Anal. Chem. **2009,** *81,* 10106–10115

Total Mass Difference Statistics Algorithm: A New Approach to Identification of High-Mass Building Blocks in Electrospray Ionization Fourier Transform Ion Cyclotron Mass Spectrometry Data of Natural Organic Matter

Erast V. Kunenkov,[†] Alexey S. Kononikhin,[‡] Irina V. Perminova,*[,†] Norbert Hertkorn,[§] Andras Gaspar,[§] Philippe Schmitt-Kopplin,[§] Igor A. Popov,[||] Andrew V. Garmash,[†] and Evgeniy N. Nikolaev[‡]

Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia, Institute for Energy Problems of Chemical Physics RAS, Moscow, Russia, Institute of Ecological Chemistry, Helmholtz Zentrum Munich for Environmental Health, Neuherberg, Germany, and Emanuel Institute of Biochemical Physics RAS, Moscow, Russia

The ultrahigh-resolution Fourier transform ion cyclotron resonance (FTICR) mass spectrum of natural organic matter (NOM) contains several thousand peaks with dozens of molecules matching the same nominal mass. Such a complexity poses a significant challenge for automatic data interpretation, in which the most difficult task is molecular formula assignment, especially in the case of heavy and/or multielement ions. In this study, a new universal algorithm for automatic treatment of FTICR mass spectra of NOM and humic substances based on total mass difference statistics (TMDS) has been developed and implemented. The algorithm enables a blind search for unknown building blocks (instead of a priori known ones) by revealing repetitive patterns present in spectra. In this respect, it differs from all previously developed approaches. This algorithm was implemented in designing FIRAN-software for fully automated analysis of mass data with high peak density. The specific feature of FIRAN is its ability to assign formulas to heavy and/or multielement molecules using "virtual elements" approach. To verify the approach, it was used for processing mass spectra of sodium polystyrene sulfonate (PSS, M_w = 2200 Da) and polymethacrylate (PMA, M_w = 3290 Da) which produce heavy multielement and multiply-charged ions. Application of TMDS identified unambiguously monomers present in the polymers consistent with their structure: $C_8H_7SO_3Na$ for PSS and $C_4H_6O_2$ for PMA. It also allowed unambiguous formula assignment to all multiply-charged peaks including the heaviest peak in PMA spectrum at mass 4025.6625 with charge state 6− (mass bias −0.33 ppm). Application of the TMDS-algorithm to processing data on the Suwannee River FA has proven its unique capacities in analysis of spectra with high peak density: it has not only identified the known small building blocks in the structure of FA such as CH_2, H_2, C_2H_2O, O but the heavier unit at 154.027 amu. The latter was identified for the first time and assigned a formula $C_7H_6O_4$ consistent with the structure of dihydroxyl-benzoic acids. The presence of these compounds in the structure of FA has so far been numerically suggested but never proven directly. It was concluded that application of the TMDS-algorithm opens new horizons in unfolding molecular complexity of NOM and other natural products.

High-resolution Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) is a powerful analytical tool, which provides a lot of information on complex mixtures including biosamples[1,2] and various kinds of natural organic matter (NOM).[3−6] NOM is ubiquitous in the environment occurring in soil, water, and air. As the products of stochastic synthesis, NOM has elemental compositions that are nonstoichiometric and structures that are irregular and heterogeneous.[7] The full scale of molecular complexity of NOM has been revealed only after application of FTICR MS. The ultrahigh-resolution FTICR mass spectrum of NOM proved to contain several thousand peaks with dozens of molecules matching the same nominal mass.[6] Such a complexity poses a significant challenge for automatic data

* To whom correspondence should be addressed. E-mail: iperm@org.chem.msu.ru.

[†] Lomonosov Moscow State University.

[‡] Institute for Energy Problems of Chemical Physics RAS.

[§] Helmholtz Zentrum Munich.

[||] Emanuel Institute of Biochemical Physics RAS.

(1) Damoc, E.; Youhnovski, N.; Crettaz, D.; Tissot, J.-D.; Przybylski, M. *Proteomics* **2003**, *3*, 1425–1433.

(2) Hemmingsen, P. V.; Kim, S.; Pettersen, H. E.; Rodgers, R. P.; Sjoblom, J.; Marshall, A. G. *Energy Fuels* **2006**, *20*, 1980–1987.

(3) Kujawinski, E. B.; Freitas, M. A.; Zang, X; Hatcher, P. G.; Green-Church, K. B.; Jones, R. B. *Org. Geochem.* **2002**, *33*, 171–180.

(4) Llewelyn, J. M.; Landing, W. M.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2002**, *74*, 600–606.

(5) Hertkorn, N.; Benner, R.; Frommberger, M.; Schmitt-Kopplin, P.; Witt, M.; Kaiser, K.; Kettrup, A.; Hedges, J. I. *Geochim. Cosmochim. Acta* **2006**, *70*, 2990–3010.

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

(7) Hayes, M. H. B.; MacCarthy, P.; Malcolm, R. L.; Swift, R. S., Eds.; *Humic Substances II: In Search of Structure*; Wiley: Chichester, U.K., 1989.

10.1021/ac901476u CCC: $40.75
Published on Web 11/11/2009

interpretation, in which the most difficult task is assigning molecular formulas to heavy and/or multielement ions.[8] Given that NOM consists of multielement (major elements are C, H, O, N, P, and S) and polymolecular compounds containing high and low molecular weight fractions, a significant portion of NOM may remain undefined during mass spectrometric analysis.[9]

Mass resolving power is known to decrease with increasing m/z,[10] while the number of possible elemental compositions increases dramatically.[11] It was shown that a mass resolution of 5 000 000 and mass accuracy of 0.1 mDa should be sufficient to unambiguously detect the true elemental composition for all theoretically possible ions made up of C, H, O, N, and S in a mass range up to 500 Da.[8] This accuracy is only reachable with high-end spectrometers available at the moment. More sophisticated algorithms of mass data processing should be employed to obtain extra information that is necessary in order to remove the ambiguity in assigning molecular formulas to high molecular weight compounds. Such algorithms utilizing techniques from computer science, graph theory, and discrete mathematics are already applied to solve many different problems related to mass spectrometry.[12]

The position of all peaks in a FTICR mass spectrum of NOM follows a certain structure. This structure may appear because of heterogeneity of the analyzed system or because of fragmentation of sample molecules. Recently, it was reported that the fragmentation of NOM cannot be minimized using soft ionization methods like electrospray ionization (ESI).[13,14] The probability of NOM fragmentation is higher for heavier ions as these require a higher cone voltage to direct them into the aperture of the mass spectrometer.[15] As a result, the relative abundance of peaks with masses above 1000 Da is usually low. Molecular formulas cannot be assigned to such heavy peaks directly by solving Diophantine equations, which is the traditional way of compound identification in FTICR mass spectra.[12] To determine molecular formulas of these peaks, extra information carried over by the data structure should be employed. This structure is usually present in the form of some repetitions observed in the mass spectrum. All published mass spectra of NOM samples exhibit remarkably regular patterns despite their stochastic compositions.[3,6,11,16,17] A lot of peak clusters differ by nominal mass of 14 (CH_2 unit) or 2 (H_2 unit or a number of double bonds). These major mass differences can be utilized for formula assignment by applying Kendrick mass defect (KMD) analysis[18] to FTICR mass spectra of NOM. This has been successfully realized for molecular formula extension to high weight multielement compounds along homologous series in the complex FTICR mass spectra of oils,[19] asphaltenes,[20] and humic substances.[21] However, the KMD analysis is limited to repetitive systems with a priori known functional groups and does not reveal other building blocks. Another way of finding different homologous series in mass spectra is van Krevelen graphic−statistical analysis, which was developed for identifying evolution patterns in coals.[22−25] Kim et al.[26] applied this method to graphic representation of the chemical compositions assigned to each peak in the FTICR mass spectrum of NOM. The same authors[26] have proposed to use the combined method which implies application of Kendrick mass defect analysis to the peaks producing trend lines in the van Krevelen diagrams. This approach facilitated identification of prevalent low mass differences corresponding to structural units like CH_2O, C_2H_2, and C_2H_4 between the peaks seen in an ultrahigh-resolution mass spectrum of NOM. However, van Krevelen analysis did not resolve the problem of high mass compounds identification, because it requires all individual ions to be previously assigned a molecular formula. Nevertheless, this approach demonstrated clearly the huge potential of mass difference analysis for data processing of high-resolution spectra.[26,27]

