Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions

Michael W. Senko,* Steven C. Beu,† and Fred W. McLafferty
Department of Chemistry, Cornell University, Ithaca, New York, USA

The coupling of electrospray ionization with Fourier-transform mass spectrometry allows the analysis of large biomolecules with mass-measuring errors of less than 1 ppm. The large number of atoms incorporated in these molecules results in a low probability for the all-monoisotopic species. This produces the potential to misassign the number of heavy isotopes in a specific peak and make a mass error of ±1 Da, although the certainty of the measurement beyond the decimal place is greater than 0.1 Da. Statistical tests are used to compare the measured isotopic distribution with the distribution for a model molecule of the same average molecular mass, which allows the assignment of the monoisotopic mass, even in cases where the monoisotopic peak is absent from the spectrum. The statistical test produces error levels that are inversely proportional to the number of molecules in a distribution, which allows an estimation of the number of ions in the trapped ion cell. It has been determined, via this method that 128 charges are required to produce a signal-to-noise ratio of 3:1, which correlates well with previous experimental methods. *(J Am Soc Mass Spectrom 1995, 6, 229–233)*

Recent advances in ionization techniques like electrospray ionization (ESI) [1–4] and matrix-assisted laser desorption–ionization (MALDI) [5, 6] have allowed the analysis of large biomolecules that previously were unamenable to mass spectral techniques. ESI in combination with quadrupole mass analyzers has allowed mass determinations of 30-kDa proteins with mass errors of less than 100 ppm [7, 8]. To obtain these low mass errors, the spectrum must either be free of any species that could interfere with peak shape (residual solvent or cation adducts) or must be recorded with high enough resolving power to separate these interfering species. For larger molecules (> 30 kDa) where interfering species cannot be resolved, the best results are obtained by using the peak maximum to calculate average molecular mass, instead of peak centroids, because this emphasizes the sample's major component [9]. Ultimately, in the absence of sample heterogeneity and instrumental error, the accuracy for determination of the average molecular mass for an unknown protein will be limited to ~8 ppm because of the natural variability of the $^{13}C/^{12}C$ isotope ratio for biological compounds [10].

By using higher performance analyzers capable of resolving isotopic peaks for larger species, the $^{13}C/^{12}C$ variability is no longer a limitation because only the abundances, and not the positions of the isotopic peaks, will change. The coupling of ESI with Fourier-transform mass spectrometry (FTMS) [11, 12] has allowed mass determinations of 17-kDa proteins with milli-Dalton mass measuring errors [13]. This level of accuracy allows for the differentiation of the two isobaric amino acids, lysine (128.095 Da) and glutamine (128.056) in a 20 kDa molecule. Similar levels of accuracy have been obtained with a combination of ESI and sector instruments [14, 15].

Resolution of the isotopic peaks presents the new question of what mass value should be reported. The chemical average mass could be reported by a calculation of the average of the isotopic peaks (weighted by abundance), but this returns to the original problem that carbon isotope variabilities limit the mass accuracy [10]. High resolution measurements of large molecules typically have reported the mass of the most abundant isotope [12, 13, 15], and although this mass may be known with milli-Dalton accuracy, the assignment of a "true" mass may be in error by ±1 Da or more because of an incorrect assignment of the number of heavy isotopes that contribute to the most abundant peak.

The most significant and accurate mass that can be reported is that of the all-monoisotopic species (all ^{12}C, ^{1}H, ^{14}N, ^{16}O, and ^{32}S), because its mass will be unaffected by isotopic variabilities and, for smaller molecules (< 5 kDa), the appearance of the monoiso-

Address reprint requests to Dr. Fred McLafferty, Cornell University, Department of Chemistry, Baker Laboratory, Ithaca, NY 14853-1301.
*Present address: National High Magnetic Field Laboratory, Florida State University, 1800 E. Paul Dirac Drive, Tallahassee, FL 32306-4005.
†Present address: 12811 Steeple Chase Drive, Austin, TX 78729-7362.

