On the Proper Use of Mass Accuracy in Proteomics*

Roman Zubarev‡ and Matthias Mann§¶

Mass measurement is the main outcome of mass spectrometry-based proteomics yet the potential of recent advances in accurate mass measurements remains largely unexploited. There is not even a clear definition of mass accuracy in the proteomics literature, and we identify at least three uses of this term: anecdotal mass accuracy, statistical mass accuracy, and the maximum mass deviation (MMD) allowed in a database search. We suggest using the second of these terms as the generic one. To make the best use of the mass precision offered by modern instruments we propose a series of simple steps involving recalibration of the data on “internal standards” contained in every proteomics data set. Each data set should be accompanied by a plot of mass errors from which the appropriate MMD can be chosen. More advanced uses of high mass accuracy include an MMD that depends on the signal abundance of each peptide. Adapting search engines to high mass accuracy in the MS/MS data is also a high priority. Proper use of high mass accuracy data can make MS-based proteomics one of the most “digital” and accurate post-genomics disciplines.  


Most analytical chemistry textbooks warn that data of unknown accuracy are useless. Yet many proteomics researchers entering the field in the last few years were raised on the idea that mass accuracy is an unimportant parameter. Until recently, much proteomics research was done on three-dimensional ion traps, very sensitive and robust but low resolution instruments. Database searches were performed with windows of ±3 Da and for several alternative charge states. Modern instruments can do more than a thousand times better and are capable of mass accuracy around 1 ppm. Speaking of masses here and throughout the text, we mean of course m/z values. Contrary to the old-fashioned magnetic sector instruments that imposed a compromise between mass accuracy and sensitivity this high accuracy now comes “for free.” That is to say modern instruments can be both extremely sensitive and capable of very high mass accuracy. In fact, the high resolution achieved by these instruments concentrates the signal into a narrow mass range, improving signal to noise of the spectra. Here we argue that proteomics thinking has not caught up with these capabilities and that consequently we are not making the best use of high mass accuracy.

WHAT IS MASS ACCURACY?

Surprisingly although the mass is the primary parameter measured in the mass spectrometric experiment, the proteomics community has not agreed on clear definitions. The proteomics literature attaches at least three different meanings to the term mass accuracy.

ANECDOTAL MASS ACCURACY

This refers to the selective reporting of mass measurements, usually to demonstrate the capabilities of the author’s instrument. The literature is full of claims of very accurate measurements made on intrinsically not-so-accurate instruments, backed up with a single figure. Even for the highest resolution instruments, of the FTICR type, premature claims had been made of low ppm or sub-ppm mass accuracy. In practice, such performance in high throughput applications had to await solution of the space charge problem in FTICR, which only happened a few years ago. It is legitimate to report anecdotal mass accuracy, but it should clearly be distinguished from routine instrument performance in day to day use. In general, when reporting a single measurement we suggest to use the term “mass deviation,” defined as measured mass minus calculated mass, instead of the term mass accuracy. In other words, one should say “the mass was measured with a deviation of x ppm” instead of “the mass was measured with an accuracy of x ppm.”

STATISTICAL MASS ACCURACY

This is the mass accuracy estimated from a statistical distribution of mass errors. For example, the root mean square deviation of a particular mass may be reported by the manufacturer of a mass spectrometer as the mass accuracy of the instrument. The average absolute mass accuracy is a parameter easily calculated for any proteomics experiment and one that we have found to be quite informative of the quality of a measurement. A problem with the statistical mass accuracy may occur if the mass error does not follow the expected normal distribution. For instance, if the error distribution is...
Lorentzian instead of Gaussian the actual data dispersion $\sigma$ is infinite, and thus any single measurement is potentially quite inaccurate. Moreover in this case statistical error cannot be reduced by averaging $n$ independent measurements because $\sigma/(n - 1)^{1/2}$ is still infinite. It is thus important to make sure (e.g. by the $\chi^2$ test) that the assumed Gaussian shape of the error distribution is valid. Plotting a graph of the error distribution and fitting to it a Gaussian distribution can be useful. Such a graph together with the absolute average deviation and dispersion $\sigma$ are excellent indicators of actually achieved mass accuracy (see Fig. 1 for an example).

**MAXIMUM ALLOWED MASS DEVIATION**

This is the cutoff value used in database search. Only peptide sequences with a calculated mass within this tolerance are reported as hits. Maximum mass deviation (MMD)\(^1\) is the only operationally important parameter related to mass accuracy in a proteomics experiment.

Inevitably in the proteomics literature anecdotal mass accuracy is the highest followed by the statistical mass accuracy, whereas MMD is usually chosen at several times the statistical mass accuracy or on a general feeling what the worst mass deviation is likely to be. We propose to use the generic term mass accuracy for the second of these three categories. Anecdotal mass accuracy is subjective, and MMD should be chosen as a multiple of the dispersion of mass errors; thus statistical mass accuracy is the only objective and independent parameter. As there are different ways to model the measured mass distribution, the statistical model should also be mentioned, for example root mean square error or absolute average deviation.

