Anal. Chem. **2007,** *79,* 6912−6921

Stable Isotope Assisted Assignment of Elemental Compositions for Metabolomics

Adrian D. Hegeman,[†,‡] Christopher F. Schulte,[†] Qiu Cui,[†] Ian A. Lewis,[†] Edward L. Huttlin,[†,‡] Hamid Eghbalnia,[†] Amy C. Harms,[‡] Eldon L. Ulrich,[†] John L. Markley,[†] and Michael R. Sussman*[,†,‡]

Department of Biochemistry, University of Wisconsin, 433 Babcock Drive, Madison, Wisconsin 53706, Biotechnology Center, University of Wisconsin, 425 Henry Mall, Madison, Wisconsin 53706

Assignment of individual compound identities within mixtures of thousands of metabolites in biological extracts is a major challenge for metabolomic technology. Mass spectrometry offers high sensitivity over a large dynamic range of abundances and molecular weights but is limited in its capacity to discriminate isobaric compounds. In this article, we have extended earlier studies using isotopic labeling for elemental composition elucidation (Rodgers, R. P.; Blumer, E. N.; Hendrickson, C. L.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* 2000, *11,* 835−40) to limit the formulas consistent with any exact mass measurement by comparing observations of metabolites extracted from *Arabidopsis thaliana* plants grown with (I) ^{12}C and ^{14}N (natural abundance), (II) ^{12}C and ^{15}N, (III) ^{13}C and ^{14}N, or (IV) ^{13}C and ^{15}N. Unique elemental compositions were determined over a dramatically enhanced mass range by analyzing exact mass measurement data from the four extracts using two methods. In the first, metabolite masses were matched with a library of 11 000 compounds known to be present in living cells by using values calculated for each of the four isotopic conditions. In the second method, metabolite masses were searched against masses calculated for a constrained subset of possible atomic combinations in all four isotopic regimes. In both methods, the lists of elemental compositions from each labeling regime were compared to find common formulas with similar retention properties by HPLC in at least three of the four regimes. These results demonstrate that metabolic labeling can be used to provide additional constraints for higher confidence formula assignments over an extended mass range.

Metabolomics seeks to characterize both the identities and quantities of metabolites in defined biological systems.[1] Unlike proteomics and genomics, in which single platform technologies can be employed within dynamic range limitations for the characterization of bulk protein or nucleic acid, metabolomics is challenged by the much larger degree of analyte chemical diversity. The high degree of chemical diversity has created the need for a careful evaluation of metabolite extraction techniques.[2−4] Although it is usually desirable to maximize information-gathering capacity and minimize analysis time and handling, the high degree of metabolite chemical diversity necessitates utilization of multiple analytical platforms.[5] To date, two general platforms, NMR and mass spectrometry (MS), provide complementary information in the analysis of complex mixtures of metabolites.[6] NMR has the capacity to characterize chemical structure and quantity but is limited to the 20−50 most abundant compounds in a given sample without isotope labeling. MS techniques are many orders of magnitude more sensitive and can detect many more compounds per a unit time, but are limited to ionizable species, have difficulties resolving isomers, and usually require standard compounds for quantification. This report focuses on a strategy for extending the analytical capabilities of MS-based metabolomics surveys.

Several MS metabolomics approaches have emerged that have different strengths and weaknesses. One that is particularly effective involves using MS/MS-based reaction monitoring strategies (either MRM or SRM) with standard compounds to quantify a limited set of specific metabolites. This technique has several advantages in that it can be used to focus on a specific subportion of metabolism, and it can provide absolute quantities for each analyte with accompanying standard compound LC and fragmentation data to confirm analyte identities. The technique has practical limitations to the number of analytes that can be characterized in a single experiment, but has nevertheless been applied to great effect in appropriate scenarios.[7,8] Broader survey approaches are attractive, not just because of the large number (often thousands) of compounds that are typically observed, but because they can be used as an exploratory tool for unanticipated species. These approaches typically involve either GC/MS, which relies on compound fragment libraries for analyte identification, or LC/MS, which relies on intact molecular ion exact mass measurements for elemental composition assignment. GC/MS is fairly mature and is quite effective for specific classes of compounds, but it is limited to those that are volatilizable, either inherently or after derivatization. Although the LC/MS strategies are more generally applicable and at least as sensitive, they are also challenged by compound assignment ambiguity on a couple of levels. First, without comparison to standard compounds, for example, by coelution, MS/MS fragmentation pattern, or both, it

* To whom correspondence should be addressed. Phone: (608) 262-8608. Fax: (608) 262-6748. E-mail: msussman@wisc.edu.

[†] Department of Biochemistry.

[‡] Biotechnology Center.

(1) Glinski, M.; Weckwerth, W. *Mass Spectrom. Rev.* **2006**, *25*, 173−214.
(2) Want, E. J.; O'Maille, G.; Smith, C. A.; Brandon. T. R.; Uritboonthai, W.; Qin, C.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 743−752.
(3) Villas-Boas, S. G.; Hojer-Pedersen, J.; Akesson, M.; Smedsgaard, J.; Nielsen J. *Yeast* **2005**, *22*, 1155−1169.
(4) Kimball, E.; Rabinowitz, J. D. *Anal. Biochem.* **2006**, *358* (2), 273−280.
(5) Ryan, D.; Robards, K. *Anal. Chem.* **2006**, *78*, 7954−7958.
(6) Dunn, W. B.; Bailey, N. J.; Johnson, H. E. *Analyst* **2005**, *130*, 606−625.
(7) Lu, W.; Kimball, E.; Rabinowitz, J. D. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 37−50.
(8) Shou, W. Z.; Magis, L.; Li, A. C.; Naidong, W.; Bryant, M. S. *J. Mass Spectrom.* **2005**, *40*, 1347−1356.

10.1021/ac070346t CCC: $37.00
Published on Web 08/21/2007

can be impossible to distinguish isomeric compounds. Second, owing to the rapid expansion of the number possible molecular composition assignments as a function of exact mass, unique formulas are typically assignable only for compounds <200−250 Da, depending on instrument mass accuracy. When thousands of compounds are observed in a single sample, it can be impractical or impossible to resolve ambiguity through analysis of standard compounds. One strategy to minimize this ambiguity is to focus only on spectral features that are changing dramatically across some biologically interesting test condition.[9] Once these changing features are singled out, resources such as MS/MS characterization and comparison with standard compounds can be brought to bear upon the much smaller number of ambiguously assigned compounds. The direct comparison of spectra from independent analyses is useful only with fairly similar samples where matrix effects are fairly consistent from sample to sample. As sample composition diverges, an increasing portion of spectral changes will be the consequence of indirect effects. Still, it is often desirable to know what you can and cannot detect and which species are unchanging across a test condition.

The primary purpose of this manuscript is to propose a means to dramatically reduce ambiguity associated with assignment of elemental composition using stable isotope metabolic labeling. The concept is not new and has been used on a small scale for the assignment of a unique formula to a 851-Da lipid molecule[10] and for additional peptide identification constraints in proteomics.[11−13] We are, however, the first to report the application of these constraints to large numbers of MS metabolite observations in an automated fashion, and we have made the resources available for general use at the following URL: http://www.bmrb.wisc.edu/metabolomics/mass_query.php.[14] The extent to which isotopic labeling can be used to constrain formula assignments has not been rigorously examined in the literature, and so the first portion of the Results and Discussion section will be devoted to this topic.