1044–0305/95/$9.50
SSDI 1044-0305(95)00017-8

Received October 18, 1994
Revised December 21, 1994
Accepted December 29, 1994

topic peak prevents any possible isotopic misassignment. The increased probability for multiple heavy isotopes as the mass of a molecule increases causes a decrease in the relative abundance of the monoisotopic peak. For bovine ubiquitin (8.6 kDa) and equine apomyoglobin (16.9 kDa), the monoisotopic peaks are less than 4% and 0.04%, respectively, of the most abundant isotopic peaks, which makes observation of the monoisotopic peak unlikely for molecules larger than 15 kDa, given the signal-to-noise limitations of current instrumentation [13, 15]. This complication necessitates a method for identification of the number of heavy isotopes for a peak by using only a measured isotopic profile. Previously it was demonstrated that the monoisotopic mass of a molecule can be estimated with fairly high accuracy by knowing only the average molecular mass [16]. The work described here extends this method by comparing high resolution spectra with model isotopic distributions to assign more accurately the monoisotopic mass.

Error levels obtained via the statistical test for assignment of the monoisotopic mass correlate with the number of molecules that contribute to the isotopic distribution. The numbers obtained from the correlation can be used to estimate the number of ions trapped in the FTMS cell and provide direct confirmation of previous experimental measurements [17]. The number of ions that produce a detected signal is important in determination of absolute sensitivities, dynamic range, dissociation efficiencies, and trapping efficiencies. This method has an advantage over previous work in that it does not assume that the ion cloud is in a tight bundle in the trap midplane or that the excitation radius is known.

Experimental

All spectra were collected by using electrospray ionization on a Fourier-transform mass spectrometer previously described [13, 18]. Frequency-to-mass calibration was performed by using the most abundant isotopic peak of bovine ubiquitin for the 7+ through 12+ charge states. A 486-based PC and Visual Basic for Microsoft Windows were used to create statistical analysis routines and isotopic abundance profiles. Theoretical isotopic distributions were created by using a previously published method [19]. For estimation of trapped-ion numbers, empirical isotopic distributions were generated with Monte Carlo methods by using the exact probabilities for each isotopic combination generated by the theoretical method (i.e., by molecule, not by atom). Comparison of isotopic profiles was done with the total abundances for the isotopic peaks included in the calculation normalized to 1. Myoglobin spectra were acquired with limited ion populations and large excitation radii (80% of cell maximum) to minimize any potential space–charge distortions of the isotopic envelope [18]. All compounds were obtained from Sigma Chemical Co. (St. Louis, MO) and were used without further purification.

Results and Discussion

A method is required to determine the number of heavy isotopes represented by a specific isotopic peak. This problem is challenging for large molecules because of the negligible abundance of the all-monoisotopic species. When working with resolved isotopic envelopes, one might calculate an average molecular mass from the isotopic envelope and then apply the previously developed method for determination of monoisotopic mass from average mass [16]. However, this method will produce a result that is obviously inaccurate, because the calculated monoisotopic mass will not match precisely a possible isotopic position, and could differ by as much as 0.5 Da (or 25 ppm at 20 kDa).

A statistical comparison of the experimental isotopic distribution and a theoretical distribution will provide more accurate results and a greater level of certainty. The chi-square test is well established in the social sciences for comparison of measured populations to expected or theoretical populations. It takes the form

$$\chi = \sum (O_i - E_i)^2 / E_i$$

where E_i is the expected number of occurrences for group i and O_i is the observed occurrences for that same group. For applications in monoisotopic mass determination, the observed population is measured by using the ESI-FTMS instrument, and the expected population easily can be determined, given a molecular formula [19]. In the case of unknown compounds, however, the molecular formula is unavailable; the only known information is the molecular mass. Therefore, a method must be developed to obtain a model molecular formula based on molecular mass.

The previous method used for determination of monoisotopic mass from average mass assumed that all amino acids have the same probability of occurrence [16], but suggested that use of a true distribution of amino acids might provide better results. By using a uniform distribution, both cysteine and methionine, which are sulfur-containing amino acids, each would appear 5% of the time. The natural abundances of these two amino acids are actually less (1.9% and 2.3%, respectively) [20], which causes an overestimation of the isotopic contribution from sulfur and shifts the average molecular mass from the monoisotopic mass more than desired. By using the statistical occurrences of the amino acids from the PIR protein data base [20], an average amino acid was developed for use in modeling isotopic distributions. This model amino acid, *averagine*, has the molecular formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ and an average molecular mass of 111.1254 Da. For example, the use of

a uniform amino acid distribution for a model compound with an average molecular mass of 20 kDa would produce a compound with a monoisotopic mass of 19986.44, whereas determination of a similar model compound via the statistical distribution of amino acids has a monoisotopic mass of 19987.26—a 41-ppm mass difference.

