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

Stefan R. Solnstev, Michael R. Shortreed, Brian L. Frey, and Lloyd M. Smith\*

Department of Chemistry and Genome Center of Wisconsin

University of Wisconsin – Madison

Madison, WI 53706

**ABSTRACT (rewrite this when paper done)**

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.

**INTRODUCTION**

We have recently described a bioinformatics tool, GPTM-D, for the discovery of new PTMs from “bottom-up” 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).

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, which we refer to as mzCal, 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.

We have also modified the search strategies to run more efficiently, by using “notch” and “comb” searches of the mass space that reduce unproductive search time and add flexibility by accounting for certain mass differences due to PTMs, adducts and monoisotopic errors. The new search strategy implementing these changes was tested in the analysis of 3 deep proteomic datasets (describe), for which the use of mzCal provided a \_\_% increase in overall peptide identification, and a \_\_% increase in the identification of post-translationally modified peptides. The improved search strategies account for a \_% decrease in search time, and an additional \_% increase in identification rate.

**EXPERIMENTAL PROCEDURES**

The mzCal software is written in C#, and works on all popular operating systems. It is released under the GPL license, and the code is available for inspection at <https://github.com/smith-chem-wisc/mzCal>.

The database search software employed is a modified version of Morpheus7. The benchmark calibration software is mzRefinery8.

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 0.005 Dalton for calibrated spectra. On-the-fly decoy database generation was used everywhere except for open and comb searches, since their purpose is discovery and not validation.

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>.

**THE ENHANCED GPTM-D WORKFLOW**

Figure 1a shows the enhanced GPTM-D workflow. The left column shows the previously described GPTM-D workflow, and the column on the right corresponds to the improved workflow incorporating the mzCal calibration and comb and notch search strategies. The input into the workflow is a standard bottom-up tandem-ms dataset obtained from a tryptically digested protein sample, 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 or not 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 wider the mass window per search element, the faster the search will run (as fewer searches are executed per mass spectrum).

|  |  |
| --- | --- |
| a)    Standard tolerance database search  mzCal calibration  Comb search  Open search  Database augmentation  Standard Tolerance Database Search  Narrow Tolerance Database search with Notches  Original New  Step 1  Step 2  Step 3  Step 4  Step 5 | b) |

Figure : a) The blue components mark the original GPTM-D workflow for identifying and localizing PTMs. The green components are the proposed enhancements to the GPTM-D workflow. b) The four different search strategies

The modified workflow consists of five 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 1%; based upon the target:decoy strategy (ref)).

**Step 2:** mzCal calibration. The identifications from step 1 are used to extract peak matches from the spectra, and a semi-parametric calibration procedure is performed. The mzCal software is described in detail in appendix.

**Step 3:** A database search only using the peptide fragment ions is performed. This search is expected to produce peptide spectrum matches that match to peptides 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, monoisotopic errors or adducts such as oxidation or metal contamination.

Introduction of the comb search instead of an open search is motivated by a careful review of the Unimod database of PTMs and Amino Acid substitutions. Known modifications with mass difference within 200 Daltons have values that are within [-0.1, 0.2] of every integer. PTM combinations also have this property. This allows us to pre-emptively filter 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.

An example output of this step is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mass Error (DA)** | **Count** | **Description** | **Assign In GPTM-D** | **Final Search Notch** |
| 0 | 10000 | Exact Match |  |  |
| 0.985 | 1021 | Deamidation | YES |  |
| 52.911 | 253 | Iron Adduct |  | YES |
| -1.003 | 104 | Missed Monoisotope |  | YES |
| -17.026 | 58 |  | YES |  |
| 15.995 | 38 | Oxidation | YES |  |
| 79.966 | 29 | Phosphorylation |  |  |
| 14.015 | 23 | Methylation | YES |  |
| -113.084 | 20 | Leucine Removed |  |  |
| 113.084 | 18 | Leucine Added |  |  |

**Step 4:** A database augmentation step that is described in the GPTM-D work.

**Step 5:** The final search with the augmented database. Newly identified localized PTMs are now an additional possibility in the search results.

The calibration step allows using a narrow precursor and product search windows, decreasing time and increasing specificity.

In addition, the introduction of a notch search capability at this final stage allows for detection of monoisotopic errors and adducts, instead of mis-assigning.

**RESULTS AND DISCUSSION**

We follow the enhanced workflow sequence in this section to provide evidence of improvement at each step.

**Step 2 numerical validation**

A 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 test the following peak exclusion strategies: **a)** Withholding random identifications **b)** Withholding all high m/z valued peaks **c)** Withholding all low m/z valued peaks.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Uncalibrated | mzRefinery | mzCal |
| **Withholding a)** | 10ppm | 3ppm | 6ppm |
| **Withholding b)** | 10ppm | 3ppm | 6ppm |
| **Withholding c)** | 10ppm | 3ppm | 6ppm |

|  |  |  |
| --- | --- | --- |
| Precursor Tolerance | Uncalibrated | mzCal |
| **10ppm** | 40% | 60% |
| **5ppm** | 40% | 60% |

We also validate the calibration results by performing a standard peptide database search. We report the fraction of spectra that is identified at 1% FDR.

The numerical results are summarized in Table 1.

**Step 3 numerical validation**

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 | 10 minutes | 5 minutes |
| Percent of spectra at 1% FDR | 50% | 60% |

The improvement in the discernability 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.

**Step 5 numerical validation**

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-uncalib | Standard Search-calib | Notch Search-calib |
| Percent of spectra under 1% FDR |  |  |  |
|  |  |  |  |

Some of the steps can be simplified in order to improve the search times, but these simplifications leave little room for discovery. Specifically, in Step 3, a notch search with notches at the most common PTMs can be used, and it would provide some entries for the augmented database, but in case of unexpected modifications it would not have the power of discovery.

1. GPTM-D.

2. Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P., A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **2015,** *33* (7), 743-9.

3. Na, S.; Bandeira, N.; Paek, E., Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics* **2012,** *11* (4), M111 010199.

4. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Hernandez, P.; Appel, R. D.; Binz, P. A.; Goodlett, D. R., On the benefits of acquiring peptide fragment ions at high measured mass accuracy. *J Am Soc Mass Spectrom* **2008,** *19* (6), 891-901.

5. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **2005,** *4* (12), 2010-21.

6. Cox, J.; Michalski, A.; Mann, M., Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom* **2011,** *22* (8), 1373-80.

7. Wenger, C. D.; Coon, J. J., A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J Proteome Res* **2013,** *12* (3), 1377-86.

8. Gibbons, B. C.; Chambers, M. C.; Monroe, M. E.; Tabb, D. L.; Payne, S. H., Correcting systematic bias and instrument measurement drift with mzRefinery. *Bioinformatics* **2015,** *31* (23), 3838-40.

9. Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Sheynkman, G. M.; Scalf, M.; Keller, M. P.; Attie, A. D.; Smith, L. M., Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search. *J Proteome Res* **2015,** *14* (11), 4714-20.

10. Cesnik, A. J.; Shortreed, M. R.; Sheynkman, G. M.; Frey, B. L.; Smith, L. M., Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. *J Proteome Res* **2016,** *15* (3), 800-8.