Published in final edited form as:

Proteomics. 2016 September ; 16(18): 2480–2490. doi:10.1002/pmic.201600137.

ProteinProcessor: A Probabilistic Analysis Using Mass Accuracy and the MS Spectrum

Jonathan A. Epstein¹, Paul S. Blank², Brian C. Searle³, Aaron D. Catlin¹, Stephanie M. Cologna⁴, Matthew T. Olson⁵, Peter S. Backlund¹, Jens R. Coorsen⁶, and Alfred L. Yergey¹

¹Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Bethesda, MD

²Section on Integrative Biophysics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Bethesda, MD

³Proteome Software Inc., Portland, OR

⁴Department of Chemistry, University of Illinois at Chicago, Chicago, IL

⁵Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD

⁶Faculty of Graduate Studies and the Departments of Health Sciences and Biological Sciences, Brock University, St. Catharines, ON, Canada

Abstract

Current approaches to protein identification rely heavily on database matching of fragmentation spectra or precursor peptide ions. We have developed a method for MALDI TOF-TOF instrumentation that uses peptide masses and their measurement errors to confirm protein identifications from a first pass MS/MS database search. The method uses MS1-level spectral data that have heretofore been ignored by most search engines. This approach uses the distribution of mass errors of peptide matches in the MS1 spectrum to develop a probability model that is independent of the MS/MS database search identifications. Peptide mass matches can come from both precursor ions that have been fragmented as well as those that are tentatively identified by accurate mass alone. This additional corroboration enables us to confirm protein identifications to MS/MS-based scores that are otherwise considered to be only of moderate quality. Straightforward and easily applicable to current proteomic analyses, this tool termed ‘ProteinProcessor’ provides a robust and invaluable addition to current protein identification tools.

INTRODUCTION

Bottom up protein identification, in the so-called ‘shotgun’ approach, is based on mass spectrometric production of fragmentation spectra from peptides that have been separated chromatographically [1]. Identification is currently implemented by searching databases consisting of proteins digested *in silico* as well as their fragmented peptides; several different algorithms can be used to assign peptide sequences to identify proteins. One

example is with MALDI-based experiments using the Mascot search algorithm. [2]. Similarly, if one defines top-down protein identification as the identification of intact proteins, these approaches generally rely on protein digests, and often follow high-resolution, two-dimensional gel electrophoresis (2DE) in order to preserve biologically critical information about the intact proteins (i.e. MW, pI, and post-translational processing). When using MALDI-based experimental platforms, such digests are typically analyzed in a single target well. The quality of the MALDI fragmentation spectra is generally such that only a small number of these spectra yield high quality Mascot identifications regardless of the fact that peptide signals arising from a single or small numbers of proteins are contained within that spot. If a protocol uses only the fragmentation spectra and the corresponding precursor masses as the basis of searching, then much information in the first-order (MS1) spectra is unused. We have developed a procedure that capitalizes on these unused data in the first-order spectra by performing a peptide mass fingerprint-like search of candidate proteins found by the fragmentation spectra-based search.

The initial approach to protein identification using mass spectrometry was Peptide Mass Fingerprinting (PMF) [3–5] in which peptide mass values in the first order spectrum were matched in a database search. In a PMF experiment, ion mass accuracy is the critical piece of information used in scoring. Known for almost 20 years, the results of database searching using PMF are frequently insufficiently robust to meet the rigorous demands of proteomics experiments due to homology of proteins and functional domains, and the lack of sequence specific information. PMF was thus supplanted by various approaches employing fragmentation spectra beginning with peptide mass tags [4] and database searching [6]. Barring methods that apply extensive re-acquisition for iterative peptide confirmation [7], each of these approaches uses only a portion of all the data available from a mass spectrometry experiment. That is, PMF uses only accurate masses of unfragmented peptides, whereas fragmentation approaches, while using accurate masses of peptide precursors, do not necessarily use all of the available peptide precursors or their equivalent, unmatched fragmented peptides.

To make optimal use of the heretofore unused first-order spectral information we have developed an analytical procedure that automatically performs (i) a traditional fragmentation database search to generate candidate protein identifications followed by a (ii) peptide mass fingerprint-like search of candidate proteins to increase the confidence of the protein assignment using additional validation. We call the cornerstone for this analysis “ProteinProcessor”. It uses the peak list derived from a first-order MALDI spectrum, coupled with a single candidate protein “C” the putative identity of which was obtained by other means (e.g. via MS/MS and an ensuing Mascot search). ProteinProcessor thus reports which peptides in the first-order spectrum also appear in the candidate protein. In principle the concepts of ProteinProcessor should also be applicable to LC-coupled electrospray ionization MS/MS using data-dependent acquisition as well but this remains to be demonstrated.

Statement of significance

This targeted peptide mass fingerprinting (PMF) approach results in an improvement in protein identification because a larger fraction of the total information available in a proteomics dataset is used. The targeted PMF method enables *validation and re-ranking* of identifications provided by typical MS/MS database searches. The combination of MS/MS and targeted PMF helps to stratify database search identifications with modest scores. More importantly, these orthogonal analytical approaches provide a mechanism to substantiate or refute many protein identifications based only on one or two peptides. Thus, ProteinProcessor increases the confidence of lower confidence MS/MS-only protein identifications. The concepts in ProteinProcessor are also applicable to LC-coupled electrospray ionization MS/MS.

