DtaRefinery, a Software Tool for Elimination of Systematic Errors from Parent Ion Mass Measurements in Tandem Mass Spectra Data Sets*

Vladislav A. Petyuk, Anoop M. Mayampurath, Matthew E. Monroe, Ashoka D. Polpitiya, Samuel O. Purvine, Gordon A. Anderson, David G. Camp II, and Richard D. Smith‡

Hybrid two-stage mass spectrometers capable of both highly accurate mass measurement and high throughput MS/MS fragmentation have become widely available in recent years, allowing for significantly better discrimination between true and false MS/MS peptide identifications by the application of a relatively narrow window for maximum allowable deviations of measured parent ion masses. To fully gain the advantage of highly accurate parent ion mass measurements, it is important to limit systematic mass measurement errors. Based on our previous studies of systematic biases in mass measurement errors, here, we have designed an algorithm and software tool that eliminates the systematic errors from the peptide ion masses in MS/MS data. We demonstrate that the elimination of the systematic mass measurement errors allows for the use of tighter criteria on the deviation of measured mass from theoretical monoisotopic peptide mass, resulting in a reduction of both false discovery and false negative rates of peptide identification. A software implementation of this algorithm called DtaRefinery reads a set of fragmentation spectra, searches for MS/MS peptide identifications using a FASTA file containing expected protein sequences, fits a regression model that can estimate systematic errors, and then corrects the parent ion mass entries by removing the estimated systematic error components. The output is a new file with fragmentation spectra with updated parent ion masses. The software is freely available. Molecular & Cellular Proteomics 9: 486–496, 2010.

A key component in modern proteomics research is peptide identification through LC coupled to tandem MS where a selected parent or precursor ion from an MS scan undergoes fragmentation by collisionally activated/induced dissociation or any other methods (1). Identification of the putative peptides corresponding to the parent ions selected for fragmentation is performed by matching the observed to the theoretical MS/MS fragmentation patterns. The first step in the data analysis process is to create a set of input files representing the fragmentation spectra. For example, for the data sets from LTQ FT and LTQ Orbitrap instruments, software tools such as extract_msn (part of BioWorks software package, Thermo Electron, San Jose, CA) or DeconMSn (2) are often used for this step, creating files in “.dta” or other formats for the fragmentation spectra. These files contain the mass and charge of the parent ion and observed fragmentation pattern in the form of a list of \( m/z \) and intensity pairs. Once created, database search tools such as SEQUEST (3), X!Tandem (4), OMSSA (5), InsPect (6), MASCOT (7), Spectrum Mill (8), RAId_DbS (9), and others are used to analyze the .dta files to associate each MS/MS fragmentation pattern with a corresponding putative peptide sequence. Therefore, MS/MS fragmentation pattern information plays a primary role and ultimately can be used as essentially the only type of information for peptide identification in LC-MS/MS experiments (4, 10). However, in this case, the lack of constraint on parent ion mass measurement error (MME) results in a high rate of incorrect peptide identifications. Conversely, improved mass accuracy helps to achieve a better discrimination between true and false peptide identifications (8, 11).

To fully utilize the high mass measurement accuracy of modern instruments, it is advantageous to eliminate systematic mass measurement errors. Eliminating the systematic MME component results in a more coherent distribution of total MME and helps to reduce the maximum allowable deviation (of the measured mass from the theoretical peptide mass) for true peptide identifications (12). Multiple sources of variation can cause systematic errors in mass measurements; for example, power supply voltage drift over time, space charge effects, differing ion compositions within the cell, ion intensity variation, and outdated calibration coefficients (for a review, see Ref. 13).
The use of internal calibrants or standards co-injected with the sample into the mass spectrometer (14–18) can help reduce such systematic errors to a certain extent but may have some practical limitations. Internal calibrants well capture scan to scan variations and correct for time and/or total ion current (TIC)-dependent systematic errors, which are associated with the entire MS scan. Technically, it is also possible to correct for intensity-related dependence of MME, which is quite prominent on a certain type of instrument (13). However, it will require an increase in the number of calibrants to cover the entire dynamic range of the mass spectrometer. In addition, in certain cases, the calibration function (MME dependence on \( m/z \) parameter) shows evidence of non-linear behavior and may not be corrected by one or even a few calibrants (for example, see Fig. 7 in Ref. 13 and Fig. 5 in Ref. 19). Thus, there still potentially can be residual systematic biases even if internal calibration is applied.

