Rapid and Accurate Global PTM Discovery (GPTM-D) Using Post-Acquisition Spectral Calibration and Defined Mass Windows

Andrew N. Other,[†,§] Fred T. Secondauthor,[†,∥] I. Ken Groupleader,[*,†,‡,§] Susanne K. Laborator,[*,¶] and Kay T. Finally[†,‡]

†*Department of Chemistry, Unknown University, Unknown Town*

‡*Department of Chemistry, Second University, Nearby Town*

¶*Lead Discovery, BigPharma, Big Town, USA*

§*A shared footnote*

∥*Current address: Some other place, Othertöwn, Germany*

E-mail: i.k.groupleader@unknown.uu; s.k.laborator@bigpharma.co

Phone: +123 (0)123 4445556. Fax: +123 (0)123 4445557

Abstract

Posttranslational-modifications (PTMs) influence many aspects of protein function in biological processes, and correctly identifying the various protein modifications in biological samples is crucial for understanding proteins. Ways of identifying and localizing PTMs are limited, but emerging techniques in the field of mass spectrometry are becoming available. GPTM-D [Journal of Very Important Results, 1, 1 (2016)] is a recently developed tool for global identification of PTMs using a single pass database search that is promising. Spectra file calibration prior to applying the tool, and algorithmic improvements in the peptide database search greatly improve the accuracy and efficiency of new PTM identification. We describe the calibration tool developed, and present numerical results that validate the proposed enhancement.

Correctly identifying protein posttranslational-modifications (PTMs) is crucial to understanding many aspects of protein function in biological processes. Methods to identify and localize PTMs from complex biological samples have been limited, but a few techniques are emerging. G-PTM-D[1] is a recently developed tool for global identification of PTMs. Spectra file calibration prior to applying the tool, We enhanced this method by using non-standard mass-window searches, and demonstrate its effectiveness in identification of numerous types of PTMs, including high-mass modifications such as glycosylations. The number of identified modifications increased by 20% and the search time decreased by an order of magnitude. We also show the advantages of using mass windows in a standard database search, when PTM discovery is not the primary goal.

Keywords

Proteomics

Introduction

Many peptide residue modification types are known, and databases containing detailed information about such modifications are readily available. Information about a modification can include the chemical or isotopic composition, average mass, monoisotopic mass, specificity to certain residues, and possible restriction of placement to certain peptide or protein termini. While this data is almost comprehensive, information about the localization of these modification to certain residues in specific proteins is scarce. We have recently described a bioinformatics tool, GPTM-D, for the discovery and localization of new PTMs from âĂIJbottom-upâĂİ tandem mass spectrometric datasets1. The GPTM-D workflow consists of three stages: 1) An open mass database search2-3 that provides spectral matches to unmodified peptides along with the mass differences between the identified peptides and the measured parent peptide masses (hypothesized to differ in mass due to the presence of a PTM). 2) For those peptides for which the mass difference corresponds to a known PTM, a database augmentation step adds plausible localized PTMs to the peptides in the search database. 3) A final standard narrow mass search of the augmented database to identify both modified and unmodified peptides subject to the standard FDR threshold (e.g. 1% FDR). We propose an extension to GPTM-D that addresses the long search times of the open mass search, and the limitations of the standard final narrow mass search. We propose a change to the standard GPTM-d workflow using a notch search in the discovery phase, using notches centered at the monoisotopic mass differences introduced by known modifications. This change improves the search time by orders of magnitude, and improves the specificity as well. Along with algorithmic improvements, we recommend using spectral calibration prior to running the GPTM-d procedure. A critical parameter for peptide identification and PTM localization is mass accuracy4. Higher mass accuracy provides increased specificity and thus confidence in peptide identifications, decreasing the false discovery rate. Instrument noise, systemic drift and miscalibration limit the mass accuracy in acquired spectra. Multiple calibration strategies to improve mass accuracy have been devised, and fall into three general categories: External calibration prior to the MS experiment (e.g. standard instrument calibration protocols); internal calibration during the MS experiment (e.g. real-time calibration using a lock mass standard5); and subsequent to the MS experiment (post-acquisition spectral calibration, REF). We use a post-acquisition calibration procedure, that builds upon the software lock mass concept6 recently reported by the Mann group. In our strategy, the m/z differences between expected and observed peaks in the peptide tandem ms spectra are compiled, and then used to recalibrate the spectra. The increased mass accuracy of the recalibrated spectra leads directly to improved identifications of both modified and unmodified peptides, as well as to increased confidence for PTM localization. The new search strategy implementing these changes was tested in the analysis of 2 deep proteomic datasets, where we saw a% increase in overall peptide identification, and a % increase in the identification of post-translationally modified peptides. Finally, we propose a carefully drafted heuristic for identifying and localizing modifications not currently present in any database.