METHODS

Materials

Analytical grade reagents were used as received unless otherwise stated. Gel electrophoresis materials and chemicals were obtained from Bio-Rad Laboratories (Hercules, CA) or Invitrogen (Carlsbad, CA). HPLC grade acetonitrile, ACN, and methanol were purchased from Sigma-Aldrich (St. Louis, MO). Sequencing grade trypsin was obtained from Promega Corporation (Madison, WI). The Universal Proteomics Standard mix was obtained from Sigma-Aldrich (St. Louis, MO).

Gel Excision, Protein Digestion and Mass Spectrometry Analysis

Protein spots were excised from Coomassie-stained preparative 2D-gels and digested with trypsin using standard protocols as previously described [8]. Briefly, gel slices were de-stained and then dehydrated in a vacuum centrifuge. A total of 200 ng of sequencing grade trypsin was added to the excised gel fragment, covered with 100 mM ammonium bicarbonate and allowed to incubate at 37° C overnight. The supernatant was collected and further extraction was performed by addition of 5% formic acid/ACN (1:1, v/v) followed by a second extraction step of 5% formic acid/ACN (5:95, v/v). The combined solvent extracts were vacuum dried and reconstituted in 0.1% TFA followed by concentration and desalting using C₁₈ ZipTips (Millipore). The resulting peptides from the digest were analyzed by MALDI-TOF/TOF using an Applied Biosystems 4800 Proteomics Analyzer. Sample preparation for analysis was as follows: tryptic peptides were mixed 1:1 (v/v) with α -cyano-4-hydroxycinnamic acid (5 mg mL⁻¹ in ACN/0.1% TFA) and spotted directly onto the MALDI target without chromatographic separation. MS spectra were calibrated using two internal standards and tryptic autolytic peaks. Typically, each MS1 spectrum was acquired in the positive-ion reflectron mode using 400 laser shots. A set of 6–12 ions, depending on the complexity of the MS1 spectrum, which had both 1) strong intensities and 2) were well resolved from neighboring ions were analyzed further. The resolution criterion, 2 above, is important in order that the timed-ion selector of the instrument unambiguously separates the post-source decay fragmentation (PSD) of the selected ion from all others. Fragmentation spectra were acquired as PSD unimolecular decompositions (collision gas off) using 1000 laser shots. MS/MS data were extracted to create .mgf files used for database searching. Figure 1 illustrates the approach used by ProteinProcessor to yield increased

protein sequence coverage. Figure 1a shows an MS1 spectrum along with the two peptides that were fragmented to yield the protein identified by a Mascot search; the identity of these two peptides are shown along with their position in the full sequence of the protein. Figure 1b illustrates how ProteinProcessor enhances protein sequence coverage following the MS1 spectrum search as described below.

Database searching was performed using either the *Mus musculus* or *Homo sapiens* subset of the SwissProt database by submitting the MS/MS spectra to Mascot v2.2 (Matrix Science); protein samples are described below. Database searching parameters were: enzyme – trypsin, one missed cleavage, fixed modification – carbamidomethylation (C), variable modification – oxidation (M), precursor mass tolerance of ± 0.15 Da and fragment ion mass tolerance of ± 0.06 Da. Only peptides with individual ion Mowse scores of ≥ 32 were used for protein identification.

Protein Samples

The proteins used to develop a statistical model of the Bayesian prior probability distribution were part of a data set arising from a study employing a murine model for Niemann-Pick Disease, Type C [9] in which homogenates of cerebellar tissue from wild-type and mutant mice were resolved by 2D gel electrophoresis. Mass errors of proteins found in 84 MALDI target wells, corresponding to 84 unique gel spots, were used to generate the distribution of “true” masses; in aggregate, 962 peptide mass errors of peptides matched to proteins in these wells were used. This is the “Model” data set. A complete description of how these errors were employed is given below. Proteins used to test the model generated using these mass errors, i.e., the “Ground Truth” data set, were obtained as a mixture of proteins termed the Universal Proteomics Standard based on the 2007 ABRF PRG Study [10]. The mixture was solubilized and aliquots were resolved in duplicate using a 1D gel electrophoresis system, fractioned into 19 separate slices, and processed as described above. Proteins identified using Mascot searches were then processed for peptide mass matches and used as the starting point for generation of decoys as described below.

Data Analysis Following Mascot Searches

A. Matching Masses in the First Order Spectrum—Peak lists for both MS and MS/MS spectra and the generation of Mascot generic format files was migrated from the 4800 Oracle database to a desktop computer for further analysis using the T2d Extractor tool (Courtesy of Phillip C. Andrews, University of Michigan). A list of proteins for each sample well is identified by the Mascot searches. Each of these proteins is digested *in silico* considering two modifications: methionine oxidation and iodoacetamide alkylation of cysteine. These modifications were chosen based on sample preparation considerations but could easily be generalized to support a more targeted Proteomics approach.

The masses of the internal standards from the peaks of first-order spectra, along with all peaks with a signal to noise ratio less than 8:1 (as determined by ABI 4800 internal software) are filtered from the peak lists. The ions remaining in the peak list are normalized to the intensity of the strongest remaining peptide signal. The mass list from the filtered spectrum is then matched against the masses predicted from the *in silico* digestion, and

errors relative to the theoretical masses are computed in PPM. Each peptide peak has a corresponding mass error that we define as the signed quantity (observed peak mass – theoretical peptide mass). We observed that a protein detected by the MS/MS database search will typically have additional peptide matches to the MS1 spectrum and that the corresponding mass errors will be relatively small, i.e., on average absolute values less than 10 ppm. Furthermore, a poorly matched or randomly selected protein will tend to have few or no peptide matches, and the mass discrepancies of those matches will tend to be large, on average greater than 20 ppm. A report is generated showing the matches, the relative intensities of the matches, the mass errors, the amino acid sequence of the matched ions and whether the match includes a modification, and the presence of up to one missed cleavage. Finally, the fraction of total ion current and percentage of sequence are reported, and a list of all unmatched masses is appended. Typical results from this process are shown in Figure 1b. This process is performed iteratively for all the proteins identified by the MS/MS analysis. Supplementary Figure S1 shows a sample output from this software for a single input protein and reflector spectrum.

