Anal Bioanal Chem (2016) 408:2461–2469
DOI 10.1007/s00216-016-9341-2

RESEARCH PAPER

Molecular formula assignment for dissolved organic matter (DOM) using high-field FT-ICR-MS: chemical perspective and validation of sulphur-rich organic components (CHOS) in pit lake samples

Peter Herzsprung[1] · Norbert Hertkorn[2] · Wolf von Tümpling[1] · Mourad Harir[2] · Kurt Friese[1] · Philippe Schmitt-Kopplin[2]

Received: 5 October 2015 / Revised: 12 January 2016 / Accepted: 15 January 2016 / Published online: 16 February 2016

Abstract Molecular formula assignment is one of the key challenges in processing high-field Fourier transform ion cyclotron resonance mass spectrometric (FT-ICR-MS) datasets. The number of potential solutions for an elemental formula increases exponentially with increasing molecular mass, especially when non-oxygen heteroatoms like N, S or P are included. A method was developed from the chemical perspective and validated using a Suwannee River Fulvic Acid (SRFA) dataset which is dominated by components consisting exclusively of C, H and O (78 % CHO). In order to get information on the application range and robustness of this method, we investigated a FT-ICR-MS dataset which was merged from 18 mine pit lake pore waters and 3 river floodplain soil waters. This dataset contained 50 % CHO and 18 % CHOS on average, whereas the former SRFA dataset contained only 1.5 % CHOS. The mass calculator was configured to allow up to five nitrogen atoms and up to one sulphur atom in assigning formulas to mass peaks. More than 50 % multiple-formula assignments were found for peaks with masses > 650 Da. Based on DBE − *O* frequency diagrams, many CHO, $CHOS_1$, $CHON_1$ and $CHON_1S_1$ molecular series were ultimately assigned to many *m*/*z* and considered to be reliable solutions. The unequivocal data pool could thus be enlarged by 523 (6.8 %) $CHOS_1$ components. In contrast to the method validation with CHO-rich SRFA, validation with sulphur-rich pit lake samples showed that formulas with a higher number of non-oxygen heteroatoms can be more reliable assignments in many cases. As an example: CHOS molecular series were reliable and the CHO classes were unreliable amongst other molecular classes in many multiple-formula assignments from the sulphur-rich pit lake samples.

Keywords DOM · FT-ICR-MS · Formula assignment · CHOS/CHON molecular series · Validation

Introduction

The application of high-field Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) has strongly advanced descriptions of molecular composition in natural organic matter (NOM) [1–9]. Possible elemental compositions are commonly computed for each detected mass peak based on predefined chemical constraints. Formula assignment implies the more or less step-by-step exclusion of less reliable elemental compositions in order to finally reach exactly one valid elemental composition for a given mass peak, as far as possible. Literature addressing this issue [10–19] was reported and discussed in Herzsprung et al. [20].

The key problem of formula assignment was reported by Koch et al. [11]: The number of possible solutions for an elemental formula increases dramatically with increasing mass if non-oxygen heteroatoms are taken into consideration for calculation. The maximum number of allowable atoms (for the mass calculation software) is generally defined by the user

Electronic supplementary material The online version of this article (doi:10.1007/s00216-016-9341-2) contains supplementary material, which is available to authorized users.

✉ Peter Herzsprung
peter.herzsprung@ufz.de

[1] UFZ Helmholtz Centre for Environmental Research, Brueckstrasse 3a, 39114 Magdeburg, Germany

[2] Helmholtz Zentrum München, German Research Center for Environmental Health; Research Unit Analytical BioGeoChemistry (BGC), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

based on his requirements and experience. To illustrate, an example is given from the dataset in Koch et al. [11]: 12 elemental formula solutions were found for the mass peak $[M - H]^-$ (molecular ion with $m/z = 567.18712$ Da) with a mass of 568.19439 Da for the neutral component—$C_{30}H_{32}O_{11}$ (I), $C_{43}H_{24}N_2$ (II), $C_{33}H_{33}O_3N_2S_1P_1$ (III), $C_{35}H_{30}N_4P_2$ (IV), $C_{25}H_{37}O_5N_4S_2P_1$ (V), $C_{27}H_{34}O_2N_6S_1P_2$ (VI), $C_{25}H_{29}O_6N_8P_1$ (VII), $C_{15}H_{36}O_{11}N_8S_2$ (VIII), $C_{17}H_{33}O_8N_{10}S_1P_1$ (IX), $C_{28}H_{28}N_{10}S_2$ (X), $C_{20}H_{34}N_{12}S_2P_2$ (XI) and $C_{28}H_{20}O_1N_{14}$ (XII). This problem was mainly circumvented in earlier studies (2001–2006) in which merely the CHO components from FT-ICR-MS datasets were used [1, 2, 4, 7]. A formula assignment procedure is required for excluding potential false formulas if non-oxygen heteroatoms such as those in the example are to be considered. Considerable progress in formula assignment has been achieved mainly by programming sophisticated software based on compositional networks (Kendrick analogous mass difference analysis) or neural networks [18, 19]. The details (and several studies on this topic) were discussed in Herzsprung et al. [20]. The principal weakness of the above-mentioned formula assignment procedures was insufficient transparency from a chemical perspective. A formula assignment method was recently established [20] from a chemical point of view using molecular class frequency versus DBE minus O diagrams (DBE, double bound equivalents; O, number of oxygen atoms) in order to overcome this deficit. The methodology was validated using a dataset from a Suwannee River Fulvic Acid (SRFA) standard, which is comparably scarce in components containing non-oxygen heteroatoms. Validation requires datasets from different sources in order to retrieve information about the application range and robustness of this methodology. Beyond this, the key question on whether formula assignments containing less non-oxygen heteroatoms are generally the more reliable solutions, as assumed by Ohno [12], will also be answered. This study will give a detailed answer on this issue using a dataset comparably rich in organic sulphur (CHOS) components. The distribution of components in the mine pit lake pore water samples used in the present validation is shown in Fig. S8 (see Electronic Supplementary Material (ESM), ESM_1), compared to the distribution of components in the recently investigated SRFA sample [20]. In short, the SRFA sample contained 78 % CHO, 18 % CHON, 2 % CHONS and 1.5 % CHOS. The pit lake dataset, however, contained on average 50 % CHO, 21 % CHON, 11 % CHONS and 18 % CHOS. Herzsprung et al. reported that mine pit lake pore waters contained many CHOS [21], CHON and CHONS components [21] (RCM_4719_sm_SupplInfo1.xls). The abundance of CHOS component peaks in mass spectra was visualized by detail expansion at the nominal mass of 319 Da [21] (Fig. 2). The content of organic sulphur in coal tailing samples was described, making about 30 % of the total sulphur [22]. The acidification process due to pyrite weathering caused by inorganic sulphur components is well understood [23]. In contrast, studies on dissolved organic sulphur components as part of the DOM in mine pit lake waters are rare [21] and the concentration of dissolved organic sulphur was not reported. FT-ICR-MS elemental formula datasets are an appropriate goal in the analysis and understanding of the dissolved organic sulphur fraction in such environments.

