analytical chemistry

Technical Note

pubs.acs.org/ac

Formularity: Software for Automated Formula Assignment of Natural and Other Organic Matter from Ultrahigh-Resolution Mass Spectra

Nikola Tolić,*[,†] Yina Liu,[†,⊥] Andrey Liyu,[†] Yufeng Shen,[†] Malak M. Tfaily,[†] Elizabeth B. Kujawinski,[‡] Krista Longnecker,[‡] Li-Jung Kuo,[§] Errol W. Robinson,[†] Ljiljana Paša-Tolić,[†] and Nancy J. Hess[†]

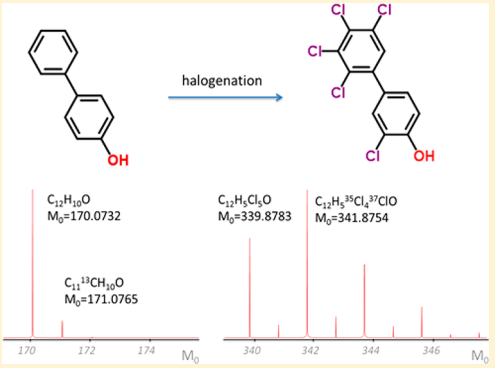
[†]Earth & Biological Sciences Division, Pacific Northwest National Laboratory (PNNL), Richland, Washington 99354, United States

[‡]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution (WHOI), Woods Hole, Massachusetts 02543, United States

[§]Marine Sciences Laboratory (MSL), Pacific Northwest National Laboratory, Sequim, Washington 98382, United States

Supporting Information

ABSTRACT: Ultrahigh resolution mass spectrometry, such as Fourier transform ion cyclotron resonance mass spectrometry (FT ICR MS), can resolve thousands of molecular ions in complex organic matrices. A Compound Identification Algorithm (CIA) was previously developed for automated elemental formula assignment for natural organic matter (NOM). In this work, we describe software Formularity with a user-friendly interface for CIA function and newly developed search function Isotopic Pattern Algorithm (IPA). While CIA assigns elemental formulas for compounds containing C, H, O, N, S, and P, IPA is capable of assigning formulas for compounds containing other elements. We used halogenated organic compounds (HOC), a chemical class that is ubiquitous in nature as well as anthropogenic systems, as an example to demonstrate the capability of Formularity with IPA. A HOC standard mix was used to evaluate the identification confidence of IPA. Tap water and HOC spike in Suwannee River NOM were used to assess HOC identification in complex environmental samples. Strategies for reconciliation of CIA and IPA assignments were discussed. Software and sample databases with documentation are freely available.



Molecular level characterization of organic matter plays an important role in understanding the fate and biogeochemical cycling of natural and anthropogenic organic moieties under different environmental conditions. Ultrahigh-resolution mass spectrometry (HR MS) such as the Fourier transform ion-cyclotron resonance mass spectrometry (FT ICR MS) has been used to characterize organic matter from different environments.[1−4] With HR MS, elemental formulas of small molecules (<500 Da) containing C, H, O, N, S, and P can be assigned based on accurate mass measurement alone.[5] A Compound Identification Algorithm (CIA) was developed by Kujawinski and colleagues at the Woods Hole Oceanographic Institution as a fully automated function to assign elemental formulas to a list of masses observed with FT ICR MS from natural organic matter (NOM) samples.[6−8] Database (DB) supporting CIA code represents universal database for NOM formula assignment, i.e., contains mathematically possible molecular formulas consisting of elements C, H, O, N, S, and P. The CIA DB consists of more than 29 million unique molecular formulas with monoisotopic mass below 1500 Da. Formula filters based on the seven heuristic golden rules[9] were included in the CIA code to ensure that the assignment is at least chemically possible. Briefly, CIA adopts established principles of formula assignment for NOM measured with HR MS:[10−14] peaks from spectra measured by an FT ICR MS are assigned with molecular formulas starting from the low m/z range searching CIA DB, and high m/z compounds are assigned using formula expansion based on CH_2, H_2, O or other homologues series building blocks, because the number of formula candidates increase substantially as mass increases.

CIA allows fine-tuning of processing parameters; however, because of the complexity of the CIA code, it is impractical for frequent changes since edits in code must be made in multiple places. As applications of HR MS in analyzing organic matter from diverse environmental matrices increase, the ability to change and keep track of CIA parameters tailored to specific sample types is crucial. In addition, as more compounds containing elements beyond C, H, O, N, S, and P are discovered, algorithms capable of assigning these compounds are becoming increasingly important. For example, it is known that organometallic and organohalides are present in different ecosystems. However, these important constituents of environmental samples[15−17] usually remain unidentified and unreported in research studies of NOM by HR MS, reducing the opportunity for deeper data dives and scientific quests.