To obtain a model molecular formula, the number of averagine units in a molecule is determined from the average molecular mass, and then this number is multiplied by the number of atoms of each element in an averagine residue. Because calculation of the theoretical isotopic profile requires integral numbers of atoms, the values obtained for C, N, O, and S are rounded to the nearest integer and the final average molecular mass is corrected by adjustment of the number of Hs. Rounding errors induced by the addition or subtraction of half a C, N, O, or S and numerous Hs do not shift the isotopic distribution a significant amount. For example, the 20-kDa model compound would be composed of 179.98 averagine units and should therefore contain 7.5 sulfur atoms. The abundances for the isotopic peaks obtained when the number of sulfurs is rounded down to 7 (while adding 16 hydrogens) differ by less than 1% relative to the isotopic abundances obtained when the number of sulfurs is rounded up to 8 (while subtracting 16 hydrogens), which is less than the typical scan-to-scan deviations of current instruments. With a model molecular formula, a theoretical isotopic profile is then created [19]. For spectra obtained via electrospray ionization, isotopic distributions can be enhanced by deconvolution of the profiles from multiple charge states [13, 21] before comparison to theoretical profiles; this reduces both statistical and systematic errors.

By using the assumption that any difference between the measured and theoretical profiles is due to statistical fluctuations and not instrumental discrimination, simple error analysis can be used to find the best match between profiles. Direct application of the chi-square test is inappropriate for comparison of isotopic profiles because it requires populations with integral constituents, and isotopic abundances are known only in relative terms. For this case, total abundances for both theoretical and experimental profiles are normalized to 1 before the chi-square formula is used. The comparison is made by initially offsetting the measured profile an integral number of mass units to the left. The error term is then recalculated after the experimental isotopic profile is shifted to the right 1 Da each time. The shift requires renormalization of the theoretical profile because each shift removes a low mass isotope and adds a high mass isotope. The offset that produces the smallest error term is the best statistical match.

The best results are obtained when the area for comparison is limited to the larger isotopic peaks in the center of the distribution. Isotopic peaks from the edge of the envelope with low expected abundances will tend to dominate the calculation and provide erratic results because small differences between the expected and observed abundances are amplified by the small E_i in the denominator of the error term. The abundances of the peaks from the edge of the isotopic envelope are also more vulnerable to distortion due to experimental error. For clean samples, it has been observed that low abundance isotopic peaks at the edges of the envelope are typically smaller than expected or even absent (unpublished results), which may be due to the recently described phenomenon of peak confluence [22]. For impure samples, residual solvent and cation adducts will affect the edge of the envelope more than the center because of relative proximity. For the best results, it has been determined empirically that only peaks of $>20\%$ abundance should be included in error term calculations.

This system was used to verify the correction by Zaia et al. [23] of the amino acid sequence for horse myoglobin. By using a measured average mass of 16,951 Da, the isotopic distribution for a 152.5-unit poly-averagine molecule with a molecular formula of $C_{753}H_{1206}N_{207}O_{225}S_6$ is constructed; this isotopic distribution compares favorably with the true molecular formula of $C_{769}H_{1212}N_{210}O_{218}S_2$ (Figure 1). Although this model molecule may miss the true average molecular mass by up to 0.5 Da, the important feature is the similarity to the true isotopic profile for myoglobin rather than its absolute position on the mass scale.

For the experimental spectrum, multiple charge states for horse myoglobin were deconvoluted to improve isotopic abundances [21]. A comparison of the isotopic profiles of the 11 central isotopic peaks generated error terms of 0.271, 0.053, **0.011**, 0.085, and 0.281 for shifts of −2, −1, **0**, 1, and 2 Da from the true value respectively (Figure 2). From the model isotopic profile, the mass of the monoisotopic peak is known to be 10 Da less than the most abundant isotopic peak, which permits a monoisotopic mass assignment of 16941.14 Da, which is a 0.18-Da (10-ppm) error. This estimation of the monoisotopic mass is possible even though the first visible isotopic peak in the deconvoluted spectrum contains three heavy isotopes. Use of the masses of all isotopic peaks instead of just the most

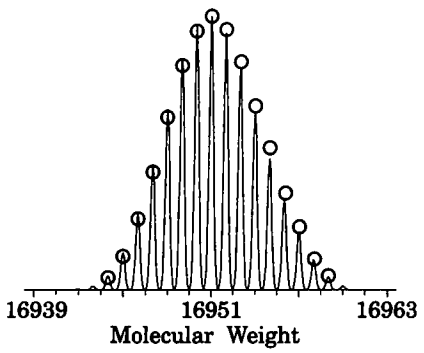

Figure 1. Theoretical isotopic distribution of equine myoglobin and overlay (O) of model distribution.