Kind and Fiehn[15] have proposed an elegant alternative strategy for extending the mass range of unique assignments by evaluating natural abundance isotopic envelope peak intensities to further constrain allowable elemental compositions. This approach has a clear advantage in that it does not require isotopic labeling and is applicable to a wider range of samples without added expense. Its effectiveness, however, has not been demonstrated for large numbers of metabolite measurements. Furthermore, it may be difficult to identify defined and distinct isotopic envelopes for each species in an automated fashion because of sample characteristics such as complexity, dynamic range, and chemical noise.

Considerable expertise has been accumulated regarding stable isotopic metabolic labeling of many different organisms. In many cases, strategies for incorporating ^{15}N and ^{13}C into bacteria and unicellular organisms have been previously described or are fairly easy to adapt from existing protocols. *Saccharomyces cerevisiae* has been both ^{15}N- and ^{13}C-labeled for quantitative proteomics applications.[12] A fair number of multicellular organisms have also been ubiquitously ^{15}N metabolically labeled for quantitative proteomics applications, including *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Rattus norvegicus*.[16,17] Plants have also been ^{15}N-labeled in cell culture for both quantitative proteomics and metabolomics.[18−21] Additionally, *Solanum tuberosum* (potato) has been ^{15}N-labeled hydroponically for NMR structural biology experiments,[22] and we have developed techniques for labeling *Arabidopsis thaliana* seedlings[13] with both ^{15}N and ^{13}C (herein). Distinctions in types of metabolic labeling between those useful for relative quantification and those used for flux analysis, for example, have been reviewed.[23]

For many organisms, including most microbes, ubiquitous metabolic labeling can be performed quite easily and at a reasonable expense. As an additional benefit, metabolic labeling can be used for quantitative comparisons of biological samples while simultaneously aiding compound identification. A number of laboratories have used ^{15}N metabolic labeling for relative quantification of metabolites from yeast[24] and plant cell cultures.[19−21] Although these studies provide quantitative information for nitrogen-containing metabolites, a significant number of metabolites do not contain nitrogen and would be absent from the analyses. By including both ^{15}N and ^{13}C in our metabolomic formula assignment routines, we are preparing to implement a relative quantification strategy that will benefit from both the extended mass range for unique formula assignment and a labeling strategy for quantification that includes the vast majority of metabolites. These efforts will draw from our experience in automating high-throughput metabolic-labeling-based quantitative proteomic analyses.[13,25,26]

METHODS AND MATERIALS

Materials. Unless specified, all chemicals were purchased from Sigma-Aldrich (St. Louis).

Plant Growth and Extraction. *Arabidopsis* seedlings (Columbia eco-type; Lehle Seeds, Round Rock, TX) were grown in liquid culture using media containing Murashige and Skoog (MS) salts (per liter of water: 6.2 mg of boric acid, 332.2 mg of $CaCl_2$, 0.025 mg of $CoCl_2{\cdot}6H_2O$, 0.025 mg of cupric sulfate × 5 H_2O, 37.26

(9) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(10) Rodgers, R. P.; Blumer, E. N.; Hendrickson, C. L.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 835−840.

(11) Zhong, H.; Marcus, S. L.; Li, L. *J. Protein Res.* **2004**, *3*, 1155−1163.

(12) Snijders, A. P. L.; de Vos, M. G. J.; Wright, P. C. *J. Proteome Res.* **2005**, *4*, 578−585.

(13) Nelson, C. J.; Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *Proteomics* **2007**, *7*, 1279−1292.

(14) Markley, J. L.; Anderson, M. E.; Cui, Q.; Eghbalnia, H. R.; Lewis, I. A.; Hegeman, A. D.; Li, J.; Schulte, C. R.; Sussman, M. R.; Westler, W. M.; Ulrich, E. L.; Zolnai, Z. *Pac. Symp. Biocomput.* **2007**, *12*, 157−168.

(15) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2006**, *7*, 234−244.

(16) Krijgsveld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H. A.; Heck, A. J. R. *Nat. Biotechnol.* **2003**, *21*, 927−931.

(17) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Matthews, D. E.; Yates, J. R., 3rd. *Anal. Chem.* **2004**, *76*, 4951−4959.

(18) Gruhler, A.; Schulze, W. X.; Matthiesen, R.; Mann, M.; Jensen, O. N. *Mol. Cell. Proteomics* **2005**, *4*, 1697−1709.

(19) Kim, J. K.; Harada, K.; Bamba, T.; Fukusaki, E.-i.; Bobayashi, A. *Biosci. Biotechnol. Biochem.* **2005**, *69*, 1331−1340.

(20) Harada, K.; Fukusaki, E.-i.; Bamba, T.; Sata, F.; Kobayashi, A. *Biotechnol. Prog.* **2006**, *22*, 1003−1011.

(21) Engelsberger, W. R.; Erban, A.; Kopka, J.; Schulze, W. X. *Plant Methods* **2006**, *2*, 14−25.

(22) Ippel, J. H.; Pouvreau, L.; Kroef, T.; Gruppen, H.; Versteeg, G.; van den Putten, P.; Struik, P. C.; van Mierlo, C. P. *Proteomics* **2004**, *4*, 226−34.

(23) Beynon, R. J.; Pratt, J. M. *Mol. Cell Proteomics* **2005**, *4*, 857−872.

(24) Lafaye, A.; Labarre, J.; Tabet, J.-C.; Ezan, E.; Junot, C. *Anal. Chem.* **2005**, *77*, 2026−2033.

(25) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *Mol. Cell Proteomics* **2007**, *6*, 860−81.

(26) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2007**, *6*, 392−398.

mg of disodium ethylenediaminetetraacetic acid, 27.8 mg of ferrous sulfate × 7 H_2O, 180.7 mg of $MgSO_4$, 16.9 mg of $MnSO_4$, 0.25 mg of Na_2MoO_4, 0.83 mg of KI, 170 mg of KPO_4 monobasic, 8.6 mg of $ZnSO_4 \cdot 7\ H_2O$, 1.65 g of NH_4NO_3, and 1.9 g of KNO_3) and MES (2-morpholinoethanesulfonic acid; 2.5 mM; pH 5.7) with 1% glucose. Media for ^{15}N, ^{13}C, or ^{15}N and ^{13}C labeling were prepared by substituting (1.69 g/L) [$^{15}N_2$]ammonium nitrate and (1.92 g/L) [^{15}N]potassium nitrate or [$^{13}C_6$]glucose (Isotech, Dayton, OH) or both for natural abundance salts, sugar, or both to the appropriate proportions. Plants were grown at room temperature (22−23 °C) with continuous illumination and orbital shaking (30 rpm). Following 12 days of growth, plants were removed from the media, rinsed briefly with distilled water, and spun in a commercial kitchen salad spinner to remove excess water prior to flash freezing in N_2(l). Ten frozen leaves were taken from frozen plant material representing each isotopic enrichment regime: (I) natural abundance, (II) ^{13}C-labeled, (III) ^{15}N-labeled, and (IV) $^{13}C/^{15}N$ double labeled. Each batch of leaves was transferred into a microcentrifuge tube and pulverized in 500 μL of 80% methanol/water using a small plastic pestle on ice. The four tubes were subjected to centrifugation to remove large particulate mater. The bright green supernatant was transferred to a new tube and evaporated to dryness. The resulting residue was resuspended in 200 μL of 0.1% formic acid in water and exhibited an accompanying color change from green to yellow/brown.

