FOCUS: INTERDISCIPLINARY BIOLOGICAL MS: RESEARCH ARTICLE

Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors

Jürgen Cox, Annette Michalski, Matthias Mann

Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

Abstract

Mass accuracy is a key parameter in proteomic experiments, improving specificity, and success rates of peptide identification. Advances in instrumentation now make it possible to routinely obtain high resolution data in proteomic experiments. To compensate for drifts in instrument calibration, a compound of known mass is often employed. This 'lock mass' provides an internal mass standard in every spectrum. Here we take advantage of the complexity of typical peptide mixtures in proteomics to eliminate the requirement for a physical lock mass. We find that mass scale drift is primarily a function of the m/z and the elution time dimensions. Using a subset of high confidence peptide identifications from a first pass database search, which effectively substitute for the lock mass, we set up a global mathematical minimization problem. We perform a simultaneous fit in two dimensions using a function whose parameterization is automatically adjusted to the complexity of the analyzed peptide mixture. Mass deviation of the high confidence peptides from their calculated values is then minimized globally as a function of both m/z value and elution time. The resulting recalibration function performs equal or better than adding a lock mass from laboratory air to LTQ-Orbitrap spectra. This 'software lock mass' drastically improves mass accuracy compared with mass measurement without lock mass (up to 10-fold), with none of the experimental cost of a physical lock mass, and it integrated into the freely available MaxQuant analysis pipeline (www.maxquant.org).

Key words: Mass accuracy, MaxQuant, Proteomics, Database search, Orbitrap, Peptide mass measurement, Lock mass

Introduction

Mass spectrometry (MS)-based proteomics [1–3] greatly benefits from high resolution and high mass accuracy measurements [4]. For example, resolving co-eluting peptides of similar mass is a prerequisite for their accurate

quantification, and high accuracy measurement of peptide masses greatly aid in their identification by providing stringent filters on possible candidates. Several definitions of mass accuracy are commonly used, and this important parameter is often only assessed anecdotally [5]. In proteomics, the operationally important definition is the best mass estimate from the MS measurement together with a statistical confidence interval. This interval can then be used as the basis for setting a permissible mass deviation window for peptide identification in databases. Such confidence intervals can be assigned to each peptide separately. They are obtained from the measured values from consecutive scans and isotope states, weighted by the signal for each data

Electronic supplementary material The online version of this article (doi:10.1007/s13361-011-0142-8) contains supplementary material, which is available to authorized users.

Correspondence to: Jürgen Cox; *e-mail:* cox@biochem.mpg.de or Matthias Mann; *e-mail:* mmann@biochem.mpg.de

Received: 16 February 2011
Revised: 16 March 2011
Accepted: 16 March 2011
Published online: 22 April 2011

point. We have previously described principles of extracting these mass values from large scale data sets and implemented the corresponding algorithms in the MaxQuant computational proteomics analysis pipeline [4, 6]. As a result of applying these computational algorithms, these peptide mass accuracies are frequently improved to the sub-ppm range. This makes the precursor mass value an important search parameter and allows a corresponding drop in the required quality of the MS/MS spectra while still maintaining a 1% false discovery rate for peptide identifications.

A precondition for the above analysis was the elimination of the systematic mass drift by using a lock mass [7, 8]. A lock mass is a defined compound of known composition that is added to the MS analysis. Some instruments feature a separate electrospray source, which is used to spray the reference compound [9, 10]. Alternatively, the reference compound could be mixed into the analyte directly, but this has disadvantages because the compound may interfere with analysis of low abundance samples or it may not be detectable in high abundance samples. In electrospray, charged droplets are formed in laboratory air and analyte ions are desorbed from them. However, these charged droplets can also absorb and ionize background chemicals that are always present in laboratory air [11, 12]. On the LTQ-Orbitrap family of instruments these ions, specifically polycyclodimethylsiloxanes, can be separately isolated in the linear ion trap and injected into the C-trap, which is an intermediate storage trap [13]. In the C-trap the lock mass ions are mixed with the MS or MS/MS ions to be analyzed and co-injected into the Orbitrap analyzer. The ion is recognized by the data system in real time and the mass scale is automatically adjusted. While this procedure is sufficiently fast to be routinely applicable in proteomic experiments, there is some time requirement for isolating the lock mass ions, adding to overall MS and MS/MS cycle times. In addition, it can often be desirable to suppress background ions in laboratory air (i.e., by the ABIRD device: www.esisourcesolutions.com). This has the side effect that the lock mass is no longer available. For these reasons, an alternative to the lock mass would be beneficial.