Received: August 16, 2017
Accepted: November 9, 2017
Published: November 9, 2017

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

Computationally, compiling a universal database for more than few additional elements would present several technical challenges: (1) the significant increase in database size, (2) complex valence rules for evaluation of chemically valid molecular formulas, and (3) likelihood of observing the monoisotopic peak for compounds containing elements with complex multi-isotopic natural abundances (for example Br, Cl, or Hg). Manual or semiautomated formula assignment and isotopic simulations using instrument manufacturer provided software yields reliable results but is limited in throughput and difficult to automate.[18] To address these challenges, Formularity software features a newly developed formula assignment function, Isotopic Pattern Algorithm (IPA). IPA is a fully automated isotopic simulation search function based on a precompiled database of predicted isotopic peaks of a target formula library. Several scoring schema for IPA were developed to assist in the validation of results.

CIA Search Function and Database. Formularity software was developed using Visual Studio (Microsoft Corporation), executables, and source code with user manual and sample databases are available from a software repository (https://omics.pnl.gov/software/formularity). Figure 1 illustrates the data analysis workflow and software main components organized by functionality. Peak alignment and CIA search function are refactored and optimized versions of original MATLAB code developed by Kujawinski and colleagues. Formularity features user-friendly interactive interface for CIA and newly developed IPA search function and database schema. Provided CIA DB was compiled from the MATLAB database updated in 2016. The least-squares regression based internal calibration function is added as well, leaving only low-level signal processing and peak picking to external tools. File formats for supporting databases and output results are described in the software manual. Formularity allows formula assignment for general high-resolution mass spectra collected in positive or negative ionization mode, with proton or electron ion physics, different molecular adducts, and charge states. In the current version, each search must be done separately, e.g., in electrospray ionization (ESI), positive mode search should to be done for H^+ and Na^+ adduct ions, results combined and evaluated, and eventual ambiguities resolved. All measurements in this work are performed in negative ESI mode with Formularity data analyses limited to deprotonated molecular ions.

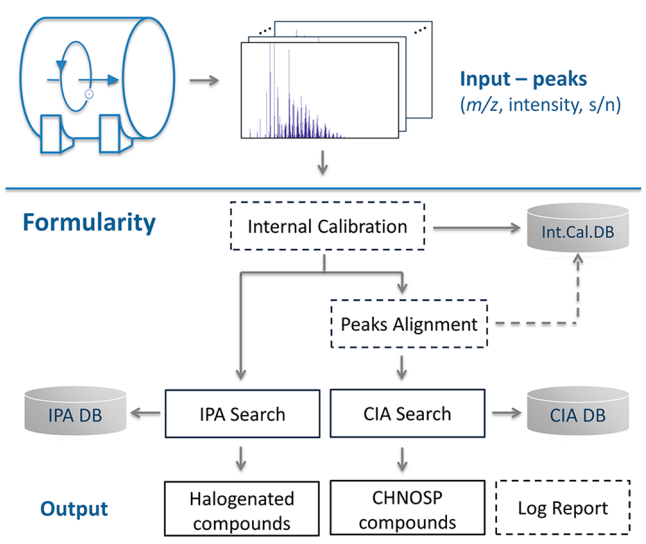


Figure 1. Formularity software flowchart; dashed lines indicate optional procedures. Input consists of single or multiple spectra in a form of list of peaks. Calibrated and optionally aligned set of peaks is submitted to CIA and/or IPA search functions, which are independent and generate separate reports. CIA DB is universal database listing molecular formula and monoisotopic mass of low-molecular-weight NOM (below 1.5 kDa). IPA DB is compiled list of predicted isotopic peaks for target set of molecular formulas. Both databases list molecular formulas as neutral (zero-charge) mass. Calibration tables, containing important validation information, are preserved in application log files.

