Anal. Chem. **2006,** *78,* 4363−4373

Automated Analysis of Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra of Natural Organic Matter

Elizabeth B. Kujawinski*,† and Mark D. Behn‡

Department of Marine Chemistry & Geochemistry, and Department of Marine Geology & Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

The advent of ultra-high-resolution mass spectrometry has revolutionized the ability of aquatic biogeochemists to examine molecular-level components of complex mixtures of organic matter. The ability to accurately assess the chemical composition, elemental formulas, or both of detected compounds is critical to these studies. Here we build on previous work that uses functional group relationships between compounds to extend elemental formulas of low molecular weight compounds to those of higher molecular weight. We propose an automated compound identification algorithm (CIA) for the analysis of ultra-high-resolution mass spectra of natural organic matter acquired by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. This approach is benchmarked with synthetic data sets of compounds cited in the literature. The sensitivity of our results is examined for different sources of error, and CIA is applied to two previously published data sets. We find that CIA works well for data sets with high mass accuracy (<1 ppm) and can accurately determine the elemental formulas for >95% of all compounds composed of C, H, O, and N. Data with lower mass accuracy must be accompanied with additional knowledge of chemical structure, composition, or both in order to yield accurate elemental formulas.

The advent of ultra-high-resolution mass spectrometry has revolutionized the ability of aquatic biogeochemists to examine the components of complex mixtures of natural organic matter (NOM) on a molecular level. Ultra-high-resolution mass spectrometers such as the Fourier transform ion cyclotron resonance (FT-ICR) analyzers resolve tens of compounds per nominal mass, thus detecting thousands of unique compounds in each mixture. The most common ionization source for organic matter characterization is electrospray ionization (ESI), which preferentially ionizes water-soluble, hydrophilic compounds. This technique has been applied to numerous natural organic mixtures such as humic and fulvic acids, marine and riverine dissolved organic matter (DOM), microbial-derived DOM, petroleum products, and forensic materials.[1−6] The data from these studies have been used to examine the structural or compositional similarity of DOM from different sources as well as to determine molecular-level changes in composition as a function of (bio)geochemical or anthropogenic processes.

A critical component of each of these studies has been the interpretation of the mass spectra and the ability of each set of investigators to accurately assess the chemical composition (and thus elemental formulas) of the compounds within the mass spectra. Two general strategies have been applied to the task of elemental formula assignment, a "brute force" approach[4] and a "formula extension" approach.[1,3,6] In the first strategy, all chemically relevant elemental formulas consisting of C, H, and O are determined for each nominal mass. Observed compound masses are then compared to masses calculated from the list of all possible elemental formulas. The closest fit between observed and calculated mass is chosen as the most likely elemental formula of the observed compound. This method has been applied to DOM from marine and terrestrial sources.[4] In the second strategy, elemental formulas are determined for a subset of compounds, generally the low molecular weight component below 500 Da. Formulas for additional compounds are found if mass differences associated with functional groups (e.g., addition or subtraction of CH_2) exist between the assigned and unassigned compounds. This method has been successfully applied to DOM,[3] fulvic acids,[6] and petroleum samples.[1]

The foundation for the "formula extension" approach is the recognition of chemical and structural relationships among compounds in each complex mixture. Kendrick mass analysis (KMA) is a special case of this approach and uses CH_2 groups as the only chemical and structural relationship among compounds

* Corresponding author. Phone: 508-289-3493. Fax: 508-457-2164. E-mail: ekujawinski@whoi.edu.

† Department of Marine Chemistry & Geochemistry.

‡ Department of Marine Geology & Geophysics.

(1) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2001**, *15*, 1186−1193.

(2) Qian, K.; et al. *Energy Fuels* **2001**, *5*, 492−498. Rodgers, R. P.; et al. *J. Forensic Sci.* **2001**, *46*, 268−279. Wu, Z.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2002**, *74*, 1879−1883. Kramer, R. W.; Kujawinski, E. B.; Hatcher, P. G. *Environ. Sci. Technol.* **2004**, *38*, 3387−3395. Kujawinski, E. B.; et al. *Org. Geochem.* **2002**, *33*, 171−180.

(3) Kim, S.; et al. *Org. Geochem.* **2003**, *34*, 1325−1335. Kujawinski, E. B.; Hatcher, P. G.; Freitas, M. A. *Anal. Chem.* **2002**, *74*, 413−419.

(4) Koch, B. P.; et al. *Geochim. Cosmochim. Acta* **2005**, *69*, 3299−3308.

(5) Stenson, A. C.; Landing, W. M.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2002**, *74*, 4397−4409.

(6) Stenson, A. C.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2003**, *75*, 1275−1284.

10.1021/ac0600306 CCC: $33.50
Published on Web 05/28/2006

within a mass spectrum. This approach was first applied to ESI FT-ICR mass spectra by Hughey et al.[1,7] KMA sorts compounds according to their deviation from a fully saturated carbon chain, and thus, all compounds with similar chemical structure but differing only in numbers of CH_2 groups fall into one chemical "family". If the elemental formula of the smallest (lowest molecular weight) compound can be determined, then the elemental formulas of all the compounds in the family can be ascertained by the simple addition of the appropriate number of CH_2 groups. This analytical simplification is important because the number of possible elemental formulas for any nominal mass increases dramatically with increasing molecular weight. In addition, the mass accuracy of the FT-ICR MS decreases with increasing molecular weight. Thus, KMA and related methods extend the possible formula assignments beyond the initial boundaries of mass accuracy and molecular weight by calculating elemental formulas only when mass accuracy is high and the number of possible elemental formulas is low.

Petroleum samples are particularly well-suited for KMA analysis due to the predominance of CH_2 series.[1,7] Elemental formulas for compounds within DOM can also be determined by Kendrick mass analysis, but the CH_2 series contain fewer compounds individually and are less numerous collectively.[3,6] This is not surprising given the variety of biogeochemical reactions that are possible in natural environments. These additional reactions could have a significant impact on the chemical composition of DOM. Thus, consideration of additional functional group relationships in studies of DOM should extend the elemental formula assignments to a larger fraction of detected compounds than consideration of CH_2 alone.

One problem with the data analysis approaches presented so far is that there has been no way to quantify the accuracy of the analysis. In other words, the approaches have not been tested with known compounds to be certain that elemental formula assignments are made correctly. Studies to date have circumvented this problem with the assumption that the "best-fit" elemental composition, or the elemental formula whose calculated mass is closest to the observed mass, is the correct formula. This assumption is also used in the approach by Koch et al.,[4] where the best-fit formula is found by comparison of the observed mass with calculated elemental formulas using only C, H, and O and the contributions of other elements such as N is neglected. Since Koch et al.[4] limit their analysis to the most abundant compounds with high signal-to-noise (S/N) ratio, and elemental analysis shows that N contributes much less on a mass basis than C, H, or O, this assumption may be fully justified.

A second problem is the lack of automation inherent in the second (and most popular) approach. Stenson et al.[6] noted that months were required to assign all elemental formulas in fulvic acid mass spectra. We have found in previous work that this time frame is not unusual and is prohibitive to the efficient analysis of one or more mass spectra. Analysis time could be substantially reduced if an automated approach could be developed for both the elemental formula assignment for low molecular weight (LMW) compounds and the extension of these formulas to high molecular weight compounds (HMW). However, with an automated approach, it will be imperative to test the analysis with known compounds to be certain that correct formulas are assigned at all molecular weight values.