Mine pit lake pore waters are characterized by high ionic strength and acidity. The pH of the investigated samples ranged from 2.6 to 7.0 (except for one sample with pH 12). DOC ranged from about 15 to 650 mg/L. Pore waters generally contained considerable amounts of sulphate (more than 200 mg/L and up to 9000 mg/L), ferrous iron (up to 5000 mg/L), aluminium, calcium, and other inorganic ions. This data is reported in [21] in the Supplementary Information, page S12 (RCM_4719_sm_SupplInfo2.doc).

The CHOS-rich pit lake pore water samples are appropriate for the validation of the recently developed formula assignment tool developed by Herzsprung et al. [20]. Here, a dataset had been used which originated from only a single sample (SRFA) dominated by CHO components. For the purpose of validation, a dataset of 21 samples was merged, making group determinations possible. Component groups could then be simultaneously determined for different samples.

The aim of the validation is to improve the chemical perspective in the formula assignment for non-oxygen heteroatom component-rich DOM samples. The question always raised is: which solution would be the most reliable, if two or more formulas were assigned to one mass peak. As an example, two results were found for the mass peak with $m/z = 649.12345$ Da ($[M - H]^-$). One is the component $C_{50}H_{18}O_2$ (mass 650.130680 Da) and the other is $C_{29}H_{30}O_{15}S_1$ (mass 650.130546 Da). With the best will in the world, it seems difficult to construct reasonable isomers (humic-like substances) for the $C_{50}H_{18}O_2$ component; however, many chemically reasonable isomer solutions can be constructed for the $C_{29}H_{30}O_{15}S_1$ component. A partly aliphatic and aromatic carbon skeleton connected with carboxylic, phenolic or one sulfonic acid group is imaginable for this component. The CHOS component would be water soluble, the CHO component likely insoluble. We will now introduce how such chemical-driven decisions can be made on a group basis.

Description of the dataset

Eighteen pore water samples from mine pit lakes and three soil water samples from the River Elbe flood plains (as listed in ESM_1, Table S1) were analysed with FT-ICR-MS. The study sites (SI_1) and the details of the experiments, i.e. sample preparation and FT-ICR-MS analysis (SI_2), are described in ESM_1. At present, only the fraction of DOM extractable by solid-phase extraction (SPE-DOM) is measured with FT-

ICR-MS. Details on this issue are addressed by Dittmar et al. [24]. The analysis of SPE-DOM depends on the ionization method and the ionization efficiency of components [8].

FT-ICR mass spectral peaks with a signal to noise ratio $S/N > 2$ were exported to peak lists. From those lists, feasible elemental formulas (mass error < 1 ppm) were computed for each peak in a batch mode by software developed in-house. The generated formulas were validated by setting sensible/reasonable chemical constraints (element counts: $C \leq 100$, $O \leq 80$, $N \leq 5$, $S \leq 1$; N rule; O/C ratio ≤ 1; H/C ratio $\leq 2n+2$; H/C ratio ≥ 0.3) in conjunction with an automated ^{13}C-isotope pattern comparison and verification.

Data management

All formula assignments with DBE < 0, non-integer values of DBE, $H/C < 0.3$ or $O/C > 1$ were excluded as described in [20]. Then, all elemental formula datasets from the 21 samples were merged in a spreadsheet. A search routine was developed to find singlets, doublets, triplets, quartets and quintets of formulas which matched the identical experimental mass. All mass peaks conforming to only one single formula were tentatively declared unequivocal components. These molecular compositions were collated in the unequivocal data pool. The remaining dataset was sorted by the number (2, 3, 4 or 5) of formulas matching one experimental mass peak, in the following denoted as equivocal components (equivocal data pool). The equivocal dataset was separated into discrete Excel tables, one each for doublets, triplets, quartets and quintets (higher multiple-formula assignments did not occur in this data evaluation). These four data tables are searched for combinations of molecular classes as shown in the database ESM_3.

Results

Size of data pools

From altogether 21 merged samples, 41,690 singlets (the so-called unequivocal data pool) and 6524 multiple-formula assignments were detected. The multiple-formula assignments consisted of 3908 doublets, 2174 triplets, 381 quartets and 61 quintets. Increasing numbers of multiple-formula assignments were found with increasing mass as shown in Table 1. The absolute mass error increases with mass whereas the relative mass error (in ppm) is considered constant over the total mass range. More formula solutions become possible with increasing absolute mass error as shown and discussed in [11].

The unequivocal and equivocal datasets were evaluated using two different parameters. The first is the total number of all elemental formulas found, allowing repeats of the same formula from different samples (N_{rep}). The second parameter is the total number of all elemental formulas found without repeats of the same formula from different samples (N_{div}). N_{div} describes the chemical diversity of the dataset.