MALDI-TOF Analysis. A small portion (1 μL) was taken from each leaf extract and combined for MALDI-TOF analysis. A portion of the combined extracts (0.5 μL) was mixed with 0.5 μL of saturated α-cyano-4-hydroxycinnamic acid matrix in 70:30 acetonitrile/water (vol:vol) on a MALDI target and air-dried prior to analysis. Data were collected on a Bruker Biflex III MALDI-TOF and averaged over 200 laser pulses.

LC/ESI-TOF Analysis. Metabolite extracts (5 μL of each) were analyzed by LC/MS using an Agilent LC/MSD-TOF equipped with an Agilent 1100 series capillary LC pump. Samples were subjected to C18 reversed-phase chromatography prior to electrospray ionization (ESI) using a 1 mm × 150 mm, Inertsil C18, 100-Å pore size, 5-μm particle size HPLC column with a constant flow rate of 25 μL/minute. Each sample was eluted over a 120-min gradient: 100% buffer A (0.1% formic acid in water) at time zero, 5% buffer B (0.1% formic acid in acetonitrile) at 5 min, 100% buffer B at 90 min, holding at 100% B for 5 min, back down to 5% B at 100 min, and isocratic at 5% B until 120 min. Blank runs with 5-μL buffer A injections were performed between each sample analysis. ESI TOF analysis was performed in positive ion mode. A reference mass solution containing purine (+1 ion exact mass = 121.050 873) and hexakis (1*H*,1*H*,3*H*-tetrafluoropropoxy)phosphazine (+1 ion exact mass = 922.009 798) in methanol were continuously introduced into the TOF via a second orthogonal ESI source and used for internal mass correction throughout on a spectrum-by-spectrum basis.

Data Processing. Features corresponding to small molecule monoisotopic masses were extracted from .wif proprietary data files using the program MassHunter (1.0.0.0) version 11 (Agilent Technologies, Santa Clara, CA). Feature extraction parameters were set to their defaults, except that the "Use all the available data" check box was selected instead of a defined data range. Briefly, the other settings included signal/noise ratio of 5 for the spectral peak detection threshold, 500 × 1000 for the maximum spectral peaks to use, NO for the peptidic isotopic envelope, multiple for the maximum charge, NO for the salt dominated; K^+ and Na^+ are possible adducts. Batch mode database searches of feature lists (both calculated and database approaches) were performed using tools that we have made available at the Biological Magnetic Resonance Data Bank (BMRB) website: http://www.bmrb.wisc.edu/metabolomics/mass_query.php.[14] The database of linear combinations of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur atoms was generated using published formula constraints.[27] Briefly, formulas were included if they (1) have a mass less than 2000 Da; (2) pass the SENIOR rules; (3) contain no more than 32 nitrogen, 63 oxygen, 6 phosphorus, or 8 sulfur atoms; and (4) fall within atom-to-carbon ratios of 0.2−3.1 H/C, 0−1.3 N/C, 0−1.2 O/C, 0−0.3 P/C, and 0−0.8 S/C. All other calculations and data processing were performed using a combination of scripts written in-house (which are available upon request unsupported) and simple data manipulation using Excel (Microsoft, Redmond, WA) or Mathematica v. 5.2 (Wolfram Research, Champaign, IL).

RESULTS AND DISCUSSION

Mass Spectrometry-Based Assignment of Molecular Formula. Subsequent to the development of high-resolution mass spectrometers capable of making high accuracy monoisotopic mass measurements, it has been possible to assign molecular composition to detectable chemical species within a certain mass range. Formula assignments may be made by comparison to known compounds or calculated by taking the sum of variable numbers of atoms of a limited set of atom types (typically C, H, N, O, S, and P for metabolites) and their known atomic masses. This can be accomplished by solving a Diophantine equation in which one can determine multiple variables for a smaller number of constraining equations as long as the variables are limited to integer solutions.[28] Monoisotopic (or exact mass) measurements are required for assignment because they do not vary with natural abundance isotopic composition, which can cause small but significant changes in average mass values. Suitable spectral resolution is required to resolve the components of isotopic envelopes over relevant charge states. Most importantly, one's capacity to assign a unique formula to a monoisotopic mass is a function of two variables: the mass value and the accuracy of the measurement. The correlation of the numbers of possible formulas to mass value is complex and will be discussed in depth below. Briefly, the number of formulas increases approximately with the magnitude of the mass but also depends on formula constraints and on where the query mass value falls in the distribution of possible values about a specific nominal mass. The consequences of mass accuracy provide simple linear constraints on the numbers of theoretically assignable formulas.

An increase in mass accuracy (typically indicated by a parts-per-million error) results in a decrease in the number of possible matching formulas. LC/TOF instrumentation, for example, using internal standards may reasonably provide 3−5 ppm mass ac-

(27) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2007**, *8*, 105−125.
(28) Hsu, C. S. *Anal. Chem.* **1984**, *56*, 1356−1361.
(29) van Breemen, R. B.; Canjura, F. L.; Schwartz, S. J. *J. Agric. Food Chem.* **1991**, *39*, 1452−1456.

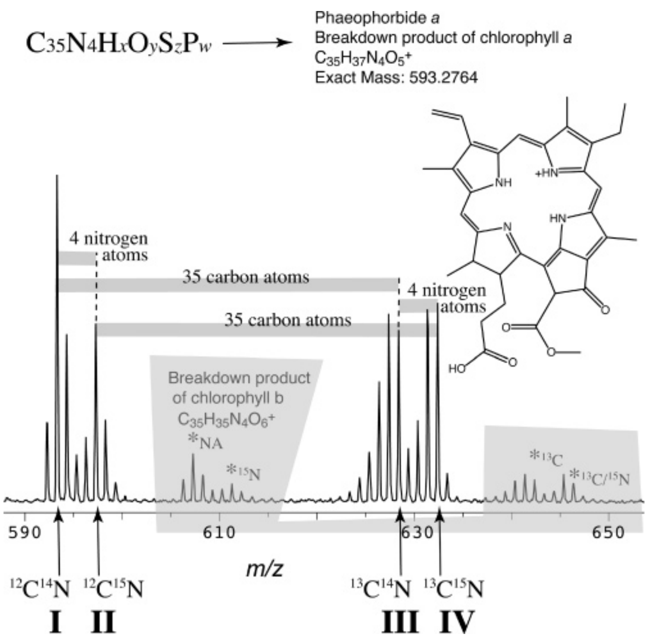

Figure 1. MS spectrum of a mixture of metabolically labeled extracts. Methanolic extracts of *Arabidopsis* leaves were pooled from (I) $^{12}C^{14}N$, (II) $^{12}C^{15}N$, (III) $^{13}C^{14}N$, and (IV) $^{13}C^{15}N$ metabolically labeled plants and analyzed by MALDI-TOF MS. The dominant species in the spectrum is phaeophorbide *a*, a well-known acid catalyzed breakdown product of chlorophyll *a*; the structure is shown in the inset.[29] The analogous and less abundant product of chlorophyll *b* is also visible. Both of these compounds give fairly intense MALDI spectra, possibly due to their selective absorption of the nitrogen laser (337 nm) and a concomitant ionization enhancement.

curacy measurements, allowing assignment of one or two molecular compositions to species under 200−250 amu. Platforms capable of higher mass accuracy, such as FTICR and possibly the Orbitrap, extend the upper mass limits for unique assignments but will still require the inclusion of additional constraints to provide unique formula assignments above 400−500 amu. Kind and Fiehn[15] have proposed an algorithm that evaluates the shape of the natural abundance isotopic envelope to provide these additional constraints. In cases that it is possible to perform metabolic stable isotopic labeling, we propose two approaches that consistently provide unique assignments up to 500−600 amu (at 3 ppm mass accuracy) and can provide unique assignments as high as 1200 amu.