A number of alternative approaches based on knowledge about the sample content have been introduced recently (13, 20–25). Instead of spiked or co-injected calibrants, they make use of putative peptide identifications as internal calibrants. Initially, such recalibration approaches have been limited mostly to peptide identifications based on either high accuracy measurements of peptide masses alone (21) or in combination with LC retention times (13, 20). Recently, we and others have proposed that partial sample knowledge can also be utilized for recalibrating parent ion masses in MS/MS data sets obtained on hybrid instrumentation (12, 13, 23–26). In one implementation, described as “postexperiment monoisotopic mass filtering and refinement” (26), the parent ion masses in the .dta files were replaced with the mass of the ion averaged over all scans in which it was observed followed by a simple recalibration. That recalibration assumes one constant value of the systematic error for the entire data set, which can be estimated by zero centering the parent ion MME distribution. There is another implementation that resembles this concept (23, 25) that is based on using putative peptide identifications as calibrants from neighboring MS/MS scans, that is scans either immediately before or after the MS scan chosen for recalibration. We reviewed this approach earlier (13) and suggested that it has a potential benefit, although it does have a few potential limitations that need to be addressed, e.g. disregarding individual ion intensity information, use of only linear calibration functions, and lack of control of the \( m/z \) range covered by putative calibrants. Recently, another MS/MS data set recalibration tool has been developed (24) that incorporates the time component into recalibration equation. However, in this case, the authors assumed linear relationships of MME and time, which is rarely the case (13) and may serve only as a rough approximation. In line with our previous report (13), we recommended using a multidimensional non-parametric recalibration, an approach that is not limited by the disadvantages mentioned above.

To derive practical benefits from our previous study on systematic MME behavior for the proteomics community, here, we developed an algorithm for eliminating systematic biases in the parent ion MME for the MS/MS data sets and implemented it into a software tool. This tool, DtaRefinery, is designed to work in tandem with either extract_msn or DeconMSn (Fig. 1). DtaRefinery first reads a set of fragmentation spectra from a concatenated .dta file (supplemental Fig. 1) produced by either extract_msn or DeconMSn. Next, it internally calls an MS/MS search engine to identify putative peptides based on matching MS/MS fragmentation patterns against an appropriate, user-specified FASTA file containing sequences of proteins expected to be present in the sample. Conceptually, it does not matter which MS/MS search engine is used, although we prefer X!Tandem as it is free, open source, and relatively lightweight. X!Tandem is included in the package: there is no need for installation of a search engine. All the database searching is done behind the scene, and the generated files with MS/MS search results exist only temporarily and are deleted after the parsing step. DtaRefinery then computes the parent ion MME based on observed masses and theoretical monoisotopic masses derived from peptide sequences.
sequences. In the next step, the DtaRefinery trains a regression-based prediction model for the systematic components of the parent ion MME. The input is a concatenated .dta file created by DeconMSn while processing a Thermo raw file. This additional information includes the MS scan at which each parent ion was detected, parent ion intensity, TIC in the Orbitrap or FTICR cell, and the automatic gain control accumulation time for the MS scan. These extra parameters can be utilized as optional explanatory variables for regression analysis to find and correct systematic error dependences. Thus, the use of DeconMSn instead of extract_ms is highly recommended because it not only accurately determines the monoisotopic masses of the parent ions but also provides additional useful information about parent ion measurement.

DtaRefinery analysis settings can be customized by editing the XML file containing processing options or using the GUI of the program. The concatenated .dta file can be produced by either DeconMSn (-XCDTA option) or concatenation of individual .dta files produced by extract_ms (using the Peptide File Extractor). Concatenated .dta files (*.dta.txt) contain each fragmentation spectrum in the .dta format separated by a header line describing the source .dta file (supplemental Fig. 1).

Once spectra are loaded, DtaRefinery internally uses X!Tandem to briefly search the MS/MS data entries from the input file to identify peptides. Note that the MS/MS search does not have to be exhaustive. Capturing the majority of the peptides should be sufficient for recalibration. For example, we suggest searching global trypic digest samples only for fully tryptic peptides and ignoring partially tryptic and non-tryptic peptides. This non-exhaustive search, restricted to the space of fully tryptic peptides, is ~1-2 orders faster than searching for all peptide types yet still captures close to 90% of all identifiable peptides. For the purpose of sample recalibration, those ~10% of semi- and non-tryptic peptide ions can be safely ignored. Another way of speeding MS/MS searches can be achieved by narrowing the parent ion MME tolerance. For example, we use a parent ion MME tolerance for preliminary peptide identification searches in the range of ±20-100 ppm for data sets obtained using high resolution hybrid mass spectrometers. This tolerance range allows for a faster search than the ±3-Da tolerance commonly used for data sets from ion trap instruments. To assess the gain in speed, we timed X!Tandem searches for a mouse brain trypic digest peptide sample that was analyzed using a 30-min-gradient LC separation and an LTQ Orbitrap mass spectrometer. With a ±3-Da parent ion MME tolerance and no enzyme rule setting, the test data set took ~9 h or ~250 ms/spectrum/megabyte of database (on a single thread). With the parent ion MME setting adjusted to ±100 ppm and a requirement that peptides be fully tryptic, the search took ~1 min or 0.5

MATERIALS AND METHODS

Software Implementation—DtaRefinery is implemented using the Python programming language and depends on the wxPython, Matplotlib, and NumPy open source Python libraries for graphical user interface (GUI), plotting images and numerical computations, respectively. For peptide identifications, the program uses X!Tandem (4), which is also freely available and open source. However, there is no need to install X!Tandem because it is shipped within DtaRefinery. Optionally, the R statistical environment and R(D)COM server application can be used for smooth spline fitting, which minimizes the sum of squared errors plus weighted squared secondary derivative (smooth.spline) function. DtaRefinery has a GUI but also can be run at the command prompt for batch automation. The software tool is available in two formats, both as a stand-alone executable version for the Microsoft Windows 32-bit platform or as a platform-independent collection of Python scripts. If installed as the stand-alone executable version, there is no need to install Python or any prerequisite libraries.