Internal calibration of spectra has immense importance and far reaching consequences for formula assignment when sub-ppm mass measurement accuracy (MMA) requirements are imposed on the analysis of complex samples with unknown elemental composition. External calibration of the mass spectrometer prior to sample measurement is considered to be one of the fundamental rules of the "best practice" guide.[19] Yet, even with external calibration performed immediately before the experiment, sub-ppm formula assignment is often achieved using internal calibration in which experimental mass measurements are adjusted based on expected m/z values of known peaks present in samples naturally or added purposefully. Table S1 in the Supporting Information lists calibration m/z values compiled from various sources used for internal calibration of NOM samples measured in ESI negative ion mode spectra. For internal calibration of a large number of spectra, compiling custom calibration peaks could increase the dynamic range and m/z coverage of assigned peaks. In Formularity, internal calibration, which is independent of the vendor platform, is an optional step; search functions could be used with a list of peaks calibrated with different tools and methods. To validate Formularity automated internal calibration calculation, we performed comparison with interactive internal calibration implemented in DataAnalysis software (Bruker Daltonic). Peak-by-peak comparison of results from both functions shows resulting m/z values agreement up to the fifth decimal place (data not shown). Calibration results, which are written in the application log file, should be inspected as a part of the validation process. Information including "before and after calibration" mass measurement errors, count, and m/z range of matched calibration peaks help assess calibration success, which is a quality that is hard to frame in a simple binary answer for evaluation of adequacy of formula assignment using sub-ppm mass tolerance. Therefore, although internal calibration is automated in Formularity software, a cautious approach with results evaluation should be performed in this sensitive stage of data analysis.

One important but often neglected result from formula assignment is the number of unassigned peaks providing information about completeness of sample characterization. Unassigned peaks could have originated from chemical or electronic contamination but could also represent important sample features being ignored by the CIA search filters and database selection. More sophisticated matching function focused on explanation of unassigned peaks is needed to increase the formula assignment rate. The same is true for ambiguous assignment; even with assumed sub-ppm MMA, only a very limited number of peaks can be identified with only one possible formula candidate. CIA function resolves this ambiguity using "lowest heteroatom count" criteria, although, in some cases, there are additional peaks in the spectrum, namely,

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

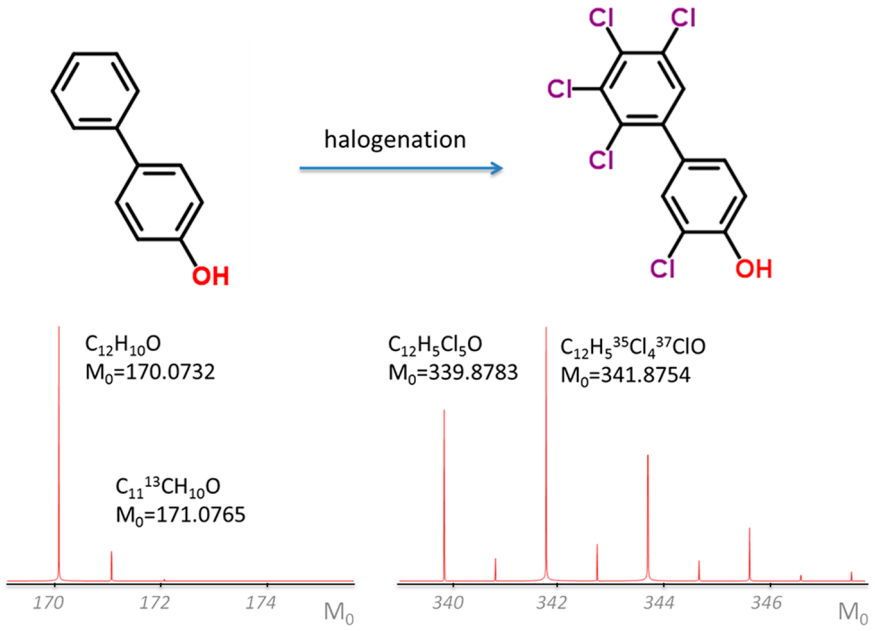


Figure 2. Simulated isotopic distributions of NOM molecule 4-biphenylol and halogenated product 2′,3,3′,4′,5′-pentachloro-4-biphenylol illustrates why the CIA method would not work well for the assignment of halogenated organic matter. The graph displays neutral molecular mass along the horizontal axis (M_0) with relative abundance as peak height calculated and visualized by Mercury software. The peak most likely to be observed in the spectrum for the singly charged ion of elemental composition C, H, N, O, P, S is a monoisotopic peak and, for a majority of formulas, a maximum of two isotopic peaks are expected to be detected above the noise level. For halogenated formulas (Cl and Br), the most abundant peak is not necessarily the monoisotopic peak and, generally, there are multiple isotopic peaks likely to be observed. This is also true for many organometallic compounds.

third isotopic peak and fine isotopic structure peaks that may be queried to resolve ambiguous formula assignments.

IPA Search Function and Database. IPA function is a targeted search function based on isotopic pattern database annotating simulated (predicted) most likely isotopic peaks for any set of stable isotope molecular formulas. The general molecular formula can have thousands of theoretically predicted isotopic peaks, yet only a small fraction is expected to be observed in the mass spectra. For the majority of low-molecular-weight NOM molecules, monoisotopic peak is most likely to be observed among all the isotopic peaks, although other isotopic peaks are often detected.[14] However, as illustrated with simulated data in Figure 2, measurement of chlorinated compound is expected to produce multiple isotopic peaks with the most abundant (MA) peak not coinciding with the monoisotopic peak. The quantifiable likelihood of the observation of additional isotopic peaks, relative to the MA peak for this type of compound, provides implementation guidance for developing new database schema, as well as search and scoring functions that facilitate the assignment of molecular formula from NOM or any other type of matter.