Recently, Kujawinski and Behn[17] presented the compound identification algorithm (CIA) for molecular formula assignment based on a predefined set of mass differences related to functional group relationships (or small building blocks) typically present in NOM (CH_2, H_2, C_2H_4O, CH_4O_{-1}, and so on). This approach is very effective for formula assignment but can only be applied to a limited set of repetitive systems with predetermined functional group relationships. To perform automated interpretation of FTICR mass spectra that would not be limited to a priori knowledge about the analyzed system, a more universal algorithm capable of identifying new building blocks (including high mass ones) should be developed.

In proteomics, such algorithms were recently developed by Pevzner et al.[28] and Zubarev et al.[29] Both methods are based on statistical analysis of mass differences and imply a blind search for unknown post-translational modifications in peptides. However, the methods were developed for analysis of tandem mass spectrometry data of peptides and cannot be transferred to FTICR MS data on NOM.

The objectives of this research are (1) to develop a universal algorithm for a blind search of unknown building blocks based on total mass difference statistics (TMDS) analysis of the repetitive

(8) Kim, S.; Rodgers, R. P.; Marshall, A. G. *J. Mass. Spectrom.* **2006**, *251*, 260–265.
(9) Hedges, J. I.; Eglinton, G.; Hatcher, P. G.; Kirchman, D. L.; Arnosti, C.; Derenne, S.; Evershed, R. P.; Kogel-Knabner, I.; de Leeuw, J. W.; Littke, R.; Michaelis, W.; Rullkotter, J. *Org. Geochem.* **2000**, *31*, 945–958.
(10) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1–35.
(11) Hertkorn, N.; Ruecker, C.; Meringer, M.; Gugish, R.; Frommberger, M.; Perdue, E. M.; Witt, M.; Schmitt-Kopplin, P. *Anal. Bioanal. Chem.* **2007**, *389*, 1311–1327.
(12) Meija, J. *Anal. Bioanal. Chem.* **2006**, *385*, 486–499.
(13) These, A.; Reemtsma, T. *Anal. Chem.* **2003**, *75*, 6275–6281.
(14) Reemtsma, T.; These, A.; Springer, A.; Linscheid, M. *Wat. Res.* **2008**, *42*, 63–72.
(15) Hunt, S. M.; Sheil, M. M.; Belov, M. B.; Derrick, P. J. *Anal. Chem.* **1998**, *70*, 1812–1822.
(16) Kujawinski, E. B.; Hatcher, P. G.; Freitas, M. A. *Anal. Chem.* **2002**, *74*, 413–419.
(17) Kujawinski, E. B.; Behn, M. D. *Anal. Chem.* **2006**, *78*, 4363–4373.
(18) Kendrick, E. *Anal. Chem.* **1963**, *35*, 2146–2154.
(19) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Qian, K.; Robbins, W. K. *Org. Geochem.* **2002**, *33*, 743–759.
(20) Rodgers, R. P.; Marshall, A. G. *Asphaltenes, Heavy Oils and Petroleomics*; Springer: New York, 2007; pp 63−93.
(21) Stenson, A. C.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2003**, *75*, 1275–1284.
(22) van Krevelen, D. W. *Fuel* **1950**, *29*, 269–284.
(23) Hatcher, P. G.; Lerch, H. E.; Bates, A. L.; Verheyen, T. V. *Org. Geochem.* **1989**, *14*, 145–155.
(24) Bostick, N. H.; Daws, T. A. *Org. Geochem.* **1994**, *21*, 35–49.
(25) Curiale, J. A.; Gibling, M. R. *Org. Geochem.* **1994**, *21*, 67–89.
(26) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336–5344.
(27) Stenson, A. C.; Marshall, A. G.; Cooper, W. T. *Anal. Chem.* **2003**, *75*, 1275–1284.
(28) Tanner, S.; Shu, H.; Frank, A.; Wang, L.-C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. *Anal. Chem.* **2005**, *77*, 4626–4639.
(29) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2006**, *5*, 935–948.

patterns present in spectra, (2) to design software based on the developed algorithm for fully automated analysis of mass data with high peak density, (3) to verify the developed data processing technique by assigning molecular formulas to FTICR mass data of synthetic polyelectrolytes, (4) to assign molecular formulas to FTICR mass data on natural organic matter using the developed software based on the TMDS approach. As far as we are aware, the proposed data processing technique has never been applied to the analysis of FTICR mass spectra of NOM.

EXPERIMENTAL METHODS

Suwannee River Fulvic Acid and Model Polyelectrolyte Samples. Suwannee River fulvic acid standard (SRFA, sample id 1S101F) was acquired from the International Humic Substances Society and was stored frozen in darkness. Separate polystyrenesulfonate (sodium salt, PSS) and polymethacrylate (sodium salt, PMA) standards of 2200 and 3290 nominal average molecular weight, respectively, were obtained from Polymer Standard Service (Mainz, Germany). The samples were dissolved immediately before use. Electrospray solvents included Milli-Q water and methanol (MeOH; Fisher, Purge and Trap-Grade). SRFA and PMA were dissolved in MeOH at concentrations of 50 and 10 mg/L, respectively. PSS was dissolved in a solution of 50:50 MeOH−H_2O at a concentration of 50 mg/L.

Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. All samples were ionized in negative-ion mode because of their abilities to form organic anions in the solutions. FTICR mass spectra of PSS 2200 and SRFA were acquired using a commercial 7 T Finnigan linear quadrupole ion trap-Fourier transform (LTQ FT) mass spectrometer (Thermo Electron Corp., Bremen, Germany) equipped with Ion Max electrospray ion source located at the facilities of the Emanuel Institute of Biochemical Physics RAS (Moscow, Russia). The following conditions were used for electrospray: flow rate 1 μL/min, negative ion mode; needle voltage 3.4 kV; no sheath and auxiliary gas flow; tube lens voltage 130 V; heated capillary temperature 250 °C. Full-scan MS spectra (m/z 200−2000) were acquired in the FTICR with resolution R = 400 000 at m/z 400. The automatic gain control (AGC) target for FTICR MS was set to 1×10^6, corresponding to the number of ions accumulated in the linear ion trap and transferred to the ICR cell. Maximum injection time to fill the linear ion trap was set to 1 s. The average FTICR mass spectrum was a sum of 100 consecutive scans. The LTQ FT tuning mix was used for external mass calibration. In the case of SRFA, the acquired spectra were internally recalibrated to mass measurement error <0.5 ppm (parts per million) using peaks of fatty acids usually occurring in the FTICR mass spectra of SRFA (e.g., $C_{12}H_{23}O_2$, $C_{14}H_{27}O_2$, $C_{16}H_{31}O_2$, $C_{18}H_{35}O_2$, $C_{20}H_{39}O_2$, $C_{22}H_{43}O_2$, $C_{24}H_{47}O_2$). In the case of polymers, the distinct peaks of monomers and oligomers were used for this purpose. FTICR data were obtained as processed mass spectra with an associated peak list using Qual Browser version 1.4 (Thermo Electron Corp., Bremen, Germany).