The dataset N_{rep} consisted of 41,690 unequivocal formulas and 16,167 equivocal formulas from the datasets used for the judgement and assessment. N_{div} was 14,037 (unequivocal formulas) and 12,246 (equivocal formulas). The possibility of an elemental formula existing in the unequivocal formula dataset (sample A) and in the equivocal formula dataset (sample B) was not considered in the calculation of N_{div}. While not considered in the calculation of N_{div}, such cases were observed in reality in the present dataset: The mass peak with $499.05116 \leq m/z \leq 499.05188$ was declared an unequivocal component (molecular ion) $C_{23}H_{15}O_{13}$ in three samples, whereas it was declared a multiple-formula assignment (doublet $C_{23}H_{15}O_{13}$/$C_{15}H_{19}O_{15}N_2S$) in five samples and as unequivocal (molecular ion) $C_{15}H_{19}O_{15}N_2S$ in one sample (Table 2).

Table 1 Number of multiple-formula assignments as function of mass range of detected molecular ions

Mass range	Singlet	Doublet	Triplet	Quartet	Quintet
	Unequivocal data pool	Equivocal data pool			
< 150 Da	13	0	0	0	0
150–250 Da	7294	0	0	0	0
250–350 Da	14,363	5	0	0	0
350–450 Da	9961	306	1	0	0
450–550 Da	6651	1579	440	0	0
550–650 Da	2041	597	467	20	0
650–750 Da	1000	737	426	65	0
750–850 Da	296	519	466	90	8
850–950 Da	51	140	300	107	25
> 950 Da	20	25	74	99	28

Table 2 Elemental formula solutions for a molecular ion in nine different pore water samples

Sample	Dataset	Formula	DBE − *O*	Decision	Experimental mass *m*/*z*	Calculated mass *m*/*z*	Relative mass deviation ppm
MTT6	*Unequivocal*	$C_{15}H_{19}O_{15}N_2S$	*−8*	*Problematic*[a]	*499.05116*	*499.05117*	*−0.02*
R12T6	Equivocal	$C_{15}H_{19}O_{15}N_2S$	−8	Unreliable	499.05144	499.05117	0.54
R12T6	**Equivocal**	$\mathbf{C_{23}H_{15}O_{13}}$	**3**	**Reliable**	**499.05144**	**499.05182**	**−0.76**
R12T2	Equivocal	$C_{15}H_{19}O_{15}N_2S$	−8	Unreliable	499.05147	499.05117	0.60
R12T2	**Equivocal**	$\mathbf{C_{23}H_{15}O_{13}}$	**3**	**Reliable**	**499.05147**	**499.05182**	**−0.70**
R12T5	Equivocal	$C_{15}H_{19}O_{15}N_2S$	−8	Unreliable	499.05155	499.05117	0.76
R12T5	**Equivocal**	$\mathbf{C_{23}H_{15}O_{13}}$	**3**	**Reliable**	**499.05155**	**499.05182**	**−0.54**
XPT1	Equivocal	$C_{15}H_{19}O_{15}N_2S$	−8	Unreliable	499.05161	499.05117	0.88
XPT1	**Equivocal**	$\mathbf{C_{23}H_{15}O_{13}}$	**3**	**Reliable**	**499.05161**	**499.05182**	**−0.42**
R11T6	Equivocal	$C_{15}H_{19}O_{15}N_2S$	−8	Unreliable	499.05165	499.05117	0.96
R11T6	**Equivocal**	$\mathbf{C_{23}H_{15}O_{13}}$	**3**	**Reliable**	**499.05165**	**499.05182**	**−0.34**
XPT5	*Unequivocal*	$C_{23}H_{15}O_{13}$	*−8*	*Reliable*	*499.05168*	*499.05182*	*−0.28*
WAT1	*Unequivocal*	$C_{23}H_{15}O_{13}$	*−8*	*Reliable*	*499.05184*	*499.05182*	*0.40*
MTT1	*Unequivocal*	$C_{23}H_{15}O_{13}$	*−8*	*Reliable*	*499.05188*	*499.05182*	*0.12*

[a] Found as unequivocal component but might be considered unreliable

This effect can be explained by the statistical spread of molecular ion mass detection. At first sight, the molecular ion $C_{23}H_{15}O_{13}$ seems to be reasonable whereas the corresponding ion $C_{15}H_{19}O_{15}N_2S$ might be regarded as incorrect. Herzsprung et al. [20] showed that *CHO* components are usually more reliable than $CHON_2S$-bearing components in multiple-formula assignments in samples scarce in non-oxygen heteroatom-bearing components. In the MTT6 sample, however, the $CHON_2S$ formula was found as an unequivocal solution for the corresponding *m*/*z*. This example shows that even inside the unequivocal dataset some problematic (unclear or unreliable) solutions might have been detected. In such cases, the maximum mass error (1 ppm, constraint for the present dataset) is too small to assign the correct formula. This might happen in the unequivocal dataset as well as the equivocal dataset. Several *m*/*z* were identified where corresponding formula assignments seemed to be incorrect, and the reasonable solution was not outputted by the formula software. Such cases were clearly (from a quantitative perspective) of minor importance. This issue may be addressed in future investigations (see "Conclusions" section).

In order to validate the observations of Herzsprung et al. [20], the pools N_{div} (unequivocal), N_{rep} (equivocal) and N_{rep} (equivocal) are further itemized as shown in Fig. S3 (ESM_1). Forty-six percent of the unequivocal dataset (N_{rep}) consisted of CHO components. This confirms the observation by Herzsprung et al. [20] derived from SRFA data that CHO is the dominant molecular class in NOM (62 % unequivocal CHO components were found in SRFA NOM). However, the NOM from the pit lakes is different from the SRFA NOM. The SRFA unequivocal data pool contains only 2.9 % CHOS components, whereas that of the pit lakes contains 19 % CHOS components.

The percentage distribution of molecular classes (pit lake NOM) is quite dissimilar in the unequivocal and equivocal data pools (N_{rep}). The percentage of $CHON_2$, $CHON_2S$ and $CHON_4S$ is elevated in the equivocal compared to the unequivocal data pool whereas the percentage of CHO and CHOS is significantly reduced. This confirms our previous observations that the molecular classes $CHON_2$, $CHON_2S$ and $CHON_4S$ were widely found as unreliable formula assignments and the corresponding *m*/*z* could be considered as CHO components [20].

