE-MMR pipeline. (B) An expanded flow chart of the dotted box of Figure 1a.
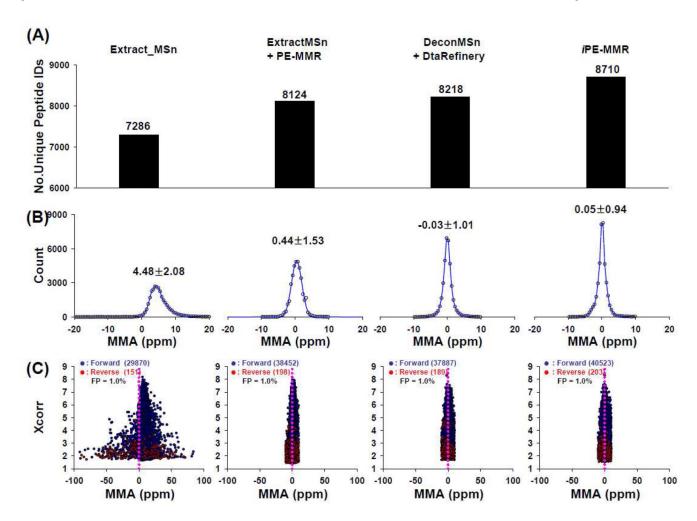
Figure 2.
Improvement in sensitivity and accuracy of peptide identifications by *i*PE-MMR. (A)
Comparison of number of unique peptide identifications from different methods. (B)
Comparison of mass measurement accuracy (MMA) distributions of peptide identifications
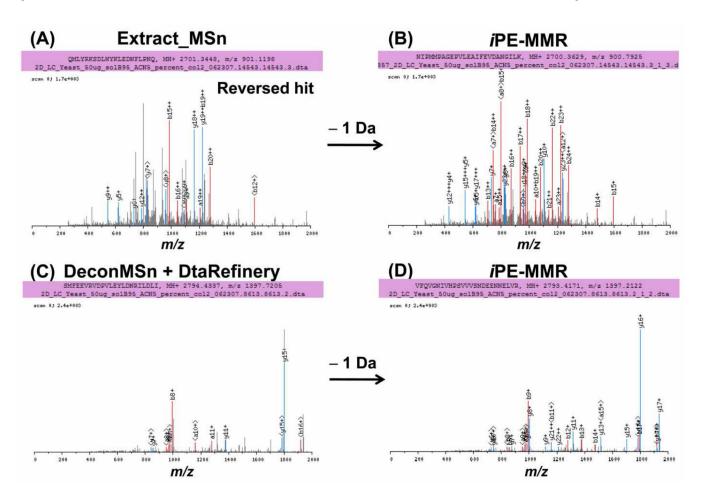from different methods. (C) Comparison of distributions of peptide-spectral matches from
different methods.

Figure 3.
Examples of improved peptide identifications after *i*PE-MMR. (A) The annotated spectrum
of Extract_MSn generated DTA. The resultant peptide identification was a reverse sequence
peptide. (B) The annotated spectrum of the *i*PE-MMR generated DTA using the same DTA.
The precursor mass was replaced by a UMC mass that was ca. 1 Da smaller. (C) The
annotated MS/MS spectrum after DeconMSn. (D) The annotated MS/MS spectrum of the
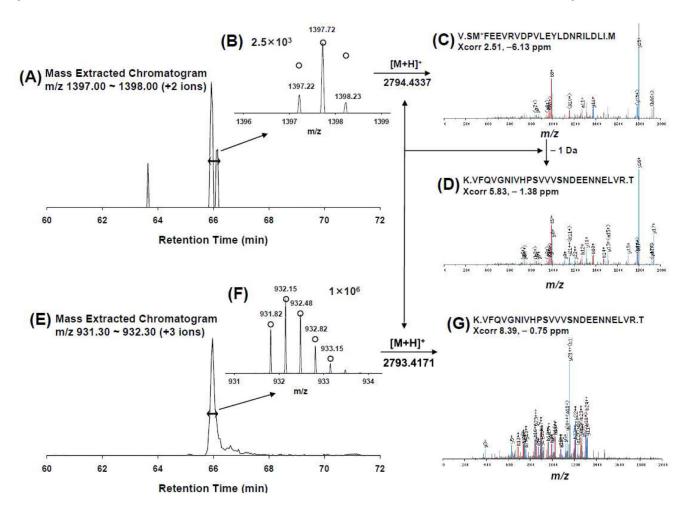same DTA but after *i*PE-MMR.