Features and Algorithm—The concept behind the parent ion MME refinement process is shown in Fig. 3. The input is a concatenated .dta file that contains MS/MS data and a FASTA file with anticipated protein sequences. Optionally, the DtaRefinery will utilize the _DeconMSn_log.txt and _profile.txt files along with a concatenated .dta file created by DeconMSn while processing a Thermo raw file. This additional information includes the MS scan at which each parent ion was detected, parent ion intensity, TIC in the Orbitrap or FTICR cell, and the automatic gain control accumulation time for the MS scan. These extra parameters can be utilized as optional explanatory variables for regression analysis to find and correct systematic error dependences. Thus, the use of DeconMSn instead of extract_ms is highly recommended because it not only accurately determines the monoisotopic masses of the parent ions but also provides additional useful information about parent ion measurement.

DtaRefinery analysis settings can be customized by editing the XML file containing processing options or using the GUI of the program. The concatenated .dta file can be produced by either DeconMSn (-XCDTA option) or concatenation of individual .dta files produced by extract_ms (using the Peptide File Extractor). Concatenated .dta files (*.dta.txt) contain each fragmentation spectrum in the .dta format separated by a header line describing the source .dta file (supplemental Fig. 1).

Once spectra are loaded, DtaRefinery internally uses X!Tandem to briefly search the MS/MS data entries from the input file to identify peptides. Note that the MS/MS search does not have to be exhaustive. Capturing the majority of the peptides should be sufficient for recalibration. For example, we suggest searching global trypic digest samples only for fully tryptic peptides and ignoring partially tryptic and non-tryptic peptides. This non-exhaustive search, restricted to the space of fully tryptic peptides, is ~1-2 orders faster than searching for all peptide types yet still captures close to 90% of all identifiable peptides. For the purpose of sample recalibration, those ~10% of semi- and non-tryptic peptide ions can be safely ignored. Another way of speeding MS/MS searches can be achieved by narrowing the parent ion MME tolerance. For example, we use a parent ion MME tolerance for preliminary peptide identification searches in the range of ±20-100 ppm for data sets obtained using high resolution hybrid mass spectrometers. This tolerance range allows for a faster search than the ±3-Da tolerance commonly used for data sets from ion trap instruments. To assess the gain in speed, we timed X!Tandem searches for a mouse brain trypic digest peptide sample that was analyzed using a 30-min-gradient LC separation and an LTQ Orbitrap mass spectrometer. With a ±3-Da parent ion MME tolerance and no enzyme rule setting, the test data set took ~9 h or ~250 ms/spectrum/megabyte of database (on a single thread). With the parent ion MME setting adjusted to ±100 ppm and a requirement that peptides be fully tryptic, the search took ~1 min or 0.5

FIG. 2. Example showing correction of highly pronounced systematic parent ion MME along dimension of scan number parameter. The example is an actual LC-MS/MS analysis on an LTQ Orbitrap instrument that is out of calibration with significant sample overloading. Because of sample overloading, the automatic gain control system was not able to properly modulate the ion population within the Orbitrap cell, resulting in space charge effects causing noticeable systematic MME. However, after applying the DtaRefinery and subtracting the systematic MME components predicted by the regression models trained in the space of all four parameters (scan number, m/z, log10 of ion intensity, and TIC), the mean of the MME distribution shifts from ~16 ppm to approximately 0 ppm, and the standard deviation contracts from 4.3 to 0.8 ppm (data not shown). A, the individual parent ion MME plotted as a function of scan number (blue circles). B, smoothing the MME residuals with Tukey’s running median (yellow circles). C, fitting a spline function into smoothed data to have a continuous function for prediction of systematic MME (red line). D, corrected parent ion MME by subtracting the systematic MME predicted by the model trained using only the scan number parameter.
The data set is analyzed by an MS/MS search engine against an expected list of protein sequences. Spectra with identified peptides go into Table 1, which will be further used for training a regression model predicting systematic MME. Table 2, which is used to store the predicted systematic MME, contains all spectra regardless of assigned peptides. After the model is trained, the parent ion masses in the original data set are corrected based on the predicted systematic MME values that are stored in Table 2 and written into an updated MS/MS data file (“*_FIXED_dta.txt”).

There are no systematic errors. The purpose of Table 2 is to store the predicted systematic MME for all of the MS/MS scan entries during iterative training of the regression model.

The rationale behind the prediction algorithm has been reported in detail in an earlier publication (13). Briefly, the MME residuals are plotted as a function of elution time, m/z, log10 of ion intensity, or other explanatory variables. After visualizing the scatter plots, it is usually apparent whether a systematic error is present or not (Fig. 2A). Although such systematic error trends can be readily modeled with non-parametric regression or scatter plot smoothing techniques, a few concerns must be taken into account. These concerns include the presence of false peptide identifications, the multidimensionality of the problem, overfitting, and finally the computation cost.