Experimental Procedures

We developed a modified version of the Morpheus software for bottom-up spectral database searching that we call MetaMorpheus, which integrates a spectral calibration procedure with the G-PTM-D workflow. Two multi-fraction mammalian datasets with deep proteome coverage were used to evaluate the performance of the methods. Uniprot XML protein databases containing only reviewed human/mouse proteins were employed. A 10ppm precursor mass tolerance was used for the initial calibration step, and then reduced to 5ppm for subsequent searches in defined narrow mass windows (notch searches). For the Global PTM Discovery (G-PTM-D) process, we allowed 120 modifications including PTMs, adducts, and chemical modifications. For the discovery of other modifications, we used a novel wide mass-window search (interval search). The database search software employed is a modified version of Morpheus7, MetaMorpheus. We incorporated the calibration procedure, and the GPTM-D workflow in this modified version for convenience. Database search parameters are 10ppm precursor mass tolerance for uncalibrated and 5ppm precursor tolerance for calibrated spectra. We use a 0.01 Dalton product mass tolerance for uncalibrated and 0.006 Dalton for calibrated spectra. On-the-fly decoy database generation was used everywhere. For both mouse and human datasets we use XML databases from uniprot acquired on 7/28/16 and containing only reviewed proteins. The datasets tested are described in detail in 9-10. You can find them at http://www.LloydSmithDatasets.com.

Results and Discussion

By employing the novel notch search strategy instead of the wide-mass search in the G-PTM-D process, using only the modifications listed in Uniprot, the number of confidently identified PTMs increased by 8%. The overall search time dropped significantly, from 30 hours to 3.5 hours for all datasets. Using an expanded modification list that includes chemical modifications such as adducts and large mass PTMs such as glycosylations, the overall identification rate increased from 164,697 to 179,345 peptide-spectral matches (PSMs), which in turn increased the number of modified peptides by an additional 20%. We identified hundreds of glycosylated peptides in these unenriched samples, with many of these modifications exceeding 1000 Daltons. Algorithmic improvements in the G-PTM-D procedure itself, such as searching for combinations of modifications and for modifications with neutral losses make it an excellent tool for PTM discovery. Previously described wide mass search strategies for discovery and localization of unknown modifications benefit from using alternative, carefully designed search modes based on interval and notch searches. We discovered that limiting mass shifts to a lower bound of -187 Da (corresponding to the largest mass difference that could be attributed to loss of a single residue, Tryptophan), and no upper bound, is an important step in eliminating spurious PSMs. Furthermore, treating highly suspect mass shifts corresponding to residue additions/substitutions/deletions in a separate notch search with individualized false discovery rate estimates is beneficial. An automated tool built into MetaMorpheus allows confidently identifying novel modifications based on these search results. These strategies for increasing confidence in identifications with unusual mass shifts can be used in any database search. optional Novel Modification Discovery step. The input into the workflow is a standard bottom-up tandem-MS dataset obtained from a tryptically digested protein sample along with a corresponding protein sequence database, and the output is a comprehensive set of peptide and PTM identifications. Before describing the workflow in detail, it is useful to review and define the different database search modes that are employed. These are illustrated in Figure 1B, which shows four variations: standard search, notch search, comb search, and open search. These differ with respect to whether the entire mass space is searched, as opposed to defined subsets of the mass space; and with respect to the narrowness of the mass window employed for each element in the search space. The smaller the fraction of the mass space, and the narrower the mass window per search element, the faster the search will run (as fewer searches are executed per mass spectrum). In the following sections we describe the various steps in the flowchart. a) b)