B. Generating Distributions and Bayesian Prior Probabilities—Within the Bayesian framework, the posterior probability is proportional to the prior probability via the likelihood function:

$$\text{Posterior} \propto \text{Likelihood} * \text{Prior}$$

The prior probability can be either ‘non-informative’ with a shape relatively uniform compared to the likelihood function, or ‘informative’ with a shape determined from prior information that when multiplied by the likelihood function leads to a posterior distribution that is different from the likelihood function. ProteinProcessor uses existing data to determine ‘informative’ priors and likelihood functions to evaluate the probability of a peptide match given an observed peptide mass error.

Peptides from the proteins identified by Mascot from the 84 samples of the Niemann-Pick dataset were used to establish a baseline for modeling the mass errors for a “true” distribution. The distribution of mass errors was based upon Mascot hits that were obtained from fragment ion spectra of high intensity and extensive sequence information coupled with the mass errors of the fragmented peptides along with those masses matched using the ProteinProcessor algorithm. The latter used only those peptides that had relative intensities 8% of the renormalized intensities, a level chosen by experience to optimize signal to noise. In theory, a uniform distribution of mass errors would be expected. However, the observed distributions were uni-modal due to the elemental composition of the peptides and their mass defect. We used a Gaussian function to characterize the shape of the two observed distributions. A Gaussian distribution was fit to these mass errors of “true peptides” yielding a “true Distribution” with mean and standard deviation, $\mu_T=6.8$ ppm and $\sigma_T=11.2$ ppm, respectively. A corresponding “false Distribution” was derived using the matches of peptides generated from proteins of similar mass to the original data. That is, for each member of the original Niemann-Pick dataset, 400 proteins of similar mass in the SwissProt database were selected. These sequences were then scrambled and digested *in silico* and “matched” to the

original peak lists, and used to model the “false Distribution” of mass errors. These data were fit to a Gaussian distribution with mean and standard deviation, $\mu_F = -1.3$ ppm and $\sigma_F = 36.9$ ppm, respectively.

We used this same Niemann-Pick dataset to estimate the Bayesian prior probability of true, $p(t)$. We computed $p(t)$ as the ratio of total observed matched peptides for all 84 samples, divided by the total number of *all* masses in the dataset, matched and unmatched. This yielded an estimate of $p(t) = 0.145$, and a corresponding prior probability of false, $p(f) = 1 - p(t) = 0.855$. While the exact prior probability estimates $p(t)$ and $p(f)$ may vary between different data sets, when implementing ProteinProcessor their values will represent ‘informative’ estimates based upon existing data.

C. Calculation of Peptide Probabilities and Protein Scores—Bayes’ Theorem was used to estimate the probability of a given peptide mass error being “true”:

$$p(t|D_i) = \frac{p(D_i|t)p(t)}{p(D_i|t)p(t) + p(D_i|f)p(f)} \quad (1)$$

where:

- $p(t|D_i)$ is the probability of a “true” peptide match given the observed peptide mass error
- $p(D_i|t)$ is the probability that this mass error fits the pre-computed “true” distribution
- $p(D_i|f)$ is the probability that this mass error fits the pre-computed “false” distribution
- $p(t)$ and $p(f)$ are the Bayesian priors described above

A Protein Score (PS) is the sum of all peptide probabilities for a given protein found by Mascot:

$$PS = \sum_{i=1}^k p(t|D_i) \quad (2)$$

Defining our metric using a summation over peptide probabilities is analogous to previously defined measures [12]. This portion of the calculations and those for the following sections were carried out using a series of Perl scripts.

D. Generation of Decoy Database for Matched Proteins—The reliability of the PS for each protein candidate was evaluated against a set of decoy proteins chosen to have masses close to that candidate, but that were not “too similar”. A decoy database of 10,000 proteins was devised for each protein candidate using the following scheme: 1. Retrieve a few hundred proteins of mass close to the candidate (± 20 Da) using Entrez; 2. Remove proteins having a strong sequence similarity to the candidate by filtering with BLAST; 3.

Randomly select 25 of the remaining proteins and generate 400 decoys for each by randomizing their sequences while maintaining Arg and Lys residues in their original positions. This latter step yielded 400 decoys with peptides of identical numbers of residues, but potentially different masses. Removing proteins that had a strong sequence similarity (homologs) to the candidate protein minimized both the possibility of biasing the results by 1) reducing the number of independent proteins with masses close to the candidate that are effectively “background” and 2) recapitulating searches based on randomized sequences of the candidate protein alone – the extreme case of sequence similarity. ProteinProcessor was run for both the candidate protein and each of the 10,000 decoys against the original reflector peak list, and the peptides matched were used to calculate protein scores for each candidate. A copy of the source code used for this study is available by request from the authors.

