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

# Introduction

The calibration of spectra files and algorithmic improvements in the peptide dictionary search improves the accuracy and efficiency of new posttranslational-modification (PTM) identification.

# Motivation

Physical measurements introduce noise, and instruments that take multiple measurements often have correlated noise between different samples. Knowledge of what some of the measurements should be, paired with an assumption of correlated error in measurement, enables us to make an intelligent guess of the error for the rest of the measurem­­­ents. This is indeed the scenario with mass spectrometry data, where the knowledge of the true mz values for some peaks comes from identified peptide sequences.

The numerical difference between a true, or *reference* value and an observed value is a sum of the *random error* and the *systematic error* of the measurement. The *random error* arises because of some inherent random variability, while the *error due to bias* is a **directed** error in the observed quantity caused by

The measurements’ *bias* (non-random or directed effects caused by a factor or factors unrelated to the independent variable) and error (random variability).

The numerical difference between a true, or *reference* value and an observed value always has a reason. This error can often be at least partially described by observable experimental conditions.

Note that the instrument *resolution* is another important measure of measurement quality, but it is unrelated to the error in an individual measurement.

The goal of the calibration process is to shift each peak in the MS and MS/MS spectra by an appropriate amount, to compensate for as much systemic error as possible. We observe that

# Previous Calibration Work

Calibrating spectra files is a necessity due to a mass spectrometer introducing a bias in measurements. On a basic level, in each spectrum a mass spectrometer captures m/z peaks with corresponding intensities. Calibrating intensity values is a more difficult task that is not the main focus of the paper. All previous work focused on calibrating m/z measurements based on knowledge of the expected location of a subset of peaks. The existing methods for calibrating spectra include constant shifts based on a chemical lock mass compound, shift based on molecules based on the digestion compound, and a recent software lock mass paper.

## Chemical Lock Mass

A compound such as EEEEE can be present everywhere in the column, and is thus seen in every MS scan. Due to the known

## Digestion Compounds

Trypsin is in itself a peptide, and high-sensitivity trypsin peaks can be observed.

## Software Lock Mass

A recent paper suggests using known identifications to create a two-dimensional model of the error in the measurement. The two variables are the Retention Time and the m/z value of each peak. The model predicts the error in the measured m/z value based on these input variables.

The differences with the work presented here are as follows:

* We do not limit ourselves to two variables, but expand to use other useful information such as observed intensity, injection time and others.
* We separate the scan-wise variables from the individual peak variables (namely m/z value and intensity). This is an important consideration, since peaks that appear in the same scan have identical retention times, thus making the distribution of retention times discrete rather than the continuous m/z distribution of peaks.
* The calibration is done on both MS and MS/MS scans, as opposed to just MS scans.
* They calculate a single mass error value for each peptide, combining multiple peaks from multiple MS scans into a single datapoint. We consider each peak separately.
* They use a mass error value calculated by MaxQuant, we use the difference between the reference and observed peaks as the errors.
* We predict the error in *m/z* values, while they predict the mass errors
* They do not shift any peaks: Instead, they run a new dictionary search with updated values for masses of MS isotope patterns. We shift the peaks: This is different, because some peaks if shifted can become a part of an isotope pattern, or fall out of one, or can create new isotope patterns. None of this can happen with their method.
* We publish our software both as a standalone tool and as a library, distributed along with its source code, in contrast to MaxQuant.

# Theoretical-Experimental Peak Matching

Every peptide sequence identification corresponds to multiple peaks in the spectra. For every identification, the MS/MS scans should include peaks corresponding to the fragment ions of the peptide produced by the dissociation method employed in the mass spectrometer. The neighboring MS scans should have evidence of the un-fragmented peptide, over the elution profile of the peptide. Each of the matches correspond to peaks at different charge states, and different isotopic peaks. All of those have a true mz value, and most of them should have corresponding peaks in the acquired spectra.

For a concrete example, assume that an identification tells us that an MS/MS spectrum corresponds to peptide sequence HVVQSISTQQEKETIAK, identified with a precursor charge 3. Since the sequence contains 17 amino acids, the total number of b and y ions that should be present in the MS/MS spectrum is 32. Each of those ions can have either 1, 2, or 3 charges, so the number of monoisotopic peaks to look for is 96. Every ion-charge state match still corresponds to multiple peaks, since every peptide has an isotope distribution. The number of peaks in the isotope distribution can be large.

Describe the differences and similarities with the original calibration paper. Explain how we extended the idea. Now we calibrate the MS2 spectra as well. Show the difference with only calibrating MS1.

# Calibration

## Multiple Constant Shifts

In order to both include more training points, and to exclude outliers that do not in reality correspond to any theoretical peak, after the initial data point search is finished, we use a simple constant-shift calibration to center our observations around zero. Then a new search of the data is done on the centered points. This helps with making the training set more symmetric with regards to outliers. Specifically, consider spectra that have all errors be of 0.01 m/z units, but we search within 0.02 m/z of zero. The number of outliers that underestimate the error are much greater than ones overestimating it, and therefore building a model based on this data would underestimate the error.

We repeat the constant shift procedure until the number of observed matches between the theoretical and experimental peaks stops increasing.

## Calibration

Once the data is collected, we have training points that correspond to matches between theoretical and experimentally observed peaks.

Instead of pre-selecting the

## Possible Improvements

Different fractions of the same experiment are expected to have overlapping identifications.

Neighboring scans, look for peaks that are repeating.

# Notch Search

Using mass differences between the observed peptide mass and the theoretical mass is an efficient method for identifying post-translational modifications. A popular approach is to consider the MS/MS spectrum in isolation, and perform a database search that identifies

## Comb Search

A search of the unimod database reveals that 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 ignore

## Notch Search

# Results

We start the section by providing general results that

## Calibration Quality

We first demonstrate the improvement in a standard protein dictionary search.

## Notch Search Time Improvement

## Mouse Data

Calibration successfully

## Jurkat Data

a. Only spectra files calibration, no additional search done. Show MSE values for different calibration functions, pick one, and say why it’s better than others

b. A new normal Morpheus run comparison: 10ppm, vs 3ppm (4 runs)

i. Show that precursor mass errors are centered at zero and have smaller variance

ii. Show that there are more PTMs identified we are confident it

iii. Show that decoy PTMs are pushed down the list

Compare with 11 cell line paper, and with Jurkat/Mouse paper

c. Notches

i. Show that now are able to discern between PTMs with similar mass errors (ones that are only different because of the mass defect)

d. Show the improvement in search time due to the new notches option