In proteomics experiments, typically hundreds or thousands of peptides are identified in every LC-MS/MS run. Many of these peptides have very information-rich MS/MS spectra, and they can be unambiguously identified even with large mass tolerances. We have previously made use of this fact by implementing a two-pass search, where the top identified peptides serve as mass references for calibration [14]. However, this recalibration was done globally for the entire LC-MS run, was only applicable to time of flight data and did not attempt to reach sub-ppm mass accuracy. For Orbitrap data, the simple mass scale adjustments [14] would not be applicable. In this paper we set out to develop algorithms to replace the physical

lock mass with a software algorithm that performs at least as well in global recalibration of Orbitrap data.

Methods

Protein Digestion

Total HeLa cell lysate was treated with a urea (6 M) and thiourea (2 M) solution followed by reduction with dithiothreitol (DTT) (1 mM) for 30 min and alkylation with iodoacetamide (IAA) (55 mM) for 20 min at room temperature. The proteins were digested with Lys-C (1 µg/50 µg protein) (Wako, Neuss, Germany) for 3 h at room temperature. The mixture was diluted with water (1:4) before incubation with trypsin (1 µg/50 µg protein) (Promega, Mannheim, Germany) for 12 h at room temperature. The digestion was stopped by addition of formic acid (3%) and the samples stored on StageTips [15].

LC-MS/MS Analysis

The peptide mixture was loaded onto a C₁₈-reversed phase column (15 cm long, 75 µm i.d.) that was packed in-house with ReproSil-Pur C18-AQ 3 µm resin (Dr. Maisch) in buffer A (0.5% acetic acid). The peptide mixture was separated with a linear gradient of 5%–60% buffer B (80% ACN and 0.5% acetic acid) at a flow rate of 250 nL/min on a nanoflow HPLC (Proxeon Easy HPLC; Thermo Fisher Scientific). On-line coupling of the HPLC system to an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was achieved using a nanoelectrospray ion source (Proxeon Biosystems, now Thermo Fisher Scientific). Data were acquired in a data-dependent ‘top5’ format, selecting the most abundant precursor ions from the survey scan (mass range 300–1650 Th) in order to isolate them in the linear ion trap and fragment them by CID with a normalized collision energy of 35 eV. Survey scans were acquired with a resolution of 60,000 at *m/z* 400 and with a target value of 10⁶ in the Orbitrap analyzer. The MS/MS scans were acquired with unit mass resolution in the LTQ using 3000 as target value. Dynamic exclusion was defined by a list size of 500 features and exclusion duration of 90 s. Early expiration was set to expiration count 3 and S/N threshold 3. The lower threshold for targeting a precursor ion in the MS scans was 1000 counts. Three technical replicates were acquired without using the lock mass option in Xcalibur. In three separate technical replicates protonated polycyclodimethylsiloxane (PCM-6) with exact *m/z* 445.1200 Th was selected as lock mass for the measurement [13].

Data was analyzed by MaxQuant [4] using the Andromeda search engine [16]. The IPI human data base was used for peptide identification in the IPI human data base (containing 87,061 entries) combined with 262 common contaminants and concatenated with the reversed versions of all sequences. Enzyme specificity was set to trypsin, allowing cleavage N-terminal to proline. Further modifications were cysteine

carbamidomethylation (fixed) as well as protein N-terminal acetylation and methionine oxidation (variable).

Computational Methods

Data were analyzed with the MaxQuant framework [4], which is written in C# in the Microsoft .NET environment. Algorithmic parts of MaxQuant are available as source code and the entire program can be freely downloaded as well from www.maxquant.org. Detailed instructions for installation and support programs are also available [17].