In this study, we develop such an automated approach for the analysis of ESI FT-ICR mass spectra of natural organic matter. Our approach looks for functional group relationships between all compounds within a given mass spectrum, determines elemental formulas for all LMW compounds (MW $\leq$ 500 Da), and then assigns formulas to HMW compounds with detected functional group relationships to LMW compounds. We have benchmarked this approach against synthetic data sets of NOM compounds cited in the literature. We show examples of this approach using two previously published data sets of refractory NOM and biologically produced DOM. This methodology can be extended easily to other data sets, other elements, and other chemical and structural relationships.

Description of Method. The compound identification algorithm (CIA) has three main functions: (1) it loads and sorts mass spectral data; (2) it looks for chemical relationships between compounds in the spectra; and (3) it assigns elemental formulas to compounds based on the observed m/z value and the functional relationships found in function 2 (summarized in the flowchart in Figure 1). Each function of the program was tested with literature-based synthetic data sets both separately and as part of CIA.

In the first function, mass spectral data are loaded into the program. For all the synthetic data set tests described below, the "data" were a single list of m/z values. However, our goal is to compare mass spectra from different samples or different experimental stages as a function of specific biogeochemical processes. Thus, we have incorporated the capability to combine two or more mass spectra into one master list of m/z values (see application to previously published spectra described below). We assume all compounds in NOM mass spectra are singly charged and so all m/z values are equivalent to molecular weight.[5] Doubly charged compounds occur quite rarely in our samples, but can be observed in simpler spectra, often those coming from cultures or other biotic experiments (unpublished data). The last step in this function is to add or subtract a H atom (and subtract or add an electron) to generate neutral masses for negative ion mode or positive ion mode spectra, respectively. In our trials with the synthetic data sets, we used a list of neutral m/z values but all work with mass spectra acquired from natural samples incorporated this correction for charged m/z values.

The premise of the second function of our program rests on the ability to identify chemically related compounds within a mass spectrum. These chemical relationships are based on functional group additions. The functional groups are chosen a priori (by the operator) and are written into the data analysis program. In this first study, we chose functional group relationships that should be common in DOM or have been observed in previous studies (Table 1). However, this approach is generally applicable to any functional group difference that is important to a given data set. Once chosen, each functional group relationship is converted into a mass difference (e.g., Δ-CH_2 = Δ14.015 65 Da). The program looks for integral multiples of these mass differences between all compounds within the mass spectrum. The number of multiples is limited to five to prohibit misassigned relationships between two compounds with a large mass difference.

(7) Hughey, C. A.; et al. *Anal. Chem.* **2001**, *73*, 4676−4681.

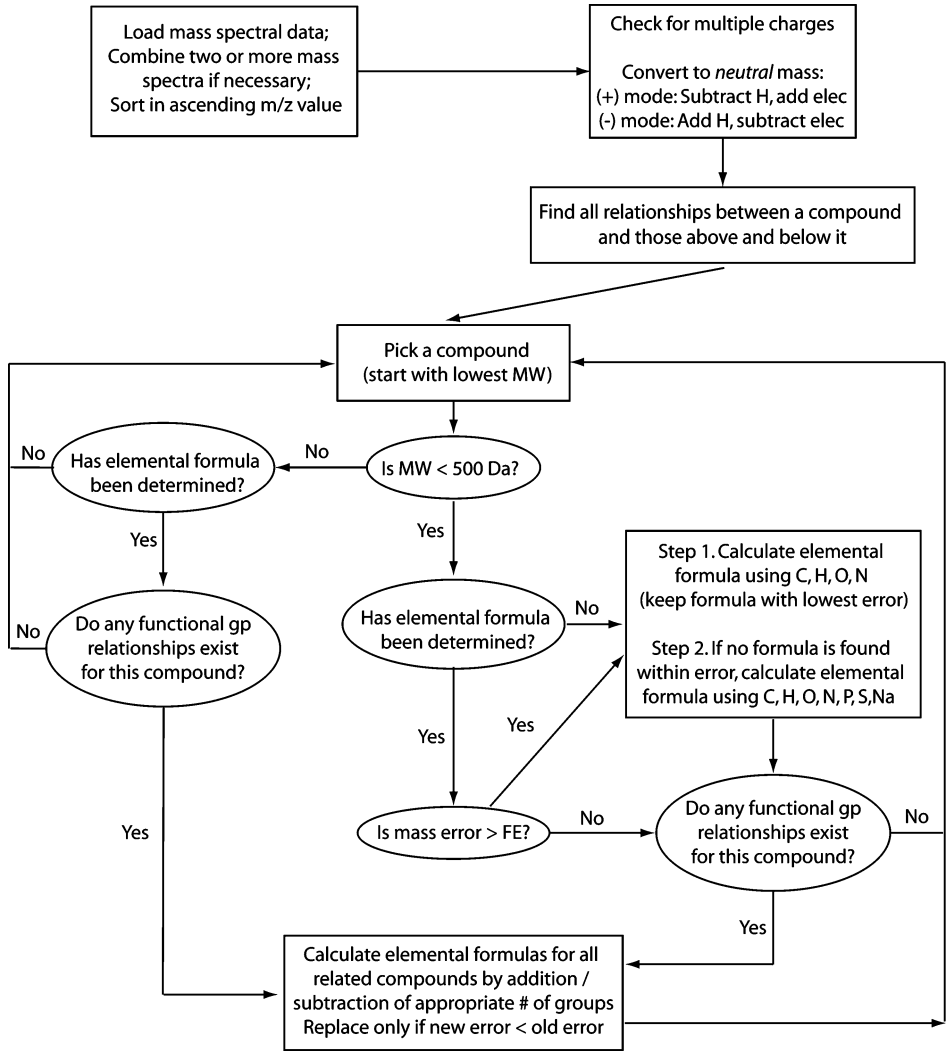

Figure 1. Flowchart of CIA.

Table 1. Functional Group Relationships Used in Generation of Synthetic Data Sets and Elemental Formula Assignment Program

functional group	MW (Da)	functional group	MW (Da)
CH_2	14.015 65	C_2H_2O	42.010 56
$CH_4 - O$	0.036 39	O	15.994 91
H_2	2.015 65	CH[a]	15.010 80
C_2H_4O	44.026 21	NH[a]	13.007 83
CO_2	43.989 83		

[a] Used in generation of synthetic data sets for trial 5.