Earlier, we suggested modeling the systematic MME using a multidimensional projection pursuit regression (27, 28) (Equation 1).

\[
MME = \sum_{j=1}^{M} \left( \beta_{ij}(m/z) + \beta_{ij}(\text{time}) + \beta_{ij}(\text{intensity}) + \beta_{ij}(\text{TIC}) \right) + \epsilon_{\text{random}} \quad (\text{Eq. 1})
\]
DtaRefinery Tool for Eliminating Errors in MS/MS Data Sets

\[ \text{MME} = \sum_{j=1}^{M_1} g_{4j}(\text{time}) + \sum_{j=1}^{M_2} g_{5j}(m/z) + \sum_{j=1}^{M_3} g_{6j}(\text{intensity}) + \sum_{j=1}^{M_4} g_{7j}(\text{TIC}) + \epsilon_{\text{random}} \]  

(Eq. 2)

where \( \beta \) denotes the coefficients of the optimal projection and \( g \) denotes the non-parametric regression functions. Error residuals are modeled as a hypersurface in space with user selected parameters and MME as dimensions. The regression is iterative; at each step, the model finds one-dimensional projections as linear combinations of the space parameters that best explain the observed data and then subtracts the residuals. Because the MS/MS searching results may contain false peptide identifications, Tukey’s running median (29) (Fig. 2B) smoothing is initially applied to mitigate the effect of outliers. Next, a non-parametric regression technique, such as smoothing splines (30) or LOWESS (31) (Fig. 2C) is applied for modeling the trend and predicting the systematic error values. Finally, the predicted systematic error values are subtracted from the actual error residuals (Fig. 2D). Finding the best one-dimensional projections involves optimization techniques, which are computationally expensive. Hence, in the DtaRefinery tool, for reasons of practicality and the speed of computation, we used a simplified additive model (Equation 2) instead of a more sophisticated projection pursuit regression (Equation 1). This approach avoids having to search for optimal projections and instead performs the regressions along the parameter dimensions. Consequently, it is much less computationally expensive, although it may not capture potential complicated interparameter dependences.

For each iteration, DtaRefinery loops through and estimates the prediction error of the regression models for each of the parameters (supplemental Fig. 2). The prediction error is estimated using one of the standard machine learning approaches, namely K-fold cross-validation as root mean squared error (RMSE) or approximately standard deviation (supplemental Fig. 3). As such, the data are split into K parts. The regression model is trained on \( K - 1 \) parts, and the systematic MME is predicted for the data points in the reserved part. The procedure is repeated \( K \) times to ensure that data points from every part contribute toward the estimated prediction error of the model. After completion of all \( K \) rounds, the RMSE is computed from the MME residual left after subtraction of the predicted systematic MME components. By default, the \( K \)-value is set to 10 but can be changed by the user. If the prediction error is higher than the original RMSE prior to the regression, then the regression fit is not considered successful because the model is distorting the data either by overfitting or some other way. When the iteration is successful, i.e. at least one of the used parameters proved to be a successful explanatory variable, a parameter providing the lowest prediction error is selected, and the corresponding predicted MME residuals are subtracted from the mass error values in both tables. The prediction error not only serves as a criteria for selecting the best explanatory variable but also as an indicator of whether any systematic trend is present that regression can remove. The iterations stop when none of the parameters provide a successful regression model for the round. Thus, although there is no limit to the number of iterations, in our experience it usually takes ~5–20 iterations to converge.

Depending on the user-specified options, the MME residuals for a single iteration can be modeled in one of two ways: either as is or using the overfitting proof mode. Use of the overfitting proof mode may be especially important for sparse data sets or sparse areas of the LC-MS/MS data sets where the danger of interpolation exists. Examples of sparse areas include the flanks of chromatograms and the high and low extremes of \( m/z \) or ion intensity values. Interpolation of the data points artificially removes not only systematic but also random errors, which can lead to undesirable effects. Overfitting can be avoided by making sure that the MS/MS scan entries used to train the regression are not in the subset for which the error is predicted. This can be achieved by using an approach quite similar to K-fold cross-validation. In particular, we split the entries for which the predictions need to be made into \( N + 1 \) partitions. The first partition contains the entries of non-identified spectra that are exclusively present in Table 2 and absent from Table 1. At this stage, overfitting is not an issue because the points used for training and prediction do not overlap. However, overfitting potentially may be an issue for scans that are present in both Table 1 and Table 2. In this case, the identified spectra entries in both tables are split into \( N \) equal partitions. To predict the systematic MME for the entries in each of the \( N \) partitions in Table 2, entries from the corresponding remaining \( N - 1 \) parts from Table 1 are used to train the regression model (Fig. 4). By default, DtaRefinery uses the overfitting proof mode with \( N \) set to 10, which splits the identified MS/MS entries into 10 parts.

