Efficient Calculation of Accurate Masses of Isotopic Peaks

Alan L. Rockwood
ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, Utah, USA

Perttu Haimi
Department of Biochemistry, Institute of Biomedicine, University of Helsinki Helsinki, Finland

This paper presents a new method for calculating accurate masses of isotopic peaks. It is based on breaking the calculation into a binary series of calculations. The molecule is built up by a series of such calculations. At each step the accurate masses are calculated as a probability weighted sum of the masses of the contributing peaks. The method is computationally efficient and accurate for both mass and relative abundance. (J Am Soc Mass Spectrom 2006, 17, 415–419) © 2006 American Society for Mass Spectrometry

This paper addresses the problem of calculating the accurate masses of isotopic peaks in an isotopic distribution of a compound of known chemical formula. For the purposes of this paper an "isotopic peak" consists of all molecules having the same number of nucleons, regardless of the isotopic fine structure of the peak. For example, the compound CO has an isotopic peak containing 29 nucleons. Although this peak has an isotopic fine structure with contributions from both $^{12}C^{17}O$ and $^{13}C^{16}O$, each of which has a slightly different mass, for the purposes of the present paper these will not be considered as two separate peaks but as a single isotopic peak with an accurate mass of 28.998297 Da.

There are several methods for calculating masses of isotopic peaks. The accuracy and speed of an isotopic calculation both depend strongly on the method used. Kubinyi [1] and Rockwood and Van Orden [2] have described fast methods that also produce semi-accurate masses. The method of reference [2] is extremely fast and generally produces results within a few millimass units [2, 3]. Although semi-accurate masses are useful for many applications, in some cases being more accurate than experimental measurements of masses [3], for other applications accurate masses would be preferable.

Accurate masses of isotopic peaks can be calculated by several methods. Polynomial-based methods work well for low and medium molecular weight compounds, but become computationally inefficient at high molecular weights. Attempts to accelerate the polynomial-based methods by applying "pruning", which is the omission of small terms from the calculation, involve a tradeoff between speed and accuracy [4, 5].

In a second method, one might start with an algorithm that calculates profile-mode isotopic distributions centered on the accurate masses [4]. Numerical values of the accurate masses of the isotopic peaks would then be extracted by numerical quadrature or by peak fitting of the nominal isotopic peaks. However, considerable computational effort would still be involved in such a method.

A third method is to perform a series of stepwise convolutions, building the full molecule by the stepwise addition of atoms to an accumulated "super atom". Masses of the individual isotopic peaks are calculated at each convolution step as a probability weighted sum of the contributions to each peak [3]. This method is accurate, but computational efficiency is not optimal.

Recently, a new algorithm was presented for the calculation of accurate masses of isotopic peaks, and application examples were discussed [6]. This method combines very high computational efficiency with very high accuracy. In it, one first calculates the isotopic composition of each isotopic peak and then uses this information, together with the accurate elemental isotopic masses, to calculate the accurate masses of the molecular isotopic peaks. Although the algorithm in reference [6] is computationally efficient, it generates "extraneous" information (isotopic compositions), and although the algorithm is not difficult to apply, the proof of the algorithm can be difficult to grasp.

Here we present an alternative algorithm. It generates both accurate masses and accurate isotopic abundances. This method is accurate and computationally efficient. Unlike the method presented in reference [6], the algorithm presented here is relatively easy to understand, and it does not generate isotopic composition information for the individual isotopic peaks.

Published online February 3, 2006

Address reprint requests to Dr. A. L. Rockwood, ARUP Institute for Clinical and Experimental Pathology, 500 Chipeta Way, Salt Lake City, UT 84108, USA. E-mail: rockwoal@aruplab.com

1044-0305/06/$32.00
doi:10.1016/j.jasms.2005.12.001

Received October 31, 2005
Revised December 2, 2005
Accepted December 2, 2005

The method presented here is related to both the Kubinyi method [1] and the Roussis-Prouix method [3]. It is based loosely on Kubinyi's general approach of subdividing the calculation into a binary series of steps. While the present paper departs somewhat from the exact sequence of steps described by Kubinyi, it shares with the Kubinyi method the computational advantage of breaking the calculation into a binary series. It shares with the Roussis-Prouix method the feature that the algorithm produces accurate masses because it uses probability-weighted sums for the mass calculation.