Once all compounds have been sorted and all relationships based on the given functional groups have been found, the program begins to assign elemental formulas. It starts by choosing a compound (in order of increasing MW). If this compound has a MW below the mass limit (here, 500 Da), the program checks if an elemental formula has already been assigned and, if so, whether the difference between the observed and calculated MW is greater than the error window (assigned by the operator based on instrument conditions and mass accuracy expectations). If an elemental formula has not been assigned or the error is too large, the program determines the elemental formula with the smallest difference between the observed and calculated masses. The elemental formula is calculated by a set of nested loops that tests all possible combinations of the chosen elements. All elemental formulas must meet basic chemical criteria: (1) the number of H atoms must be at least 1/3 the number of C atoms[4] and cannot exceed 2C + N + 2; (2) the sum of H and N atoms must be even (the "nitrogen rule"); and (3) the number of N or O atoms cannot exceed the number of C atoms. These criteria ensure that the elemental formulas generated can exist (at least chemically if not geochemically). The elements C, H, O, and N are considered for all compounds in the mass spectra. If a formula cannot be found within the error window, the additional elements of P, S, and Na are also considered but their number is limited to two atoms per compound. Most compounds within DOM contain only C, H, and O with minor contributions from N and trace contributions from P and S. Inclusion of all elements (C, H, O, N, P, S, Na) for all compounds leads to substantial increases in analysis time (due to the computational requirements of nested *for* loops in Matlab). Once the elemental formula for the chosen compound has been found, the program checks for any functional group relationships between this compound and others and calculates the elemental formulas for these related compounds by adding or subtracting

the appropriate number of functional groups. Previously determined elemental formulas are replaced by new formulas only if the error between the observed and newly calculated masses is lower than the original formula.

The program repeats the above procedure with all compounds below 500 Da. When the MW of compounds exceeds 500 Da, no elemental formulas are calculated by the program. This is due to the larger number of possible elemental formulas and the decrease in the mass accuracy above this m/z value. However, if an elemental formula was determined earlier for one of the HMW compounds, functional group relationships are sought for this compound and the elemental formulas are calculated for these additional related compounds. The elemental formulas assigned by this program are checked many times throughout the procedure since functional group differences are sought above and below a compound of interest. In general, this redundancy is sufficient to accurately assign the large majority of elemental formulas (see examples below).

There are three types of error that must be considered in mass spectral data and in the approach described here: (1) instrument error (IE)—the error associated with the data set as collected from the instrument; (2) formula error (FE)—the maximum error allowed by the elemental formula determination algorithm; and (3) relation error (RE)—the maximum error on the functional group relationship when formulas are extended above the MW cutoff. All errors are expressed in units of ppm (1×10^{-6}). The IE is a function of the instrument and conditions used to generate the data. FT-ICR MS analyses with magnet strengths of 7 T and above generally achieve mass accuracies on the order of 1 ppm or lower.

Synthetic Data Set. A synthetic data set was formulated in order to benchmark the CIA. The synthetic data set was composed of compounds chosen from primary and secondary literature sources to reflect the observed composition of dissolved organic matter and cell components (Supporting Information Table 1). The data set was separated into two categories—compounds composed only of C, H, O, and N (104 compounds) and those with additional elements such as P, Na, and S (27 compounds). We examined the placement of our chosen compounds in the van Krevelen diagram[8,9] to test the applicability of this diagram to predict the composition of DOM molecules from elemental information alone (Figure 2A). In general, different compound classes with substantial differences in chemical structure should occur in different regions on this plot (for example, Kim et al.[8] and Hedges[10]). Small changes in atomic composition can shift a compound's position on this graph, and so, accurate elemental formula determination is essential if the van Krevelen diagram is to be used. We were surprised by the lack of overlap between the predicted and observed positions for some compound classes, in particular, the lipid and protein classes. Very few of our "lipid" compounds were observed in the predicted range for this class. This is most likely due to the inconsistent definition of this compound class across different scientific disciplines. Thus, a common chemical composition should be more difficult to define accurately. Proteins, on the other hand, are well-defined as a compound class but their heteroatom composition can vary significantly depending on amino acid composition, thus yielding a variable O/C ratio. Furthermore, the length of the peptide may be important since covalent linkages of free amino acids yield one H_2O for each amide bond formed. This can have a significant impact on the H/C ratio. In the case of carbohydrates, the predicted and observed positions were very similar. This is probably due to the largely invariant elemental composition of monosaccharides. Our conclusion from this work is that the van Krevelen diagram should be used cautiously to assess the composition of compounds detected within ESI mass spectra. In fact, it is best used to elucidate chemical changes between spectra that can be related to changes in atomic ratios (e.g., oxidation/reduction reactions, methylation/demethylation reactions, etc.).

For each trial described below, 60−100 compounds were chosen randomly from the initial data set (the "parent data set"). Additional compounds were then generated by addition of functional groups to parent compounds. We used the same seven functional group relationships as in our formula assignment program to extend our trial data sets (see Table 1). To generate the synthetic data sets, we used the following approach. First, a compound was chosen randomly from the parent data set and a functional group was chosen randomly from the possible list (Table 1). A new compound mass was generated by adding a random integral multiple (between 1 and 5) of the functional group to the initial one. The new compound was then integrated into the data set and the process was repeated until the "test" data set had 1000 components. Ten unique test data sets were generated in this fashion for each trial of CIA. The composition of each test data set covered the possible compositions of DOM and everything between (Figure 2B). It is likely that not every compound in our data sets has an extant counterpart in DOM. However, the broad compositional range of the compounds shown in Figure 2B ensures that CIA is not biased for one particular compound type or for a specific DOM composition, but instead will be able to determine elemental formulas for a variety of LMW compounds within DOM mass spectra.

RESULTS

Trial Runs with Synthetically Generated Data Sets. Below, we examine the relative importance of the three errors (IE, FE, RE) for accurately assigning elemental formulas within synthetic data sets. Each test of our program (CIA) was conducted with 10 unique data sets generated from the literature-based data set. Unless otherwise stated, each parent data set contained only neutral compounds composed of C, H, O, and N and had molecular weights less than 500 Da. The test data sets contained compounds with higher MW values due to the addition of multiple functional groups. For each trial, we determined the total number of compound identifications (out of 1000 possible), the number of correct identifications, and the number of incorrect identifications. Quantitative assessments of the accuracy of CIA were made by comparison of the known formulas of compounds with those generated by CIA. After conducting initial trials with CHON-only compounds with MW < 500 Da, we examined the ability of CIA to identify test data sets that included (1) literature-derived compounds with additional elements (P, S, Na) and (2) literature-derived compounds with MW > 500 Da. Applications of CIA to

(8) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336−5344.
(9) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2004**, *76*, 2511−2516.
(10) Hedges, J. I. In *Organic Acids in Aquatic Ecosystems*; Perdue, E. M., Gjessing, E. T., Eds.; John Wiley & Sons: New York, 1990; pp 43−63.