Compilation of IPA DB for use with Formularity is an external process. Fast exact, combinatorial calculation algorithm,[20] which is available as software implementation of ecipex[21] or numerical algorithms represented in tools such as Mercury[22] or Deuterium,[23] could be applied to generate simulated isotopic distributions using the natural abundance of stable isotopes of multi-isotopic elements. IPA DB can include any number of isotopic peaks for each molecule, and different classes of compounds could require more or less peaks for confident identification. IPA DB of $k \times l$ order consists of molecular formula (MF) records, organized into matrices of $k \times l$ major (M_i) and $k \times l$ minor peaks ($m_{i,j}$), where $i = 1, ..., k$ and $j = 1, ..., l$. Major peaks are simply the most probable peaks at each isotopic nominal mass, and minor peaks are other calculated isotopic peaks around each major peak, i.e., peaks of fine isotopic structure. Each peak of an MF record is assigned a pseudo-probability of observation, assuming that MF has been detected $(P_i)(p_{i,j})$ $(i = 1, ..., k; j = 1, ..., l)$, where P_i is the probability of major peaks, and $p_{i,j}$ is the probability of minor peaks. The MA peak is assigned a probability of 1, because, for formula assignment, the MA peak must be observed, and probabilities of other peaks are calculated as the MA peak abundance normalized values. This database structure allows more comprehensive search of isotopic peaks and exploration of the isotopic fine structure given sufficiently revealing mass measurement. In IPA DB, the number of isotopic peaks is the same for all MF; if too few isotopic peaks are predicted to fill the MF record, extra positions are filled with 0 values. This custom structured IPA DB over the same set of molecular formulas could be used for different purposes, for example, 12 × 2 IPA DB could be used for more conservative formula identification, while a 6 × 4 database could resolve ambiguous assignments based on more peaks of isotopic fine structure. To evaluate formula and peak assignments, several different scores are reported that should be used in combination for discrimination of true and false matches. A very important scoring element for IPA formula assignment is MMA which is provided as "tma_err" in the results from IPA formula (see Tables S3A and S4 in the Supporting Information). Statistical evaluation of distribution of mass errors for calibration peaks allows elimination of assigned molecular formulas with outlier MMA. This method, which is not automated in the current version, can be applied to both CIA and IPA search functions. There are two other scoring types. First, the presence/absence (pa) scores are measuring observed peak pattern agreement

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

(more is better) with predicted isotopic pattern. Second, distance (d) scores measure how much observed and simulated patterns differ, with 0 being perfect score.

Two pa scoring schemes are included: (1) the presence of each peak is awarded peak predicted pseudo-probability and (2) ad-hoc pa score is used, where present major peaks are awarded a value of 1 and minor peaks are assigned a value of 0.1. For both pa scores, the relative (pa_{rel}) score is obtained by normalization of the pa score with the maximum score (pa_{max}) for a particular formula. Formally, for a spectrum consisting of experimentally observed set of peaks $(ME, AE)_i$ (where ME is the measured peak m/z, AE is the measured peak abundance, and $i = 1, ..., n$; predicted peak notations are as presented above), the matching set of experimental peaks (MF_E) for molecular formula MF is obtained through comparison of predicted and measured peaks based on specified MMA tolerance. MF_E consists of measured peaks $(ME,PE)_i$ $(me,pe)_{i,j}$ $(i = 1, ..., k; j = 1, ..., l)$ matching predicted major $(M,P)_{MF}$ and minor peaks $(m,p)_{MF}$ for formula MF. A value of (0, 0) is assigned where no match is found. Pseudo-probabilities PE, pe for measured major and minor peaks, respectively, are calculated as peak relative abundance within the MF_E. The normalization procedure is performed the same way for database peaks, which allows distance-based scoring functions to describe the fit between simulated and measured isotopic profiles.