The different percentage distribution of molecular classes in N_{rep} and N_{div} from the unequivocal data pool is a new observation. To summarize, the 21 samples contain comparably more different components containing N and/or S than CHO components. That means that the diversity of CHOS, CHON or CHONS seems to be higher compared to that of CHO in the dataset of investigated samples. We calculated how many molecular class combinations can be expected for multiple-formula assignments [20]. A balance of the frequency of doublets, triplets, quartets and quintets specific to molecular class combinations was presented and discussed for the SRFA dataset.

A similar balance is presented in Table S2 (multiple-formula assignments derived from molecular ions with equal mass; see ESM_1) and Table S3 (multiple-formula assignments derived from molecular ions with odd mass; see ESM_1) for the pit lake dataset. Some molecular class combinations and the associated frequencies of multiple-formula assignments are listed in Table 3. Table 3 shows which

Table 3 Comparison of the frequencies of the molecular class combinations in a SRFA sample [11, 20] and a merged dataset of 21 pit lake samples

Molecular classes combinations	SRFA Number of combinations, N_{SRFA}	Pit lake samples Number of combinations, $N_{pit\ lake}$	$N_{SRFA}/N_{pit\ lake}$
CHO/$CHON_2$	383	765	0.50
CHON/$CHON_3$	91	127	0.72
CHO/$CHON_2S$	171	519	0.33
CHON/$CHON_3S$	24	127	0.19
CHO/CHOS	28	172	0.16
CHON/CHONS	0	58	0
CHOS/$CHON_2S$	2	119	0.017
CHO/$CHON_2$/$CHON_2S$	1199	705	1.70
CHON/$CHON_3$/$CHON_3S$	82	73	1.12
$CHON_2$/$CHON_2S$/$CHON_4S$	46	391	0.12
CHO/CHOS/$CHON_2S$	2	270	0.007
CHON/CHONS/$CHON_3S$	0	106	0
CHO/$CHON_2$/$CHON_2S$/$CHON_4S$	296	120	2.47

molecular class combinations are frequent in both datasets and where the frequency is different from a qualitative perspective. The class combinations CHO/$CHON_2$, CHON/$CHON_3$, CHO/$CHON_2S$, CHO/$CHON_2$/$CHON_2S$, CHON/$CHON_3$/$CHON_3S$ and CHO/$CHON_2$/$CHON_2S$/$CHON_4S$ are frequent in both datasets. Typically, the class combinations CHO/CHOS, CHOS/$CHON_2S$ and CHO/CHOS/$CHON_2S$ are only highly frequent in the pit lake dataset. This is due to the high content of CHOS components in the pit lake samples compared to that of the SRFA sample.

Validation of decisions

Herzsprung et al. [20] derived rules for group decisions in which the molecular class from multiple-formula assignment groups is reliable. A reliable component (group) means that this would be the most probable component in comparison to other components (groups), which are considered less reliable or unreliable. In the graphical abstract, we used the synonym "reasonable" for the most reliable component group. Reasonable means that humic-like isomers can be easily constructed from the corresponding elemental formula. Such isomers should be water soluble (from a chemical perspective) but not water mixable (like polyols or similar extreme oxygen-rich components). "Unthinkable" is a term which can be used as an attribute for components for which no reasonable isomers can be constructed at any stretch of the imagination. This is the case for components with DBE − O values > 20. All such hypothetical isomers would have extreme aromaticity and hydrophobicity and would be not humic - like. Component groups with such high DBE − O values can be easily determined in order to consider them unreliable. In some decision diagrams, the attribute "unclear" is used. A decision or a corresponding component group is unclear if all objective and subjective (depending on the user) criteria do not seem sufficient in order to find out the most reliable component group. This is limited by the fact that all these decisions are unquantifiable and are of somewhat subjective nature depending on the user, as already discussed in [20]. Nevertheless, the user has received appropriate tools in order to facilitate assumptions and ultimately make decisions [20].

In the following, molecular class combinations are presented where the decision was identical for the SRFA and the pit lake datasets. The triplet group CHO/$CHON_2$/$CHON_2S$ is shown in Fig. 1. In the SRFA sample, it was possible to assign all m/z to CHO components (Fig. 1a). In the pit lake samples, the majority of m/z was assigned to CHO components (Fig. 1b). The triplet group CHON/$CHON_3$/$CHON_3S$ of the pit lake samples (Fig. 1c) shows a similar distribution of DBE − O values compared to the triplet group CHO/$CHON_2$/$CHON_2S$ with considerably smaller frequencies.

For comparable decisions, further examples are shown (Fig. S4–S7; see ESM_1). From the doublet group CHO/$CHON_2$, similar decisions can be generated for both datasets, e.g. SRFA and pit lake samples (ESM_1, Fig. S4 a) and b)). However, in the pit lake dataset, a subgroup of the molecular class group CHO/$CHON_2$ can be identified (ESM_1, Fig. S4 b)), for which the decision of formula assignment is unclear. The CHON/$CHON_3$ group (ESM_1, Fig. S4 c)) also shows two subgroup distributions. A major fraction of CHON from the pit lake sample dataset was considered reliable as shown by Herzsprung et al. [20] (ESM_2) for the SRFA dataset. Figure S5 a) and b) (see ESM_1) thus shows the doublet group CHO/$CHON_2S$. In both datasets of SRFA and pit lake samples, the CHO components were considered reliable. Some more small subgroups in the pit lake dataset compared to the SRFA dataset were identified with unclear formula