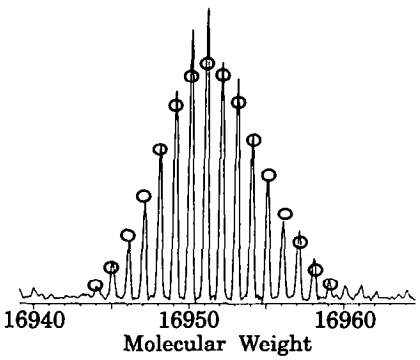

Figure 2. Experimental isotopic distribution of myoglobin obtained by deconvolution of multiple charge states and overlay (O) of model distribution.

abundant peak does not improve results in this case because the mass measurement error is systematic and is caused by ion populations that differ during measurement and calibration. The use of an internal calibrant [13, 15] in combination with the chi-square test should allow the determination of monoisotopic masses for molecules in excess of 15 kDa with errors of less than 1 ppm.

This statistical test provides an opportunity to estimate the number of ions observed for a spectrum if the relationship between ion numbers and error rates can be determined. Estimates for the number of molecules measured in the FTMS trapped-ion cell can be made by using the fact that the error for an isotopic profile varies inversely with the number of contributing molecules. To determine the exact relationship, Monte Carlo methods were used to generate isotopic profiles of myoglobin that included 200, 400, 600, 800, or 1000 molecules. Average error rates for the comparison of the 11 central peaks from the empirical and exact isotopic profiles were determined from 1000 simulations performed at each level. Linear regression was then performed ($r^2 = 1.00$) to obtain the equation

$$\text{molecules} = 10.0/\text{error} - 0.8$$

Error rates were determined for the measured isotopic distributions for the 15+ charge state from five different spectra of myoglobin. For an undamped signal of 0.827 s produced by ions at a radius of 80% of the cell maximum by using an open elongated cylindrical cell [18, 24], 4.92 ± 3.67 ions (or 73.8 ± 55.1 charges) were necessary to obtain a signal-to-noise ratio of 3, as defined by Limbach et al. [17].

For a comparison to the previous work, the value is adjusted to compensate for cell geometry, cell radius, and acquisition time. The elongated cylindrical geometry of this trap produces approximately 20% greater signal than a cubic trap [25], the excitation radius of $0.8r$ should produce 60% greater signal than a radius of $0.5r$ [25], and the 0.827-s acquisition time produces 11% less signal than the specified 1-s acquisition [17]. No corrections are made for cyclotron frequency differences (83 kHz here versus 600 kHz for Limbach et al.) because above 10 kHz, the signal-to-noise ratio should be independent of frequency [17]. Thus, to obtain a signal-to-noise ratio of 3:1 from a 1-s signal at an ion radius of 50% of cell maximum would require 128 singly charged ions. This statistical method assumes that no systematic error occurs in measurement of the isotopic profiles, so the value of 128 ions should be considered a lower limit for obtaining a signal-to-noise ratio of 3:1. This value compares favorably with Limbach's value of 177 ions, as well as a more recent estimation by Bruce et al. [26] of ≈ 100 ions obtained by observations of large, multiply charged single ions.

This method benefits from the fact that it makes no assumptions about the shape and position of the ion cloud nor display any consequences of the shape and position due to radial excitation prior to detection. This is a benefit in that the number of ions in the cell can be measured for any ion cloud in any initial position. An additional advantage of this method is that no knowledge of excitation radius is necessary for calculate of the number of ions in the cell.