While CIA search function loops through a list of peaks and assigns formula for each peak, IPA search is database-oriented, i.e., it loops through all database records, matches peaks, and scores each record independently. To formalize, for each IPA DB record, the MF max score is given with

$$pa_{max}(MF) = \sum_i \left(P_i + \sum_j p_{i,j} \right)$$

and the absolute score, using matched experimental peaks, is given as

$$pa_{abs}(MF) = \sum_{ME_i>0} \left(P_i + \sum_{me_{i,j}>0} p_{i,j} \right)$$

The relative score is then given as

$$pa_{rel}(MF) = \frac{pa_{abs}(MF)}{pa_{max}(MF)}$$

Peak pseudo-probability thresholds are applied in calculating the ad-hoc presence/absence absolute score (pa_{abs}), with a consequence that the relative score (pa_{rel}) could be more than 1, with interpretation of "better than expected" measurement as an attempt to highlight exceptional agreement with predicted peaks. Score using pseudoprobability weighting does not implement threshold limits and, therefore, cannot assume a value of >1.

Scores describing the fit of expected abundance profile for MF with a matching collection of peaks assumed to form isotopic envelope use distance function on predicted and measured isotopic patterns without penalizing for peaks not observed. With the established notation reported, the d_2 score is calculated as

$$d_2 = \sqrt{\sum_{ME_i>0} (P_i - PE_i)^2}$$

Infinity and taxicab distance scores are included and, especially, the infinity score should be useful to spot IPA matches, showing a large disparity between the relative abundance of the observed and predicted MA peaks.

Software Demonstration Using Standard Mixture Spike and Tap Water Analysis. We used pure standards and municipal tap water (MTW) samples to test formula assignment of halogenated compounds with Formularity. Municipal tap water is known to contain many halogenated organic compounds (HOC) as disinfection byproducts introduced during chlorination treatment.[24−27] This choice of analyte allows simultaneous testing of both search functions and reveals cases of ambiguous peak assignments and strategy for consolidation of results. We used established solid-phase extraction protocol, following Dittmar et al.,[28] to obtain organic matter from MTW for analysis by FT ICR MS. A total of 27 pure HOC standards with a mass range from 128.0029 Da to 665.6982 Da were used for preliminary tests and benchmarking HOC assignment in MTW (Table S2 in the Supporting Information). Standard compound identification was performed in two different matrices: (1) standard mix in organic solvent and (2) standard mix spiked into Suwannee River NOM (SRNOM). SRNOM is a standard NOM reference material purchased from the International Humic Substances Society. All HOC standards were purchased from Sigma−Aldrich in their highest purity.

All samples were analyzed on a 12 T Bruker SolariX FT ICR MS with ESI in negative mode. Details of experimental settings and peak processing are listed in the Supporting Information. The resulting list of peaks from spiked samples were processed using Formularity internal calibration and IPA search functions. Small IPA database with 26 molecular formulas, structured as 12 major and 5 minor peaks, was searched with a mass tolerance of 1 ppm and formula with at least 2 major peaks matched were reported. In the pure standard mix, IPA search assigned peaks for 21 out of 26 formulas, peaks for the 5 formula IPA has not assigned could not have been confidently matched, even with a manual inspection of spectra. Total of 171 predicted isotopic peaks for 21 assigned molecular formulas were matched with mean and standard deviation of mass measurement error 0.051 and 0.127 ppm, respectively (root mean-square (rms) error of 0.137 ppm). Full output from the IPA search is listed in Tables S3A and S3B in the Supporting Information, with an example of the detected and identified peaks of fine isotopic structure annotated in Figure S1 in the Supporting Information. For SRNOM standard mix spike, IPA search assigned peaks for 14 out of 26 formulas, with a total of 94 peaks matched with an average MMA of 0.035 ppm and a standard deviation of 0.327 ppm (rms error of 0.327 ppm). As expected, all 14 formulas assigned in SRNOM spike were assigned in the organic solvent spike. Charge competition in the significantly more complex matrix of SRNOM explains the lower count of assigned standard mix peaks yet complex isotopic pattern annotations are found (see Figure S2 in the Supporting Information). Evaluation of various scores from the pure standard mix experiments was used to set initial cutoff values in IPA search for MTW measurements. IPA matches with relative pa scores of <0.35, d2 scores of >0.75, and mass errors of the MA peak that were >0.5 ppm were discarded.