Stable Isotopic Labeling Provides Constraints for Enhanced Formula Assignment. In the course of investigating ^{15}N and ^{13}C metabolic labeling of *A. thaliana* for NMR- and MS-based metabolite quantification, we found that we could obtain carbon and nitrogen counts fairly easily by examining spectra of pooled metabolites of different isotopic labeling regimes. This was first made apparent in examining the MALDI-TOF spectrum in Figure 1. Because the isotopes ^{15}N and ^{14}N, and ^{13}C and ^{12}C each differ by a single nominal mass unit, one can directly compare the monoisotopic masses for the four labeling regimes (I, II, III, and IV) to derive the numbers of nitrogen and carbon atoms in the chemical formula. With these additional atom constraints, one may effectively remove both C and N variables from the Diophantine equation, significantly decreasing the number of possible solutions for a given mass input.

Similar strategies to take advantage of atom number constraints have been used previously, none of which have been applied to bulk metabolites. Rodgers and co-workers[10] reported a 3-fold enhancement in the mass range for unique formula assignments using an FTICR (10 ppm mass accuracy) to characterize a 851-Da membrane lipid from ^{13}C-labeled *Rhodococcus rhodochrous*. They reported that the carbon constraint allowed the number of assignable formulas to be reduced from 394 to 1 for the single species. Others[15] have indicated, as elaborated below, why single-compound characterizations may be misleading for the evaluation of atom count constraint effectiveness. More recently, ^{15}N metabolic labeling has been used for proteomic studies as a check or additional constraint for peptide identifications.[11−13] None of the previous reports attempted to utilize multiple atom constraints or double-labeled materials, although the idea was suggested by Rodgers and co-workers.[10] Double labeling allows one to derive the N and C counts from two different sets of peaks. Of the four labeling regimes, characterization of three monoisotopic masses would provide the numbers of nitrogen and carbon atoms. This has advantages in the analysis of complex mixtures in which coanalyte interference can occur.

Formula Assignment Specificity as a Function of Mass. As indicated above, the correlation of the numbers of possible formulas to mass is complex and has two subcomponents, the magnitude and the "mass error". Mass error refers to the extent to which a mass value deviates from its nominal mass value and depends on the number and type of non-carbon atoms in the formula (^{12}C = 12.000 000 amu exactly by definition). The elements hydrogen and nitrogen have positive contributions to mass error (monoisotopic atomic weights: ^{1}H = 1.007 825 amu; ^{14}N = 14.003 074 amu), and oxygen, phosphorus, and sulfur have negative contributions to mass error (monoisotopic atomic weights: ^{16}O = 15.994 914 amu, ^{31}P = 30.973 761 amu; ^{32}S = 31.972 070 amu).[30] Depending on the balance of H and N to O, P, and S in the elemental composition, a mass falls somewhere in a roughly normal distribution about the nominal mass value. It is possible to calculate all possible combinations of a set of atoms using their exact mass values to theoretically constitute the complete set of formulas and masses within a nominal mass distribution. For these calculations to be meaningful, the list of formulas must also be restricted by using rules of chemical connectivity, such as those of Lewis and Senior. Briefly, these rules stipulate that (1) the total number of odd valence atoms must be even, (2) the sum of the valences is greater than or equal to twice the maximum valence, and (3) the sum of valences is greater than or equal to twice the number of atoms minus 1.[15,31] It is informative to examine how masses calculated from a significant population of formulas are distributed. For example, Figure 2 shows the distributions for (panel A) the masses calculated for all formulas with nominal mass value 519 and (panel B) nitrogen and carbon counts across those distributed formulas from panel A. Note that the number of nitrogen atoms (B) is always odd, as is stipulated by the "nitrogen rule".[32] The subdistributions for nitrogen and carbon (panel C) do not mirror the larger unconstrained distribution (see how the top of the $N = 3$ distribution is

(30) De Laeter, J. R.; Böhlke, J. K.; De Bièvre, P.; Hidaka, H.; Peiser, H. S.; Rosman, K. J. R.; Taylor, P. D. P. *Pure Appl. Chem.* **2003**, *75*, 683−800.
(31) Morikawa, T.; Newbold, B. T. *Chemistry (Romania)* **2003**, *12* (6), 445−450.

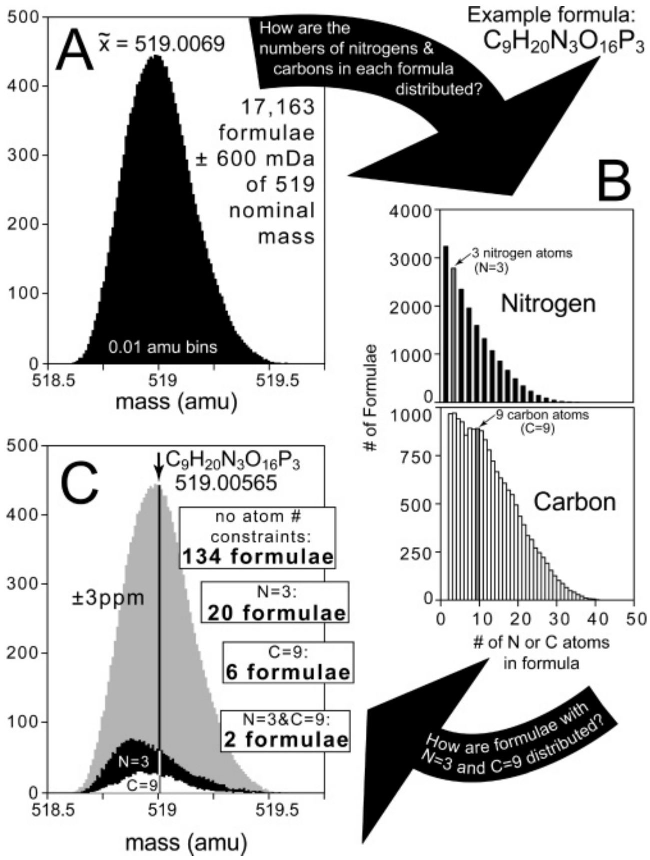

Figure 2. Distributions of calculated elemental compositions around nominal mass 519. Panel A shows the distribution of masses calculated for formulas with an even number of electrons that obey the Lewis and Senior rules, composed of C, H, N, O, P, and S (2−2000 C, 3−3000 H, 0−500 N, 0−500 O, 0−50 P, and 0−50 S) about the nominal mass value 519. Here, 17 163 formulas fall within ±0.6 Da of the nominal mass, and there is no significant overlap with adjacent nominal mass distributions. Panel B shows the distribution of nitrogen and carbon counts across the formulas from panel A. Triphosphoserine (formula $C_9H_{20}N_3O_{16}P_3$ and monoisotopic mass 519.005 65 amu) falls close to the top of the 519 nominal mass distribution and thus serves as a "worst case" example for mass-based formula assignment. The subset of formulas containing the same numbers of nitrogen (N = 3) and carbon (C = 9) atoms as triphosphoserine are indicated by the gray bars in panel B, and their distributions are shown in panel C. A search for formulas within ±3 ppm of the mass of triphosphoserine (519.005 65 amu) yielded 134 formulas with no atom constraints, 20 formulas with 3 nitrogen atoms, 6 formulas with 9 carbon atoms, and 2 formulas with both atom constraints.

significantly less than 519), and so the effectiveness of the atom constraints will vary significantly depending on the individual formulas.