Figure 4.
(A) Mass extracted chromatogram of +2 state precursor ions. (B) A spectral summed MS spectrum of the +2 charge state. (C) An annotated MS/MS spectrum of the +2 charge state ion using a DTA mass of 2792.4337. (D) An annotated MS/MS spectrum of the +2 charge state ion after correcting the DTA mass by −1.00235 Da. (E) Mass extracted chromatogram of the +3 state precursor ions. (F) A spectral summed MS spectrum of the +3 charge state. (G) An annotated MS/MS spectrum of a +3 charge state ion using DeconMSn generated DTA. ○ denotes theoretical abundances of the corresponding peptide.
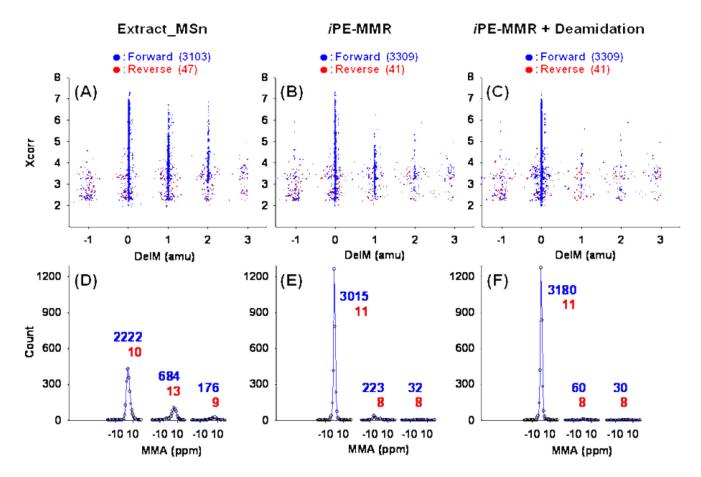
Figure 5.
Comparison of plots of mass differential (DelM) versus SEQUEST Xcorr for filtered
peptide identification from Extract_MSn generated DTAs (A), *i*PE-MMR generated DTAs
(B), and *i*PE-MMR generated DTAs with consideration of deamidation (C). Static filters
used: ΔCn ≥0.1, Xcorr≥1.9 for +1 peptides, Xcorr≥2.2 for +2 peptides, Xcorr≥3.2 for ≥+3
peptides. Comparison of mass measurement accuracy distributions around each unit bin of
DelM of Extract_MSn generated DTAs (D), *i*PE-MMR generated DTAs (E), and *i*PE-MMR
generated DTAs with consideration of deamidation (F). The numbers are the amount of non-
redundant peptide identifications within ±10 ppm of DelM = 0, 1, 2. Peptide identifications
around other DelM were not considered as they exhibited complete randomness. For clarity,
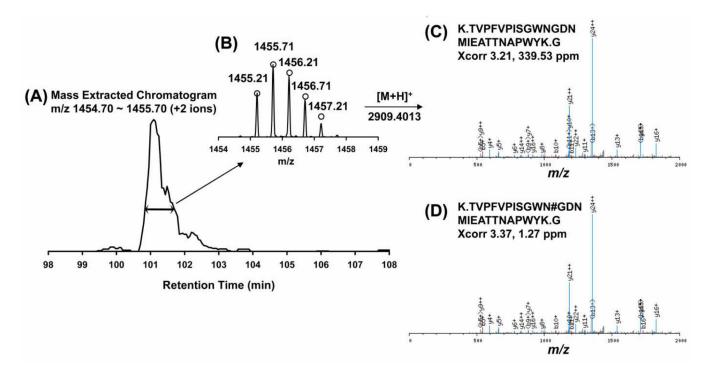MS/MS data from the 50% salt fraction was used for this figure.

Figure 6.
(A) Mass extracted chromatogram of a +2 precursor ion. (B) An averaged isotopic distribution of the corresponding precursor ion. ○ denotes the theoretical isotopic distribution of the predicted peptide (R.NMIIVPEMIGSVVGIYN#GK.A, #=deamidation). (C) Peptide identification without consideration of deamidation. (D) Peptide identification with consideration of deamidation.