Algorithm Description

Consider two types of clusters, which we will refer to as "super atoms", terminology borrowed from Roussis and Prouix [3]. In the present paper, a "molecular super atom" refers to a fictitious chemical compound whose formula is a partial composition of the target compound. An example is C_2H, which is a partial composition of C_2H_5Br. An "elemental super atom" refers to an elemental cluster, which in our treatment is always selected from a binary series, such as H_1, H_2, H_4, $H_{8,}$. . .

In the present algorithm we iteratively update the compositions of the molecular super atom and the elemental super atom, always building the composition of the molecular super atom toward the final composition of the target compound. In addition, the probabilities and accurate masses of the isotopic peaks of both super atoms are updated at each step. At the end of the calculation, the molecular super atom has the chemical formula of the target compound, and the probabilities and accurate masses of the molecular super atom are those of the target compound.

In this process, we build the chemical composition of the target compound by systematically joining smaller pieces of the molecule. For example, at one point in the process of generating C_2H_5Br we may join the molecular super atom C_2H with the elemental super atom H_4 to make new molecular super atom C_2H_5.

The algorithm uses two kinds of vectors, probability vectors and mass vectors. The indices into the vectors represent nucleon number. The elements of the probability vectors represent isotopic probabilities. The elements of a mass vector represent the corresponding accurate masses. Each nominal isotopic peak is thus characterized by two parameters, a probability and an accurate mass.

Let $f_p(j)$ represent the *jth* element of the probability vector for one of two fragments that will be joined together, and let $f_m(j)$ represent the *jth* element of the mass vector for the same fragment. (Here, "element" is used in the mathematical sense, i.e., element of a vector, not in the chemical sense, i.e., chemical element.) Let $g_p(i)$ represent the *ith* element of the probability vector for the second fragment. This second fragment will be joined with the first to make a more complex molecule, which we have previously referred to as a super atom. Let $g_m(i)$ represent the *ith* element of the mass vector for the second fragment. Let $h_p(k)$ and $h_m(k)$ represent the *kth* elements of the probability and mass vectors, respectively, for the resulting combination. The values of $h_p(k)$ are given by the following equation [7].

$$h_p(k) = \sum_i g_p(i) f_p(k-i) \tag{1}$$

where we have changed the notation slightly from reference 7 by the addition of the subscript *p*. The accurate masses, $h_m(k)$, are given by the following expression

$$h_m(k) = \left(\sum_i g_p(i) f_p(k-i)\right)^{-1} \sum_i g_p(i) f_p(k-i)(g_m(i) + f_m(k-i)) \tag{2}$$

Although this equation is presented without proof, it is based on the idea that the mass of a nominal isotopic peak is a probability-weighted sum of the masses of the isotopic peaks of the fragments that are being combined. This is similar to an approach briefly described elsewhere [3].

As mentioned above, at each step in the calculation, we generate a new molecular super atom and a new elemental super atom. This entails updating the compositions, probabilities, and masses of both super atoms. In the process, we use the binary representation of the chemical formula of the target compound. For example, the binary representation of $C_2H_5Br_1$ is $C_{\mathit{10}}H_{\mathit{101}}Br_{\mathit{1}}$, where italicized subscripts represent binary numbers and non-italicized subscripts represent normal base ten numbers. Unlike conventional notation we have explicitly included a subscript for the single bromine atom in the compound.

To update the super atoms, we systematically work through binary representation of the chemical composition of the target compound, starting with the least significant digit of the first element in the chemical formula. If the current binary digit is *0*, then we do nothing to the molecular super atom, but we do update the elemental super atom using a procedure described later. If the value of the current binary digit is *1*, then we convolute the chemical composition of the current elemental super atom with the chemical composition of the current molecular super atom. We update the probabilities and masses for the molecular super atom by applying eqs 1 and 2. We also update the elemental super atom using a process described later. We repeat this process until we have examined all digits in the chemical formula of the current element. Then we start on the next element in the chemical formula of the target compound.

To update the elemental super atom we generally double its composition and update its probabilities and masses using eqs 1 and 2. However, if we have finished with the current element, we update the elemental super atom by replacing it with one atom of the next element in the chemical formula of the target com-

Table 1. Effect of pruning to mass accuracy and intensity accuracy for DNA oligomer $(ACGT)_{1000}$