To understand the extent of the advantages associated with atom number constraints in the assignment of formulas, it is desirable to create a mathematical model. By picking a mass for a formula from or near the *mode* of the nominal mass distribution, we can model a "worst case scenario" in which the given mass value will result in the largest possible number of formulas. Unfortunately, we cannot simply search for the number of solutions for the mode mass value, but must have the mass of a specific formula so that we can apply the atom constraints in the calculation of the formulas. We found that the masses of mono- and poly(phosphoserine) (n = 1, 2, 3, ..., 10, 15) fall very close to

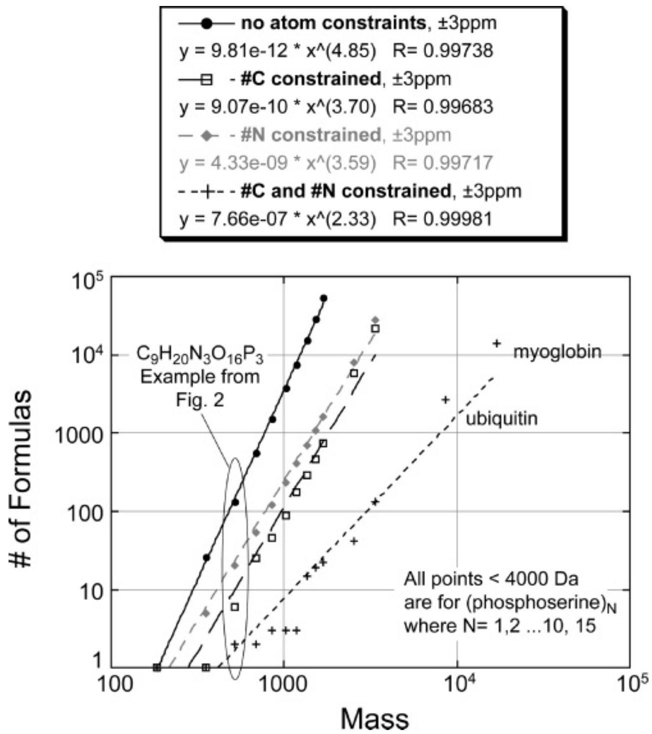

Figure 3. Estimates of maximum numbers of formula as a function of mass. Four sets of points are plotted representing the numbers of formulas calculated from the masses of mono- and poly(phosphoserine)s (n = 1, 2, 3, ..., 10, 15), which were found to reside near the modes for each nominal mass distribution queried. Numbers of formulas were compiled from (1) the unconstrained, (2) the carbon-constrained, (3) the nitrogen-constrained, and (4) the carbon- and nitrogen- doubly constrained solution sets. For the doubly constrained sets, the masses and formulas for the proteins myoglobin (16 941 Da) and ubiquitin (8559 Da) were included to extend the solutions beyond $\sim 10^2$ formulas. Although the protein masses do not represent nominal mass modal values as well as phosphoserine at low mass, these considerations are less important at higher mass values, where the distributions broaden and the amplitude of the variation in the numbers of formulas per mass is dampened. Each solution set was fitted to the following continuous function: no. of formulas = $N \times (\text{mass})^{\text{Exp}}$; where N and Exp are parameters that depend on the numbers of atom types allowed, as well as other constraints. The function maps to a line in the log/log plot with the parameters listed in the key. The continuous function loses meaning as the number of formulas decreases past 1 because the problem requires discrete integer solutions.

the nominal mass distribution modes and that the numbers of formulas can be used to model the maximum number of formulas per unit mass with and without atom constraints (see example for n = 3 in Figure 2). Figure 3 shows our attempt to provide continuous functions that model the behavior of these data. The numbers of formulas = 1 intercept for each of the functions are

(32) The nitrogen rule is not really a rule but a guideline that states that odd nominal mass compounds contain an odd number of nitrogen atoms, and even nominal mass compounds contain an even number of nitrogen atoms. This is a consequence of nitrogen's being the only even nominal mass atom with an odd valence (among C, H, N, O, P, and S). All other atoms have both even (C, O, S) or both odd (H, P) nominal mass and valence. It should be noted that ^{15}N labeling negates the rule, because ^{15}N has both an odd mass and odd valence, as does ^{13}C labeling, since it introduces an odd nominal mass atom with an even valence. But ^{15}N, ^{13}C double labeling makes ^{13}C the sole odd mass, even valence atom such that odd nominal mass compounds contain an odd number of ^{13}C atoms, and even nominal mass compounds contain an even number of ^{13}C atoms.

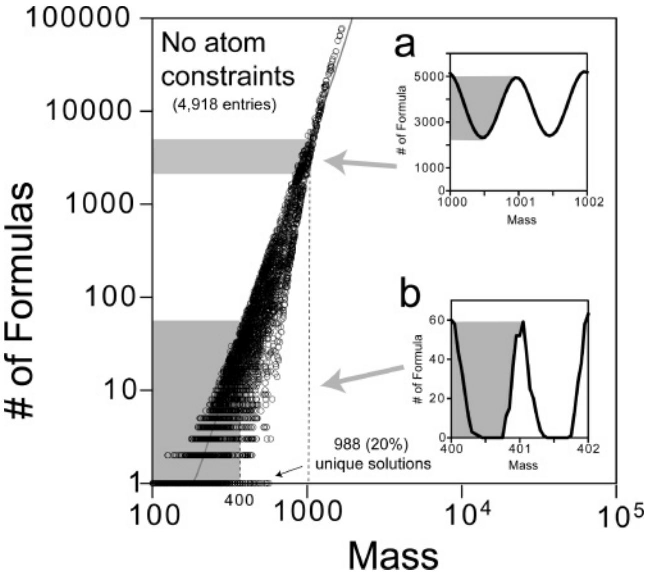

Figure 4. Numbers of formulas calculated for the masses of ~5000 metabolites plotted vs mass. All of the metabolite formulas from the BMRB metabolite database (see methods for URL) were filtered to provide unique formulas composed of carbon, hydrogen, nitrogen, oxygen, phosphorus, sulfur, or a combination thereof. Masses were calculated for the resulting list of 4918 formulas and were plotted against the number of formulas calculated within 3 ppm. Numbers of formulas for each mass value were calculated in batch using a Mathematica script to within ±3 ppm with the following atom number constraints: 2−2000, carbon; 3−3000, hydrogen; 0−500, nitrogen; 0−500, oxygen; 0−50, phosphorus; 0−50, sulfur. Only 988 out of 4918 (20%) of the solutions were unique. The insets show how a lower boundary for the solution set appears as the mass increases due to coalescence of the nominal mass distributions. Though somewhat obscured by the data points, the fitted line for the unconstrained continuous function in Figure 3 is shown for comparison.