Figure 1: a) The blue components mark the enhanced GPTM-D workflow for identifying and localizing PTMs. The green components are the steps in the novel Modification discovery heuristic. b) The four different search strategies CALIBRATION The spectral calibration procedure consists of two steps: Step 1: A standard tolerance (e.g. 10 ppm) database search is performed on the dataset. This yields a set of peptide identifications subject to a desired false discovery rate (e.g 1A simple test of the calibration quality is to withhold some known peak matches from the inputs of the calibration software, and to determine if they have been shifted closer to the theoretically correct value. We withheld 30a) b)

Figure 2: a) b)

ENHANCED GPTM-D The purpose of the Notch Search is to identify peptides that have a mass that is different from the selected ion mass by a mass corresponding to a known modification. This search is expected to produce some peptide spectrum matches with a significantly different mass than the corresponding extracted mass from the MS spectrum. This difference is hypothesized to arise from the presence of PTMs, adducts such as oxidation or metal contamination, or combinations thereof. Even though adducts arising from experimental artefacts are not ultimately interesting to biologists, not including them as an option decreases the total number of identified proteins and PTMS. Using the notch search capability, we can perform this search an order of magnitude faster than the equivalent open mass search in GPTM-D. A curated set of notches and the corresponding modifications is given in Appendix A. Using an off-the shelf database of modification is not advisable, due to shortcomings in each examined database, see next section. The database augmentation step is like the procedure described in the original GPTM-d work: an option to have the modification with the appropriate mass is added to the protein database, so that a subsequent search may include it as an option. An important difference is the ability to deal with combinations of modifications. Since the XML database is annotated with PTMs from the UniProt PTM list, it is a natural first idea to use it in the augmentation step as well. It has a few significant drawbacks: Some mass values do not correspond to actual mass shifts, lability of modifications is not taken into account, and many things are missing, including adducts and glycosylations. The final search with the augmented database is conducted with notches: the final notch search is specific to each dataset: we only recommend including the common missed monoisotopic mass values in the final search: see the following section for justification. Newly identified localized PTMs are now readily identified The effects of the increased mass accuracy were demonstrated in the numerical results for step 2, so to make the comparison fair we use the same precursor and product mass tolerances for both the standard and the notch mass search. The search time is understandably slightly higher for the Notch search, but it is not a big deal since the search times for both are not long, and orders of magnitude shorter than the searches in step 3. Standard Search-Calibrated Notch Search-Calibrated Time (jurkat) 22.1 hrs 2.9 hrs Time (Mouse) 13.5 hrs 1.2 hrs