RESULTS & DISCUSSION

Our approach combines features of both PMF and peptide fragmentation approaches to mass spectrometry-based protein identifications. The combined approach notably enables critical cross-validation of the output of fragmentation-only (i.e. shotgun) analyses. That is, we use the results of the fragmentation spectra database search and employ them in a *targeted* PMF, that is a search on a list of candidate proteins detected from a search based on fragmentation spectra, and thus we specifically search for peptides derived only from proteins identified in the database search of fragmentation spectra. This iterative, bidirectional approach uses the full complement of available data from each MS analysis, thereby enabling the best possible protein identifications based on maximal sequence coverage. The present approach differs from seemingly similar alternatives such as the Mascot Error Tolerant Search. The latter is based on refining results from a search of fragmentation spectra only and cannot be used with PMF. Post-translational modifications are currently included in this approach only to the extent of oxidized methionine or intentionally modified cysteine. More specific targeting of particular post-translational modifications will be the subject of future work.

We used the Model data set of 962 peptides from mutant and wild-type mice to calculate the mass error distributions for these peptides. This distribution is shown in Figure 2 along with the distribution of mass errors for the decoy proteins generated for these same data. Interestingly, we observe a positive mean for the true distribution. The marked differences in the width of these two distributions is an expected characteristic of the data used to produce them; one expects to see a broader distribution for decoy data than for true. The reason for the positive mean is most likely due to an instrumental bias that appears to be invariant with time. While ProteinProcessor implicitly accounts for mass biases, it is important to recalibrate the tool each time the instrument has been serviced. These types of biases serve as a cautionary marker for other investigators and demonstrate the need to routinely characterize the instrument response.

If no prior information or expectation exists, the value for the Bayesian prior true, $p(t)$, can be set to one of several classes of “non-informative” priors. However, and most importantly, we are able to calculate an ‘informative’ prior from first principles using a known, validated data set, the Model data set; the priors used in this study are based upon observations of

existing data. This again is an important point to consider when undertaking such estimations in that it is preferable to use actual data to calculate a prior probability distribution rather than assuming an arbitrary or ‘non-informative’ one.

The results from Mascot searches of the spectra from the duplicate gel lanes of the Ground Truth Dataset are summarized in Table 1. In summary, a total of 33 unique proteins were found, most of them in both replicates with 6 of them in only a single replicate. There were no false positive identifications; that is, no proteins were found that were not expected to be present. These results are entirely consistent with the fraction of proteins found by this laboratory in the 2007 ABRF study [11]. The proteins found are listed in Table 1 using a coding that corresponds to the MALDI sample well from which they were identified. For example, three proteins were found in the well A01, while only one protein was found in well A03; since there were 18 wells used for each replicate lane of the 1D gel, the fact that there is no listing for well A02 means that no protein was identified from that well. When multiple proteins were identified from a single well, they are listed in the Table in order of decreasing Mascot scores. Note that some proteins were identified in multiple wells of the same lane, e.g., Glutathione S-transferase A1 in wells B10 and B11 as well as Glutathione S-transferase P in wells B11 and B12; likely owing to the reduced precision of band excision, homology of protein families (i.e. isoforms, splice variants, and so forth) or differences in post-translational and processing modifications.

Figure 3a shows PS values for the Ground Truth Dataset along with box and whisker plots for the PS of the decoys for each true protein; up to 10 peptide mass errors were used to produce this Figure. All of the PS values for true proteins are well separated from the distributions of their respective decoys PS values as indicated by the blue diamonds in the Figure. Strikingly, there exists a fraction of decoy scores that have median values of zero. Thus, even though a PS for any given protein may have a low value, it can be differentiated clearly from its corresponding decoys. This aspect of the plot is summarized in Figure 3b showing the percentage of all decoy scores that have values that exceed the PS for a true protein. For the 60 true proteins, less than 1% of them have decoy scores that exceed the true protein PS values. PS values for the true proteins as well as the mean ($\mu_{\text{Decoy PS}}$) and standard deviation ($\sigma_{\text{Decoy PS}}$) of the PS values for the 10,000 decoy proteins are summarized in Table 2. Proteins for which only a single peptide was matched by ProteinProcessor are noted by an ‘*’ in the last column; there are four of these. The 6th column of the Table shows the Z-score of the true protein PS values ($\text{PS}_{\text{True Protein}} - \mu_{\text{Decoy PS}} / \sigma_{\text{Decoy PS}}$) using the distribution properties ($\mu_{\text{Decoy PS}}$ and $\sigma_{\text{Decoy PS}}$) of the decoy scores; this score is a measure of the degree to which the PS of a true protein differs from the mean of the decoy PS in units of $\sigma_{\text{Decoy PS}}$. Table 2 also shows that for the four proteins that had only a single peptide matched by ProteinProcessor, three of them yielded Z-scores that are more than 3 standard deviations from the mean; the sole exception is Glutathione S-transferase P in well B12. This can also be seen by inspection of these four examples in Figure 3a. The z-score analysis assumes a normal distribution that is only approximate for decoy distributions populated with identically zero PS. Furthermore, the probabilities associated with large z-scores, PS values located in the tails of the normal distribution, may be best calculated using an extreme value analysis. However, if all that is required is a relative ordering comparison, then the simple Z-score analysis may be

sufficient. Fitting the tails of these decoy distributions using statistical modeling of extreme values is ongoing. The Ground Truth dataset encompasses proteins ranging in molecular weight from <15 to >75 kD, and is applicable to a wide range of proteins. Specifically, at both low and high molecular weights the method added increased sequence coverage to what was obtained from MS2 experiments alone.