Pruning limit	Weighted RMS mass difference (ppb)	RMS intensity difference	Running time(s)
1.00×10^{-6}	1.31×10^{-1}	1.75×10^{-4}	0.02
1.00×10^{-7}	5.77×10^{-2}	2.23×10^{-5}	0.02
1.00×10^{-8}	2.28×10^{-2}	1.52×10^{-6}	0.02
1.00×10^{-9}	1.09×10^{-2}	6.89×10^{-8}	0.02
1.00×10^{-10}	5.93×10^{-3}	2.06×10^{-8}	0.02
1.00×10^{-12}	3.20×10^{-4}	1.76×10^{-10}	0.02
1.00×10^{-14}	5.79×10^{-6}	1.50×10^{-12}	0.02
1.00×10^{-16}	5.71×10^{-8}	1.14×10^{-14}	0.03
1.00×10^{-18}	6.04×10^{-10}	1.06×10^{-16}	0.03
1.00×10^{-20}	0	8.26×10^{-18}	0.03
1.00×10^{-25}	0	6.33×10^{-26}	0.03
1.00×10^{-30}	0	0	0.03
1.00×10^{-40}	0	0	0.04
1.00×10^{-50}	0	0	0.04
1.00×10^{-100}	0	0	0.04
0	N/A	N/A	0.26

The mass differences are the differences of calculated mass between corresponding peaks at subsequent pruning limits. Weighted RMS difference is the root mean square summary statistic of the mass differences, weighted by the square root of peak intensity. RMS intensity difference is the root mean square from the intensity differences between corresponding peaks. Running times are the average of five measurements. Program was compiled with gcc 3.3.5 with command line parameter -O3. The tests were run on a computer with Intel 3.4 GHz Pentium 4 processor.

pound, including substitution of the relevant probabilities and masses.

Other than loading dummy vectors for an initialization step and terminating the algorithm, the procedure given in the previous two paragraphs presents the complete algorithm. However, the description given above is somewhat misleading because apparently one has to take explicit account of the chemical compositions of the super atoms at each step. It can be shown that this is not necessary because the computational bookkeeping involved in examining the digits of the binary form of the target compound composition implicitly tracks the chemical compositions.

The algorithm above can be pictured as a ladder with some missing rungs. The left rail of the ladder represents a series of molecular super atoms, and the right rail represents a series of elemental super atoms. The rungs indicate the updating of the molecular super atom. If a rung is present, the updating step consists of convoluting the elemental super atom with the molecular super atom. If a rung is absent, the molecular super atom is carried forward without modification. The presence or absence of rungs is determined by digits in the binary representation of the composition, *1* representing the presence of a rung and *0* representing the absence of a rung. The updating of the elemental super atom (the right rail) is either a doubling of the elemental super atom or the replacement of the super atom, depending on whether we have examined all the binary digits for a given element or not.

To avoid unwieldy array sizes, we recommend that a form of pruning be applied at intermediate steps in the calculation. The basis of this strategy has been described elsewhere [1, 3]. Briefly, at each stage in the calculation, one prunes small peaks from the super atoms based on the application of a pruning threshold, i.e., peaks below or equal to a certain threshold are removed. We suggest a pruning threshold in the range of 1×10^{-25} to 1×10^{-30} as a good compromise between speed and accuracy (see Table 1), though this threshold might be raised or lowered, depending on the needs of the calculation. It can be shown that this form of pruning is relatively benign in its effect on peak intensities, unlike pruning in polynomial calculations, which can lead to large errors in calculated probabilities for nominal isotopic peaks [4]. Interestingly, even a pruning threshold of zero will result in some pruning in the case of very high molecular weight compounds as a result of floating point underflow.

An important programming detail is to always use a pruning threshold, even if it is set to zero. Otherwise, peaks that have underflowed will continue to be carried through the calculation at a considerable cost of computational time and memory usage.

Benchmarks and Discussion

Other than pruning and roundoff error, the algorithm described above is exact. To benchmark the algorithm, we have implemented the method in a computer program called emass. Masses of isotopic peaks calculated by emass for $C_2Br_3Cl_3$ agree with those using a previously published algorithm [6] to well under ppb levels. Interestingly, the semi-accurate masses produced by a double precision version of an algorithm described in [2] have only a 4.3 ppm weighted RMS average difference from the present method, quite good considering the fact that the semi-accurate mass algorithm described in reference [2] makes no pretense at being a true accurate mass algorithm.