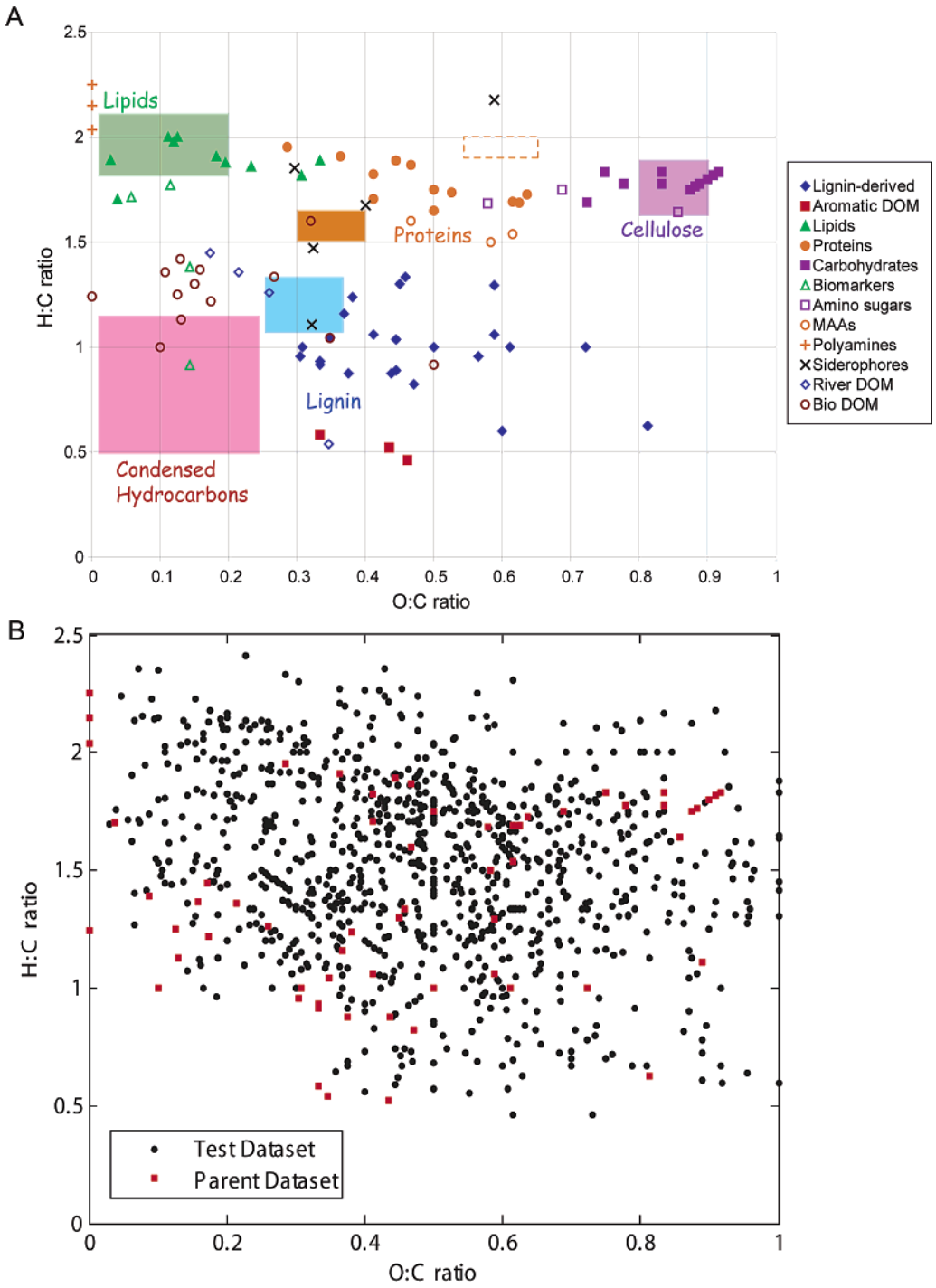

Figure 2. Van Krevelen[14] diagrams for literature-based data set (A) and parent (red squares) and test (black circles) data sets for trial 1 data set 1 (B). Elemental ratios are calculated from the known elemental formulas of each compound. Compound class assignments in A were taken from Hedges[10] and Kim et al.,[8] with preference given to the Hedges assignments when the two references did not match. The predictions for protein molar ratios are separated into two categories: covalently linked peptides (solid box) made of 1000 amino acids (from Reuter and Perdue[15]) and free amino acids (dotted box). Both ranges were calculated from the number average and weight average of amino acid composition in freshwater.[16]

previously published NOM data were also conducted (see below). These data were externally and internally calibrated prior to CIA application, and all charged m/z values were converted to neutral masses (described below).

Formula Error. The FE represents the window within which the elemental formula's calculated mass must fit when determined analytically with the formula subroutine. Two extreme cases were examined first—one in which the program should make no elemental formula assignments (the *null* trial) and one in which the program should get identify all compounds correctly (the *perfect* trial). The null trial was the case where each compound in the test data set had a 1 ppm offset from the true molecular weight but the formula subroutine was limited to identifications with 0 ppm error. In contrast, the perfect trial was the case where each compound had no offset from the "true" value and the formula subroutine was limited to identifications with a 0.5 ppm error. As