While Figure 3 is generated from the sum of peptide probabilities, calculated using Equation 2, an alternative approach would use the product of these peptide scores. That is:

$$P=1-\prod_{i=1}^k(1-p(t|D_i)) \quad (3)$$

We have explored the use of this approach and have found that summing yields a more robust result despite intuition that would suggest using peptide score products (Equation 3). However, using the peptide score products, we observed that the score reaches a plateau as a function of the number of peptides matched by ProteinProcessor in the reflector spectrum peak list for all proteins found in the Ground Truth dataset. That is, the plateau in the product score observed when matching more than five or six peptide masses using our targeted PMF approach does little to alter the product score of a true protein. This suggests that there is a limit and optimal amount of additional information that can be extracted using the targeted PMF approach in a peptide score product. Independent of the peptide score used, it is clear that “one hit wonders” no matter what their origin, either MS/MS database searching or targeted PMF, are of little value in yielding confident protein identifications. That said, since ProteinProcessor assigns Bayesian probabilities to peaks identified from the MS1 spectrum, these peptide probabilities could be combined using a more robust ProteinProphet-like protein probability model [12]. Indeed, if MS/MS peptide identifications were assessed with other Bayesian probability generating tools, such as PeptideProphet [13] or Percolator [14], one could conceivably combine MS/MS and MS1-only peptide identifications into a single protein probability. We are investigating this extension of ProteinProcessor in a future study.

We considered the degree to which ProteinProcessor and the MS1 spectrum can guide the ranking of candidate Mascot hits. In Table 3, we show that 12 of the 26 total samples for which either the PS Z-score yields a different ranking than that provided by Mascot. Note in particular that the proteins found in Well A12 are both Mascot “one-hit wonders” for which not only is the Z-scores ordering the reverse of that found in Mascot, but also the peptides matched by ProteinProcessor lead to confident identifications. Similarly, sample A04 shows a different ordering where none of the proteins found by Mascot are one-hit wonders. By utilizing additional information present in an MS dataset, Mascot and similar identifications can be further validated using complementary hypothesis tests derived from an extreme value analysis of decoy distributions.

CONCLUSIONS AND FUTURE DIRECTIONS

This targeted PMF approach is used to *validate and re-rank* the identifications made in MS/MS database searches. We have found that the combined approach helps to stratify identifications with modest scores based on only database searches of peptide sequences. Perhaps even more importantly, these complimentary analytical approaches provide a mechanism to substantiate or refute many protein identifications currently classified as ‘one-hit wonders’ (i.e. based only on one or two peptides). ProteinProcessor allows elevation of lower confidence MS2-only protein identifications to give a more confident identification. These improvements in protein identification are a consequence of using a larger fraction of the information present in the total dataset available to the experimentalist. Furthermore, as ProteinProcessor reports the mass errors between the observed peaks and the theoretical peptide peaks, we also use these peptide mass matches and associated errors to develop a model of protein identification that can be used to substantiate the fragmentation spectra database search.

We have demonstrated the advantages of using additional information present in MS data sets to further validate and re-rank MS/MS database search identifications. A concern about possible routine use of this approach is whether or not an MS/MS identified candidate that is incorrect, for example, a member of a decoy data set generated during the search, would be elevated to a validated candidate by ProteinProcessor thereby generating a false discovery. In the context of the standard protein mixture, i.e., the “Ground Truth Data Set”, even though we did not identify all of the proteins known to be present, every protein that was identified was a known member of the mixture. This indicates an effective false discovery rate of 0%. Developing a measure of the false discovery rate may be possible by looking at a large set of ‘one-hit wonders’ for which database identification is made on a very limited number of proteins.

The concepts in ProteinProcessor are applicable to LC-coupled electrospray ionization MS/MS using data-dependent acquisition as well. Here one would need to consider the pool of MS1 peaks across an LC gradient to confirm protein identifications made using traditional MS/MS database searching. While the confidence in any individual MS1-only match would be relatively lower due to multiple testing issues from having a larger search space, as well as far more peaks, it is likely that additional peptides implied by MS1 peaks not selected for MS/MS sequencing will improve confidence in marginal protein identifications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been supported by the intramural program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services.

References

1. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*. 2001; 19:242–247.
2. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*. 1999; 20:3551–3567. [PubMed: 10612281]
3. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. *Proc Natl Acad Sci U S A*. 1993; 50:5011–5015.
4. Mann M, Højrup P, Roepstorff P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*. 1993; 22:338–345. [PubMed: 8329463]
5. Pappin D, Højrup P, Bleasby A. Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Current Biology*. 1993; 3:327–332. [PubMed: 15335725]
6. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*. 1994; 5:976–989. [PubMed: 24226387]
7. Pascal BD, West GM, Scharager-Tapia C, Flefil R, Moroni T, Martinez-Acedo P, et al. Software Analysis of Uncorrelated MS1 Peaks for Discovery of Post-Translational Modifications. *Journal of the American Society for Mass Spectrometry*. 2015; 25:2133–2140.
8. Jiang XS, Backlund PS, Wassif CA, Yerger AL, Porter FD. Quantitative proteomics analysis of inborn errors of cholesterol synthesis: identification of altered metabolic pathways in DHCR7 and SC5D deficiency. *Molecular and Cellular Proteomics*. 2010; 9:1461–1475.
9. Cologna SM, Jiang XS, Backlund PS, Cluzeau C, Dail MK, Yanjanin NM, et al. A Quantitative Proteomic Analysis in Niemann-Pick Disease, type C1 Reveals Protein Markers and Pathological Insight. *PLoS Genetics*. 2012; 7:e47845.
10. Andrews, PC.; Arnott, P.; Gawinowicz, MA.; Kowalak, JA.; Lane, WS.; Lilley, KS.; Martin, TL.; Stein, SE. ABRF-sPRG2006 Study: A Proteomics Standard. 2006. <http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPsters/ABRFsPRGStudy2006poster.pdf>
11. Turck CW, Falick AM, Kowalak JA, Lane WS, Lilley KS, Phinney BS, et al. The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 Study. *Molecular and Cellular Proteomics*. 2007; 6:1291–1298. [PubMed: 17513294]
12. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry*. 2003; 75:4646–4658. [PubMed: 14632076]
13. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry*. 2002; 74:5383–5392. [PubMed: 12403597]
14. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*. 2007; 4:923–925. [PubMed: 17952086]