The FTICR mass spectrum of PMA 3290 was acquired using a 12 T Apex Qe (Bruker Daltonics, Billerica, MA) Fourier transform ion cyclotron resonance mass spectrometer equipped with an Apollo II electrospray ion source located at the facilities of the HelmholtzZentrum Munich for Environmental Health (Germany). The following conditions were used for electrospray: flow rate 2 μL/min; negative ion mode; spray shield 3 kV; capillary voltage 3.5 kV. Ions were externally accumulated in the hexapole for 1 s. The spectra were acquired with a time domain size of 4 mega words with a mass range of 200−2000 m/z. Acquisition time was set to 1.4 ms. Arginine clusters were used for external mass calibration. All acquired spectra were internally recalibrated to mass measurement error <0.1 ppm using monomers and oligomers distinctly seen in the spectrum of PMA. A total of 100 time-domain data were coadded, baseline zeroed, zero-filled once, and fast Fourier transformed using ApexControl 1.2 and Data Analysis (Bruker Daltonics, Billerica, MA).

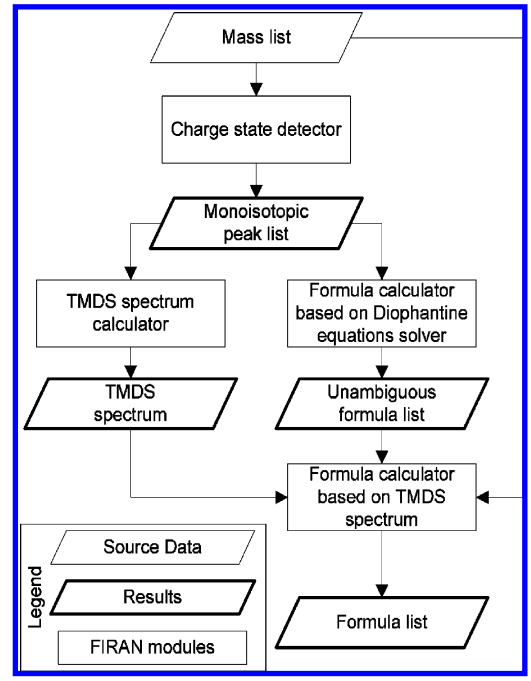

Figure 1. Flowchart of FIRAN software.

Data Analysis. The FTICR mass spectra of PSS 2200 and SRFA acquired using the LTQ FT spectrometer were converted into mass lists using Qual Browser (Thermo Electron Corp., Bremen, Germany). The software does not allow a use of any signal-to-noise (S/N) or relative intensity criteria upon generating a mass list from the FTICR mass spectrum. As a result, all peaks were used as an input data set for the FIRAN program. The FTICR mass spectrum of PMA 3290 (acquired on the Apex Qe) was converted into a mass list with S/N = 3 and relative intensity threshold (base peak) = 0.01% using Data Analysis software (Bruker Daltonics, Billerica, MA). Further mass list processing was performed offline using self-designed FIRAN software (Figure 1). It included calculation of total mass difference statistics (TMDS), chemical building blocks identification, and assigning molecular formulas to individual peaks in mass spectra. The FIRAN software was written in REXX (restructured extended executor) programming language developed by IBM.

RESULTS AND DISCUSSION

TMDS Algorithm. *Identification of Monoisotopic Peaks and Charge State Determination.* Mass difference appearance probabilities were calculated using the following algorithm. First, monoisotopic peaks were identified and their charge states

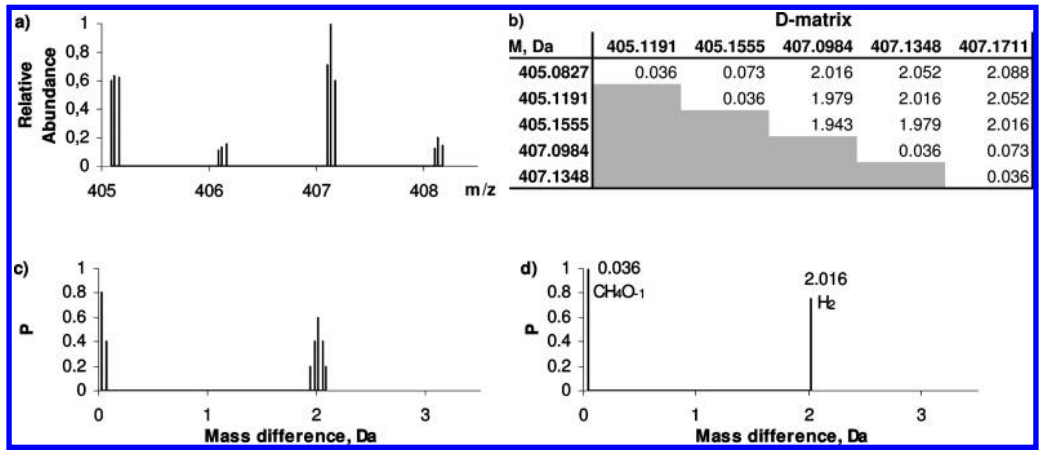

Figure 2. Expanded region of Suwannee river fulvic acid mass spectrum in negative ionization mode (a). D-matrix calculated from monoisotopic ions found in this fragment (b). Total mass difference statistics (TMDS) spectrum calculated from this D-matrix before filtering (c) and after filtering (d).

determined by mass differences between monoisotopic peaks and corresponding ^{13}C isotopologue peaks as described by Stenson et al.[30] All peaks without isotopologue peaks found in the neighborhood were excluded from further consideration. An amount of the excluded peaks for PSS 2200 and PMA 3290 accounted for 10 and 7%, respectively, whereas for SRFA it reached almost 70%. The peak exclusion was undertaken only at the stage of TMDS calculation, but the original peak list was kept intact for formula assigning. With the knowledge of the charge states, ion masses were calculated from the corresponding m/z values.

Calculation of Mass Difference Matrix. The mass difference matrix (D-matrix) is composed of all pairwise mass differences between monoisotopic peaks observed in the spectrum. Figure 2b shows the D-matrix calculated for a narrow window of the Suwannee River fulvic acid mass spectrum (Figure 2a). Each row and column in the obtained D-matrix corresponds to a peak in the mass spectrum. Then, the probabilities of mass difference appearance were calculated. For a given mass difference d, the appearance probability $P(d)$ is defined as the probability of finding a peak at mass m for a randomly selected peak at mass $m \pm d$. It may be estimated as $P(d) = n(d)/(N_{peaks} - 1)$, where $n(d)$ is the total count of appearances of a mass difference equal to d in the D-matrix and N_{peaks} is a total number of monoisotopic peaks in the mass spectrum. Calculation of the D-matrix is quite a time-consuming procedure because it requires the calculation of $(N_{peaks}^2 - N_{peaks})/2$ mass differences. This is the reason why the calculation is limited only to monoisotopic peaks. To minimize the calculation time, the maximum value of mass differences can also be limited. In our work, we have set the maximum value of mass differences equal to 300 amu. It should be specifically noted that this value should be a factor 3−4 less as compared to the width of ion mass distribution in the source mass spectrum. Another important limitation is calculation precision. The calculated mass differences may have low precision, especially when calculated from higher masses that cannot be determined accurately. However, mass differences as a rule are much lower as compared to the source masses,

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

Table 1. Amount of Mass Differences ≤300 amu Calculated from the Mass Spectrum of Sodium Polymethacrylate Containing 1684 Peaks of Monoisotopic Ions

	no. of decimal digits after decimal point		
	3	4	5
no. of mass differences calculated	115 624	221 419	254 506

and even in a wide range of errors the amount of acceptable formulas obtained through solving the Diophantine equation[12] cannot be very high. Therefore, a high-precision calculation of mass differences in the D-matrix is useless. The influence of the number of decimal digits taken for calculations is shown in Table 1. It can be seen that addition of the fourth digit into calculated mass differences doubles their total amount, while addition of the fifth digit has a much smaller effect because there are only a few equal mass differences at this precision level, all of them differ from one another due to random errors only.