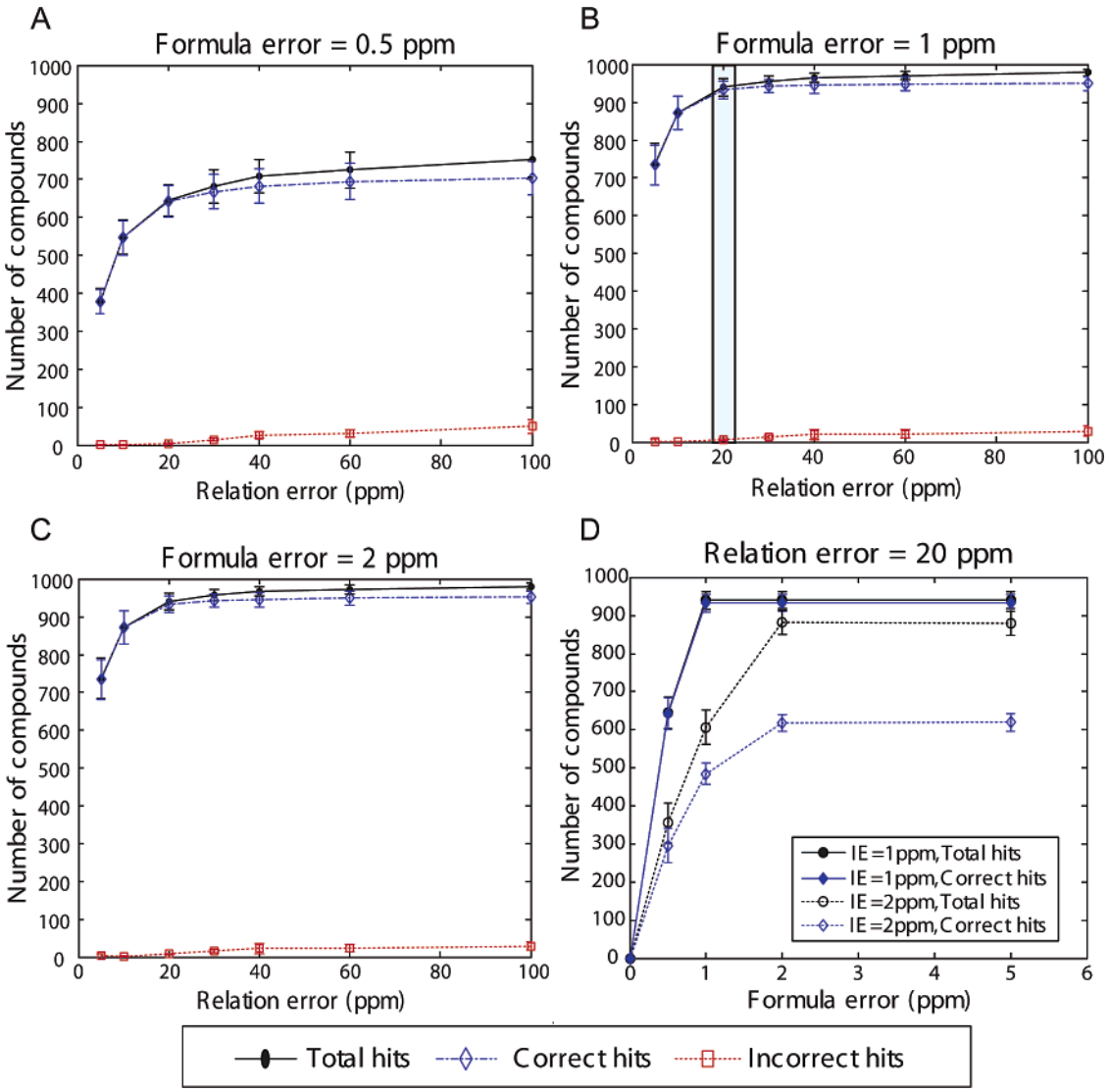

Figure 3. Effect of formula error in trial 1. Panels A−C show the total number of compounds with assigned elemental formulas (black dots and solid black lines), the number of correct assignments (blue open diamonds and dash−dot lines), and the number of incorrect assignments (red open squares and dashed lines). Each point is the average of 10 data sets with error bars representing one standard deviation from the mean. The instrument error for all three FE trials was 1 ppm. The gray box in (B) shows the best error parameters that are used in all subsequent trials: IE = FE = 1 ppm and RE = 20 ppm. (D) shows the effect of formula error on total (black circles) and correct (blue diamonds) identifications for two instrument errors, 1 and 2 ppm, when the relation error is 20 ppm.

expected, no compound identifications were made in the null trial, but all compounds were identified correctly in the perfect trial.

For all subsequent trials, the formula error varied between 0.25 and 5 ppm. When the formula error was less than the instrument error, the total number of identifications was low but most of the identifications were correct (Figure 3A). As the formula error approached the instrument error, both the total and correct number of identifications increased with a small increase in incorrect identifications. Once the formula error exceeded the instrument error, the total number of identifications continued to increase, but the number of incorrect identifications increased significantly. In fact, once FE > IE, most of the additional identifications were incorrect. Thus, we conclude that setting the formula error larger than the instrument error yields too few additional correct identifications, and thus, all remaining trials were conducted with FE = IE. This result was shown experimentally by Yanofsky et al.[11] Many elemental formula programs that come with mass spectral data acquisition and analysis software encourage/allow formula error windows that are significantly higher than the inherent instrument or data error. This is required (and appropriate) during protein identification from peptide spectra since large error windows accommodate unforeseen posttranslational modifications. However, application of this technique to DOM spectra is only appropriate when the operator has additional knowledge about the chemical composition or structure of the unknown compound. This approach is commonly used during calibration of a data set, and the correct formula may lie outside the uncalibrated error window. In other circumstances, use of this approach may lead to erroneous identifications of DOM

(11) Yanofsky, C. M.; Bell, A. W.; Lesimple, S.; Morales, F.; Lam, T. T.; Blakney, G. T.; Marshall, A. G.; Carrillo, B.; Lekpor, K.; Boismenu, D.; Kearney, R. E. *Anal. Chem.* **2005**, *77*, 7246−7254.

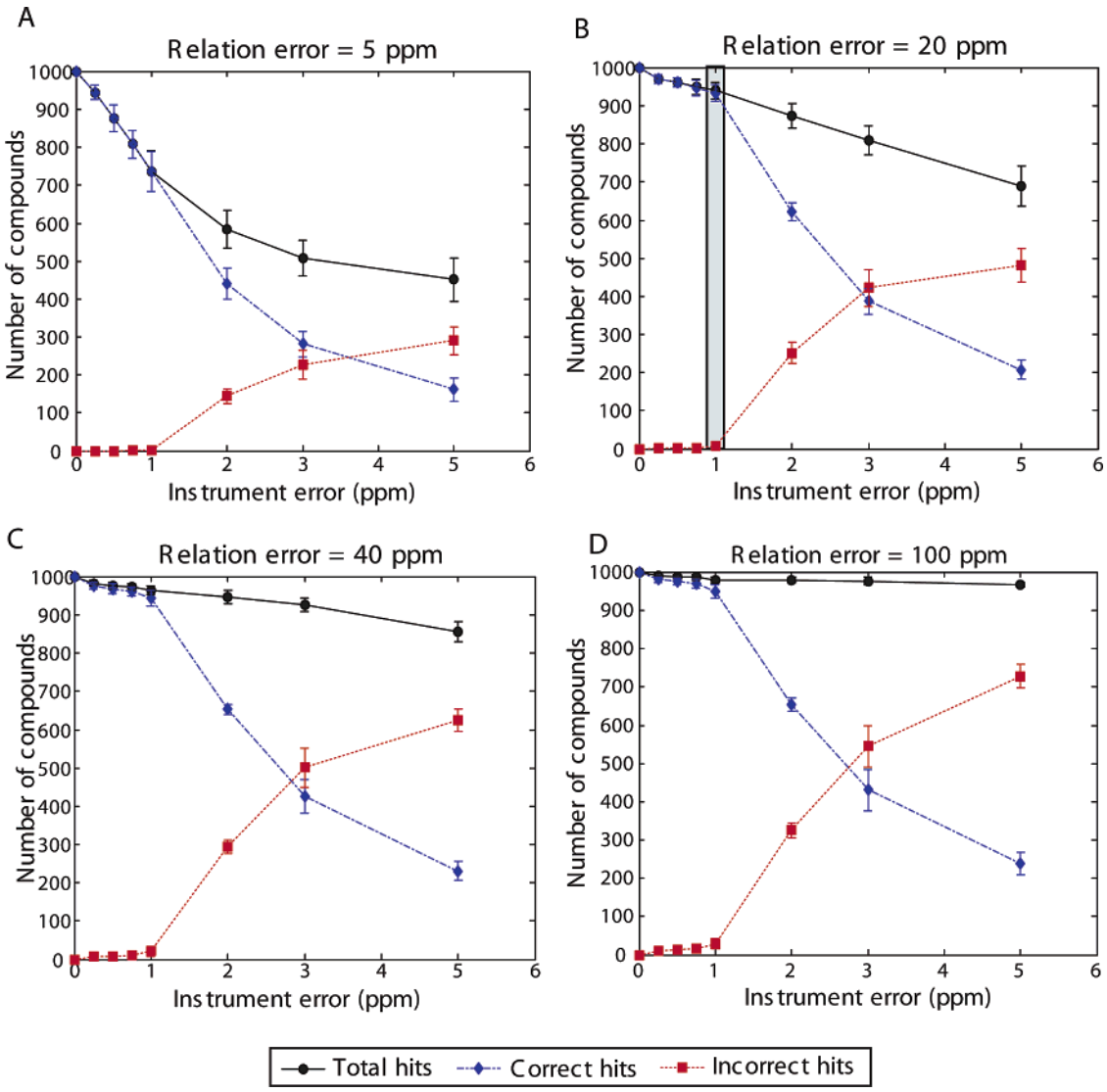

Figure 4. Effect of relation error in trial 1. The results from trial 1 for four relation error values: RE = 5 ppm (A); RE = 20 ppm (B); RE = 40 ppm (C); RE = 100 ppm (D). For each of these panels, the total number of compounds with assigned elemental formulas (black dots with solid line), the number of correct assignments (blue diamonds with dot−dash line), and the number of incorrect assignments (red squares with dotted line) are shown. Each point is the average of 10 data sets when FE = 1 ppm, and the standard deviation is indicated. The gray box in (B) shows the best error parameters that are used in all subsequent trials: IE = FE = 1 ppm and RE = 20 ppm.

compounds. This effect is exacerbated when the instrument error exceeds 1 ppm (Figure 3D).