Figure 1a

Mascot Results

m/z	Sequence	Δm (Da)	Score
1081.565	YDLTVPFAR	-0.0022	54
991.550	IFSIVEQR	-0.0067	61

Candidate Protein Sequence
(Ions Fragmented shown in blue)

1 MAERAAL EEL VKLQGERVRG LKQKASAE L IEEVAKLLK LKAQLGPDES
 51 KQKFLV LKTPK GTRDYSPRQM AVREKVFDVI IRCFKRHGAE VIDTPVFELK
 101 ETLMGKYGED SKLIYDLKDQ GGELLSLR **YDLTVPFAR** LA MNKLTNIKRY
 151 HIAKVYRRDN PAMTRGRYRE FYQCDFDIAG NFDPMIPDAE CLKIMCEILS
 201 SLQIGDFLVK VNDRRILDGM FAICGVSDSK FRTICSSVDK LDKVSWEEVK
 251 NEMVGEKGLA PEVADRIGDY VQQHGGVSLV EQLLQDPKLS QNKQALEGLG
 301 DLKLLFEYLT LFGIDDKISF DLSLARGLDY YTGVIYEAVL LQTPAQAGEE
 351 PLGVGSVAAG GRYDGLVGMF DPKGRKVPCV GLSIGVER **IFSIVEQR** LEAL
 401 EEKIRTTETQ VLVASAQKKL LEERLKLVE LWDAGIKAEL LYKKNPKLLN
 451 QLQYCEEAGI PLVAIGEQE LKDGVIKLR VTSREEVDVR REDLVEEIKR
 501 RTGQPLCIC

Figure 1b.

ProteinProcessor Results
(Candidate Protein Sequence Coverage)

```

1  MAERAALEEL VKLQGERVRG LKQQKASAEI IEEVAKLLK LKAQLGPDES
51  KQKFVLKTPK GTRDYSPRQM AVREKVFQVI IRCFKRHGAE VIDTPVFELK
101 ETLMGKYGED SKLIYDLK DQ GGELLSLRYD LTVPFARYLA MNKLTNIKRY
151 HIAKVYRRDN PAMTRGRYRE FYQCDFDIAG NFDPMIPDAE CLKIMCEILS
201 SLQIGDFLVK VNDRRILDGM FAICGVSDSK FRTICSSVDK LDKVSWEVK
251 NEMVGEKGLA PEVADRIGDY VQQHGGVSLV EQLLQDPKLS QNKQALEGLG
301 DLKLLFEYLT LFGIDDKISF DLSLARGLDY YTGVIYEAVL LQTPAQAGEE
351 PLGVGSVAAG GRYDGLVGMF DPKGRKVPCV GLSIGVERIF SIVEQRLEAL
401 EEKIRTETQ VLVASAQKKL LEERLKLVE LWDAGIKAEL LYKKNPKLLN
451 QLQYCEEAGI PLVAIGEQE LKDGVIKLS VTSREEVDVR REDLVEEIKR
501 RTGQPLCIC

```

Figure 1.

Schematic representing the approach used by ProteinProcessor.

a. A typical MS1 spectrum of a given well showing tryptic autolysis product ions, labeled 'X'. In addition, the two peptide ions that were fragmented to yield the protein found by a Mascot search with a score of 115, labeled "‡" (in blue); the identity of those two peptides are at their position in the full sequence of the protein (bold italic blue letters). Finally, the additional peptides detected using ProteinProcessor are noted with a red *. b. Results of searching the MS1 spectrum of Figure 1a for peptides using the ProteinProcessor methodology with sequence locations shown in bold italic red letters. *Note: Ultimately three proteins were detected in the MS1 spectrum of Figure 1a. Details of the other two proteins are not shown, but their presence accounts for the peaks not otherwise explained in the Figure.*

True and False Fitted Distributions of Mass Errors

Figure 2.

True and False distribution of mass errors. The true distribution is derived from a dataset of mouse proteins and the false distribution is derived using decoys for each of the true proteins. Each distribution is described by a Gaussian with respective mean, μ , and standard deviation, σ ($N(\mu, \sigma^2)$).

Figure3a.

Figure 3b.

Figure 3.

a. Validation of ground truth dataset scoring 56 (Protein Candidate, MS1) pairs versus decoy databases run against the same MS1. All 56 ground truth Protein Scores (blue diamonds) are separated from the decoy data's Protein Scores by a Z-score of at least 1.98 or two standard deviations from the mean. The central bars represent median values, the extremes of the boxes are 25th and 75th percentiles of the decoy scores and the whiskers adjacent values of 1.5-fold greater than the 75th percentile and of 1.5-fold less than the 25th percentile; for the sake of clarity, outliers, i.e., scores beyond the adjacent values, are not shown. b. The percentage of all decoy scores that have values that exceed the true protein score is <1% with all but one, <0.2%.