As mentioned earlier, the iterations are stopped when further regression using any of the parameters does not improve the MME residuals distribution, i.e. the RMSE after trial regressions is not smaller than the RMSE before the regression attempt. At this point, the estimated contributions of the systematic MME in Table 2 are considered final. DtaRefinery uses these final estimates of the systematic MME to correct the parent ion masses and output the refined MS/MS data to form a new concatenated .dta file. Additionally, DtaRefinery outputs the mass error distribution histograms before and after parent ion mass refinement plus scatter plots that show original and final dependences of MME on parameters selected for building the additive regression model. The mean and S.D. for the MME distribution are estimated in two alternative ways. The first estimate is computed using the expectation maximization approach, which models errors as a mixture of a normal distribution of true identifications and a uniform distribution of false peptide identifications. This approach is more appropriate for data that contain a significant amount of false positive identifications because they can be explicitly modeled as a uniform distribution. Note that the uniform distribution approximates the density of false identifications well enough only within an MME window of 20–30 ppm or less, although this is sufficient for high resolution instruments like the LTQ FT and LTQ Orbitrap. The second approach treats false peptide identifications as outliers and estimates the mean as the median value and the standard deviation using the median absolute deviation. This approach tends to be more robust for cases in which the histogram of true identifications noticeably deviates from a normal distribution. Also note that DtaRefinery can optionally output scatter plots corresponding to regression fits along each of the parameters for each of the iterations.

Sample Preparation and LC-MS/MS Analysis—The tryptic peptides from four different mouse brain regions (cortex, striatum, cerebellum, and the rest of the brain) were prepared and fractionated by strong cation exchange chromatography as described elsewhere (32). Each of 32 fractions and unfractionated samples were analyzed by a liquid chromatography system with ~200 full-width half-maximum peak capacity. The 75-μm-inner diameter × 15-cm-long columns were packed with 3-μm Jupiter C18 particles (Phenomenex, Torrance, CA). The mobile phase solvents consisted of 0.1% formic acid in water (A) and 0.1% formic acid in 90% acetonitrile (B). An exponential 35-min gradient was used for the separation, starting with 100% A and gradually increasing to 60% B over 35 min at a constant pressure of 5,000 p.s.i. The column was coupled with an LTQ Orbitrap (Thermo Fisher Scientific Inc., Waltham, MA) mass spectrometer. The instrument method was as follows: MS survey scan having 100,000 full-width half-maximum resolution and 1 × 10^6 automatic gain control target followed by five MS/MS scans analyzed by the ion trap (~1,000 resolution). All 136 LC-MS/MS data sets are available upon request.
RESULTS AND DISCUSSION

Here, we evaluated the effect of improving parent ion mass measurement accuracy in the context of a “cataloguing” type of proteomics study wherein the final product or result is a non-redundant list of the peptides and proteins observed in a given biological sample. To assess the performance of such proteomics analyses, it is important to estimate two quality metrics: false discovery rate (FDR) and false negative rate (FNR) both for peptide and protein identifications. For the sake of simplicity, we will limit ourselves here with estimation of FDR and FNR for non-redundant peptide identifications only. For this type of study, FDR should be defined as the proportion of non-redundant false peptide entries within the non-redundant peptide identification entries, whereas FNR is the ratio of the number of non-redundant true peptides not passing the threshold criteria required for confident identification to the total number of non-redundant true peptides.

As a demonstration, we used the results from 136 LC-MS/MS analyses of a tryptic digest mouse brain sample pre-fractionated with strong cation exchange chromatography. To refine these data sets, we selected a non-linear, LOWESS-based, additive regression model with the overfitting proof option enabled. Fig. 5 demonstrates an example of a quality control output of the DtaRefinery tool, showing MME residuals as a function of multiple parameters before and after refinement for a typical data set. For this particular LC-MS/MS data set, prediction followed by subtraction of the systematic errors effectively shifted the overall systematic bias from ~2.9 to practically 0 ppm and reduced the standard deviation ~1.5-fold from ~1.0 to ~0.65 ppm as evident from another refinement quality control plot (Fig. 6). Note (supplemental Figs. 4 and 5) that linear recalibration is not as efficient as one that is LOWESS-based because there are still some obvious residual trends in the scan number and m/z domains. The resulting decrease in the standard deviation, although also being quite significant, is not as large. For example, according to Expectation Maximization estimates, the standard deviation decreases from 0.96 to 0.85 ppm in linear recalibration mode and decreases to 0.69 ppm in the case of LOWESS.

As noted earlier, a stricter requirement for the maximum allowable parent ion mass deviation serves to improve discrimination between true and false peptide identifications (8, 11, 12, 18). Thus, we expect that preprocessing the MS/MS data sets by removing the systematic component of the MME of the precursor ions should appreciably improve the quality of peptide identifications. For example, in the demonstrated case (Fig. 6) on the original data set, one needs to use no less than a ±6-ppm threshold to retain most of the true peptide identifications. However, after refinement, the maximum allowable deviation may be decreased to ±2 ppm, providing the potential for an ~3-fold decrease in the FDR of peptide identifications with almost no loss of true identifications. The other benefit of reducing the maximum allowable tolerance for parent ion MME may be a reduction in the FNR by including poorly fragmenting peptides that receive MS/MS fragment match scores below a specified score threshold. These peptides have the potential to be matched by loosening MS/MS match score thresholds and simultaneously applying more stringent requirements for parent ion mass deviation.