Instrument Error. We next examined the influence of instrument error on the accuracy of formula assignments within CIA. In each trial, the mass of each compound in a data set was assigned an error, chosen randomly from an even distribution of errors above and below 0, with maximums at the positive and negative values of the "instrument error". In these trials, we examined IE values between 0 and 5 ppm. For each IE value, trials were run for a series of RE values between 0 and 100 ppm and FE = IE. For all RE values, we found that IE values above 1 ppm resulted in a sharp decrease in the number of correct identifications (Figure 4). In fact, with IE ≥ 2 ppm, the number of correct identifications was only 3 times greater than the number of incorrect identifications, and this ratio decreased to 0.5 when IE = 5 ppm. Thus, our data analysis method will only work with data with sufficiently low instrument error, i.e., whose mass accuracy is sufficiently high (≤1 ppm). This mass accuracy can only be achieved for complex mixtures such as DOM with ultra-high-resolution mass spectrometers such as the FT-ICR analyzer. Mass spectrometers such as the qTOF and ion trap mass spectrometers often quote high mass accuracies such as 1−2 ppm, but this usually applies to simple mixtures of two to five components with high S/N ratios. Since mass accuracy of a compound depends on the S/N ratio and thus its relative abundance, only those compounds with sufficiently high S/N ratios will have the requisite precision in m/z value to allow the unique determination of an accurate elemental formula.[12] Our algorithm could thus be applied to data from lower resolution mass spectrometers if only compounds with sufficient mass accuracy were included. It is possible to circumvent this mass accuracy limitation if additional information about the chemical structure is obtained and boundaries can be placed on the possible elemental formulas. We are examining the possibility of including

(12) Chen, L.; et al. *Chemom. Intell. Lab. Syst.* **1986**, *1*, 51−58.

Table 2. Summary of Results for All Trials[a]

trial	elements in test data set; MW limit	total IDs	correct IDs	% accurate	incorrect IDs	% inaccurate
1	CHON MW < 500 Da	967 ± 10	967 ± 10	100	0	0
2	P, Na, S (low %) MW < 500 Da; all elements in analysis	915 ± 22	849 ± 26	93	67 ± 25	7
	analysis without P, Na, S	*888 ± 23*	*844 ± 25*	*95*	*44 ± 23*	*5*
2A	P, Na, S (high %) MW < 500 Da; all elements	941 ± 17	734 ± 101	78	207 ± 91	22
	analysis without P, Na, S	*866 ± 48*	*737 ± 103*	*85*	*129 ± 92*	*15*
3	CHON MW > 500 Da	841 ± 36	837 ± 43	99.5	3 ± 7	0.5
4	P, Na, S (low %) MW > 500 Da; all elements	778 ± 43	755 ± 46	97	23 ± 22	3
	analysis without P, Na, S	*790 ± 37*	*758 ± 45*	*96*	*32 ± 23*	*4*
5	Fn group test: CHON MW < 500 Da; all elements	541 ± 101	516 ± 125	95	25 ± 37	5

[a] All values are calculated for instrument error (IE) = 1 ppm, formula error (FE) = 1 ppm, and relation error (RE) = 20 ppm. In trial 5, more functional groups were used in the generation of the synthetic data sets than were available in the elemental formula assignment program.

fragmentation spectra (i.e., MS/MS data) or stable isotopic patterns (e.g., the use of ^{13}C or ^{34}S to confirm elemental formula assignment) in CIA, but this is beyond the scope of this manuscript.

Relation Error. The last form of error examined was the RE, or the window allowed on the identification of functional group relationships between two masses. We considered relation errors ranging from 0 to 100 ppm. The errors around the functional groups are inherently larger due to the relatively low mass of each individual functional group. In other words, a ±0.0001 Da error would represent a 1 ppm error at 400 Da but a 10 ppm error at 40 Da. In the current version of CIA, only one relation error value is used for all functional group relationships. This is generally appropriate for functional groups with similar mass differences. However, it is possible that we overestimate the number of functional group assignments for small mass differences (e.g., $CH_4 + O = \Delta 0.036\ 39$ Da). Relation errors that are too small miss some relationships between compounds with large molecular weight differences, but relation errors that are too big will assign nonexistent functional group relationships. In our study, the best relation error value was 20 ppm. This value gave the largest number of total identifications with the highest ratio of correct to incorrect identifications (Figure 3B and 4B). To make CIA more generally applicable, we will modify future versions of CIA to utilize the relation error values that vary as a function of the mass difference of the functional group. This will limit errors associated with over- or underestimates of functional group differences among compounds within the spectra. Furthermore, we recognize that some functional groups will be more prevalent in certain types of NOM than in others (e.g., more CH_2 series in petroleum rather than in humic NOM). Assigning a higher priority to prevalent functional groups may further speed up our analyses and yield an increased number of accurate elemental formula assignments.

Addition of P, S, and Na. For this and all subsequent trials, we used only the most appropriate error parameters found in the above trials, namely, IE = ≤1 ppm, FE = IE, and RE = 20 ppm. We ran two trials with low molecular weight compounds containing non-CHON elements, differing in the percent of these compounds in the parent data set. The results from these trials were compared to the benchmark trial 1, containing parent compounds with MW < 500 Da and composed of C, H, O, and N. In trial 2, only 10% (approximately) of the compounds contained non-CHON elements. The results in terms of the total and correct numbers of identifications were very similar to those found in trial 1 (Table 2). There is a slightly bigger offset between the total and correct numbers of identifications when non-CHON elements are included (i.e., lower percent accuracy). This is probably due to the larger number of possible elemental formulas within a given error window. This offset increases with increasing instrument error, implying that instruments with higher mass accuracy are better able to resolve small mass differences between different element combinations. In trial 2A, ~50% of the parent compounds contained non-CHON elements. A bigger offset between total and correct identifications is observed than in trial 2. Some data analysis software exploits incremental mass differences in elemental isotopes (such as ^{13}C, ^{15}N, or ^{34}S) to decide between two (or more) elemental formula possibilities. This type of information qualifies as "additional" information such as MS/MS data that would be helpful in determining the accurate elemental formula. Without this additional information, our data suggest that an operator would be more likely to identify compounds correctly if trace elements are not included in the analysis or if increasingly high-resolution mass spectrometers can be used.