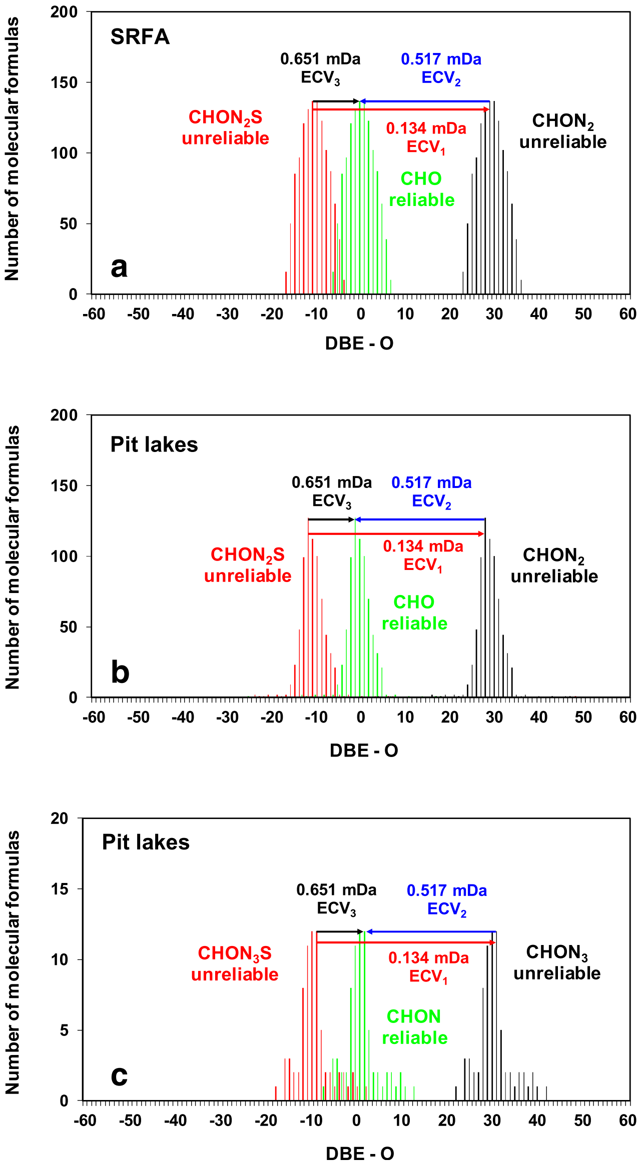

Fig. 1 Frequency versus DBE − *O* diagrams to derive decisions from the triplet group *CHO*/*CHO*$\mathbf{N_2}$/*CHO*$\mathbf{N_2S}$ in the SRFA dataset (**a**) and in the pit lake sample dataset (**b**) and from the triplet group *CHO*N/*CHO*N_3/*CHO*$\mathbf{N_3S}$ in the pit lake sample dataset (**c**)

assignment decisions. In Fig. S5 c) in ESM_1, the doublet group CHON/$CHON_3S$ is shown (pit lake dataset). A subgroup of CHON components could be determined as reliable (as shown in the SRFA dataset in [20]; ESM_2), whereas a bigger subgroup was unclear. The quartet group CHO/$CHON_2$/$CHON_2S$/$CHON_4S$ in Fig. S6 a) and b) (see ESM_1) shows similar decisions for both compared datasets in principle. CHO were the most reliable components here. Again, some minor subgroups were identified in the pit lake dataset with unclear decisions (ESM_1, Fig. S6 b)). A quartet group with an odd number of nitrogen (CHON/$CHON_3$/$CHON_3S$/$CHON_5S$) was observed (ESM_1, Fig. S6 c)), corresponding to the quartet group CHO/$CHON_2$/$CHON_2S$/$CHON_4S$ (equal number of nitrogen atoms). The quartet group with the odd number of nitrogen is of minor relevance, however, due to the small number of reliable assigned components.

The quintet group CHO/$CHON_2$/$CHON_4$/$CHON_2S$/$CHON_4S$ (ESM_1, Fig. S7) shows CHO to be the reliable formula assignments in the SRFA sample (ESM_1, Fig. S7 a)), whereas in the pit lake samples, several subgroups were identified (ESM_1, Fig. S7 b)). Some CHO were assigned to the *m*/*z* as "reliable"; other *m*/*z* remained unclear.

The triplet group CHO/CHOS/$CHON_2S$ is shown in Fig. 2a, b. Most of the CHOS components here can be considered reliable in the pit lake samples (Fig. 2b), whereas the

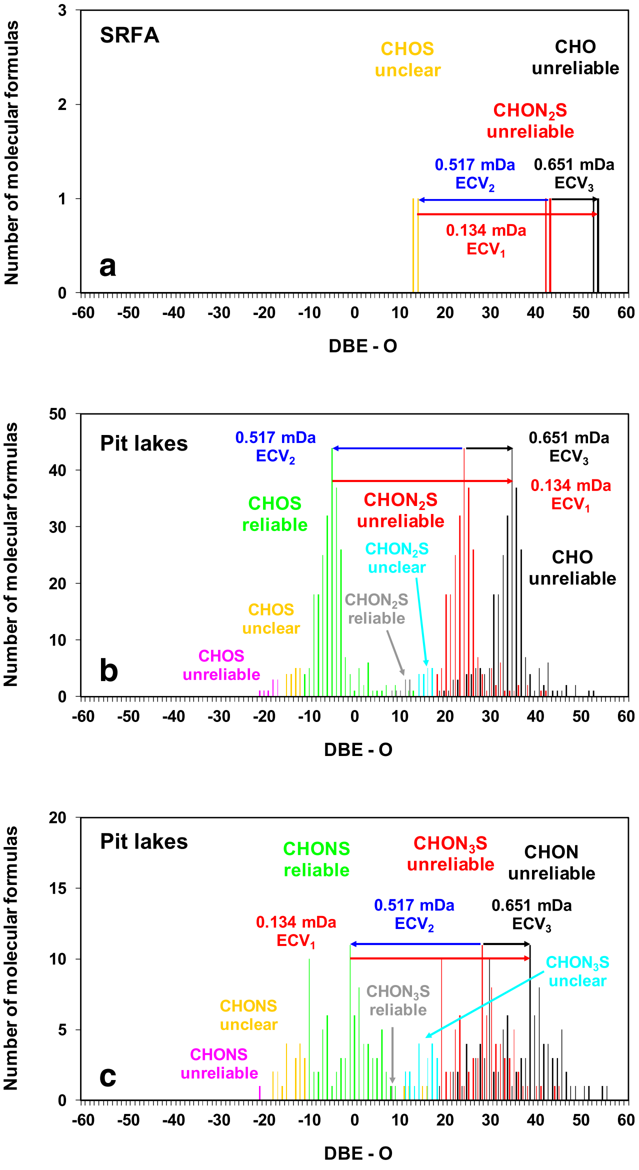

Fig. 2 Frequency versus DBE − *O* diagrams to derive decisions from the triplet group *CHO*/*CHO*$\mathbf{S}$/*CHO*$\mathbf{N_2S}$ in the SRFA dataset (**a**) and in the pit lake sample dataset (**b**) and from the triplet group *CHO*/*CHO*$\mathbf{NS}$/*CHO*$\mathbf{N_3S}$ in the pit lake sample dataset (**c**)

corresponding CHO solutions are far outside the DBE − *O* range −10 to 10. This is the first documented example in which the formula solution which contains more non-oxygen heteroatoms can be considered more reliable; CHOS is more reliable than CHO in this case. The $CHON_2S$ group (some minor frequent exceptions) can also be considered unreliable.