To assess the effect of elimination of the systematic MME component, we searched all the above mentioned 136 LC-MS/MS data sets against mouse International Protein Index
database version 3.51 combined with sequences of human keratin proteins and pig trypsin peptides. For assessing the number of true and false identifications after the search is completed, we created a decoy database by concatenating the combined FASTA file with itself but with reversed protein sequences (17). All of the identifications from reversed sequences were considered false, whereas identifications from forward sequences could be either true or false. False identifications were assumed to be distributed equally between matches to forward and reversed sequences. The data sets were preprocessed by DeconMSn to generate the concatenated .dta files. For peptide identification analysis, we used both the original concatenated .dta files as well as files processed by DtaRefinery. The peptide identification was accomplished using the SEQUEST search engine. However, qualitatively, the results and conclusions should stay the same regardless of the MS/MS search engine utilized. The SEQUEST searches were performed both ways: restricting peptide sequences only to the ones that satisfy trypsin cleavage specificity and with no restriction to cleavage specificity. Fig. 7 shows the parent ion MME distribution histogram for all the spectra with assigned peptides for all 136 data sets collated together. Note that this is a rather typical data analysis arrangement, i.e. where all the data sets are processed at once with a fixed maximum allowable parent mass deviation. However, sequentially “one-by-one” processing all the data sets...
and applying individual tolerance criteria would provide better results but would require specialized software (such as DtaRefinery). Without such correction of the systematic errors, one would have to allow up to \(-10\) ppm mass measurement error for identified peptide identifications from about 10 (left, blue) to about 3 (right, green) ppm.

Fig. 7. Mass error distribution histograms for all peptide to spectrum assignments with 2+ charge produced by SEQUEST searches for fully tryptic peptides within 136 LC-MS/MS data sets. No XCorr or \(\Delta Cn\) filtering criteria have been applied. The bin width is 0.5 ppm. The DtaRefinery tool has noticeably reduced the width of the MME distribution histogram and thus maximum allowable deviation of the parent ion mass for true peptide identifications from about 10 (left, blue) to about 3 (right, green) ppm.

Fig. 8. Estimates of maximum number of unique peptides with 2+ charge that can be identified by SEQUEST within allowed FDR values. The green curve represents the results of SEQUEST searches of 136 LC-MS/MS data sets preprocessed by DtaRefinery. The results of non-preprocessed data sets are shown in blue. The maximum number (max. num.) of peptide identifications (ids) was obtained by searching the space of \(\Delta M\), XCorr, and \(\Delta Cn\) parameters within the ranges from 1 to 10 ppm \(\Delta M\) mass (with a 0.1-ppm step size), from 0 to 8 XCorr (0.1 steps), and from 0 to 0.4 \(\Delta Cn\) (0.01 steps), giving 288,000 combinations in total. Note that for any given FDR value the results from refined data set searches provide more true peptide identifications. It is also true that for a given number of true peptide identifications it is always possible to achieve noticeably lower FDR by preprocessing the LC-MS/MS data sets with DtaRefinery.

and reduce the overall spread for the two other types of peptides (supplemental Fig. 6).

Fig. 8 shows the estimate of the maximum number of non-redundant true peptide identifications within 136 data sets that can be achieved with a given allowed maximum FDR. The maximum number of non-redundant peptide identifications within the given FDR was selected by searching for the best combination of \(\Delta M\), XCorr threshold, and \(\Delta Cn\) within the ranges from 1 to 10 ppm \(\Delta M\) mass (with a 0.1-ppm step size), from 0 to 8 XCorr (0.1 steps), and from 0 to 0.4 \(\Delta Cn\) (0.01 steps), giving 288,000 combinations in total. Clearly, in most of the cases and especially in the range of reasonably low FDRs (up to 5%), searches of the refined data sets deliver better results. In other words, for any given number of true peptide identifications, it is possible to achieve significantly lower FDRs. Conversely, for any given FDR limit, improvements in mass measurement accuracy allow us to obtain more peptide identifications. The MS/MS fragmentation spectra of those rescued peptides probably did not contain enough information to meet the elevated MS/MS fragment matching criteria because of physicochemical properties un-
favorable for fragmentation, low intensity, or various other reasons. Table I lists several specific values of true and false peptide identifications shown on Fig. 8. As shown, the average reduction of FDR is about 2-fold. The actual FDR decrease varies from 1.3- to 3-fold and is based on search type and the absolute FDR values. For example, in the case of non-restricted SEQUEST searches, after applying 8.9 ppm, 2.3, and 0.17 for $/H9004M$, $/H9004XCorr$, and $/H9004Cn$ criteria, respectively, it is possible to achieve the maximum number of 13,380 true peptide identifications within the 1% FDR limit. Notably, after refining the data sets and tightening the $/H9004M$ criteria to 2.9 ppm with simultaneous slight relaxation of $/H9004Cn$ to 0.16, the number of false peptide identifications goes down almost 3-fold from 132 to 48 while maintaining the same or even larger number of true identifications.