**HOW SHOULD STATISTICAL MASS ACCURACY AND MMD BE RELATED?**

If mass errors were normally distributed, three standard deviations would capture 99.7% of the peptides. For example, a root mean square error of 2.5 ppm would lead to an MMD of 7.5 ppm. However, for low signal to noise peaks the distribution is broader, and mass accuracy is worse than the average value would suggest. For these reasons and because the systematic error is not eliminated, many researchers set the MMD very high, in effect degrading a high accuracy instrument into a medium accuracy one. By this they achieve maximum sensitivity. Alternatively in an attempt to maximize specificity they may decide to narrow the mass tolerance window to an unreasonably small value, but this only achieves the illusion of high mass accuracy at the expense of losing true positives. For determining the appropriate size of the window, the graph of the mass error distribution is indispensable.

**ELIMINATION OF SYSTEMATIC MASS ERRORS**

Mass measurement error is partly made up of the systematic component caused by factors such as miscalibration, temperature drift of the calibration, or space charge. For decades internal standards, in the form of an added compound of known mass, have been used in MS to eliminate the systematic error and to obtain very high mass accuracy. For example, “peak matching methods” were used to reduce the error to a few ppm, a requirement of organic chemistry journals to prove identity of a synthesized compound.

In proteomics experiments the internal standard comes for free because many of the thousand of MS and MS/MS measurements unambiguously identify peptides by their very high scores or by the fact that they are commonly occurring background peptides from keratins or trypsin autolysis products. We have used this procedure for a number of years in the open source program MSQuant (1) and find that it often improves mass accuracy severalfold, especially for TOF instruments (2). Because these internal standards are present in virtually every proteomics experiment there is no excuse for data sets with systematic error.

\(^1\) The abbreviation used is: MMD, maximum mass deviation.
A SIMPLE PROCEDURE TO OBTAIN THE BENEFITS OF HIGH MASS ACCURACY

To implement the above principles, we suggest routinely following the following steps.

(i) Search the data with a permissive MMD chosen to retain essentially all correct hits.

(ii) Recalibrate the mass scale by least square fitting using a few hundred hits from the highest scoring peptides and/or known contaminants.

(iii) Plot the mass errors of all identified peptides. Check whether error distribution is normal. If not, remove outliers and recalibrate the mass scale using only peptides with small mass deviations and repeat (iii).

(iv) Choose the MMD appropriate for the experiment from the distribution of mass errors in the plot. If the mass accuracy changed dramatically, such as more than a factor of 2 or 3, the search should be repeated with the new MMD.

(v) Retain or discard peptide hits based on this MMD.

These steps can readily be automated with the help of scripts implemented into proteomic data processing pipelines. Nonetheless we encourage search engine developers to implement them directly into their software. If the experiment is performed with different MS conditions or on different days, the procedure should be applied to each data set separately. The elimination of systematic error and the generation of the mass error plot can be performed completely automatically and by default. The software could let the user choose the desired trade-off between sensitivity and specificity by setting the MMD in terms of a certain multiple of the standard deviation or percentage of discarded true positives. To further help proteomics researchers embrace this quality-enhancing procedure, journals publishing proteomics results could encourage routine presentation of the plots of the distributions of mass deviations with corresponding statistical mass accuracy and the chosen MMD.

HOW ACCURATE IS ACCURATE ENOUGH?

High mass accuracy does imply high, at the very least isotopic, resolution and thus enables correct charge state determination and identification of the monoisotopic mass. This already means a specificity increase by a factor of 3–5 compared with unresolved isotopic distributions. Moreover as a rule of thumb the molecular mass of a peptide determined with 1 ppm mass accuracy rules out 99% of amino acid compositions possible for a given integer or nominal mass (3).

Thus total specificity improvement K (proportion of false possibilities filtered out) is 300–500 for ±1 ppm measurements compared with low resolution measurements. However, this estimate depends upon the number of possible alternatives, and for limited databases it can be much lower. An MMD of a few ppm is possible in routine practice (4–7), and consistent achievement of this value would eliminate the vast majority of false positive identifications currently reported in the literature.

A very accurately measured monoisotopic molecular mass can completely specify the elemental composition of the molecule. The abundances of the isotopic peaks may be used to achieve some filter for possible composition at lower mass accuracy, but high sensitivity measurements must rely first of all on monoisotopic masses. For peptides of 1 kDa an MMD of 0.2–0.3 ppm is required (3, 8), and this may become available in the future. At higher masses, even this MMD is not sufficient to determine the composition, but the database is more sparsely populated with tryptic peptides in this mass range. Thus a high mass accuracy measurement may suggest a single fully tryptic peptide for a large peptide.