186, 278, 214, and 422 for the unconstrained, carbon-constrained, nitrogen-constrained, and doubly constrained cases, respectively. These intercepts estimate the maximum molecular weight at which unique solutions are generated. They do not, however, tell the complete story, because the exponents (4.85, 3.70, 3.59, and 2.33 for I, II, III, and IV, respectively) drop significantly for the constrained sets. This means that the expansion in the number of formulas is ~2.5 orders of magnitude more gradual in the doubly constrained than in the unconstrained set. It is still quite likely that single or double formula solutions will occur up to and above 1000 Da.

To characterize a data set more representative of existing analytes that includes masses that sample points across each nominal mass distribution, we took the masses from ~5000 unique formulas containing only C, H, N, O, P, and S and calculated the number of formulas with and without atom constraints. The unconstrained calculations are plotted in Figure 4; the atom-constrained calculations are plotted on separate graphs in Figure 5. It is apparent from these plots that the continuous functions described in Figure 3 fairly approximate the trends in the maximal edge. This agreement degrades on the high mass end as the deviation from the nominal mass distribution mode is compounded (with phosphoserine polymer length) and on the low mass end where the discrete datasets are not well represented by a

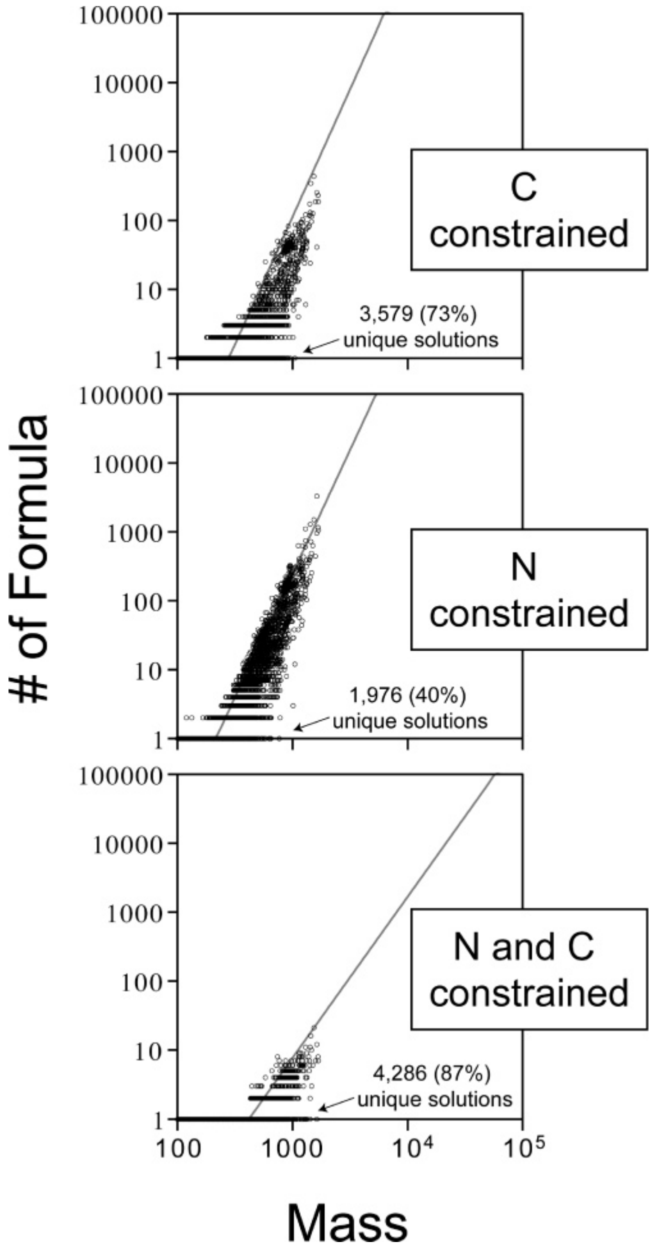

Figure 5. Numbers of formulas for the masses of ~5000 metabolites calculated with atom constraints. The numbers of formulas were calculated from the masses of the 4918 formulas from Figure 4 with nitrogen and carbon atom constraints applied separately or together. The numbers of unique solutions increase dramatically to 73, 40, and 87%, respectively, for the carbon, nitrogen, and doubly constrained calculations. The fitted lines for the respective constrained continuous functions in Figure 3 are shown for comparative purposes.

continuous functions. The apparent tapering and discrete nature of the solution set are accentuated by the log/log plot.

Using Isotope Assisted Formula Assignment for Metabolomics. Next, we wanted to try to apply isotopic formula constraints to the analysis of bulk metabolites. Our approach is given schematically in Figure 6.

Feature Extraction. LC/MS data were processed by a computer program called MassHunter, which is under development by Agilent technologies, to identify "features" consisting of monoisotopic mass measurements and estimates of intensity and