A larger molecule, oligomer $(ACGT)_{1000}$ has 0.06 ppb weighted average RMS mass difference when com-

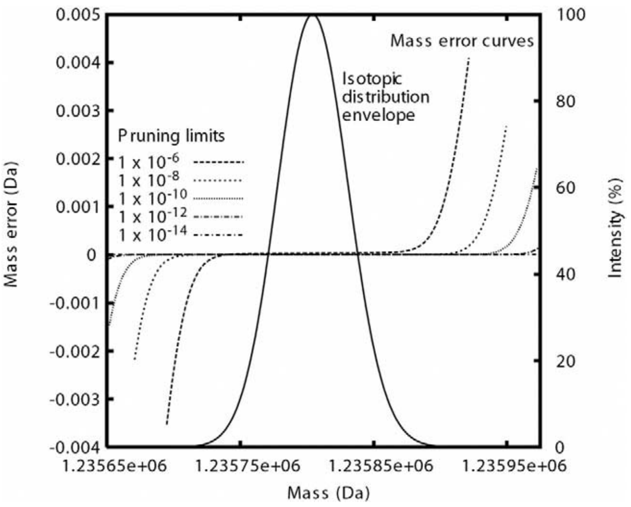

Figure 1. Effect of pruning on mass accuracy. The isotopic distribution for $(ACGT)_{1000}$ was calculated with pruning limit 0 (solid line indicating isotopic envelope). Mass differences between calculations made with pruning limit 0 (assumed to be essentially exact) and other pruning limits are plotted with non-continuous lines.

pared to results using the algorithm described in reference [6] and 0.5 ppb weighted average RMS mass difference when comparing to the semi-exact algorithm. In these comparisons, pruning threshold 1×10^{-30} was used, and peaks smaller than 5×10^{-7} percent of the largest peak were ignored.

The peak intensities calculated by both methods agree well, having 2.3×10^{-15} RMS difference for $C_2Br_3Cl_3$ and 2.6×10^{-13} RMS difference for $(ACGT)_{1000}$.

As mentioned above, in this algorithm the effect of pruning on mass accuracy is relatively benign. Table 1 illustrates this for $(ACGT)_{1000}$. The error entries in the table are incremental values, i.e., they represent the differences between successive calculations as the pruning threshold is decreased. Even at a relatively aggressive (i.e., high) pruning threshold of 1×10^{-6}, the calculation is already very accurate, with sub ppm errors in masses and small errors in intensities. By the time the threshold reaches 1×10^{-30}, there is no further improvement, indicating that for practical purposes the calculation is error free. The computation time is weakly dependent on the pruning threshold, as shown in the last column of Table 1, so there is very little penalty in setting a very conservative pruning threshold.

Figure 1 illustrates a different aspect of accuracy as a function of pruning threshold. The salient point is that mass errors are concentrated in the tails of the isotopic distribution. Even at a relatively aggressive pruning threshold of 1×10^{-6} it is only in the extreme tails of the distribution that the mass error exceeds 0.001 Da, and these peaks would be difficult to see in an experimental mass spectrum, so the practical effect of even rather aggressive pruning is minimal. Also obvious in the figure is that fact that errors drop even further when the pruning threshold is decreased to more conservative values.

Comparison to the algorithm described in reference [6] confirms the excellent accuracy. The algorithm from reference [6], including error correction strategy 2 described in that reference, was implemented as a computer program named qmass and compared against the emass program described above. For the $(AGCT)_{1000}$, the calculated masses for all peaks with abundance greater than 0.001% of the base peak agreed to better than one part per trillion between the two methods, and for peaks greater than 1×10^{-6} percent of the base peak the agreement was better than 50 parts per billion.

When pruning is applied, the number of peaks in the isotope distributions grows approximately as the square root of the number of atoms in the molecular formula. This leads to linear time dependency as the number of atoms in the formula increases, which was also confirmed by benchmarks (Figure 2). If no pruning is done, the time dependency becomes quadratic. Notable in the benchmark is the overall speed of the method. Even for the largest molecules in the benchmark (molecular weight $>1.23 \times 10^8$ Da) the calculations ran in well under 1 s.

The time dependency of the algorithm described in reference [6] is theoretically better: $O(nk) = \sqrt{n}\log(n)k$, where n is the number of atoms in the molecular formula and k is the number of different elements. However, in our implementation (qmass), the advantage is not seen in molecules of realistic sizes. In both methods the computational time scaled approximately linearly with the number of atoms, and emass ran slightly (~40%) faster than qmass.