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

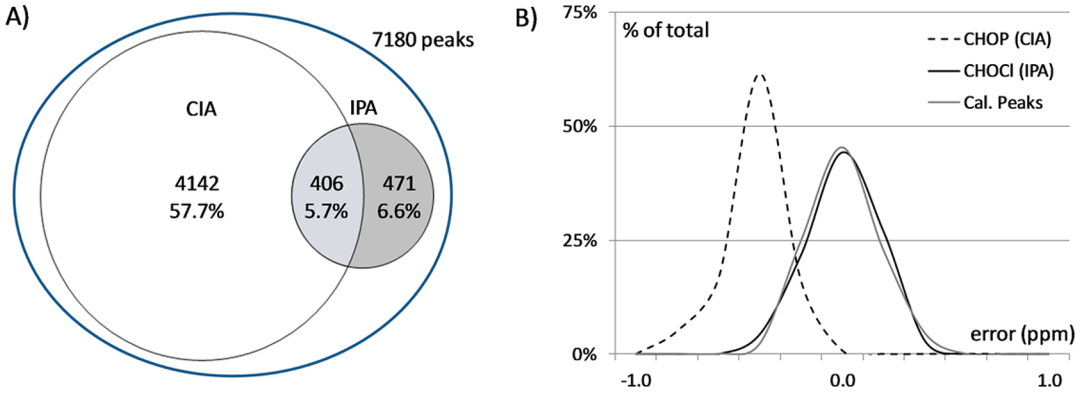


Figure 3. (A) Formula assignment counts for 7180 peaks measured for MTW samples. In total, ~70% of the peaks were assigned molecular formula by combined CIA and IPA functions. IPA contributed with 6.6% assignment of previously unassigned peaks, as well as a putative reassignment of 5.7% of CIA matches. (B) HR MS allows ambiguity resolution for many molecular formulas with close m/z based on MMA alone. Error distribution for 138 peaks assigned by IPA shows better agreement with error distribution of calibration peaks compared with error distribution of CHOP formulas assigned by CIA for the same set of peaks (Table S5 in the Supporting Information).

Peaks from three technical replicates of MTW extract were aligned using an alignment function with a m/z tolerance of 1 ppm to obtain the list of 7180 consensus peaks selected for formula assignment. CIA search function was performed using a mass error tolerance of 0.5 ppm, CIA DB with enforced "golden rules" check, and formula expansion using CH_2, H_2, and O as building blocks. Ambiguity in formula assignment was resolved using the criteria of lowest count of heteroatoms and lowest mass measurement error in the case of equal heteroatom counts. The CIA function also employed a user-defined filter requiring at least 1 O atom and a maximum of 3 N atoms, 2 S atoms, and 1 P atom. For IPA search, we compiled a small database of molecular species with elemental composition CHClO using snapshot of ChemSpider database and a list of compounds from the supplemental tables of Zhang et al.,[26] resulting in a database with 8875 formulas. Ecipex software (version 1.0) was used to compile 8 × 3 IPA DB used in search. Results of these searches are summarized in Figure 3A. Using the IPA function yielded a 6.6% increase in a total number of assigned peaks. IPA function matched total of 877 peaks mapping to 279 distinct formulas with elemental composition CHClO, 134 of which were reported in tap water analysis by Zhang et al. A van Krevelen plot of assigned CHClO species (Figure S3 in the Supporting Information) shows that most of the chlorinated compounds occur in the "lignin-like" region, suggesting chlorination in MTW may be targeting phenolic compounds with a nature similar to that of lignin-derived compounds.

It is important to note that some peaks were assigned by both search functions and, thus, conflicting results from CIA and IPA searches may occur. These peaks with conflicting assignment were marked for more careful inspection. Comparing results of search by CIA and IPA functions, there were 406 peaks with conflicting assignments, of which 198 peaks were assigned by CIA as elemental composition CHNO, 138 peaks were assigned as CHOP, 56 peaks were assigned as CHNOPS, and 7 peaks each were assigned as CHNOP and CHOPS. CIA assigned 76 out of 406 ambiguous peaks as ^{13}C peaks, 68 of them were also ^{13}C matching isotopic peaks for ambiguous monoisotopic peaks of elemental composition CHOP. All these peaks were used by the IPA function to assign formula with elemental composition CHClO with at least 2 other major isotopic peaks, including the most abundant peak. Careful examination of mass measurement error and IPA scores can help resolve ambiguities for many peaks as illustrated in Figure 3B, using the analysis of mass measurement errors. For CIA-assigned formula with elemental composition CHOP, ambiguity is described through substitution in molecular formula of OP in CIA-assigned formula by CCl in IPA-assigned formula, with respective masses of 46.96868 and 46.96885. This mass difference accounts for a relative mass error of ~0.35 ppm at m/z = 500, which is sufficient in most cases for distinguishing between formulas based on expected MMA obtained from internal calibration. On the other hand, for 184 out of 198 peaks assigned by CIA as CHNO formula, ambiguity is described through formula substitution of NO_3 with C_2H_2Cl, where the monoisotopic peak of the NO_3-containing formula is also assigned as the ^{13}C peak of the C_2H_2Cl-contained formula with respective masses of 61.98782 and 61.98786. This mass difference at m/z = 500 corresponds to a difference in relative error of only 0.08 ppm, making the choice between ambiguous formula assignments using mass error as the only criteria problematic. By evaluating the abundance profile of peaks assigned as ^{13}C peaks of formula with elemental composition CHO, we can estimate the detection limit for ^{13}C peaks of CHNO formulas. For 81 out of 184 cases, where the predicted ^{13}C peak was above the detection limit, the absence of ^{13}C peak was used to reject the formulas with an elemental composition of CHNO. Analysis of mass measurement error variation between ^{12}C and ^{13}C peaks for ambiguously assigned formula could be used for further discrimination of correct and incorrect assignments. A full list of assignments by both functions including ambiguous assignment can be found in Table S4 in the Supporting Information. Table S5 in the Supporting Information shows a mass error evaluation for ambiguously assigned peaks with an elemental composition of CHOP, as shown in Figure 3B.