The DtaRefinery software has been extensively evaluated at our laboratory for processing LC-MS/MS data sets where parent ion mass has been acquired with high accuracy, including the LTQ Orbitrap and LTQ FT instruments, and has consistently reduced overall MME deviations. However, the extent of the mass accuracy improvement varies due to several factors, e.g. the time between instrument calibrations, the amount of sample loaded for analysis, the length of the LC gradient, the automated gain control settings, and complexity of the sample itself. Regardless, the final values for maximum allowable deviation are typically about ±2 ppm for individual data sets obtained using either LTQ Orbitrap or LTQ FT instruments.

To date, we have used DtaRefinery in our proteomics data processing pipeline (Fig. 1) in tandem with DeconMSn prior to X!Tandem, InsPect, or SEQUEST MS/MS search engines. The fact that DtaRefinery utilized a concatenated .dta file as an input and outputs the concatenated .dta file in the same format allows its ready incorporation into various MS/MS data processing pipelines. The concatenated .dta file can be used for input into a search engine without modification or can be split into individual .dta files or converted into MASCOT generic format with currently available tools. Such files can be further converted into a number of other formats if necessary (33). In the future, we foresee other preprocessing approaches, e.g. more sophisticated peak picking in MS/MS spectra, MS/MS spectra recalibration (as implemented in VEMS (34) or in MS-Dictionary (26)), and others put together as additional components in the MS/MS data processing pipeline.

Based on our literature survey of studies utilizing LTQ Orbitrap or LTQ FT instruments for shotgun proteomics to date, the maximum allowable MME deviation is typically set between ±5 and ±10 ppm (e.g. Refs. 35 and 36). DtaRefinery clearly could be useful for these studies as it can generally help reduce the parent ion MME tolerance severalfold (typi-
DtaRefinery Tool for Eliminating Errors in MS/MS Data Sets

DtaRefinery can benefit the most include identification of peptides with post-translational modifications, identification of peptides resulting from nonspecific proteolysis, and searches using exhaustively translated genomes in all six reading frames from stop-to-stop codons as a set of putative protein sequences.

Acknowledgments—Proteomics analyses were performed in the Environmental Molecular Sciences Laboratory, a United States Department of Energy (DOE) national scientific user facility located at the Pacific Northwest National Laboratory (PNNL) in Richland, WA. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

* This work was supported, in whole or in part, by National Institutes of Health Grant RR18522 (to R. D. S.) from the National Center for Research Resources.

[5] The on-line version of this article (available at http://www.mcponline.org) contains supplemental Figs. 1–6.

† To whom correspondence should be addressed: Biological Sciences Division, Pacific Northwest National Laboratory, P. O. Box 999, MSIN: K8-98, Richland, WA 99352. Fax: 509-376-7722; E-mail: rds@pnl.gov.