The SRFA dataset (Fig. 2a) showed only a small frequency of that triplet group, and the reliability of the CHOS components could be called into question due to their DBE − *O* values > 10. In Fig. 2c, some CHONS components were shown to be reliable in contrast to the CHON components, which contain less non-oxygen heteroatoms. The doublet subgroup CHO/CHOS in Fig. 3b (pit lake dataset) confirms that the CHOS components can be the reliable components and not CHO. In the same manner, CHONS components are reliable and not CHON as shown in Fig. 3c. Characteristically, this doublet subgroup is altogether rare in the SRFA dataset, since SRFA contains only few CHOS components.

The doublet group CHO/CHOS shows two different elemental composition vectors (ECVs, explained in [20]; ECM_1). If Δ*m* was 0.134 mDa, the CHOS components were reliable. If Δ*m* was 1.001 mDa, then the CHO components were reliable and the corresponding CHOS unreliable, as shown in Fig. 4b. This CHO/CHOS doublet subgroup (with Δ*m* = 1.001 mDa) is also frequent in the SRFA dataset, and the CHO molecular class is reliable, as in the pit lake dataset.

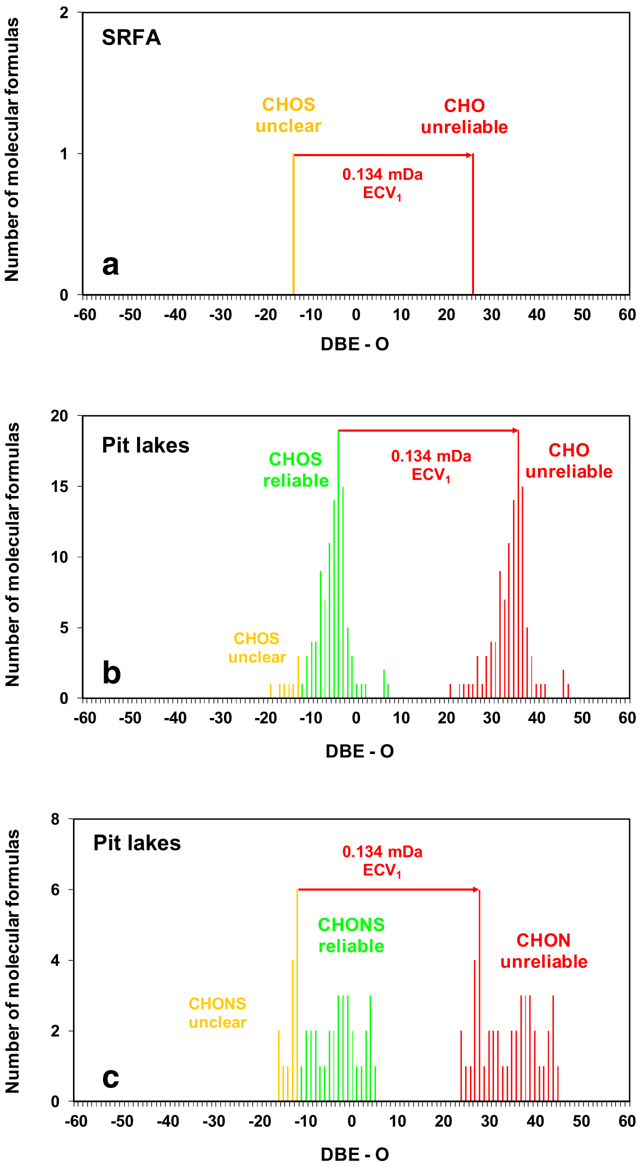

Fig. 3 Frequency versus DBE − *O* diagrams to derive decisions from the doublet subgroup *CHO*/*CHO***S** in the SRFA dataset (**a**) and in the pit lake sample dataset (**b**) and from the doublet subgroup *CHO*/*CHO***NS** in the pit lake sample dataset (**c**); Δ*m* = 0.134 mDa