Once the probabilities of mass differences have been calculated, the TMDS spectrum can be derived by plotting mass differences versus the appearance probabilities as shown in Figure 2c. The most abundant peaks in this plot may reflect the "building blocks" of the sample. These blocks are characteristic to sample genesis and represent homologous series. However, peaks in such plots are wide and noise-affected (Figure 3) because of random factors. So, even if only three decimal digits are used for mass difference calculation, we still have to convert the obtained TMDS spectrum into discrete form to remove noise from the peaks.

Conversion of TMDS Spectrum into Discrete Form. The complete algorithm of peak conversion and filtration consists of a great number of steps which can be combined into the following principal blocks: (1) a priori (based on magnitude of errors for mass difference measurement) estimation of mass difference ranges corresponding to a single peak, (2) integration of peaks within those ranges, (3) estimation of the fraction of "noisy" peaks in the source mass spectrum which do not come from the original compound but appear because of noise, solvent impurities, etc., (4) recalculation of peak probabilities, and (5) filtering off peaks

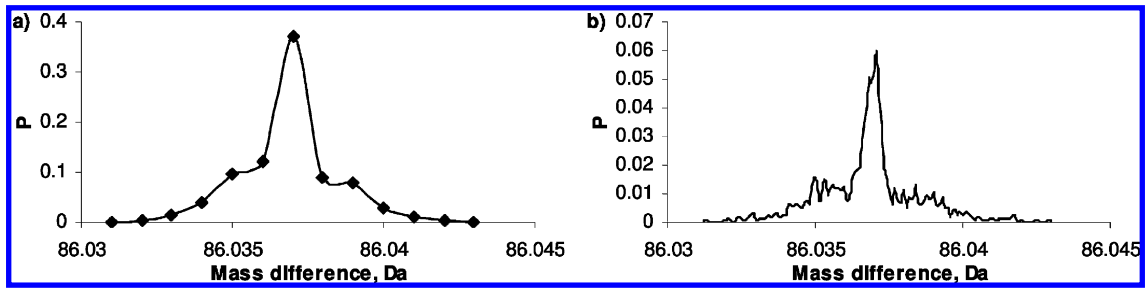

Figure 3. Single peak in the TMDS spectrum (before conversion) calculated from the mass spectrum of sodium polymethacrylate with M_w = 3290 Da given in Figure 5a with 3 (a) and 4 (b) digits after the decimal point used during the D-matrix creation.

that appeared in the TMDS spectrum because of unsystematic combinations of more mass differences with higher probabilities. Operations 3−5 are based on comparison of probabilities $P(|d_1 \pm d_2|)$ and corresponding products $P(d_1)P(d_2)$, where d_1 and d_2 are two mass differences in the TMDS spectrum.

In practice, it is impossible to consider all combinations of all mass differences in the filtering procedure, because it requires $\sim N_{peaks}{}^4$ operations, where N_{peaks} is the number of monoisotopic ions found in the original mass spectrum. As a result, "low probability cutoff" P_{low} must be introduced to exclude all mass differences with lower probabilities. The reasonable value for P_{low} seems to be 0.2. It was chosen as an acceptable compromise for the high-density spectra used in this study: it helps to reduce the total number of operations but still keep important mass differences in consideration. For less populated spectra, P_{low} might be set to the lower values; and for the higher density spectra, it might be set even to the higher values.

An example of a filtering procedure result is presented in Figure 2d. It shows that all monoisotopic peaks from the original spectrum (Figure 2a) differ only by hydrogenation and substitution between O and CH_4.

Assigning Formulas. Assigning stoichiometric formulas to mass differences is performed in the same way as to the peaks of the original spectrum, i.e., by solving corresponding Diophantine equations.[12] The key difference between assigning formulas to mass differences and to peaks in the mass spectrum is that mass difference formulas may have negative indexes (e.g., sodium polystyrenesulfonate tends to exchange sodium with hydrogen in the solutions, so mass difference NaH_{-1} may appear). The calculated mass differences with formulas assigned can be used for assigning formulas to peaks from the original mass spectrum in two ways. The first way ("formula extension"[17] method) is acceptable for highly heterogeneous mixtures with very wide and continuous mass distribution that includes low masses, where some peaks can be assigned to formulas unambiguously. In this case, calculated mass differences may be used as stairs between ions with low molecular weight (with formulas assigned unambiguously) and heavier ions: the difference in formulas of two ions is expected to be the same as the formula assigned to their mass difference. The formula assignment algorithm implemented in FIRAN takes unambiguously determined formulas, finds all peaks with masses $m \pm d$ (where m is the mass of a peak with known formula and d is one of the mass differences, for which the stoichiometric formula has been determined), assigns formulas to these peaks, and repeats these steps until at least one new peak can be assigned. This method is based on the same principle as the CIA method developed by Kujawinski and Behn,[17] but it uses mass differences found by the TMDS algorithm instead of a predefined set of functional group relationships.

Another way of using mass difference formulas ("virtual elements" method) may be suitable if there is a lack of unambiguously identified peaks (all ions are multielement or have high masses). In this case, identified mass differences can be included into the Diophantine equation as masses of "virtual" chemical elements instead of masses of individual chemical elements. The corresponding Diophantine equation may look like the following:

$$m_{CcHhXxYyZzWw} = cm_C + hm_H + xm_X + ym_Y + zm_Z + wm_W$$

where c, h, x, y, z, and w are integer numbers (stoichiometric coefficients), m_C and m_H are masses of individual chemical elements (atoms of carbon and hydrogen), whereas m_X, m_Y, m_Z, and m_W are masses of "virtual elements" corresponding to mass difference formulas identified using the TDMS algorithm.

As it can be seen from the given example, the true elements' masses should also be present in the Diophantine equation, but the range of correspondent variables' values (stoichiometric coefficients) may be limited to low numbers because a common molecule in a repetitive system may be expected to consist of some "building blocks" or relatively heavy structural units identified by corresponding mass differences and a few extra atoms. This approach can reduce the number of Diophantine equation solutions significantly, as it is shown below.

Polystyrenesulfonate. To test the developed method, a sodium salt of polystyrenesulfonate with M_w = 2200 Da (PSS 2200) was used as a model compound. The PSS was chosen because it is widely used as a calibration standard for molecular weight determination of humic substances and NOM using size exclusion chromatography (SEC).[31] Similar to HS, it is negatively charged at neutral pH since it consists of an aromatic backbone substituted with strongly acidic functional groups. So, we expected PSS to produce heavy multielement multiply-charged ions, most of which cannot be identified unambiguously without extra information (they may have a number of possible formulas in the range of mass measurement accuracy, 0.5 ppm).