Described findings lead to an important point to keep in mind when analyzing chemically complex environmental samples: if search space is sufficiently large, the ambiguity in peak assignment is a common occurrence, even at sub-ppm MMA.[29] With advances in FT ICR MS instrumentation,[30,31] one can expect that using ultrahigh-resolution mass precision and careful examination of results leads to novel insights and methods for systematic resolution of ambiguities in formula assignment.

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

CONCLUSION AND FUTURE DIRECTIONS

We have developed the publicly available software Formularity, featuring an updated and user-friendly version of previously described CIA software for identification of NOM species using HR MS. A graphical user interface and implemented filters facilitate throughput and application for characterization of NOM in environmental matrices.

To enable detection and identification of molecular species comprising elements other than C, H, N, O, S and P, Formularity is equipped with secondary fully independent search function IPA probing database of isotopic peaks. Software flow and utility were demonstrated through the analysis of pure halogenated standards and municipal tap water using both CIA and IPA search functions. Initial evaluation of search results demonstrated that the new function allows not only detection of molecular species undetectable by CIA but also labeling of potentially incorrect assignments. Results also point to ambiguities in the formula assignment of complex samples using two functions, some of which could be addressed through incremental resolving power of FT ICR instrumentation or separation methods focusing on sample complexity reduction. Future software releases will focus on three major technical improvements: (1) validation of internal calibration outcome, (2) automation of multipass searches, and (3) evaluation of introduced scoring schema, in an effort to provide insights into false discovery rates for CIA and IPA assignments.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b03318.

List of NOM sample calibration peaks for negative mode ESI (Table S1); list of compounds in standard mix used in this study; (Table S2); IPA function output for HOC standard mix spiked in organic solvent (MeOH) (Table S3); List of peaks and assigned molecular formula by CIA and IPA search functions in MeOH extract of Sequim municipal tap water (Table S4); analysis of mass error distribution for ambiguously assigned CHOP peaks by CIA and CHOCL by IPA functions (Table S5); measured isotopic pattern of compound $C_{12}H_3Cl_7O$ from pure standard mixture spiked into organic solvent (MeOH) (Figure S1); isotopic pattern of compound $C_{12}H_3Cl_7O$ from pure standard mixture spiked into NOM matrix (Figure S2); Van Krevelen plot of assigned molecular formula from three technical replicates of municipal tap water (Figure S3) (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: nikola.tolic@pnnl.gov.

ORCID

Nikola Tolić: 0000-0003-3950-9130

Yina Liu: 0000-0002-3485-7542

Li-Jung Kuo: 0000-0003-1319-7394

Present Address

[⊥]Geochemical and Environmental Research Group, Texas A&M University, College Station, TX 77845, USA.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research was performed using Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory, Richland, WA. A portion of this study was conducted under the Laboratory Directed Research and Development Program (project # 204059) at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

GLOSSARY OF NON-STANDARD ACRONYMS

CIA = Compound Identification Algorithm
HOC = halogenated organic compounds
IPA = Isotopic Pattern Algorithm
MF = molecular formula
MA = most abundant
MTW = municipal tap water

REFERENCES

(1) Koch, B. P.; Witt, M.; Engbrodt, R.; Dittmar, T.; Kattner, G. *Geochim. Cosmochim. Acta* **2005**, *69* (13), 3299−3308.

(2) Altieri, K. E.; Seitzinger, S. P.; Carlton, A. G.; Turpin, B. J.; Klein, G. C.; Marshall, A. G. *Atmos. Environ.* **2008**, *42* (7), 1476−1490.

(3) Ohno, T.; He, Z.; Sleighter, R. L.; Honeycutt, C. W.; Hatcher, P. G. *Environ. Sci. Technol.* **2010**, *44* (22), 8594−8600.

(4) Tfaily, M. T.; Chu, R. K.; Tolić, N.; Roscioli, K. M.; Anderton, C. R.; Paša-Tolić, L.; Robinson, E. W.; Hess, N. J. *Anal. Chem.* **2015**, *87* (10), 5206−5215.