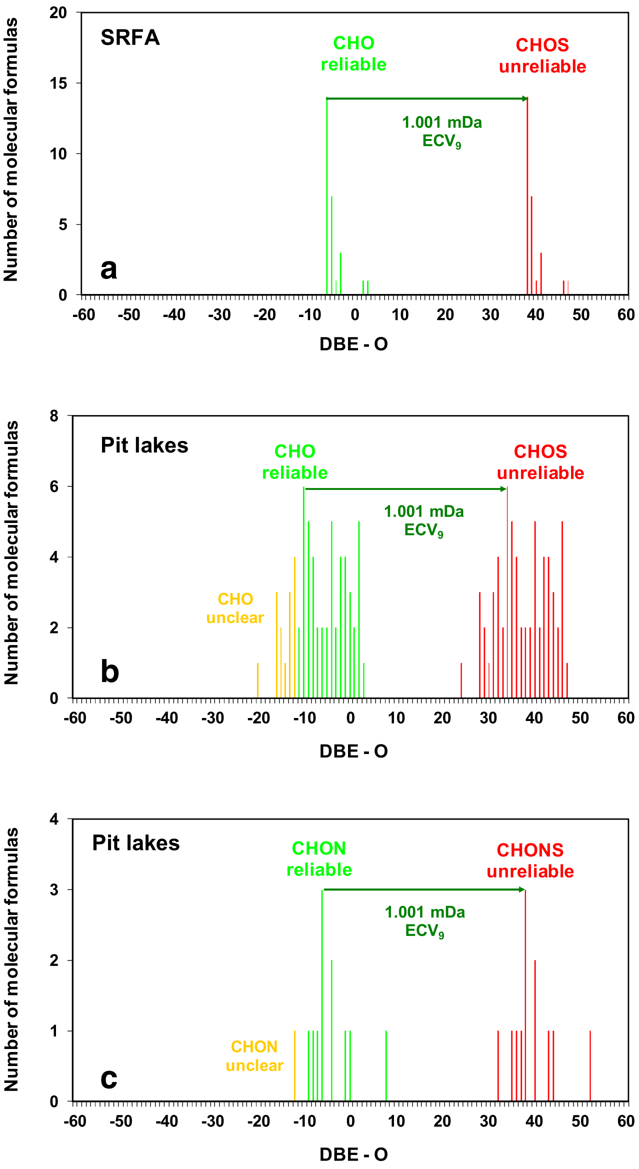

Fig. 4 Frequency versus DBE − *O* diagrams to derive decisions from the doublet subgroup *CHO*/*CHO***S** in the SRFA dataset (**a**) and in the pit lake sample dataset (**b**) and from the doublet subgroup *CHO*/*CHO***NS** in the pit lake sample dataset (**c**); Δ*m* = 1.001 mDa

Table 4 Benefit of the multiple-formula assignment decision and assessment procedure specific to the presented dataset: transfer of molecular classes which were decided to be reliable from the equivocal data pool to the unequivocal data pool

Component	Unequivocal	Equivocal, considered reliable	% contribution to the component-specific unequivocal data pool
CHO	19,301	1811	9.4
CHON	2627	264	10.0
CHOS	7737	523	6.8
CHONS	1538	215	14.0
$CHON_2$	1668	109	6.5
$CHON_2S$	1091	57	5.2
$CHON_3$	1443	31	2.1
$CHON_3S$	715	40	5.6
$CHON_4$	2414	437	18.1
$CHON_4S$	1549	500	32.3
$CHON_5$	1099	97	8.8
$CHON_5S$	508	39	7.7

Quantitative degree of dataset correction by addition of components determined to be reliable

The change in the number of present components can be balanced after adding all the components (from the equivocal dataset) determined to be reliable to the unequivocal dataset (Table 4). The unequivocal dataset primarily contained 41,690 components, and 4123 components were added by the reported decision procedure [20]. The dataset was enlarged by 9.4 % CHO (assignment of 1811 components in addition to 19,301 unequivocal components), by 6.8 % CHOS (assignment of 523 components in addition to 7737 unequivocal components), by 10 % CHON (assignment of 264 components in addition to 2627 unequivocal components) and by 14 % CHONS (assignment of 215 new components in addition to 1538 unequivocal components). Some more components containing more than one N atom were also assigned (Table 4). The comparably (unusual) high number of $CHON_4$ and $CHON_4S$ assignments in both the unequivocal data pool and the equivocal data pool may be questionable (for example, see decision diagrams 5 (D6 odd, D7a odd), 8 (D9 odd), 16 (T6 odd), 19 (T10 odd), 20 (T11 odd), 27 (Q2a odd), 28 (Q3a odd) in ESM_2). Additional investigation is required in order to determine the plausibility of reliable $CHON_4$ and $CHON_4S$ components at such comparably high frequencies in NOM. Herzsprung et al. [20] reported the possible limitations of the presented decision procedure. Additional observations from FT-ICR-MS datasets might elucidate this issue. Nevertheless, the assignment of many CHO and CHON could be demonstrated for the pit lake dataset. This is in good agreement with our previously described SRFA dataset [20].

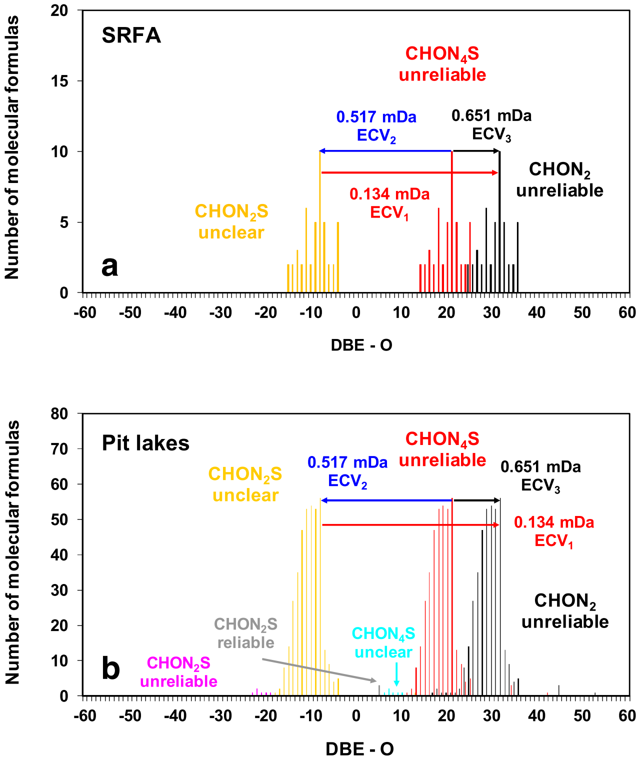

Fig. 5 Frequency versus DBE − *O* diagrams to derive decisions from the triplet group *CHO*N_2/*CHO*N_2S/*CHO*N_4S in the SRFA dataset (**a**) and in the pit lake sample dataset (**b**)

Conclusions

The described formula assignment procedure was successfully validated by a dataset of 21 pit lake/river floodplain samples. Similar multiple-formula assignments (doublets, triplets, quartets and quintets) were observed, and similar decisions for the group assignment of molecular classes were derived from frequency versus DBE−O diagrams. Pit lake samples were rich in CHOS components compared to the SRFA sample. The validation is also accordingly successful. In the pit lake dataset, many CHOS (and CHONS) were assigned (besides many CHO) using the described procedure [20], which produced mainly CHO and CHON assignments for the SRFA dataset. By far, not all multiple-formula assignments were clarified and many of them remained unclear (as shown in ESM_2). One should anticipate that in any multiple-formula assignments, the correct solution (molecular class) may be not recorded. As an example, in the triplet group $CHON_2$/$CHON_2S$/$CHON_4S$, too few assignments lie in the DBE $-O$ range from −10 to 10 (Fig. 5). Other solutions such as CHO should be taken into account. Due to statistical reasons and analytical uncertainty, the allowed mass error (<1 ppm) might be not sufficient for correct formula assignment. This issue should be addressed in future investigations.