The mass spectrum of PSS 2200 contains a lot of ions and has only few visible regularities (Figure 4a). Most of the ions are

(31) Perminova, I. V.; Frimmel, F. H.; Kovalevskii, D. V.; Abbt-Braun, G.; Kudryavtsev, A. V.; Hesse, S. *Wat. Res.* **1998**, *32*, 872–881.

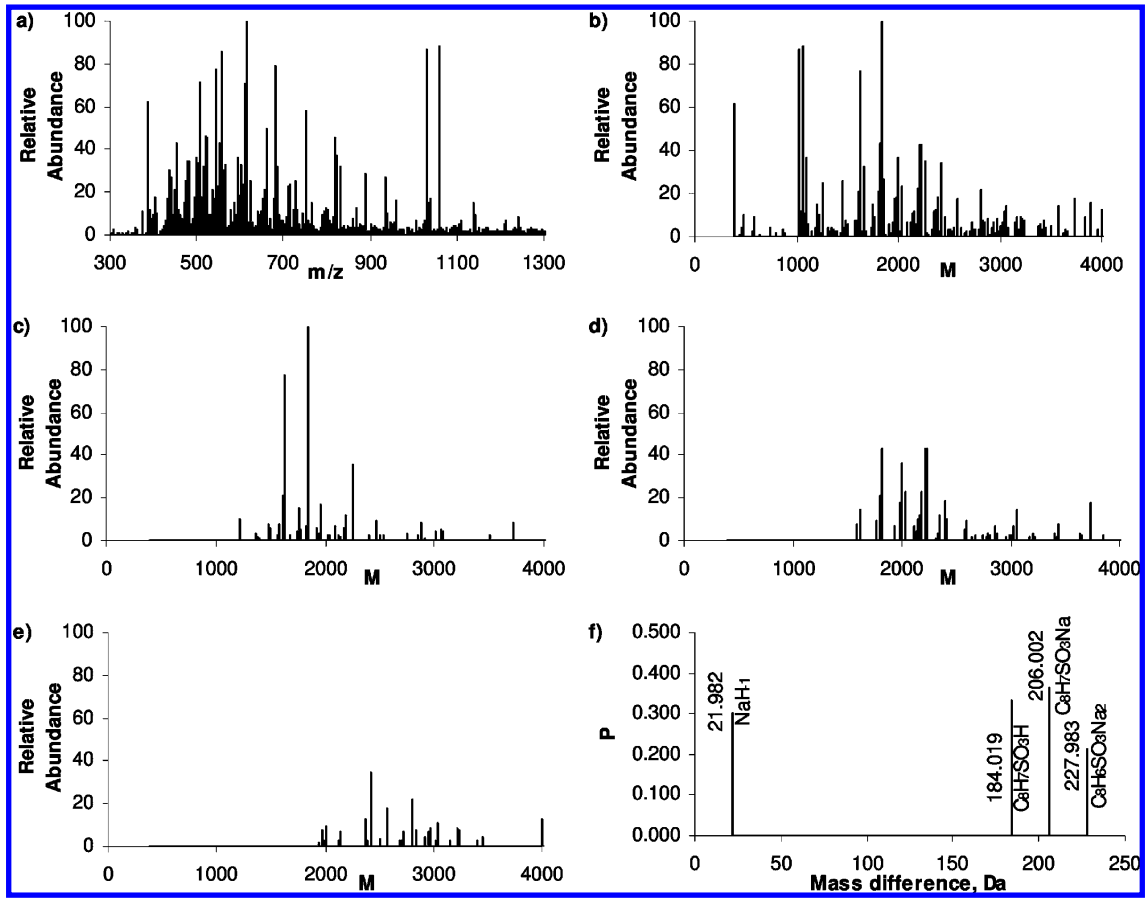

Figure 4. Mass spectrum (a) of sodium polystyrenesulfonate with M_w = 2200 Da in the negative ionization mode. Ion mass distribution calculated from this mass spectrum for all ions (b), for ions with charges 3− (c), 4− (d), and 5− (e). TMDS spectrum (f) calculated from this mass spectrum.

Table 2. Results of Direct Formula Assignment to Most Abundant Peaks in FTICR Mass Spectrum of Sodium Polystyrenesulfonate with M_w = 2200 Da in Negative Ionization Mode[a]

mass	charge state	no. of possible formulas that fit given limits
389.01350	1−	8
1637.12627	3−	51
1843.12827	3−	37

[a] Limitations used for formula assignment: H ≤ 80, O ≤ 30, S ≤ 10, Na ≤ 8, DBE ≥ 0, unpaired electrons disallowed, and mass measurement accuracy 0.5 ppm. No thresholds values were set for C atom content.

Table 3. Results of Formula Assignment to Most Abundant Mass Differences in TMDS Spectrum from Figure 4f[a]

mass difference	no. of possible formulas that fit given limits	formula	error, ppm (theoretical mass difference minus the measured one)
184.020	1	$C_8H_7SO_3H$	+3.18
206.002	1	$C_8H_7SO_3Na$	+3.12
227.983	1	$C_8H_6SO_3Na_2$	−1.31

[a] Limitations used for formula assignment: 0 ≤ C ≤ 80, H ≤ 80, O ≤ 30, S ≤ 10, Na ≤ 8, DBE ≥ 0, unpaired electrons disallowed, and mass difference measurement accuracy 3.5 ppm.

multiply charged (z = 3− to 5−, Figure 4c−e). As it was expected, their high molecular weights (Figure 4b) and complex composition (five elements) make direct determination of exact formulas impossible (Table 2).

At the same time, application of the TMDS approach simplifies the spectrum significantly by visualizing regularity of the PSS structure and indicating the presence of major structural units (Figure 4f). The three most intensive mass differences can be identified much more distinctly than ions with the most abundant peaks themselves (see Table 3).

Determination of formulas corresponding to most abundant differences gives an opportunity to identify the monomer unit of the polymer analyzed ($C_8H_7SO_3Na$). This information was used to remove ambiguity during formula assignment to heavy ions using the virtual element approach. For doing so, the found mass differences were used as "virtual elements" in the formula assignment process to construct a molecule like $C_mH_nX_pY_qZ_rW_s$, where C and H are atoms of carbon and hydrogen, whereas X = $C_8H_7SO_3H$, Y = $C_8H_7SO_3Na$, Z = $C_8H_6SO_3Na_2$, and W = NaH_{-1}. Three ions from Table 2 were identified using the following limitations: 0 ≤ C ≤8, −3 ≤ H ≤ 8, 0 ≤ S ≤1, 0 ≤ O ≤3, 0 ≤ X ≤10, 0 ≤ Y ≤10, 0 ≤ Z ≤10, 0 ≤ W ≤10. All formulas found in the range of mass measurement accuracy 0.5 ppm were converted to a common form of the molecular formula to remove identical formulas (e.g., $C_8H_3X_9W_3 = C_8H_3X_8YW_2$ and so on). After this, only one possible formula remained for each of these

Table 4. Results of Formula Assignment to the Most Abundant Peaks in FTICR Mass Spectrum of Sodium Polystyrenesulfonate with M_w = 2200 Da in Negative Ionization Mode Using "Virtual Elements" Method[a]

mass or mass difference	formula assigned using "virtual elements" method	no. of converted formulas which fit given limits	converted formula	error, ppm (theoretical mass minus measured mass)
389.01350	X_2W^-	1	$C_{16}H_{14}S_2O_6Na^-$	+0.01
1637.12627	$C_6H_3X_8W_3^{3-}$	1	$C_{72}H_{62}S_8O_{24}H_2Na_3^{3-}$	+0.02
1843.12827	$C_6H_3X_9W_3^{3-}$	1	$C_{80}H_{69}S_9O_{27}H_2Na_4^{3-}$	+0.37

[a] Mass measurement accuracy is 0.5 ppm.

ions, even for the heaviest one at 1843 amu, and the correspondent formula $C_{80}H_{69}S_9O_{27}H_2Na_4^{3-}$ was an exact match of the polymer structure expected (see Table 4).

