Bioinformatics, 37(17), 2021, 2768–2769
doi: 10.1093/bioinformatics/btab056
Advance Access Publication Date: 4 February 2021
Applications Note

Gene expression

mzRecal: universal MS1 recalibration in mzML using identified peptides in mzIdentML as internal calibrants

Rob Marissen * and Magnus Palmblad *

Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on October 29, 2020; revised on December 31, 2020; editorial decision on January 22, 2021; accepted on January 26, 2021

Abstract

Summary: In mass spectrometry-based proteomics, accurate peptide masses improve identifications, alignment and quantitation. Getting the most out of any instrument therefore requires proper calibration. Here, we present a new stand-alone software, mzRecal, for universal automatic recalibration of data from all common mass analyzers using standard open formats and based on physical principles.

Availability and implementation: mzRecal is implemented in Go and freely available on https://github.com/524D/mzRecal.

Contact: r.j.marissen@lumc.nl or n.m.palmblad@lumc.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Obtaining a high number of confidentially identified peptides is advantageous in mass spectrometry-based proteomics. Accurate instrument calibration is important, as most peptide identification algorithms use the precursor (MS1) mass to select candidate peptides for comparison with the tandem mass spectrum (MS2) (Haas *et al.*, 2006). Accurate MS1 data also limits the search space, reducing CPU time.

Even with poor calibration, some peptides can be confidently identified from MS2 data with a wide error tolerance window. The exact masses for these identified peptides can be calculated and the peptides used as internal calibrants in MS1, where the recalibration improves mass measurement accuracy for all peptides. In a previously described a software for automated internal calibration using identified peptides we used an FTICR calibration function (Palmblad *et al.*, 2006). This software only works for FTICR data in the older mzXML and pepXML formats for liquid chromatography-mass spectrometry data and peptide identifications.

Recalibration has been incorporated in other software, including the popular MaxQuant (Cox and Mann, 2008) and the MSFragger pipelines (Kong *et al.*, 2017). However, these and other efforts do not calibrate a wide range of mass analyzers based on physical principles. Nor do they output recalibrated data in the current community standard mzML. mzRecal is first to do all of this, becoming a universal MS1 recalibration tool. As mzRecal consumes and produces the same data types and formats, it can be inserted into any proteomics data analysis workflow using these formats. mzRecal is compatible with identification results from many different search engines.

2 Materials and methods

mzRecal takes peptides identified above a specified confidence threshold as potential calibrants along with alternative charge states in the *m/z* range and common polydimethylcyclosiloxane background ions. For all potential calibrants, the theoretical *m/z* is calculated from the elemental composition and charge. The potential calibrants for each MS1 spectrum are obtained by selecting those that fall within a user-selected elution window. Multiple peptides having the same *m/z* are considered one calibrant. mzRecal fits the calibration function to each MS1 spectrum independently to correct for instrument drift and varying ion abundance.

A recalibration function based on the physical principles (Supplementary Section S2) of the instrument makes the procedure more robust than a generic polynomial function, which struggles to extrapolate beyond the calibrants. By default, mzRecal deduces the instrument type from the CV term in the mzML file and applies the appropriate function (Table 1). Recalibration is performed for each

Table 1. Recalibration functions for different instrument types

MS instrument	Recalibration function
Orbitrap	$m' = \frac{A}{(\sqrt{m}-B)^2}$
TOF	$m' = Am + B\sqrt{m} + C$
FTICR	$m' = \frac{A}{1/m-B}$

Note: m is short for *m/z*. Constants A-C are optimized for each spectrum by the procedure. Other, similar, functions have been used for all of these analyzers, but these were found to work well with the mzRecal recalibration algorithm. .