Table 1

Ground Truth Dataset: Proteins Detected

CODE	Protein Name and Accession Number	Accession	Mascot Score	PROTEIN LENGTH (Number of AAs)
A01a	Serotransferrin	P02787	168	698
A01b	Lactotransferrin	P02788	77	710
A01c	Gelsolin	P06396	60	782
A03	Serum albumin	P02768	325	609
A04a	Catalase	P04040	244	527
A04b	Antithrombin-III	P01008	154	464
A04c	Histidyl-tRNA synthetase, cyto	P12081	90	509
A06	Creatine kinase M-type	P06732	101	381
A07	Small ubiquitin-related modifier 1	P63165	81	101
A08	Annexin	P08758	177	320
A09a	NAD(P)H dehydrogenase	P15559	83	274
A09b	Carbonic anhydrase 1	P00915	33	261
A11a	Glutathione S-transferase A1	P08263	118	222
A11b	GTPase Hras	P01112	69	189
A12a	BH3-interacting death agoist	P55957	94	195
A12b	Peroxiredoxin-1	Q06830	38	199
A13a	Cytochrome b5	P00167	152	134
A13b	Gamma-synuclein	O76070	81	127
A13c	Interferon gamma	P01579	62	166
A13d	Myoglobin	P02144	31	154
A14a	Fatty acid BP, heart	P05413	176	133
A14b	Lysozyme C	P61626	120	148
A14c	Cytochrome c	P99999	49	105
A16a	Ubiquitin-40S ribosomal protein S27a	P62979	114	156
A16b	Beta-2-microglobulin	P61769	47	119
A17a	NEDD8	Q15843	91	81
A17b	IGF-II	P01344	36	180
B01a	Serotransferrin	P02787	128	698
B01b	Lactotransferrin	P02788	117	710
B01c	Gelsolin	P06396	59	782
B03	Serum albumin	P02768	208	609
B04a	Catalase	P04040	116	527
B04b	Antithrombin-III	P01008	112	464
B04c	Histidyl-tRNA synthetase, cyto	P12081	140	509
B04d	Microtubule-associated protein tau	P10636	62	758
B06	Creatine kinase M-type	P06732	72	381
B07	Small ubiquitin-related modifier 1	P63165	123	101
B08	Annexin	P08758	223	320
B09a	NAD(P)H dehydrogenase	P15559	182	274

CODE	Protein Name and Accession Number	Accession	Mascot Score	PROTEIN LENGTH (Number of AAs)
B09b	Carbonic anhydrase 1	P00915	49	261
B09c	Carbonic anhydrase 2	P00918	41	260
B10a	Ribosylidihydronicotinamide dehydrogenase	P16083	214	231
B10b	Glutathione S-transferase A1	P08263	43	222
B11a	Glutathione S-transferase A1	P08263	102	222
B11b	GTPase Hras	P01112	76	189
B11c	Glutathione S-transferase P	P09211	60	210
B12a	Ubiquitin-conjugating enzyme E2 C	O00762	96	179
B12b	BH3-interacting death agoins	P55957	69	195
B12c	Peroxiredoxin-1	Q06830	53	199
B12d	Glutathione S-transferase P	P09211	43	210
B13a	Cytochrome b5	P00167	118	134
B13b	Interferon gamma	P01579	75	166
B13c	Myoglobin	P02144	52	154
B13d	Tumor necrosis factor	P01375	47	233
B13e	Gamma-synuclein	O76070	45	127
B14a	Lysozyme C	P61626	118	148
B14b	Fatty acid BP, heart	P05413	59	133
B14c	Cytochrome c	P99999	31	105
B16a	Ubiquitin-40S ribosomal protein S27a	P62979	76	156
B16b	Beta-2-microglobulin	P61769	53	119

Table 2

Protein Scores (PS) for Ground Truth Dataset and Decoy Datasets

CODE	Protein Name and Accession Number	PS (True)	μ (PS Decoys)	σ (PS Decoys)	Z-score	One Peptide Matched
A01a	Serotransferrin	3.55	0.466	0.380	8.11	
A01b	Lactotransferrin	3.54	0.564	0.399	7.47	
A01c	Gelsolin	2.13	0.545	0.417	3.81	
A03	Serum albumin	2.92	0.090	0.157	18.07	
A04a	Catalase	3.09	0.198	0.241	11.99	
A04b	Antithrombin-III	2.88	0.189	0.231	11.68	
A04c	Histidyl-tRNA synthetase, cyto	3.45	0.179	0.222	14.70	
A06	Creatine kinase M-type	3.53	0.250	0.282	11.62	
A07	Small ubiquitin-related modifier 1	0.70	0.010	0.051	13.49	
A08	Annexin	3.48	0.065	0.136	25.19	
A09a	NAD(P)H dehydrogenase	2.41	0.117	0.185	12.42	
A09b	Carbonic anhydrase I	2.18	0.109	0.180	11.55	
A11a	Glutathione S-transferase A1	1.29	0.066	0.137	8.94	
A11b	GTPase Hras	1.06	0.052	0.119	8.46	
A12a	BH3-interacting death agoinst	1.59	0.109	0.184	8.07	
A12b	Peroxiredoxin-1	2.82	0.117	0.187	14.47	
A13a	Cytochrome b5	1.65	0.032	0.095	17.03	
A13b	Gamma-synuclein	0.32	0.027	0.085	3.42	*
A13c	Interferon gamma	0.64	0.030	0.090	6.81	
A13d	Myoglobin	0.67	0.036	0.100	6.29	
A14a	Fatty acid BP, heart	1.62	0.063	0.134	11.66	
A14b	Lysozyme C	1.39	0.058	0.126	10.57	
A14c	Cytochrome c	1.02	0.050	0.117	8.28	
A16a	Ubiquitin-40S ribosomal protein S27a	0.89	0.028	0.089	9.63	
A16b	Beta-2-microglobulin	1.06	0.027	0.088	11.73	
A17a	NEDD8	1.40	0.026	0.085	16.16	
A17b	IGF-II	0.63	0.049	0.119	4.91	
B01a	Serotransferrin	3.55	0.430	0.376	8.31	