DISCOVERY OF MODIFICATIONS NOT IN DATABASE The proposed notch search strategy can be used for discovery of novel modifications not present in a database. We describe the heuristic approach for augmenting the modification database and the final notch list. The discovery heuristic starts with conducting a modified comb interval search of a calibrated file. The highly non-uniform shape of the allowable mass differences to search is motivated by the Open Search Case Study in the following section. This allows us to pre-emptively restrict the database search to only look for peptides in intervals that satisfy the condition of being within the specified search windows. This alternative to the open search improves search times and increases specificity. After the conclusion of the search, the PSMs within 11. No mass error. Exact match (within a certain noise tolerance) between the peptide corresponding to the MS2 fragmentation and the reported precursor monoisotopic mass. 2. Monoisotopic error. Peaks around values such as 1.003, 2.006, 3.009 (multiples of the difference in mass between C12 and C13). These correspond to having misidentified the monoisotopic peak in the preprocessing deconvolution step. In this case, the identified peptide is still correct. 3. Peaks corresponding to PTMs or adducts, e.g. 42.010 for Acetylation or 21.981 for Sodium adduct. In this analysis, there is no conceptual difference between PTMs and adducts: both are observed simply as mass differences between the identified peptide and the precursor mass. These peaks are further classified into a. Localizeable âĂŞ e.g. Methylation. There is evidence for the modification in the MS2 spectrum: the corresponding peaks are shifted by an appropriate amount. b. Labile âĂŞ e.g. Sulfonation. There is no evidence of the modification in the fragmentation spectrum. c. Either Localizeable or Labile, e.g. Phosphorylation. Sometimes there is evidence of it in the fragmentation spectrum, and sometimes there is not. 4. Amino acid removals, additions or substitutions. 5. Modification dependent mass shifts âĂŞ these peaks occur only in presence of certain modifications in the identified peptide. E.g. -15.995 in presence of an identified oxidation, or -79.966 in presence on an identified phosphorylation. 6. Combinations of any of the above, e.g. 1.987 for combination of a monoisotopic error and a deamidation. Some of these combinations occur frequently, and it is crucial to account for them. An automated script analyzes each identified peak, and provides clues to the nature of the peak. For every peak, a profile is automatically generated. It includes the total number of unique peptides associated with the mass shift, the fraction of decoys, mass match with any known entry in the Unimod or UniProt database, mass match to an amino acid addition/removal combination, mass match to a combination of higher frequency peaks, fraction of localizable targets, localization residues and/or termini, and presence of any modifications in the matched peptides. We follow the following procedure to classify mass shift peaks. 1. A z score is computed for the fraction of decoys in the peak, comparing with 12.

An attempted localization of the mass difference is automatically performed on every peptide-spectrum match. We carefully examine the mass differences with a localization fraction greater than 0.2, and attempt to deduce the chemical formula, specificity sites and positions within a peptide. Once a determination has been made, we add the new modification type to the modifications list. An example output of this step is:

The improvements of using a Comb search instead of an Open search are two-fold. We summarize the differences in Table 2. Open Search Comb Search Search Time 54.36 hrs 29.66 hrs

The improvement in the discernibility of PTMs is apparent when running a comb search on calibrated vs uncalibrated spectra files. Specifically, the discernibility of PTMs with similar mass errors (ones that are only different because of the mass defect). This was not really possible previously. In the current run, Sulfation and Phosphorylation are readily distinguishable. We present the numbers for the example below in the step 5 numerical validation section. OPEN SEARCH CASE STUDY Introduction of the comb search instead of an open search is motivated by a careful review of the Unimod database. Known modifications with mass difference within 200 Daltons have values that are within [-0.1, 0.2] of every integer. This interval choice stems from the mass defect in the primary isotope of the common elements. PTM combinations also have this property. We tested multiple search strategies in a search for one that gives the smallest number of false positives. KNOWN MODIFICATION CONFIRMATION CASE STUDY Confirmation of the properties of Sodium Adducts, Phosphorylation and Sulfonation. GLYCOPROTEINS CASE STUDY Proteins that contain modifications such as Hexose, HexNAC, and similar ones are interesting. The uniprot database only lists some possible sites for such modifications, but does not specify the actual type âĂŞ thus creating ambiguity regarding the type of the modification. The new modification discovery workflow is well suited to identify such proteins. NOVEL MODIFICATION CASE STUDY

Acknowledgement

The authors thank . . .

Supporting Information Available

A listing of the contents of each file supplied as Supporting Information should be included. For instructions on what should be included in the Supporting Information as well as how to prepare this material for publications, refer to the journal's Instructions for Authors.

The following files are available free of charge.

- Filename: brief description

- Filename: brief description

References

(1) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V.; Scalf, M.; Smith, L. M. Global Post-translational Modification Discovery. *Journal of Proteome Research* **0**, *0*, null.

Graphical TOC Entry

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct.
Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.
The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.
This box and the associated title will always be printed on a separate page at the end of the document.