Results and Discussion

Time and m/z Dependence of the Mass Error

We start by illustrating global features of the mass error distributions in liquid chromatography-tandem mass spectrometry runs. We performed 2 h LC-MS/MS runs of a HeLa lysate acquired on an LTQ Orbitrap without enabling the lock mass feature. Peptides were identified with a suitably large tolerance for the peptide mass, sufficient to include possible deviations due to instrumental drift (20 ppm in this case). We then calculated the mass error for each peptide. Figure 1 shows the results of these measurements for the elution profiles (MS-level isotope patterns) of four peptides. Indeed, all four peptides in Figure 1 are shifted by approximately 5 to 6 ppm because of the lack of calibration and because the lock mass feature was not applied. These mass deviations are far in excess of the sub-ppm accuracy

that the instrument is capable of [4]. It would now be interesting to determine if the masses are off due to a global shift, due to statistical fluctuations, or if the mass error is a function of either retention time or mass, or if it depends on both. The two doubly charged peptides in Figure 1a and b have approximately the same mass but differ in their retention times. One finds that the mass errors of the two peptides differ from each other by more than 1 ppm, suggesting a time dependence of the mass error. Likewise, we can compare the two triply charged peptides in Figure 1c and d, which have similar retention time but differ in mass. Again, the difference in the mass errors is more than 1 ppm, indicating an m/z dependence of the mass error.

To investigate time dependence of the mass error in a systematic manner we plotted the ppm mass error as a function of retention times in Figure 2a (red data points). Clearly, there are systematic effects in the mass error distribution. There is a tendency for the error to be slightly greater at larger elution times. In addition there is a wave-like pattern on the timescale of 10 to 20 min. Figure 2b shows a zoom of the same data into a smaller retention time interval. This figure reveals that the red curve has systematic structures on smaller timescales of about 1 min as well. The blue data points in Figure 2a and b are mass errors from the corresponding LC-MS/MS run in which the lock mass feature was enabled. As expected, the average mass error is now much closer to zero. However, also here larger deviations on smaller timescales can be seen. For instance, at $t=80$ min and $t=$

Figure 1. Three-dimensional views of three MS isotope patterns corresponding to peptides. Peptides (a) and (b) have similar mass but different retention times. Their mass errors differ by more than one ppm. Peptides (c) and (d) have similar retention time but differ in m/z . They also require different mass recalibrations

Figure 2. (a) Mass error in ppm for the peptides identified in two LC-MS runs as a function of retention time. Blue points originate from an LC-MS run in which the lock mass feature has been used while the red points are from an LC-MS run without lock mass. (b) Same data zoomed in the time window from min 76 to 90 min

83.5 min, the mass error rises for short times to 4 ppm. Inspection of the data files reveals that this is due to loss of the lock mass in these time intervals.

Figure 3a is a plot of the mass error as a function of m/z instead of elution time. Again a systematic nonlinear dependence can clearly be seen. These systematic variations seem to be only on larger scales of 100 Da without an indication of systematic effects on lower m/z scales. Histograms of these mass deviations are shown in Figure 3b for the data with lock mass (blue) and without lock mass (red). The lock mass helps in keeping the deviations near zero but does not completely center them there. This is partially due to the mass dependence of the error, which is not eliminated

by the lock mass. The tail to the right of the distribution is mainly derived from the time intervals where the lock mass has not been found. The absolute average mass deviation of the lock mass data is 0.53 ppm. In the data that were acquired without lock mass the errors are centered at ~6 ppm. The full width half maximum of both distributions is similar and, in both cases, around 1 ppm.

The Software Lock Mass Optimization Problem

As we have seen the mass error is a function of (at least) two variables, time and m/z , and projections onto each of them display rich structure and clear functional dependencies. It is