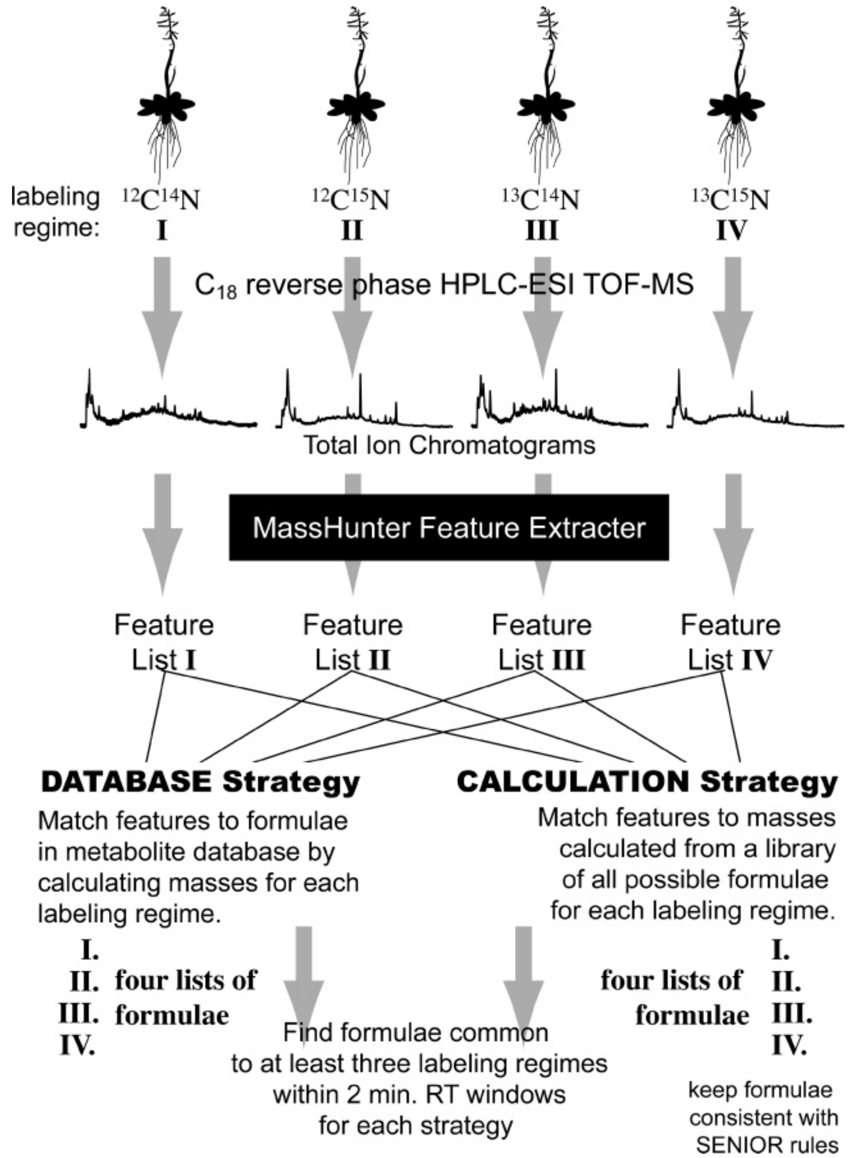

Figure 6. Isotope-assisted metabolomics approach schematic. Four batches of *A. thaliana* were grown: one in a natural abundance medium (I), two in either a ^{13}C- (II) or ^{15}N-labeled (III) medium, and one in a doubly ^{13}C- and ^{15}N-labeled (IV) medium. Dried methanolic plant extracts were generated and reconstituted in 0.1% formic acid/water, and each was analyzed once by C_{18} reversed-phase LC/TOF MS over a 90-min linear gradient. The total ion chromatograms were similar in overall appearance and varied in intensity within a factor of 2. No single species was observed with signal intensity greater than 10^6, the range in which point mass accuracy is perturbed by detector saturation phenomena. Two independent approaches were employed for formula assignment. The first approach, which we denote as the database strategy, involves searching a database of known metabolites using the appropriate mass definitions for the four isotopic labeling regimes. The four resulting lists of compounds are then compared to find matching formulas within 2-min retention time windows. The second approach, which we denote as the calculated strategy, involves searching a list of all of the masses from the complete set of formulas containing carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur with the constraints described in the Methods and Materials section. Similar to the database approach, the four resulting lists of formulas were compared to find matches in at least three cases, also within 2-min retention time windows. The metabolite database and search tools utilized for these approaches have been made publicly available at the BMRB website URL: http://www.bmrb.wisc.edu/metabolomics.[14]

elution time. The program attempts to combine adducts, isotopic peaks, and multiple charge states and then integrates these related components over the LC dimension to provide feature intensity estimates. Combining these different components has not yet been demonstrated to accurately represent relative feature compositions from sample to sample. Otherwise, the program seems to do a good job of identifying and combining related ions within natural abundance samples. The version of the software lacked sufficient flexibility within the LC peak-fitting algorithm to provide usable intensity information for all but the most ideal chromatographic scenarios and tended to split single eluting species into multiple species within the chromatographic domain. Elution time assignment also suffered from this shortcoming. Because MassHunter does a reasonable job of defining the mass of the defined features, we used it primarily as a mass list generator and were reluctant to constrain elution time values much beyond 2 min in our initial comparisons. The software was also evaluated for its capacity to assign mass values to isotopically labeled species. MassHunter's

Table 1. Numbers of Features Input and Formulas Output by the Database and Calculated Strategies

	sample used for MS data collection								
	labeling pattern	I	II	III	IV	combined	av RT standard deviation[a]	mass accuracy[b]	mass precision[c]
	isotopes	natural abundance	^{13}C-labeled	^{15}N-labeled	$^{13}C,^{15}N$ double labeled				
	no. of features	2430	2498	1277	1568				
data analysis strategy									
database strategy alone	no. of formulas	2394	1921	1517	1316	144	±8.1 s	±2.61	±3.27
calcd strategy alone	no. of formulas	290 411	258 679	172 576	84 375	330	±6.3 s	±2.70	±3.13
combined	no. of unique formulas					373			

[a] Average of the standard deviations in retention time for LC/MS features used to identify each formula. [b] Root-mean-square (rms) of the average mass errors (in ppm) for the LC/MS features used to identify each formula. [c] rms of the average absolute values of the mass errors (in ppm) for the LC/MS features used to identify each formula.

deisotoping routine (which cannot be deactivated) succeeded in collapsing natural abundance and ^{15}N-labeled envelopes into neutral monoisotopic masses, but atypical envelope shapes, such as those observed for ^{13}C single- and $^{13}C^{15}N$ double-labeled species, resulted in features' being assigned to multiple isotopic envelope peaks. In these cases, we found that isotopic envelope peaks would be assigned as multiple features, but because the true monoisotopic peaks tended to be one of the two, or three, most intense peaks of their respective envelopes, the bona fide monoisotopic peaks were typically included. For examples, refer to the Supporting Information.

Bulk Feature Elemental Composition Assignment. Formulas were assigned to LC/MS features by matching their calculated masses to feature mass lists from at least three of the four isotopic labeling regimes (I, natural abundance; II, ^{13}C; III, ^{15}N; and IV, $^{13}C^{15}N$). Three observations is the minimum required (and it does not matter which three) to effectively convey carbon and nitrogen count information. Matches were also constrained by chromatographic retention time. Two independent approaches for formula assignment were used as outlined in Figure 6. Web-based formula assignment tools for both the database and the calculated strategies are available at the BMRB website listed above.

Additional Constraints. Table 1 summarizes the total number of features extracted for each labeling regime as well as the resulting numbers of formulas assigned using the database and calculated strategies. Variation in the numbers of features reveals differences in the amount of sample analyzed as well as the consequences of universally applied deisotoping routines. Both approaches were performed initially with ±10 ppm mass error and 2-min elution time windows but were constrained further once formulas common to three or more labeling regimes were identified. The combined lists for each technique include only formulas that match features with elution times within 1 min. To address the feature extractor's tendency to split single eluting species into multiple features, identical formulas eluting within 2 min of each other on the combined list were condensed into a single species. Hundreds of instances of this phenomenon occurred with the calculated strategy and 11 with the database

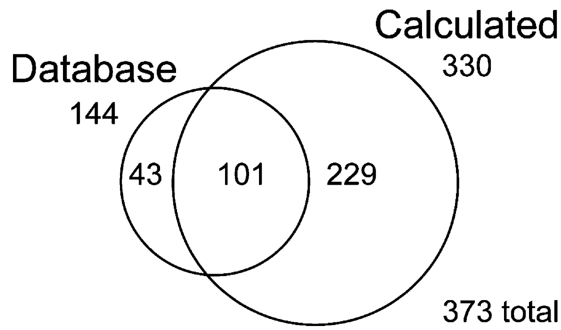

Figure 7. Extent of overlap between the database and calculated approach results. A Venn diagram shows the extent (~30%) of overlap between the database and calculated approach results.

strategy. In 112 of 330 cases in the calculated strategy, a single feature was assigned to multiple formulas, and the higher mass error assignments were removed. Larger numbers of ambiguous feature assignments with the calculated strategy are likely a function of the dramatically greater formula space sampled in this approach than for the database strategy. This likely introduces an additional stochastic element to the calculated approach that will require further characterization. The numbers of formulas on the resulting constrained lists are shown in Table 1 with associated average retention times, root-mean-square (rms) of the average exact mass errors (addressing accuracy), and the rms of the average of the absolute values of the exact mass errors (addressing precision) over the two sets of feature compositions identified.