Addition of HMW Compounds. The addition of high molecular weight (>500 Da) compounds to the parent data set should result in a decrease in the number of total identifications. The

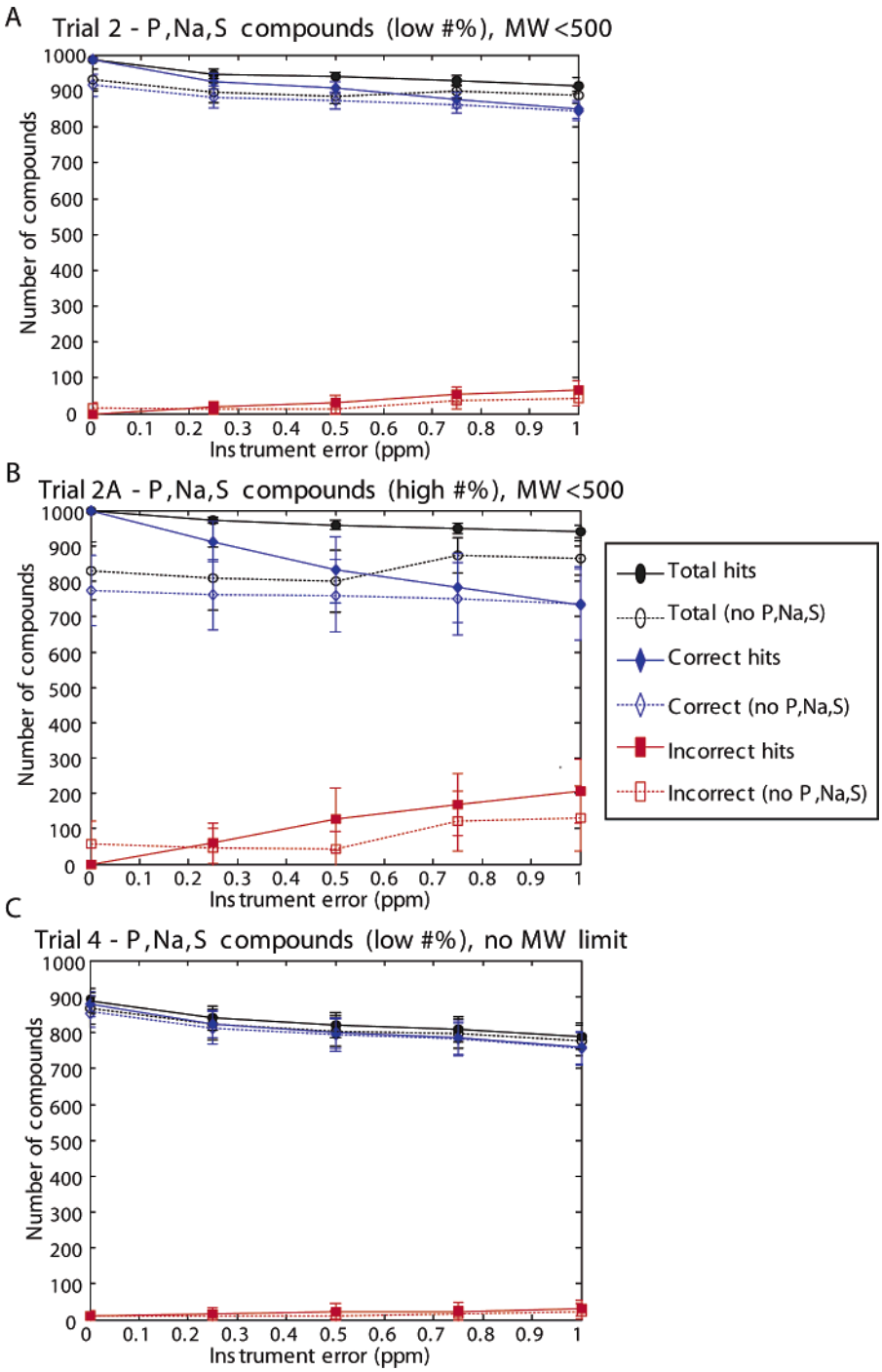

Figure 5. Effect of inclusion of P, Na, or S in elemental formula determination for trials 2 (A), 2A (B), and 4 (C).Each graph shows the total identifications (black circles), correct identifications (blue diamonds), and incorrect identifications (red squares) for CIA analyses with (solid lines) and without (dotted lines) consideration of P, Na, and S. Each point represents the average of 10 test data sets run under the same conditions and the error bars represent one standard deviation from the mean.

reliance on elemental formula determination for peaks with $m/z \leq 500$ Da means that any compounds with MW > 500 Da (that are not related to LMW compounds by some combination of functional groups) will not be identified by CIA. The relative number of correct identifications should remain high, however, since all other compounds should be identified in the same manner as the LMW trials described above. Our trials with both CHON-only (trial 3) and non-CHON-containing compounds (trial 4, Figure 5C) support this hypothesis. In both trials, fewer compounds were identified but the identifications were mostly correct. In fact, the percent accuracy of the LMW and HMW trials were statistically the same (Table 2).

Application to DOM Data Sets. Our data analysis program was applied to two previously published data sets:[13] (1) Suwannee

(13) Kujawinski, E. B.; et al. *Mar. Chem.* **2004**, *92*, 23−37.
(14) Van Krevelen, D. W. *Fuel 29*, 269−284.
(15) Reuter, J. H.; Perdue, E. M. *Mitt. Geol.-Palaont. Inst. Univ. Hamburg* **1984**, *56*, 249−262.
(16) Perdue, E. M.; Ritchie, J. D. In *Treatise on Geochemistry: Surface and Ground Water, Weathering and Soils*; Drever, J. I., Ed.; Elsevier Ltd.: New York, 2003; Vol. 5, pp 273−318.

River dissolved organic matter and (2) DOM derived from a bacterial control and a protozoan/bacterial grazing culture. Extraction and analysis of DOM from these aqueous samples has been described previously.[13] ESI FT-ICR MS data were externally and internally calibrated before conversion to neutral masses. The instrument was calibrated daily with a proprietary mixture of compounds (HP mix) in the appropriate ionization mode (positive or negative). For internal calibration, a CH_2 series of four to six compounds was identified and elemental formulas were assigned to each compound in the series. Accurate masses were then calculated for each compound in the series, and the spectrum was internally calibrated to these values. Where possible, the same series of compounds were used in all spectra to ensure uniform calibration. In the previously published study, elemental formulas were generated for as many compounds as possible using a two-step "manual" method. First, elemental formulas of compounds with m/z values below 500 Da were determined by manually entering the mass and an acceptable error into a web-based formula generator (http://medlib.med.utah.edu/masspec/elcomp.htm). Second, elemental formulas were extended into the higher molecular weight range using nine functional group relationships, but in an Excel spreadsheet. Full analysis of these data took approximately two months to accomplish and generated elemental formulas for >90% of SRDOM and 77% of the bacterial and protozoan DOM. However, the accuracy of these assignments could not be determined because no check with synthetic data sets was possible.