Figure 3. (a) Same data as in Figure 2a but plotted as a function of m/z . Red and blue points originate from LC-MS runs without and with lock mass, respectively. (b) Histograms of the mass errors shown in (a)

a reasonable assumption that the mass error depends on these two parameters in an additive way, i.e., the non-linearities in the mass scale should be independent of elution time. This assumption makes sense from a physical point of view. The m/z -dependent error is determined by static properties of the mass spectrometer that do not vary with time; for instance, imperfections in the geometry of the Orbitrap cell. In addition to this static error, a dynamic component is superimposed that is caused by any kind of disturbance that happens during the chromatographic time scale, with temperature shifts being a prominent example. This means that it should be possible to parameterize the mass error function as the sum of two terms, f and g , each depending on only one of the two variables and each being parameterized by sets of parameters θ_f and θ_g :

$$\Delta m = f(t, \theta_f) + g(m/z, \theta_g). \quad (1)$$

Note that this equation does not assume linearity in any of the variables or parameters, but only that the contribution

of each variable can be represented as a sum of nonlinear terms. The explicit form of the parameterization of the functions f and g are described below.

To determine the functions f and g in the above equation (equation 1), we generate number triples $(\Delta m_j, t_j, m/z_j)$ by performing a first peptide database search with the Andromeda search engine [16], which is integrated into the MaxQuant software package (see Figure 4). For this purpose, we allow a large tolerance of the peptide mass of 20 ppm. This initial tolerance can be set by the user. While we have found 20 ppm to be a good value for routine operation on this instrument class, this number can and should be increased in cases where the calibration is off by more than 20 ppm. All peptide identifications that have an Andromeda score of at least 80 are accepted. The mass error is then calculated based on the elemental composition of the identified peptides and the experimentally measured masses. Note that in MaxQuant the measured peptide mass is derived from the entire three-dimensional isotope pattern that the MS/MS spectrum was associated with [4]. MS/MS spectra not associated with a three-dimensional MS isotope pattern

Figure 4. Algorithmic steps of the software lock mass workflow

are discarded. The retention time is estimated as the intensity-weighted time average over the elution profile of the peptide. The computational task is now to determine the functions f and g in such a way that their sum best approximates the calculated mass error. To achieve this, we minimize the sum of squares of the residual errors (equation 2).

$$\sum_j \left(\Delta m_j - f(t, \theta_f) - g\left(\frac{m}{z}, \theta_g\right) \right)^2 \quad (2)$$

For this purpose, the functions f and g have to be parameterized in a suitable way. We use piecewise linear functions for f and g . First, the x-positions of these functions are adapted to the data, and they are then treated as constant during the minimization. The number of x-positions and their exact location are chosen such that the number of degrees of freedom adjusts itself to the complexity of the data. Roughly speaking, the more data are available the more complex the parameterizations of the functions can be. The x positions are chosen such that there are at least 80 data points per x position in the m/z direction and 50 data points

per x position in the time direction. Furthermore, the x positions have to be at least 50 Th apart in the m/z direction. The numbers that are being determined during the optimization are the y-values at these fixed x positions, typically several dozens or hundreds of coefficients. The functions f and g are linearly interpolated between these positions. One of the y-values has to be fixed to an arbitrary value since the system otherwise has a zero mode. The numerical solution of this minimization problem is obtained by the Levenberg-Marquard method (see, e.g., reference [18] for an introduction). After the parameters of f and g have been determined, we can subtract the systematic mass error from the measured mass of each MS isotope pattern in the LC-MS run. Subsequently, the actual database search (second pass search) is performed with individualized peptide mass tolerances inside the MaxQuant framework as before.

Performance of the Software Lock Mass

Figure 5 depicts the mass error distribution after recalibration and second pass Andromeda search. Figure 5a and b show the dependence on m/z while Figure 5c and d show the

Figure 5. Mass errors after recalibration. (a) Lock mass; plotted against m/z . (b) No lock mass; plotted against m/z . (c) Lock mass; plotted against retention time. (d) No lock mass; plotted against retention time

time dependence. For the data with and without lock mass our algorithm has removed all systematic effects from the data. Figure 6 shows histograms of the mass error after recalibration. The absolute average mass deviations are 0.29 ppm for the data with lock mass and 0.27 ppm for the data without lock mass. The corresponding mass standard deviations are 0.42 and 0.39 ppm, respectively. This indicates that when using the software lock mass workflow, the mass accuracy is as good as for data that were acquired with lock mass. While shown here for a particular example, we have found this to be true in general. As an example, Supplemental Figure 1 shows a very challenging LC-MS/MS run acquired with lock mass feature in which the lock mass was lost and found again several times. Panel a shows

Figure 6. Histograms of the mass errors after recalibration for data acquired with lock mass (blue), and without lock mass (red)

the time-dependence of the mass error before recalibration. The time series of the mass error is bi-stable, flipping back and forth between zero and 7 ppm. Nevertheless, this very difficult case is reliably recalibrated by the software lock mass (Suppl. Figure 1).