The presence of Na in the determined formula can be explained by the polyelectrolyte properties of PSS displayed in a drastic increase in charge density of the polyanion along with its dissociation degree that causes Na ions to produce adducts with highly charged PSS anions.[32]

Polymethacrylate. Sodium polymethacrylate with M_w of 3290 Da (PMA 3290) was used as another model compound for verification of the TMDS algorithm. As in the case of PSS, it is used as a calibration standard for molecular weight determination of HS and NOM using size exclusion chromatography (SEC).[31] As opposed to PSS, PMA has no aromatic rings in its structure, but its functional groups are presented by carboxyls which are most abundant in the structure of HS and NOM. ESI of PMA yields a lot of ions (Figure 5a) with different charge states.

All charge states from 1− to 7− are present in the spectrum. Charges from 1− to 5− are present in comparable amounts (Figure 5c−g), charges 6− and 7− are much less abundant but can be observed as well. The spectrum in general is characterized with a wide molecular mass distribution up to 4 kDa (Figure 5b). The conversion of regular MS data into the TMDS spectrum enables visualization of the monomer unit $C_4H_6O_2$ (see Figure 5h) which represents the most abundant mass difference: measured mass difference of 86.037, theoretic mass difference of 86.0368, no ambiguity in the range of ±0.005 amu).

Formula assignment to the most abundant peaks requires a substitution of ^{12}C by ^{13}C to be taken into account because PMA molecules may contain a lot of carbon atoms, which contradicts our previous assumption that the most abundant peak in each isotopologue series is monoisotopic. The results of direct formula assignment are presented in Table 5.

The information obtained from TMDS was used to remove ambiguity in formula assignment. Unlike the case of PSS, low mass ions can be identified unambiguously, so high mass ions can be identified by referring them to lighter ones using mass differences ("formula extension" method). The "virtual elements" method was also applied to ions with one virtual element $X = C_4H_6O_2$ and the following limitations: $0 \le {}^{12}C \le 4$, $-6 \le H \le 6$, $0 \le O \le 2$, $0 \le {}^{13}C \le 1$, $0 \le X \le 40$. The "virtual elements" method provides unambiguous formulas in all cases. A total of 1901 peaks have been identified using the "formula extension" method. The heaviest identified peak at mass 4025.6625 with charge state 6− has been assigned the formula $C_{190}H_{272}O_{92}^{6-}$ with mass bias of −0.33 ppm.

Humic Substances. The next object for testing capacities of the TMDS algorithm was a sample of standard fulvic acid isolated from the Suwannee River (SRFA). Fulvic acids compose water-soluble fractions of humic substances (HS) and possess low molecular weight and highly oxidized structure. This is confirmed by elemental composition of SRFA determined by elemental analysis (%, mass):[33] C, 52.55; H, 4.40; O, 42.53; N, 1.19. The sample of SRFA is prepared and distributed by the International Humic Substances Society as a reference material of fresh water FA. It was chosen for analysis as, probably, the best studied HS sample in the world that would facilitate comparison of the data obtained using the TMDS algorithm with the results of other researchers.

Numerous studies have demonstrated that HS have very complex mass spectra containing thousands of peaks.[34−36] As it can be seen from Figure 6, the results of our study are consistent with those findings. Nevertheless, thorough inspection of the spectra obtained have shown some regular patterns in the peaks of SRFA (Figure 6a). There are some visible "peak clusters" differing one from another by 14.016 ($-CH_2-$ unit, Figure 6b) and 2.016 (H_2, Figure 6c).

The TMDS approach was used to reveal repetitive structures in FA. Several building blocks with a wide range of masses from 2.016 amu (H_2) to 154.027 amu ($C_7H_6O_4$) were found (Figure 7). Some of these building blocks (CH_2, H_2, C_2H_2O, O) have been already used as "functional group relationships" by Kujawinski and Behn[17] for assigning molecular formulas using the CIA algorithm, while others with higher molecular weights have never been used before (C_3H_2O, $C_7H_6O_4$).

The $C_7H_6O_4$ fragment seems to contain an aromatic ring and may be assigned, for example, to dihydroxyl-benzoic acids. The latter might be formed as a result of lignin degradation. Despite generally accepted knowledge of lignin as one of the major precursors of humic substances,[37] identification of the $C_7H_6O_4$ unit is the first repetitive building block of this type identified in Suwannee River FA.

The information obtained from TMDS was used to remove ambiguity in formula assignment. A total of 3016 formulas were automatically assigned to peaks using the "formula extension" method (mass measurement bias 0.5 ppm was used for preliminary unambiguous formula assignment; the bias was increased to 1 ppm for formulas assigned by the "formula extension" method). The preliminary formula assignment in the 0.5 ppm mass window was carried out using the mass calculator option of Qual Browser 1.4 (The Diophantine equation was solved using the following parameters: $0 \le C \le 80$, $0 \le H \le 160$, $0 \le O \le 40$, $0 \le N \le 1$, DBE ≥ 0, unpaired electrons disallowed, the formula with

(32) Manning, G. S. *J. Phys. Chem.* **1975**, *79*, 262–265.
(33) International Humic Substances Society. http://ihss.gatech.edu/ihss2/elements.html.
(34) Solouki, T.; Freitas, M. A.; Alomary, A. *Anal. Chem.* **1999**, *71*, 4719–4726.
(35) Brown, T. L.; Rice, J. A. *Anal. Chem.* **2000**, *72*, 384–390.
(36) Leenheer, J. A.; Rostad, C. E.; Gates, P. M.; Furlong, E. T.; Ferrer, I. *Anal. Chem.* **2001**, *73*, 1461–1471.
(37) Stevenson, F. J. *Humus Chemistry: Genesis, Composition, Reactions*; Wiley: New York, 1982.