High mass accuracy is also beneficial for finding related peaks, that is peptides that cover the same amino acid sequence but that have different mass due to modifications or missed enzymatic cleavage sites. For example, the program ModifComb correlates peptide masses and their fragmentation spectra to a “base peptide” that is unmodified (9), and the program MS-Alignment groups peptides with related MS/MS spectra (10). The better the mass accuracy, the better the correlation and grouping. More generally, the better the mass accuracy, the less the need for peptide sequencing. An extremely accurate mass, perhaps combined with the elution time of the peptide, could be sufficiently characteristic of a peptide to identify it. In our experience, this is not realistic when using proteomics in a “discovery mode” but becomes quite practical in a “remeasurement mode.” We have extensively used this concept in the “protein correlation profiling” technique in which thousands of peptides are quantified across protein fractions to distinguish true members of organelles from co-migrating background proteins (11, 12). Complete avoidance of peptide fragmentation is the basis of the “accurate mass and time” strategy (13), which will become much more reliable with the very high mass accuracy of the current generation of instruments. However, modern hybrid instruments can perform peptide fragmentation and acquisition of the tandem mass spectrum simultaneously with the acquisition of the survey mass spectrum, so it is always desirable to sequence at least a subpopulation of peptides.

MASS ACCURACY IN THE MS VERSUS MS/MS MODE

Several types of instruments achieve different mass accuracy in the MS versus the MS/MS mode. For example, orthogonal TOF instruments generally have good mass accuracy in the MS/MS mode but are often limited by ion statistics and consequently diminished mass accuracy for weak peaks and in the MS/MS mode. However, ever improving performance of TOF instruments makes them strong contenders for high mass accuracy proteomics. In hybrid instruments the survey spectrum is typically obtained in the high resolution part at the same time that the MS/MS spectrum is obtained at high sensitivity in the low resolution part of the instrument.
high mass accuracy in MS/MS mode is clearly desirable, especially for the identification of modified peptides or de novo sequencing. However, this often comes at the expense of sensitivity. Furthermore because the peaks are not statistically independent but are linked by the limited number of possible amino acid masses and the molecular mass, the total specificity improvement \( K \) in database searches for \( n \) peaks is smaller than \( K^n \). In general, low resolution measurements are more advantageous for small databases and only few possible modifications, whereas high mass accuracy in the MS/MS mode may be absolutely required in discovery of new or unexpected sequences and modifications. For peptides not in the database, high accuracy MS/MS data furthermore have the advantage of distinguishing between lysine and glutamine.

A further challenge to the proper use of high mass accuracy data is the fact that database searching algorithms do not yet make full use of accurate MS/MS data. For example, there is a threshold in mass accuracy beyond which the Mascot program fails to assign higher score to higher accuracy data. Likewise the multiple Fourier transformations associated with cross-correlation techniques become computationally prohibitive with high resolution data. Thus proteomics researchers do not obtain commensurately higher scores from higher resolution data, and as a result they are reluctant to acquire data in high resolution mode even when this is easily possible. Therefore we issue an urgent call to adapt peptide identification programs to high accuracy data.

MORE SOPHISTICATED USES OF MASS ACCURACY

The current use of mass accuracy is still quite crude. We usually set a fixed acceptance boundary for the MMD and completely accept anything that is within the boundary, whereas we completely reject everything outside. Furthermore peaks with a very abundant signal and good signal to noise are treated exactly the same as peaks close to noise level. We propose that mass accuracy should be treated in a more nuanced way by a flexible mass deviation window. Well defined peptide peaks must be treated much more stringently in terms of MMD than peaks close to the level of detection. Ideally each peptide peak should have its own mass tolerance. This could be based on repeat measurement over an LC peak (5); this also has the advantage that the mass taken from the LC apex where it can be determined best is given larger weight compared with the spectrum in which the peak was “picked” for sequencing and where it is often very weak. Furthermore peptide molecular masses should be determined from the whole available data set or data sets and not only from a single peak in a single mass spectrum.

In the future, mass deviation should be part of the overall database search score with increasing weight for peptides measured with lower mass deviation. Such a weighting has already been incorporated into at least one composite peptide identification score (14). The advantage of such composite scores, in addition to taking explicit account of the mass accuracy, is that they can relatively easily incorporate additional criteria, such as the shape of the isotopic distribution, the presence of satellite ions, deviation of the retention time of the peptide from the expected value, and so on.

In conclusion, MS-based proteomics can inherently be extremely precise and accurate. It is a strength of our field that our data rest on an objective and unambiguous parameter of each substance, its mass, and we should capitalize on this strength. This is especially true now when we can determine masses to many decimal places, bringing us close to the ultimate goal of determining the elemental composition. This high accuracy is in stark contrast to many other areas of biology, which have large intrinsic errors and gray zones of interpretation. It potentially makes MS-based proteomics one of the most “digital” and error-free of the life sciences at least as concerns peptide identification. It is up to us to make the most of this good fortune by taking the best possible advantage of mass accuracy.

Acknowledgment—We thank Jesper V. Olsen for helpful discussion and for the figure.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence may be addressed. E-mail: Roman.Zubarev@bmms.uu.se.

‡ To whom correspondence may be addressed. E-mail: mmann@biochem.mpg.de.

REFERENCES