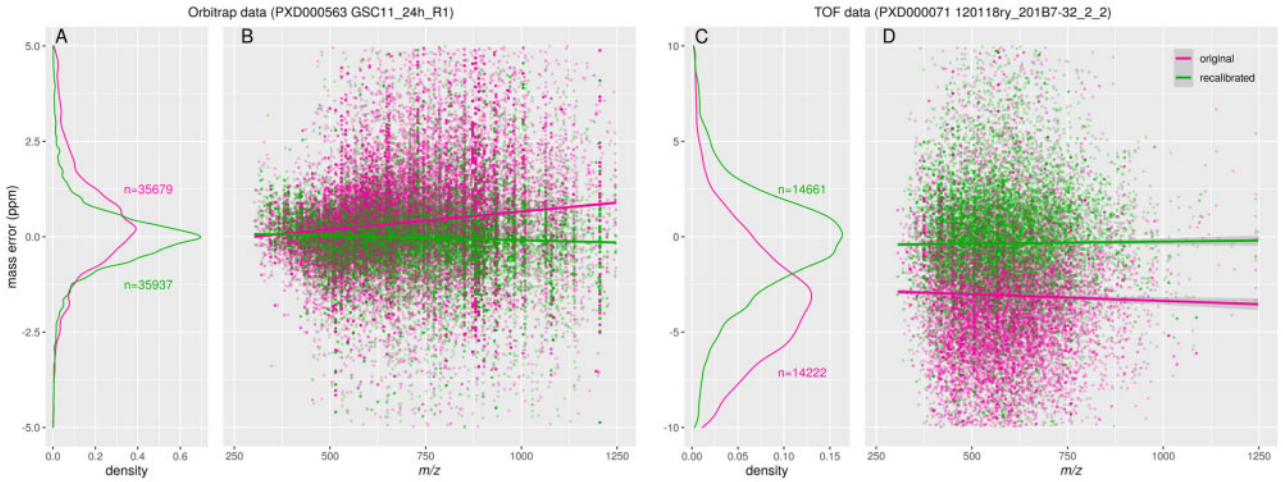

Fig. 1. Mass measurement errors in Orbitrap (**A** and **B**) and TOF (**C** and **D**) LC-MS/MS data before (magenta) and after (green) mzRecal recalibration. Only confident (expectation value < 0.01) PSMs without isotope error are shown. Accuracy and precision both improved (A and C), as the mean error and standard deviation went from 0.397 ± 1.407 ppm to -0.034 ± 0.966 ppm (Orbitrap) and from -3.079 ± 3.372 ppm to -0.356 ± 3.007 ppm (TOF data). Linear regression (B and D) show that error varies less with *m/z* after recalibration of these datasets

MS1 spectrum independently by finding the parameters that minimize the RMS error of the calibrants in the spectrum. Further refinement of the calibration is certainly possible, but we have prioritized keeping all calibration functions simple to avoid overfitting when few suitable calibrants exist.

A robust method must account for false matches between MS1 peaks and potential calibrants. By default, mzRecal removes outlier calibrants according to the definition of mzQC (The HUPO-PSI Quality Control Working Group, 2020). After outlier removal, the recalibration is repeated until all calibrants are accepted or there are too few left to recalibrate the spectrum without overfitting.

To demonstrate mzRecal, we analyzed public datasets from different instruments and vendors, including the same Orbitrap and TOF datasets used previously (Holl *et al.*, 2015). The datasets were originally published by Lichti *et al.* (2014) and Yamana *et al.* (2013), and are available on PRIDE with accession numbers PXD000563 and PXD000071. We also tested mzRecal on several other PRIDE datasets from different instruments (Supplementary Information). Comet (Eng *et al.*, 2013) version 2019.01 rev. 5 was used for peptide identification before and after recalibration. A *k*-fold cross-validation was used in the validation of mzRecal, combining the resuts from 10 recalibation runs, where in each the peptide-spectrum matches (PSMs) corresponding to 90% of unique theoretical peptide masses were used for recalibration and the remaining 10% for validation.

3 Results

Figure 1 shows the results of recalibration of Orbitrap and TOF data using tenfold cross-validation. Though only PSMs with high confidence and without isotope error (the triggering of the MS2 event on peaks other than the monoisotopic) are visualized, the masses of all peptides improved, regardless of the quality of the MS2 data. mzRecal improved both accuracy and precision in MS1 data. Residual bias was reduced to very close to zero ($\ll$1 ppm). As can be seen in Figure 1, precision is also improved, focusing mass errors around zero, resulting in a higher, more narrow, peak in the error distribution. This is what one expects from a recalibration procedure that does more than simply shifting the masses according to average error at a given *m/z*. We have tested the recalibrated mzML files with a multitude of software and did not experience any format-related issues reading, plotting or analyzing the recalibrated data, suggesting mzRecal can be plugged into most proteomics data analysis workflows based on mzML.

Funding

mzRecal was developed as part of the ELIXIR Implementation Study 'Crowdsourcing the annotation of public proteomics datasets to improve data reusability'.

Conflict of Interest: none declared.

References

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.

Eng,J.K. *et al.* (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics*, **13**, 22–24.

Haas,W. *et al.* (2006) Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell Proteomics*, **5**, 1326–1337.

Holl,S. *et al.* (2015) Scientific workflow optimization for improved peptide and protein identification. *BMC Bioinform.*, **16**, 284.

Kong,A.T. *et al.* (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods*, **14**, 513–520.

Lichti,C.F. *et al.* (2014) Integrated chromosome 19 transcriptomic and proteomic data sets derived from glioma cancer stem-cell lines. *J. Proteome Res.*, **13**, 191–199.

Palmblad,M. *et al.* (2006) Automatic internal calibration in liquid chromatography/Fourier transform ion cyclotron resonance mass spectrometry of protein digests. *Rapid. Commun. Mass. Spectrom.*, **20**, 3076–3080.

The HUPO-PSI Quality Control Working Group. (2020) mzQC: reporting and exchange format for mass spectrometry quality control data. https://github.com/HUPO-PSI/mzQC/blob/bulk-cvterms/cv/qc-cv.obo (4 February 2021, date last accessed).

Yamana,R. *et al.* (2013) Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC–MS/MS analysis with meter-scale monolithic silica columns. *J. Proteome Res.*, **12**, 214–221.