(5) Kim, S.; Rodgers, R. P.; Marshall, A. G. *Int. J. Mass Spectrom.* **2006**, *251* (2), 260−265.

(6) Kujawinski, E. B.; Behn, M. D. *Anal. Chem.* **2006**, *78* (13), 4363−4373.

(7) Kujawinski, E. B.; Longnecker, K.; Blough, N. V.; Del Vecchio, R.; Finlay, L.; Kitner, J. B.; Giovannoni, S. J. *Geochim. Cosmochim. Acta* **2009**, *73* (15), 4384−4399.

(8) Kujawinski, E. B.; Longnecker, K. *Compound Identification Algorithm*, 2016; available via the Internet at: https://github.com/KujawinskiLaboratory/findformula.

(9) Kind, T.; Fiehn, O. *BMC Bioinf.* **2007**, *8* (1), 105.

(10) Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Energy Fuels* **2001**, *15* (2), 492−498.

(11) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Energy Fuels* **2001**, *15* (5), 1186−1193.

(12) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75* (20), 5336−5344.

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

(14) Koch, B. P.; Dittmar, T.; Witt, M.; Kattner, G. *Anal. Chem.* **2007**, *79* (4), 1758−1763.

(15) Richardson, S. D. *Anal. Chem.* **2008**, *80* (12), 4373−4402.

(16) Gonsior, M.; Schmitt-Kopplin, P.; Stavklint, H.; Richardson, S. D.; Hertkorn, N.; Bastviken, D. *Environ. Sci. Technol.* **2014**, *48* (21), 12714−12722.

(17) Roullier, C.; Guitton, Y.; Valery, M.; Amand, S.; Prado, S.; Robiou du Pont, T.; Grovel, O.; Pouchus, Y. F. *Anal. Chem.* **2016**, *88* (18), 9143−9150.

(18) Gonsior, M.; Mitchelmore, C.; Heyes, A.; Harir, M.; Richardson, S. D.; Petty, W. T.; Wright, D. A.; Schmitt-Kopplin, P. *Environ. Sci. Technol.* **2015**, *49* (15), 9048−9055.

(19) Webb, K.; Bristow, T.; Sargent, M.; Stein, B. *Methodology for Accurate Mass Measurement of Small Molecules: Best Practice Guide*; LGC, Ltd.: London, 2004.

(20) Ipsen, A. *Anal. Chem.* **2014**, *86* (11), 5316−5322.

(21) Ipsen, A. *ecipex: Efficient calculation of fine structure isotope patterns via Fourier transforms of simplex-based elemental models*, 2014;

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665

available via the Internet at: https://cran.r-project.org/web/packages/ecipex/index.html.
(22) Rockwood, A. L.; Van Orden, S. L. *Anal. Chem.* **1996**, *68* (13), 2027−2030.
(23) Fernandez-de-Cossio, J. *Anal. Chem.* **2010**, *82* (15), 6726−6729.
(24) Shiraishi, H.; Pilkington, N. H.; Otsuki, A.; Fuwa, K. *Environ. Sci. Technol.* **1985**, *19* (7), 585−590.
(25) Lavonen, E. E.; Gonsior, M.; Tranvik, L. J.; Schmitt-Kopplin, P.; Köhler, S. J. *Environ. Sci. Technol.* **2013**, *47* (5), 2264−2271.
(26) Zhang, H.; Zhang, Y.; Shi, Q.; Ren, S.; Yu, J.; Ji, F.; Luo, W.; Yang, M. *Water Res.* **2012**, *46* (16), 5197−5204.
(27) Zwiener, C.; Richardson, S. D. *TrAC, Trends Anal. Chem.* **2005**, *24* (7), 613−621.
(28) Dittmar, T.; Koch, B. P.; Hertkorn, N.; Kattner, G. *Limnol. Oceanogr.: Methods* **2008**, *6*, 230−235.
(29) Kind, T.; Fiehn, O. *BMC Bioinf.* **2006**, *7* (1), 234.
(30) Hendrickson, C. L.; Quinn, J. P.; Kaiser, N. K.; Smith, D. F.; Blakney, G. T.; Chen, T.; Marshall, A. G.; Weisbrod, C. R.; Beu, S. C. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (9), 1626−1632.
(31) Shaw, J. B.; Lin, T. Y.; Leach, F. E.; Tolmachev, A. V.; Tolić, N.; Robinson, E. W.; Koppenaal, D. W.; Paša-Tolić, L. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (12), 1929−1936.

DOI: 10.1021/acs.analchem.7b03318
Anal. Chem. 2017, 89, 12659−12665