Conclusion

A method has been developed for determination of the monoisotopic mass of large molecules by using only an experimental isotopic profile. Although the monoisotopic peak is not visible in the high-resolution spectra of equine myoglobin, its value can be determined with an error of 10 ppm via this method, which eliminates previous concerns regarding a ± 1-Da error due to an incorrect assignment for the number of heavy isotopes that contribute to a specific isotopic peak. To improve the accuracy of the method, any prior knowledge of amino acid composition could be incorporated into the model, which should prove particularly beneficial for much larger (> 40-kDa) molecules that incorporate an unusual number of cysteines or methionines. An additional benefit of the statistical test used is the ability to estimate the number of molecules that produce a specific isotopic profile, and thus the number of ions needed to create a specific signal in the FTMS. This method for determination of the number of ions is preferable to previous experimental methods in that no knowledge of ion cloud shape or position are necessary.

Acknowledgments

We gratefully acknowledge the valuable advice and assistance of K. D. Henry, P. A. Limbach, P. B. O'Connor, J. P. Quinn, and T. D. Wood, and the financial support of the National Institutes of Health (Grant GM16609) and the Society of Analytical Chemists of Pittsburgh (summer fellowship for MWS).

References

1. Teer, D.; Dole, M. *J. Polym. Sci.* **1975**, *13*, 985.
2. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64.

3. Huang, E. C.; Henion, J. D. *J. Am. Soc. Mass Spectrom.* **1990,** *1,* 158.
4. Smith, R. D.; Loo, J. A.; Ogorzalek Loo, R. R.; Busman, M.; Udseth, H. R. *Mass Spectrom. Rev.* **1991,** *10,* 359–451.
5. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. Int. *J. Mass Spectrom. Ion Processes* **1987,** *78,* 53–68.
6. Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T. *Anal. Chem.* **1991,** *63,* A1193–A1203.
7. Loo, J. A.; Udseth, H. R.; Smith, R. D. *Anal. Biochem.* **1989,** *179,* 404–412.
8. Smith, R. D.; Loo, J. A.; Edmonds, C. G.; Barinaga, C. J.; Udseth, H. R. *Anal. Chem.* **1990,** *62,* 882–899.
9. Loo, J. A.; Edmonds, C. G.; Smith, R. D. *Anal. Chem.* **1991,** *63,* 2488–2499.
10. Beavis, R. C. *Anal. Chem.* **1993,** *65,* 496–497.
11. Henry, K. D.; Williams, E. R.; Wang, B. H.; McLafferty, F. W.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci., U.S.A.* **1989,** *86,* 9075.
12. Henry, K. D.; Quinn, J. P.; McLafferty, F. W. *J. Am. Chem. Soc.* **1991,** *113,* 5447–5449.
13. Beu, S. C.; Senko, M. W.; Quinn, J. P.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1993,** *4,* 190.
14. Cody, R. B.; Tamura, J.; Musselman, B. D. *Anal. Chem.* **1992,** *64,* 1561–1570.
15. Dobberstein, P.; Schroeder, E. *Rapid Commun. Mass Spectrom.* **1993,** *7,* 861–864.
16. Zubarev, R. A.; Bondarenko, P. V. *Rapid Commun. Mass Spectrom.* **1991,** *5,* 276.
17 Limbach, P. A.; Grosshans, P. B.; Marshall, A. G. *Anal. Chem.* **1993,** *65,* 135.
18. Beu, S. C.; Senko, M. W.; Quinn, J. P.; Wampler, F. M.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1993,** *4,* 557–565.
19. Yergey, J. A. *Int. J. Mass Spectrom. Ion Phys.* **1983,** *52,* 337.
20. Protein Identification Resource, National Biomedical Research Foundation.
21. Labowski, M.; Whitehouse, C. M.; Fenn, J. B. *Rapid Commun. Mass Spectrom.* **1993,** *7,* 71.
22. Naito, Y.; Inoue, M. *J. Mass Spectrom. Soc. Jpn.* **1994,** *42,* 1–9.
23. Zaia, J.; Annan, R. S.; Biemann, K. *Rapid Commun. Mass Spectrom.* **1992,** *6,* 32.
24. Beu, S. C.; Laude, D. A. *Int. J. Mass Spectrom. Ion Processes* **1992,** *112,* 215–230.
25. Grosshans, P. B.; Shields, P. J.; Marshall, A. G. *J. Chem. Phys.* **1991,** *94,* 5341–5352.
26. Bruce, J. E.; Cheng, X.; Bakhtiar, R.; Wu, Q.; Hofstadler, S. A.; Anderson, G. A.; Smith, R. D. *J. Am. Chem. Soc.* **1994,** *116,* 7839–7847.