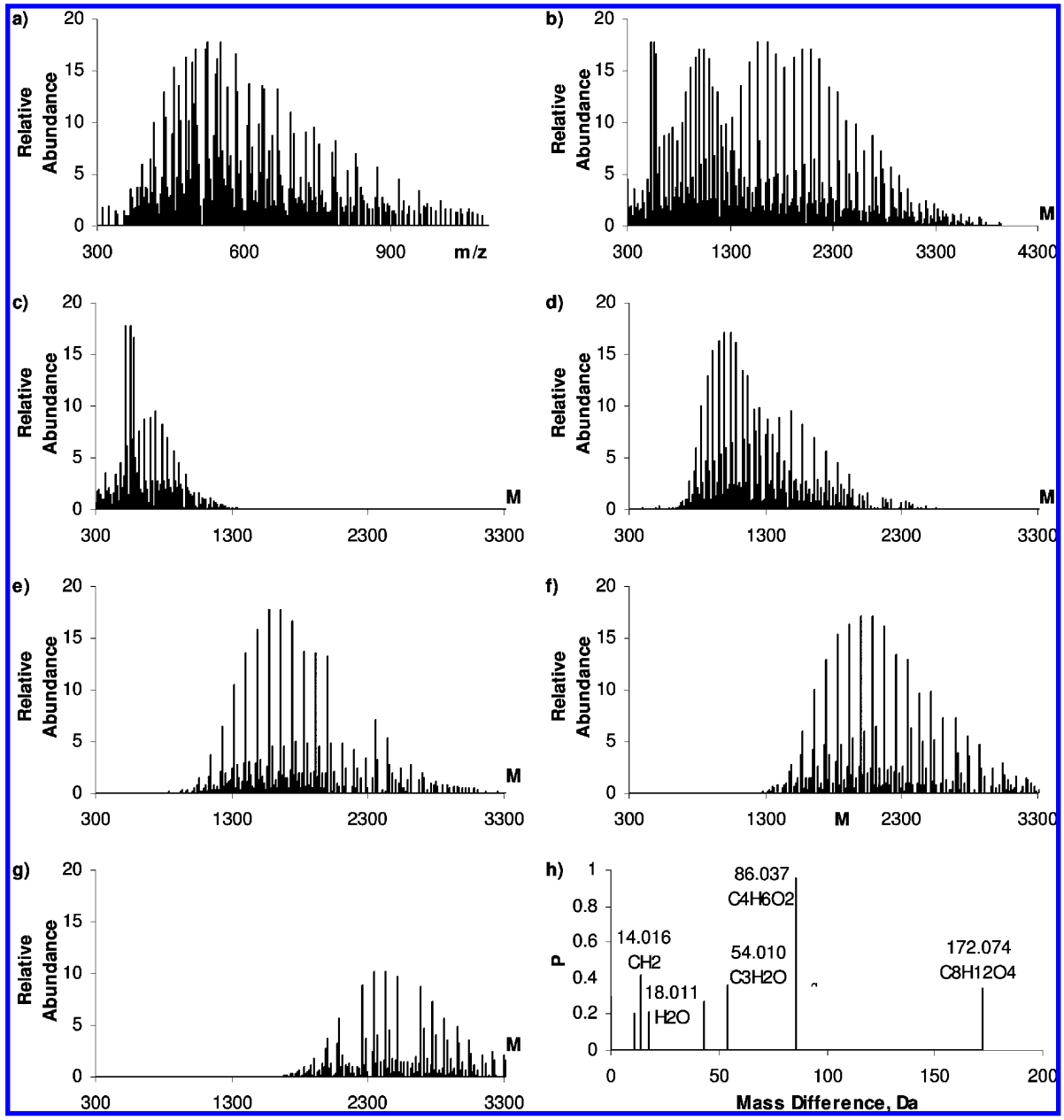

Figure 5. Mass spectrum (a) of sodium polymethacrylate with M_w = 3290 Da in negative ionization mode. Ion mass distribution calculated from this mass spectrum for all ions (b), for ions with charges 1− (c), 2− (d), 3− (e), 4− (f), and 5− (g). TMDS spectrum (h) calculated from this mass spectrum (nonlabeled mass differences are 12 (C) and 42.011(C_2H_2O).

minimal mass bias was selected in case of ambiguity) and gave 2832 formulas in the mass range from 295 to 1000 Da. The sulfur was not included into the formula assignment due to its very low content in the sample (0.44% mass). Hence, identification of sulfur-containing structures would provide only little additional information but drastically complicate the assigning procedure. It should be noted that some of the determined formulas looked rather unlikely (e.g., $C_{77}H_{10}ON^-$). It is difficult to use isotopologue peaks, like it was suggested by Koch et al.,[38] for validation of these formulas because many of these peaks have a poor signal-to-noise ratio and corresponding isotopologue peaks were not always observed (only 1160 monoisotopic peaks with ^{13}C isotopologues in their neighborhood were found by FIRAN, so not more than 1160 formulas could be validated using these method). However, none of these unrealistic formulas were present in the formula list generated using the TMDS approach. A total of 3016 formulas were automatically identified using the "formula extension" method within the nominal mass range from 295 to 1666 Da, the peak at mass 1666.2415 with charge state 2− has been assigned the formula $C_{78}H_{58}O_{42}^{2-}$ with mass bias of 0.07 ppm. Therefore, the use of the TMDS approach enabled substantial improvement of the robustness of molecular formula assignments as well as an extension of the mass range of peaks available for formula assignment and identification of a set of larger repetitive units characteristic to the SRFA sample. We have also recently reported that application of this software tool for identifying multiply-charged species in the FTICR MS spectra of SRNOM[39] allowed for substantial improvement in their formula assignment. The performed differentiation between charge states performed with a use of the TMDS

Table 5. Results of Direct Formula Assignment to Most Abundant Peaks with Different Charge States from FTICR Mass Spectrum of Sodium Polymethacrylate with M_w = 3290 Da in Negative Ionization Mode[a]

exact mass	charge state	no. of possible formulas which fit given limits
1575.6872	3−	1
1661.7245	3−	3
1832.7902	4−	2
1918.8271	4−	3
2349.0070	5−	3
2435.0438	5−	3
2778.1837	6−	4
2864.2206	6−	4

[a] Limitations used for formula assignment: $^{12}C \leq 135$, $^{13}C \leq 1$, $H \leq 200$, $O \leq 70$, DBE ≥ 0, unpaired electrons disallowed, and mass measurement accuracy 0.5 ppm.

algorithm was supported by ion mobility mass spectrometry measurements that implies separation on the basis of size/charge ratios.

We believe that the successful application of the TMDS approach for identifying multiply-charged ions in the spectrum of SRFA will allow us to move forward toward resolving substantial discrepancies in the reported molecular weight estimates of HS measured with a use of SEC and ESI MS.[30,35,36,40] The problem of much higher molecular weight estimates measured by SEC (dozens of kilodaltons) as opposed to those measured for the same samples by ESI MS (hundreds of daltons) has been greatly discussed in the literature, but its solution is still to be found. To contribute into resolving this problem, we currently have work in progress on processing FTICR MS data acquired for a set of humic and fulvic acid samples isolated from different terrestrial sources (water and peat) and fractionated into narrow molecular weight fractions using SEC. By applying the TMDS algorithm, we could already identify a lot of multiply-charged peaks in the positive-ion ESI spectra of humic acid fractions. We hope that comparison of the molecular weight estimates calculated from the TMDS-processed data on humic fractions with the corresponding SEC estimates might provide a clue to the nature of bias observed between ESI MS and SEC estimates and facilitate unanimous opinion on the molecular weights of HS.