REFERENCES

1. Wysocki, V. H., Resing, K. A., Zhang, Q., and Cheng, G. (2005) Mass spectrometry of peptides and proteins. Methods 35, 211–222
2. Mayampurath, A. M., Jaitly, N., Purvine, S. O., Monroe, M. E., Auberry, K. J., Adkins, J. N., and Smith, R. D. (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics 24, 1021–1023
3. Yates, J. R., 3rd, Eng, J. K., McCormack, A. L., and Schietz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal. Chem. 67, 1426–1436
4. Fenyö, 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
5. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. J. Proteome Res. 3, 958–964
6. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPeCT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. 77, 2871–2882
7. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567
8. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (1/–10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. 71, 2871–2882
9. Alves, G., Oughtov, A. Y., Kwok, S., Wu, W. W., Wang, G., Shen, R. F., and Yu, Y. K. (2008) Detection of co-eluted peptides using database search methods. Biol. Direct 3, 27
10. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res. 2, 137–146
11. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol. Cell. Proteomics 3, 608–614
12. Zubarev, R., and Mann, M. (2007) On the proper use of mass accuracy in proteomics. Mol. Cell. Proteomics 6, 377–381
13. Petyuk, V. A., Jaitly, N., Moore, R. J., Ding, J., Metz, T. O., Tang, K., Monroe, M. E., Tolmachev, A. V., Adkins, J. N., Belov, M. E., Daubney, A. R., Qian, W. J., Camp, D. G., 2nd, and Smith, R. D. (2008) Elimination of systematic mass measurement errors in liquid chromatography-mass spectrometry based proteomics using regression models and a priori partial knowledge of the sample content. Anal. Chem. 80, 693–706
14. Palmer, M. E., Clench, M. R., Teter, L. W., and Little, D. R. (1999) Exact mass determination precision and accuracy for proteome measurements using an orthogonal acceleration time-of-flight mass spectrometer. Rapid Commun. Mass Spectrom. 13, 256–263
15. Belov, M. E., Zhang, R., Strittmatter, E. F., Prior, D. C., Tang, K., and Smith, R. D. (2003) Automated gain control and internal calibration with external ion accumulation capillary liquid chromatography-electrospray ionization Fourier transform ion cyclotron resonance. Anal. Chem. 75, 4195–4205
16. Henniman, J. M., Bristow, T. W., O’Connor, G., Jarvis, J., and Langley, G. J. (2004) Improved precision and accuracy for high-performance liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry exact mass measurement of small molecules from the simultaneous and controlled introduction of internal calibrants via a second electrospray nebuliser. Rapid. Commun. Mass Spectrom. 18, 3035–3040
17. Haas, W., Faherty, B. K., Gerber, S. A., Elias, J. E., Beausoleil, S. A., Bakalarski, C. E., Li, X., Villen, J., and Gygi, S. P. (2006) Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol. Cell. Proteomics 5, 1326–1337
18. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Honing, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol. Cell. Proteomics 4, 2010–2021
19. Cox, J., and Mann, M. (2008) Computational principles of determining and improving mass determination precision and accuracy for proteome measurements in an Orbitrap. J. Am. Soc. Mass Spectrom. 20, 1477–1485
20. Tolmachev, A. V., Monroe, M. E., Jaitly, N., Petyuk, V. A., Adkins, J. N., and Smith, R. D. (2006) Mass measurement accuracy in analyses of highly complex mixtures based upon multidimensional recalibration. Anal. Chem. 78, 8374–8385
21. Yanofsky, C. M., Bell, A. W., Lesimple, S., Morales, F., Lam, T. T., Blakney, G. T., Marshall, A. G., Carrillo, B., Lekrop, M., Belov, M. E., Dabney, K., Koistinen, D., and Kearney, R. E. (2005) Multicomponent internal recalibration of an LC-FTICR-MS analysis employing a partially characterized complex peptide mixture: systematic and random errors. Anal. Chem. 77, 7246–7254
22. Becker, C. H., Kumar, P., Jones, T., and Lin, H. (2007) Nonparametric mass calibration using hundreds of internal calibrants. Anal. Chem. 79, 1702–1707
23. Komblum, M., Bindschedler, L. V., Gibson, T. M., and Cramer, R. (2006) Automatic internal calibration in liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry of protein digests. Rapid Commun. Mass Spectrom. 20, 3076–3080
24. Zhang, J., Ma, J., Dou, L., Wu, S., Qian, X., Xie, H., Zhu, Y., and He, F. (2009) Mass measurement errors of Fourier-transform mass spectrometry (FTMS); distribution, recalibration, and application. J. Proteome Res. 8, 645–659
25. Dandekar, R. M., Ouny-Patat, S. A., Scarlett, C. O., Speir, J. P., and Borchers, C. H. (2008) Data Self-Recalibration and Mixture Mass Fingerprint Searching (DASER-MMF) to enhance protein identification within complex mixtures. J. Am. Soc. Mass Spectrom. 19, 1914–1925
26. Shin, B., Jung, H. J., Hyung, S. W., Kim, H., Lee, D., Lee, C., Yu, M. H., and Lee, S. W. (2008) Postexperiment monotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometry data increases accuracy of protein identification in LC/MS/MS. Mol. Cell. Proteomics 7, 1124–1134
27. Friedman, J. H., and Stuetzle, W. (1981) Projection Pursuit Regression. J. Am. Stat. Assoc. 76, 817–823
28. Härdle, W. (1990) Applied Nonparametric Regression, Cambridge University Press, Cambridge, UK, 425–430
29. Tukey, J. W. (1977) Exploratory Data Analysis, Addison-Wesley Publishing Co., Reading, MA
30. Hastie, T., Tibshirani, R., and Friedman, J. H. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 127–134
31. Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74, 829–836
32. Wang, H., Qian, W. J., Chin, M. H., Petyuk, V. A., Barry, R. C., Liu, T., Gritsenko, M. A., Mottaz, H. M., Moore, R. J., Camp, D. G., 2nd, Khan, A. H., Smith, D. J., and Smith, R. D. (2006) Characterization of the mouse brain proteome using global proteomic analysis complemented with cysteinyl-peptide enrichment. J. Proteome Res. 5, 361–369
33. Falkner, J. A., Falkner, J. W., and Andrews, P. C. (2007) ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats. Bioinformatics 23, 262–263
34. Matthiesen, R., Trelle, M. B., Hejrup, P., Bunkenborg, J., and Jensen, O. N. (2005) VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J. Proteome Res. 4, 2338–2347
35. Zanivan, S., Gnädig, F., Wickström, S. A., Geiger, T., Macek, B., Cox, J., Fassler, R., and Mann, M. (2008) Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. J. Proteome Res. 7, 5314–5326
36. Ballif, B. A., Carey, G. R., Sunyaev, S. R., and Gygi, S. P. (2008) Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain. J. Proteome Res. 7, 311–318