CODE	Protein Name and Accession Number	PS (True)	μ (PS Decoys)	σ (PS Decoys)	Z-score	One Peptide Matched
B01b	Lactotransferrin	3.53	0.451	0.371	8.30	
B01c	Gelsolin	2.04	0.457	0.384	4.12	
B03	Serum albumin	3.42	0.051	0.116	29.05	
B04a	Catalase	3.20	0.243	0.270	10.94	
B04b	Antithrombin-III	2.91	0.210	0.245	11.06	
B04c	Histidyl-tRNA synthetase, cyto	3.53	0.211	0.247	13.44	
B04d	Microtubule-associated protein tau	3.12	0.347	0.315	8.79	
B06	Creatine kinase M-type	3.55	0.126	0.195	17.55	
B07	Small ubiquitin-related modifier 1	0.70	0.003	0.029	24.04	
B08	Annexin	3.40	0.033	0.097	34.49	
B09a	NAD(P)H dehydrogenase	2.44	0.078	0.149	15.85	
B09b	Carbonic anhydrase 1	1.43	0.067	0.137	9.99	
B09c	Carbonic anhydrase 2	1.29	0.070	0.140	8.76	
B10a	Ribosylidihydronicotinamide dehydrogenase	3.38	0.082	0.152	21.75	
B10b	Glutathione S-transferase A1	1.63	0.074	0.145	10.74	
B11a	Glutathione S-transferase A1	1.26	0.053	0.121	10.00	
B11b	GTPase Hras	1.15	0.041	0.107	10.46	
B11c	Glutathione S-transferase P	0.62	0.038	0.102	5.70	
B12a	Ubiquitin-conjugating enzyme E2 C	1.93	0.094	0.165	11.13	
B12b	BH3-interacting death agoinst	2.03	0.089	0.161	12.07	
B12c	Peroxioredoxin-1	2.41	0.095	0.172	13.43	
B12d	Glutathione S-transferase P	0.30	0.071	0.147	1.54	*
B13a	Cytochrome b5	1.33	0.023	0.079	16.45	
B13b	Interferon gamma	0.35	0.023	0.080	4.05	*
B13c	Myoglobin	0.60	0.023	0.082	7.00	
B13d	Tumor necrosis factor	1.35	0.035	0.103	12.84	
B13e	Gamma-synuclein	0.35	0.018	0.071	4.62	*
B14a	Lysozyme C	1.29	0.047	0.116	10.75	
B14b	Fatty acid BP, heart	1.54	0.051	0.122	12.24	
B14c	Cytochrome c	0.89	0.042	0.109	7.76	
B16a	Ubiquitin-40S ribosomal protein S27a	0.76	0.024	0.083	8.83	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

CODE	Protein Name and Accession Number	PS (True)	μ (PS Decoys)	σ (PS Decoys)	Z-score	One Peptide Matched
B16b	Beta-2-microglobulin	0.57	0.019	0.072	7.66	

Table 3

Comparison of Mascot and PS Z-scores Showing Reordering by PS Z-score

Well	Order of MascotSum (pep scores)	MascotSums (pep scores)	# Mascot Peptides	Order of PS Z-Scores	PS Z-Scores
A04	a,b,c	244,154,90	5,4,90	*c,a,b	14,704,11,993,11,676
A12	a,b	94,38	1,1	*b,a	14,474,8,067
A16	a,b	114,47	2,1	*b,a	11,732,9,632
B04	b,c,a,d	112,140,116,62	2,4,116,2	*c,b,a,d	13,438,10,942,11,056,8,791
B11	a,b,c	102,76,60	2,2	*b,a,c	10,458,9,995,5,700
B12	a,b,c	96,69,53	3,1,1	*c,b,a	13,432,12,072,11,131
B14	a,b,c	118,59,31	3,59,1	*b,a,c	12,236,10,754,7,756
A01	a,b,c	168,77,60	5,3,2	a,b,c	8,113,7,467,3,809
A03	s	325	9	s	18,075
A06	s	101	5	s	11,618
A07	s	81	2	s	13,488
A08	s	177	6	s	25,19
A09	a,b	83,33	2,2	a,b	12,421,11,547
A11	a,b	118,69	3,3	a,b	8,938,8,458
A13	a,c,d	152,62,31	3,1,2	a,c,d	17,035,6,815,6,294
A14	a,b,c	176,120,49	176,2,1	a,b,c	11,661,10,572,8,282
A17	a,b	91,36	2,2	a,b	16,163,4,913
B01	a,b,c	128,117,59	3,4,2	a,b,c	8,306,8,305,4,118
B03	s	208	5	s	29,051
B06	s	72	3	s	17,546
B07	s	123	2	s	24,039
B08	s	223	6	s	34,492
B09	a,b,c	182,49,41	4,3,1	a,b,c	15,847,9,993,8,756
B10	a,b	214,43	4,2	a,b	21,746,10,738
B13	a,d,c	118,26,52	2,1,1	a,d,c	16,452,7,003,12,836
B16	a,b	76,53	1,2	a,b	8,829,7,665