Computational efficiency is always desirable, but

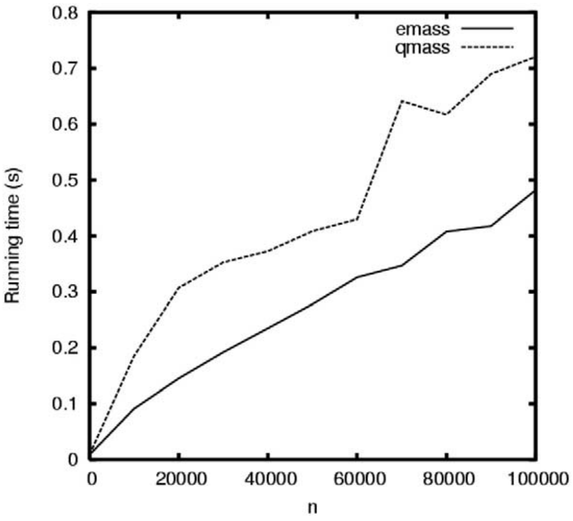

Figure 2. Effect of molecular complexity on running time. The isotopic distribution for DNA oligomers of form $(ACGT)_n$ with various values of n were calculated by emass and algorithm described in [6] (qmass). Pruning limit 1×10^{-30} was used. Programs were compiled and run as described in Table 1.

becomes especially important if one must perform a series of calculations. An application of this type discussed elsewhere [6] is to use accurate masses of isotopic peaks to deduce elemental composition. This requires that one perform a series of accurate isotopic calculations on many different chemical compositions.

Here we propose strategies that might be used to improve the efficiency of such calculations. These could be used either individually or in combination. They use the concept of screening, using fast computational methods before the application of accurate mass algorithms for the full isotopic distribution. One might first screen the compounds by comparing the measured average molecular weight against the calculated average molecular weight for a trial composition and then comparing the experimental isotopic profile against a semi-accurate mass isotopic calculation, such as one using the algorithm described in reference [2]. If these results are within an acceptable error tolerance (perhaps 20 ppm weighted RMS error), then one would progress to an accurate mass method, such as the one described in the present paper or the one described earlier [6]. We believe that these strategies could be almost as fast as mono-isotopic calculations, but would have the advantage of using information from the full isotopic distribution to help deduce chemical composition.

Computer Code Availability

Source code for the programs emass and qmass is released under BSD license and is available as a Supplementary Material section of this paper (which can be found in the electronic version of this article.). The Supplementary Material section also contains documentation for the programs, including instructions for compiling the program under both Linux and Microsoft Windows operating systems, as well as compiled versions of the programs for computers running under Microsoft Windows operating systems. The programs may also be found at the website http://www.helsinki.fi/science/lipids/software.html. This website will generally contain the most current versions of the programs. Questions and discussions about the programs are best directed to Perttu Haimi at perttu.haimi@helsinki.fi, and questions and discussions about the algorithm itself are best directed to Alan Rockwood at rockwoal@aruplab.com.

Summary

By basing the calculation on a binary series of compositions, and by calculating accurate masses at each step in the method using a weighted sum approach, one can calculate isotopic probabilities and accurate masses with high efficiency and accuracy.

Acknowledgments

Support for this work was provided by the ARUP Institute for Clinical and Experimental Pathology and in part by grant number 55081 from the Finnish Academy to the laboratory of PH.

References

1. Kubinyi, H. Calculation of Isotope Distributions in Mass Spectrometry. A Trivial Solution for a Non-Trivial Problem. *Anal. Chim. Acta* **1991,** *247,* 107–109.
2. Rockwood, A. L.; Van Orden, S. L. Ultrahigh-Speed Calculation of Isotope Distributions. *J. Am. Soc. Mass Spectrom.* **1996,** *68,* 2027–2030.
3. Roussis, S. G.; Prouix, R. Reduction of Chemical Formulas from the Isotopic Peak Distributions of High-Resolution Mass Spectra. *Anal. Chem.* **2003,** *75,* 1470–1482.
4. Rockwood, A. L.; Van Orden, S. L.; Smith, R. D. Rapid Calculation of Isotope Distributions. *Anal. Chem.* **1995,** *67,* 2699–2704.
5. Yergey, J. A. A General Approach to Calculating Isotopic Distributions for Mass Spectrometry. *Int. J. Mass Spectrom. Ion Phys.* **1983,** *52,* 337–349.
6. Rockwood, A. L.; Van Orman, J. R.; Dearden, D. V. Isotopic Composition and Exact Masses of Single Isotopic Peaks. *J. Am. Soc. Mass Spectrom.* **2004,** *15,* 12–21.
7. Rockwood, A. L.; Kushnir, M. M.; Nelson, G. J. Dissociation of Individual Isotopic Peaks: Predicting Isotopic Distributions of Product Ions in MS^n. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 311–322.