Acknowledgments We thank two anonymous reviewers for very constructive comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

1. Stenson AC, Landing WM, Marshall AG, Cooper WT. Ionization and fragmentation of humic substances in electrospray ionization Fourier transform-ion cyclotron resonance mass spectrometry. Anal Chem. 2002;74:4397–409.
2. Stenson AC, Marshall AG, Cooper WT. Exact masses and chemical formulas of individual Suwannee River fulvic acids from ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectra. Anal Chem. 2003;75:1275–84.
3. Kujawinski EB, Del Vecchio R, Blough NV, Klein GC, Marshall AG. Probing molecular-level transformations of dissolved organic matter: insights on photochemical degradation and protozoan modification of DOM from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Mar Chem. 2004;92:23–37.
4. Koch B, Witt M, Engbrodt R, Dittmar T, Kattner G. Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Geochim Cosmochim Acta. 2005;69:3299–308.
5. Reemtsma T, These A, Venkatachari P, Xia XY, Hopke PK, Springer A, et al. Identification of fulvic acids and sulfated and nitrated analogues in atmospheric aerosol by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Anal Chem. 2006;78:8299–304.
6. Grannas AM, Hockaday WC, Hatcher PG, Thompson LG, Mosley-Thompson E. New revelations on the nature of organic matter in ice cores. J Geophys Res. 2006;111:(D4).
7. Hertkorn N, Benner R, Frommberger M, Schmitt-Kopplin P, Witt M, Kaiser K, et al. Characterization of a major refractory component of marine dissolved organic matter. Geochim Cosmochim Acta. 2006;70:2990–3010.
8. Hertkorn N, Frommberger M, Witt M, Koch BP, Schmitt-Kopplin P, Perdue EM. Natural organic matter and the event horizon of mass spectrometry. Anal Chem. 2008;80:8908–19.
9. Yassine MM, Harir M, Dabek-Zlotorzynska E, Schmitt-Kopplin P. Structural characterization of organic aerosol using Fourier transform ion cyclotron resonance mass spectrometry: aromaticity equivalent approach. Rapid Commun Mass Spectrom. 2014;28:2445–54.
10. Marshall AG, Blakney GT, Chen T, Kaiser NK, McKenna AM, Rodgers RP, et al. Mass resolution and mass accuracy: how much is enough? Mass Spectrom. 2013;2(S0009):1.
11. Koch BP, Dittmar T, Witt M, Kattner G. Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. Anal Chem. 2007;79:1758–63.
12. Ohno T, Ohno PE. Influence of heteroatom pre-selection on the molecular formula assignment of soil organic matter components determined by ultrahigh resolution mass spectrometry. Anal Bioanal Chem. 2013;405:3299–306.
13. Kujawinski EB, Longnecker K, Blough NV, Del Vecchio R, Finlay L, Kitner JB, et al. Identification of possible source markers in marine dissolved organic matter using ultrahigh resolution mass spectrometry. Geochim Cosmochim Acta. 2009;73:4384–99.
14. Ohno T, He Z, Sleighter RL, Honeycutt CW, Hatcher PG. Ultrahigh resolution mass spectrometry and indicator species analysis to identify marker components of soil- and plant biomass-derived organic matter fractions. Environ Sci Technol. 2010;44:8594–600.
15. Kunenkov EV, Kononikhin AS, Perminova IV, Hertkorn N, Gaspar A, Schmitt-Kopplin P, et al. 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. Anal Chem. 2009;81:10106–15.
16. Grinhut T, Lansky D, Gaspar A, Hertkorn N, Schmitt-Kopplin P, Hadar Y, et al. Novel software for data analysis of Fourier transform ion cyclotron resonance mass spectra applied to natural organic matter. Rapid Commun Mass Spectrom. 2010;24:2831–7.
17. Roach PJ, Laskin J, Laskin A. Higher-order mass defect analysis for mass spectra of complex organic mixtures. Anal Chem. 2011;83:4924–9.
18. Tziotis D, Hertkorn N, Schmitt-Kopplin P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. Eur J Mass Spectrom. 2011;17:415–21.
19. Kilgour DPA, Mackay CL, Langridge-Smith PRR, O'Connor PB. Appropriate degree of trust: deriving confidence metrics for automatic peak assignment in high-resolution mass spectrometry. Anal Chem. 2012;84:7431–5.
20. Herzsprung P, Hertkorn N, von Tümpling W, Harir M, Friese K, Schmitt-Kopplin P. Understanding molecular formula assignment of Fourier transform ion cyclotron resonance mass spectrometry data of natural organic matter from a chemical point of view. Anal Bioanal Chem. 2014;406:7977–87.
21. Herzsprung P, Hertkorn N, Friese K, Schmitt-Kopplin P. Photochemical degradation of natural organic sulfur compounds (*CHO*S) from iron-rich mine pit lake pore waters-an initial understanding from evaluation of single-elemental formulae using ultra-high-resolution mass spectrometry. Rapid Commun Mass Spectrom. 2010;24:2909–024.
22. Kotelo LG. Characterising the acid mine drainage potential of fine coal wastes. PHD thesis, University of Cape Town. 2013.
23. Geller W, Klapper H, Schultze M. Natural and anthropogenic sulfuric acidification of lakes. In: Geller W, Klapper H, Salomons W, editors. Acid mining lakes—acid mine drainage, limnology and reclamation. Berlin, Germany: Springer; 1998. p. 3–14. Environ. Sci. Series.
24. Dittmar T, Koch B, Hertkorn N, Kattner G. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. Limnol Oceanogr Methods. 2008;6:230–5.