In each of these two data sets, multiple spectra were compared to examine the role of a particular process in the molecular-level modification of DOM (photochemistry in the case of SRDOM and grazing metabolism in the case of Bac/Pro DOM). The two peak lists from each experiment were combined prior to elemental formula generation. Information regarding the presence or absence of a specific compound within each mass spectrum is retained in a matrix of relative magnitude values (no. of rows = no. of unique compounds in all spectra and no. of columns = no. of mass spectra). Once combined, the resulting master list is sorted in ascending m/z value. We assume all compounds in DOM mass spectra are singly charged, and so all m/z values are equivalent to molecular weight. However, compounds with m/z < 500 and mass defects between 0.7 and 0.9 are usually doubly charged. CIA takes this into account by doubling the mass of the compound and adding or subtracting a H atom for negative ($M - 2H^{-2}$) and positive ($M + 2H^{+2}$) ion modes, respectively. The compound is then inserted into the master list at its singly charged value. Doubly charged compounds occur quite rarely in our samples, but can be observed in simpler spectra, often those coming from cultures or other biotic experiments (unpublished data). Last, a H atom is added or subtracted to generate neutral masses for negative ion mode or positive ion mode spectra, respectively. This can be problematic if a large number of sodiated complexes is present (more relevant for positive ion mode spectra than for negative ion mode). The generation of these matrixes allows us to retain information regarding the composition of each individual sample but also increases the efficiency of the automated analysis by negating the need for identifying common compounds more than once and by further extending the identifications of compounds in each spectrum by building on all compounds at once.

SRDOM. The elemental composition of SRDOM and its primary component, Suwannee River fulvic acids, are well known to contain primarily C, H, and O with a minor contribution from N. The contribution from P or S is always negligible, and little Na should be present when DOM is extracted with C18 resin instead of the classical NaOH extraction technique for fulvic and humic acids. Furthermore, Na adducts occur primarily in positive ion mode, and our SRDOM spectra were collected in negative ion mode. We applied CIA to this data set using only C, H, O, and N at first and then reran the analysis with the additional elements of P, Na, and S. In the first case, 2352 of 2866 (82%) of the compounds were identified (i.e., had elemental formulas assigned to them). In the second case, 2858 of 2866 (>99%) were identified, but 653 compounds (23% of IDs) were found to contain P, Na, or S. The large fraction of compounds containing trace elements is likely incorrect and highlights the difficulty in using the best-fit approach for compound identification. Higher instrument resolution would be necessary to confirm the presence of these additional elements to our satisfaction. Nonetheless, the agreement between the previously published results and our new data is remarkably good when we use the data from the CHON-only analysis, implying that we have not generated new unsubstantiated results. Only 20 min was needed for this analysis, showing that we have substantially increased our analytical efficiency with this approach.

Bacterial/Protozoan DOM. We analyzed the Bact/Prot DOM with the same two approaches as SRDOM. The inclusion of P, Na, and S is relevant due to the biological availability and reactivity of P and S and the fact that these samples were run in positive ion mode (and extracted from seawater). In the case of CHON-only, 191 of 622 compounds were identified (31%), and when P, Na, and S were included, 275 compounds were identified (44%). In this second case, 89 (32% of total ID's) compounds contained P, Na, or S. Many fewer compounds were identified with our automated approach than had been identified with the manual approach in the previous publication. Two additional functional groups (NH, CH) were used in the previous work, but inclusion of these two groups gave only 221 IDs (36%) when P, Na, or S is excluded and 309 IDs (50%) when P, Na, or S is considered (122 formulas have P, Na, or S). Just as in SRDOM, we are concerned about the large fraction of compounds containing P, S, or Na within our results.

DISCUSSION

Implications for Different Sample Types. The automated approach described above is generally applicable to any sample type, provided the inherent chemical reaction pathways are known. The functional group relationships in this initial work were chosen due to their observed predominance in previous work with highly degraded or refractory samples from terrestrial and aquatic environments such as humic or fulvic acids. Thus, between 80 and 90% of the compounds in the SRDOM sample were identified by our program. However, significantly fewer compounds in the biological DOM sample (Bac/Pro DOM) were identified by CIA than by the manual approach used earlier. This is probably due to two major factors: (1) possible operator bias in formula extension and (2) a larger error window for formula identification

(2 ppm was used in our previous work). We now know that formula errors above 1 ppm result in a significant number of incorrect identifications, and we must conclude that our previous work misassigned many elemental formulas.

An additional factor in this analysis was the choice of functional group relationships that are put into the CIA program prior to data analysis. To test the impact of functional group choice, we generated a fifth trial of 10 test data sets where compounds were generated using two more functional groups than are used in CIA as currently written (CH, NH). Many fewer compounds were assigned overall (60−70% identified relative to 100% in other trials: see trial 5 in Table 2). In addition, more incorrect assignments are made than in previous trials where the functional groups used to build the data set exactly matched the functional groups used to determine elemental formulas. Perhaps not surprisingly, the effect of additional chemical relationships is much greater than the effect of additional elements in the compounds. Thus, the use of functional groups that are particularly relevant to a specific sample is critical to identification of the majority of compounds. However, inclusion of metabolically relevant functional group relationships in biological DOM samples will require an iterative process because too few reactions are currently recognized. Given the paucity of information on relevant molecular-level reactions within organic matter in different environments, this type of work will yield important insights into the dominant metabolic reactions within microbial communities as they recycle organic matter.

Implications for OM Characterization by ESI FT-ICR MS. The analytical approach described here is a substantial step forward in the automation of elemental formula determination for components of natural organic matter. Our ability to compare two or more mass spectra efficiently opens the door for molecular-level comparative analyses in a manner that was not previously possible. We have utilized chemical relationships based on functional group differences to assign elemental formulas to a substantial fraction of compounds within our complex NOM mixtures. However, much more work remains to be done to accurately assign elemental formulas to all compounds within NOM mixtures. For example, CIA does not currently assign elemental formulas that contain heavy isotopes of common elements such as C (^{13}C), N (^{15}N), or S (^{34}S). Due to the relatively low abundance of compounds with one heavy isotope (e.g., $^{12}C_x{}^{13}C_1$ relative to $^{12}C_{x+1}$), we have neglected their contribution in our analyses of NOM. However, we are currently working on modifications to CIA that will incorporate these isotopes into elemental formula assignments.

Determination of elemental formulas, however, is simply one step toward full characterization of organic matter. For example, elemental formulas do not provide chemical structure due to the numerous isomeric forms that are possible, especially for (relatively) high molecular weight compounds. Additional information will always be needed to constrain possible chemical structures and therefore, MS/MS analysis will play an important role in structural characterization. We anticipate using our CIA protocol to identify those compounds within mass spectra that are most critical for a given scientific question and then performing MS/MS analysis on this subset of compounds. As data-dependent MS/MS techniques evolve, we will be able to combine fragmentation patterns with our algorithm to constrain the possible elemental formula for a compound as well as its chemical structure.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding sources for this work: NSF Chemical Oceanography (CAREER-OCE-0529101 and OCE-0443217), NSF Chemistry (CHE-9909502 at the NSF National ICR Users' Facility at the National High Magnetic Field Laboratory, Tallahassee, FL) and WHOI startup funds. Drs. Alan Marshall, Ryan Rodgers, and Carol Nilsson were invaluable in data acquisition and discussions of the evolution of this algorithm. The work was substantially improved by discussions with Drs. Neil Blough, Boris Koch, and E. Michael Perdue and by comments from an anonymous reviewer.

SUPPORTING INFORMATION AVAILABLE

Synthetic data set. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review January 5, 2006. Accepted April 10, 2006.

AC0600306