The demonstrated capability of TMDS to identify repetitive units in the SRFA spectra allows us to speculate on a big promise this algorithm holds for search and identification of different

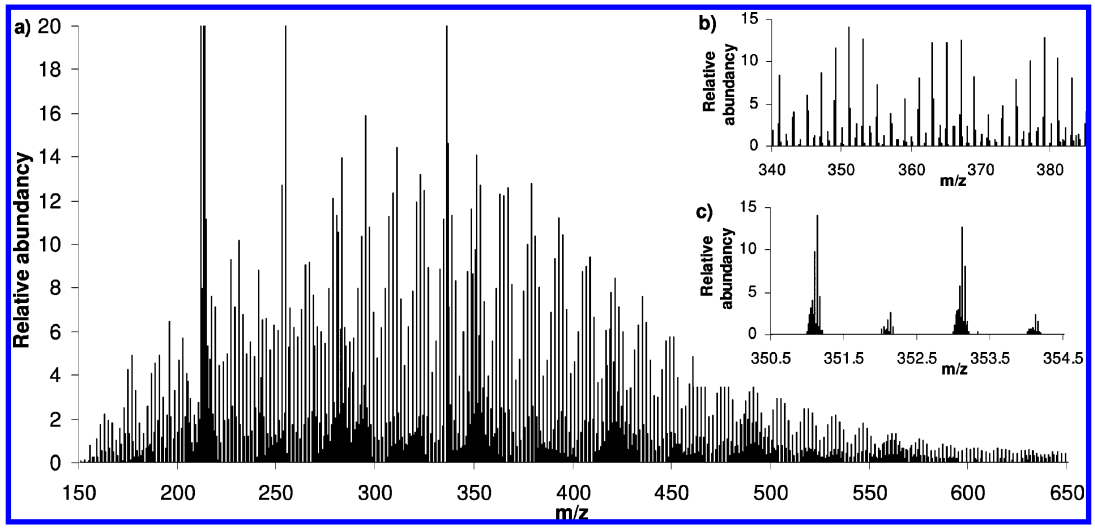

Figure 6. Mass spectrum of Suwannee River fulvic acid in negative ionization mode: (a) broadband spectrum for *m*/*z* between 150 and 650 amu; (b) negative ions between nominal masses 340 and 385; (c) negative ions between nominal masses 351 and 354.

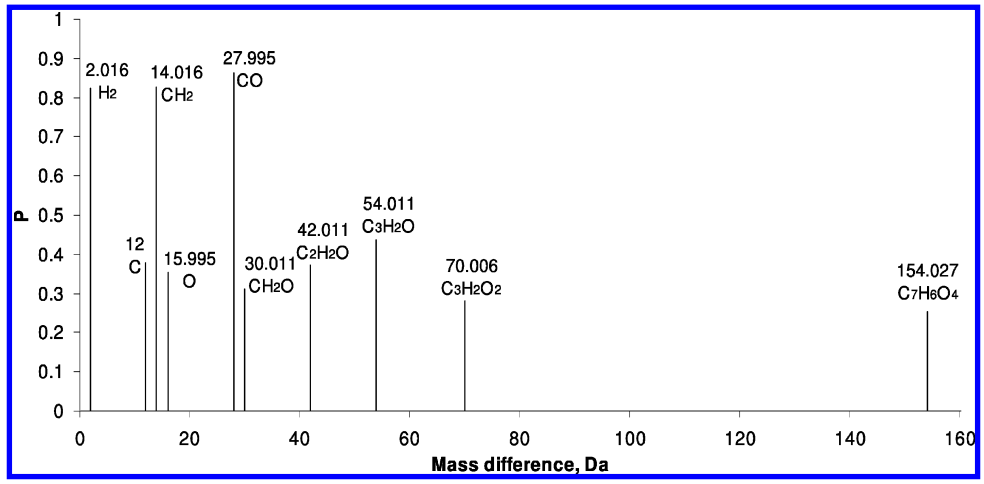

Figure 7. TMDS spectrum of Suwannee River fulvic acid calculated from the mass spectrum given in Figure 6a.

Table 6. Results of Formula Assignment to Most Abundant Peaks with Different Charge States from FTICR Mass Spectrum of Sodium Polymethacrylate with M_w = 3290 Da in Negative Ionization Mode Using "Virtual Elements" and "Formula Extension" Methods[a])

exact mass	formula assigned using "virtual elements" method	formula assigned using "formula extension" method	error, ppm (theoretical mass minus measured mass)
1575.6872	$C_2H_3X_{18}^{3-}$	$C_{74}H_{111}O_{36}^{3-}$	0.04
1661.7245	$C_2H_3X_{19}^{3-}$	$C_{78}H_{117}O_{38}^{3-}$	0.35
1832.7902	$C_2H_2X_{21}^{4-}$	$C_{86}H_{128}O_{42}^{4-}$	0.00
1918.8271	$C_2H_2X_{22}^{4-}$	$C_{90}H_{134}O_{44}^{4-}$	0.06
2349.0070	$C^{13}CHX_{27}^{5-}$	$C_{110}H_{163}O_{54}^{5-}$[b]	0.02
2435.0438	$C^{13}CHX_{28}^{5-}$	$C_{114}H_{169}O_{56}^{5-}$[b]	0.03
2778.1837	$C^{13}CX_{32}^{6-}$	$C_{130}H_{192}O_{64}^{5-}$[b]	0.04
2864.2206	$C^{13}CX_{33}^{6-}$	$C_{134}H_{198}O_{66}^{5-}$[b]	0.09

[a] X in the formulas stand for the $C_4H_6O_2$ fragment. [b] Formulas were assigned to the isotopologue peaks containing one ^{13}C atom.

environmental markers. For example, its application for data interpretation on the large set of humic substances from different sources will facilitate identification of major repetitive units characteristic to metabolic processes taking place in the specific environments which will shed light on the humification mechanisms and set forth a use of humic substances as markers of changing environmental processes including global climate change. Another promising development of the TMDS algorithm is its adaptation for identification of repetitive patterns in fractal systems which might set the grounds for the new formula assigning approach to humic substances and other natural products consisting of complex mixtures of similar but not identical molecules.

CONCLUSIONS

The presented results on application of the proposed TMDS approach both to synthetic and natural polyelectrolytes provide a convincing proof on the large potential of the TMDS algorithm in interpretation of complex mass spectra. It was capable of identifying monomer units in both model polymers as well as of assigning unambiguous molecular formulas to multiply-charged and heavy ions. Application of the TMDS algorithm to mass data on Suwannee River fulvic acid greatly studied by other researchers has allowed for the first time the identification of a new repetitive structural block with a formula $C_7H_6O_4$ assigned to a mass of 154.027 amu. Of particular importance is that the identified repetitive unit may be assigned to dihydroxyl-benzoic acid, which is consistent with the expected structural blocks of fulvic acids stemming from oxidized lignin structures. Identification of this new repetitive unit in the structure of SRFA got feasible thanks only to the major feature of the TMDS algorithm, it is a lack of a priori assumptions on the possible structures present in the investigated compounds. This makes the TMDS algorithm an indispensable tool both in molecular formula assignment to complex mixtures of unknown compositions and in identifying multielement and multiply-charged ions produced by macromolecular compounds. Hence, implementation of the TMDS algorithm in data processing allows for expanding a set of objects that can be successfully analyzed using FTICR mass spectrometry. Currently we seek a much deeper understanding on the nature and role of repetitive structures within humic systems by processing FTICR MS data acquired for a large set of humic substances isolated from different sources (water, soil, peat) and fractionated into fractions with different molecular weights using size exclusion chromatography. In addition to identifying building blocks in oxygen-rich humic heteropolymers, we plan to apply the TMDS algorithm for exploring structure of tholins which are nitrogen rich organic heteropolymers formed in the non-Earth atmosphere. To make this algorithm publicly available, we are currently working at creating a Windows-compatible interface for the developed FIRAN-software, which can be accessed from the Web site of the MGUMUS-research group (http://www.humus.ru/).

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

(39) Gaspar, A.; Kunenkov, E. V.; Lock, R.; Desor, M.; Perminova, I. V.; Schmitt-Kopplin, Ph. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 683–688.

(40) Phillips, S. L.; Olesik, S. V. *Anal. Chem.* **2003**, *75*, 5544–5553.

ACKNOWLEDGMENT

This research was supported by ISTC (Grant KR-964), Russian Foundation for Basic Research Grant 09-03-92500, and the Program of Innovative Education within Green Chemistry Center of Lomonosov Moscow State University.

Received for review July 4, 2009. Accepted October 26, 2009.

AC901476U