Comparison of Database and Calculated Approaches. As shown in Table 1, 144 and 330 formulas, respectively, were identified by the database and calculated strategy. Figure 7 shows that ~27% of the total formulas were identified by both approaches. Roughly two-thirds of the database strategy formulas were found by the calculated strategy, whereas only one-third of the formulas identified by the calculated strategy were also identified by the database strategy. Figure 8 shows the features identified for the natural abundance LC/MS analysis. The mass

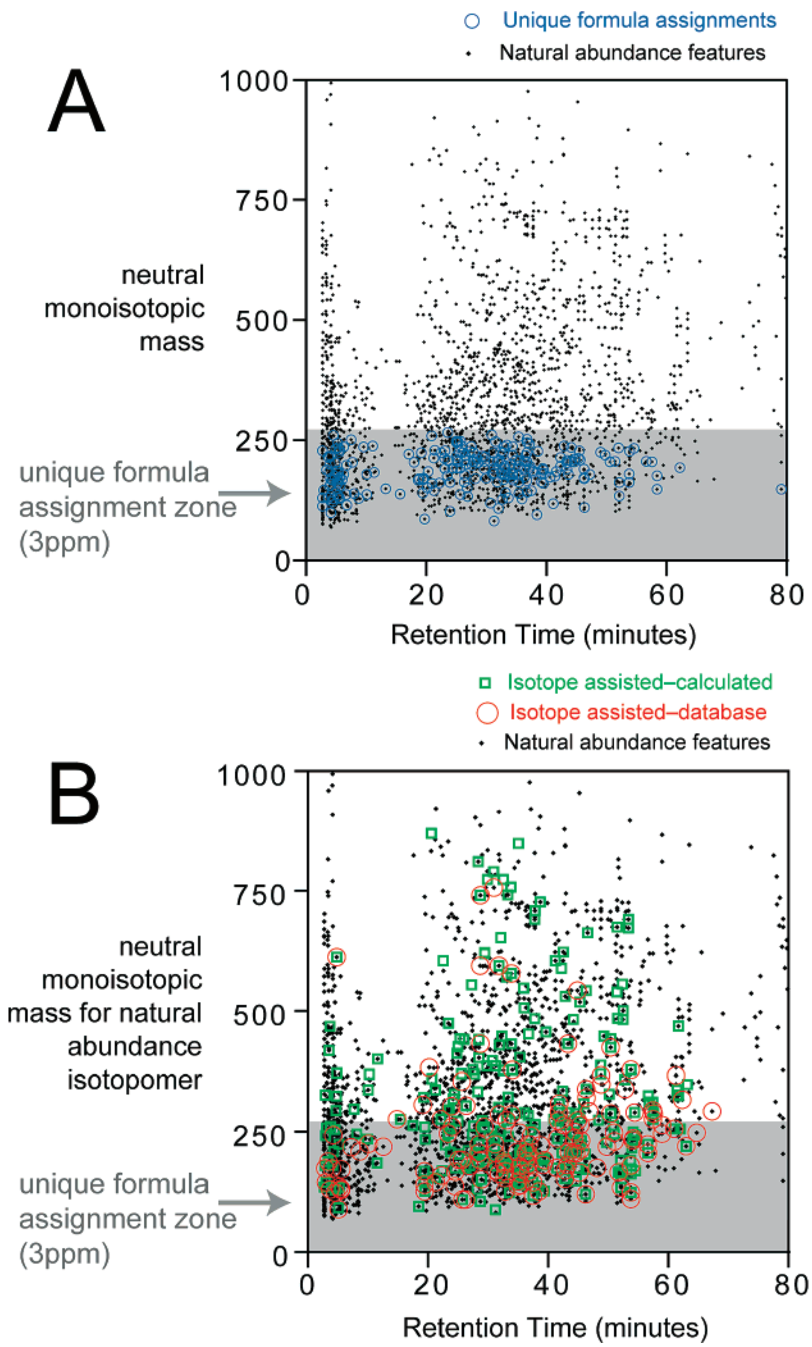

Figure 8. Elution profiles showing identified compositions. Shown are two identical plots (mass vs retention time) of features extracted from the natural abundance ESI-TOF analysis. Panel A shows the unique formula assignments (±3 ppm from the calculated strategy) circled in blue within a gray zone of unique formula assignment. Panel B shows the same features and zone with the unique assignments indicated for the calculated (green squares) and database (red circles) isotope-assisted approaches. The extent to which the isotope-assisted techniques enhance formula assignment is made clear by the number of uniquely identified features above the gray zone.

limitations that would normally constrain unique formula assignment are illustrated in panel A and contrasted with the features uniquely identified using the two isotope-assisted strategies in panel B.

Although this pilot experiment has demonstrated the applicability of our approaches in principle, the numbers of compounds identified suffered from limitations in data processing and feature extraction. We anticipate that improvements in these steps will dramatically enhance the performance of these approaches.

CONCLUSIONS

Isotopic labeling has been employed previously as a tool for constraining the number of formulas assignable to exact mass measurements.[10] Although prior studies have demonstrated the advantages of the approach in assigning formulas to single compounds, none have attempted to characterize the theoretical benefit of its application over a broad range of compounds with diverse masses. Here, we have modeled the numbers of assignable formulas as a continuous function of mass with and without labeling-derived atom constraints. We compared these models with calculations of formula counts for thousands of formulas from a bona fide metabolite database. Of roughly 5000 formulas, double isotopic labeling (^{15}N and ^{13}C) atom constraints resulted in unique assignments in 87% of the instances vs 20% with the typical unconstrained approach. To extend these advantages to metabolomic studies, we have instituted two isotope-assisted strategies for enhanced assignment of formulas that are amenable to high-throughput MS analysis. The database strategy uses web-based search tools that are currently available to search a database of existing metabolites. The calculated strategy, which uses calculated masses from a contiguous defined set of atom combinations, can be used to identify previously uncharacterized species. The approaches have a number of distinct advantages over currently used formula assignment strategies, the most important of which is the dramatic extension of the mass range for which a unique formula can be assigned. The strategies also provide some capacity for avoiding isobaric interference in complex mass spectra because of mass shifts associated with each labeling regime. Finally, metabolic labeling serves as an internal check that proves that the species characterized are derived from the organism and not from contamination introduced in processing.

ACKNOWLEDGMENT

The authors thank Dr. Gregory Barrett-Wilt, James Brown, and Grzegorz Sabat from the University of Wisconsin−Madison Biotechnology Center Mass Spectrometry facility for technical advice and helpful conversations throughout this project. This work was supported by Grants 4 R33 DK070297 and P41 LM05799 (BMRB) from the National Institute of Health and MCB-0448369 from the National Science Foundation.

SUPPORTING INFORMATION AVAILABLE

The complete lists of formulas assigned by the database and calculated strategies are provided in Table S1.xls, and Table S2.xls, respectively, and examples of feature assignments to isotopic envelopes by MassHunter are provided in Sample_features.pdf. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review February 19, 2007. Accepted June 29, 2007.

AC070346T