Conclusion

Here we have investigated the concept of a software lock mass, a replacement for its physical version, which is integrated into the MaxQuant/Andromeda computational proteomics workflow. We have demonstrated that it performs as least as well as the physical lock mass on typical complex proteome data. Even data that were acquired with a lock mass may benefit from the application of our recalibration workflow, especially in cases where the lock mass performance was not optimal. In contrast to the hardware lock mass option, the software lock mass can correct nonlinearities in the mass scale. Here, we have demonstrated the method on an Orbitrap instrument. However, we speculate that other instrument types would also benefit from the software lock mass approach. For instance, mass calibration drift typically is an issue of practical importance for time of flight instruments. Furthermore, while shown here for MS spectra, the benefits of the software lock mass also carry over to high-resolution MS/MS spectra.

Importantly, use of the software lock mass is completely free from an experimental point of view. All it requires is a peptide mixture of sufficient complexity. In contrast, a physical lock mass, even if derived from laboratory air, always has some experimental cost, such as additional hardware, influence on the spectra, or a slight increase in cycle time. Since the software lock mass is an unmitigated benefit, it can be adopted for all proteomics experiments, as we have done for some time in our laboratory.

Acknowledgments

The authors thank all the other members of the Proteomics and Signal Transduction group for help and discussions. This work was partially supported by the European Union 7th Framework Program (HEALTH-F4-2008-201648/PROSPECTS).

Open Access

This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction

in any medium, provided the original author(s) and source are credited.

References

1. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003)
2. Yates, J.R.I.I.: Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 297–316 (2004)
3. Yates III, J.R., et al.: Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6**, 702–714 (2005)
4. Cox, J., Mann, M.: MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008)
5. Zubarev, R., Mann, M.: On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteom.* **6**, 377–381 (2007)
6. Cox, J., Mann, M.: Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* **20**(8), 1477–1385 (2009)
7. Muddiman, D.C., Oberg, A.L.: Statistical evaluation of internal and external mass calibration laws utilized in fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **77**(8), 2406–2414 (2005)
8. Williams Jr., D.K., Muddiman, D.C.: Parts-per-billion mass measurement accuracy achieved through the combination of multiple linear regression and automatic gain control in a Fourier transform ion cyclotron resonance mass spectrometer. *Anal. Chem.* **79**(13), 58–63 (2007)
9. Hannis, J.C., Muddiman, D.C.: A dual electrospray ionization source combined with hexapole accumulation to achieve high mass accuracy of biopolymers in Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11**(10), 876–883 (2000)
10. Nepomuceno, A.I., Muddiman, D.C., Bergen, H.R. 3rd, Craighead, J. R., Burke, M.J., Caskey, P.E., Allan, J.A.: Dual electrospray ionization source for confident generation of accurate mass tags using liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **75**(14), 3411–3418 (2003)
11. Fenn, J.: Proceedings of the 34th ASMS, Cincinnati, OH; p. 507–509 (1986)
12. Schlosser, A., Volkmer-Engert, R.: Volatile polydimethylcyclsiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. *J. Mass Spectrom.* **38** (5), 523–525 (2003)
13. 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. Proteom.* **4**(12), 2010–2021 (2005)
14. Mortensen, P., Gouw, J.W., Olsen, J.V., Ong, S.E., Rigbolt, K.T., Bunkenborg, J., Cox, J., Foster, L.J., Heck, A.J., Blagoev, B., Andersen, J.S., Mann, M.: MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* **9** (1), 393–403 (2010)
15. Rappsilber, J., Ishihama, Y., Mann, M.: Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**(3), 663–670 (2003)
16. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M.: Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**(4), 1794–1805 (2011)
17. Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J.V., Mann, M.: A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.* **4**(5), 698–705 (2009)
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge, 683–688 (2007)