Article

# Mass-Remainder Analysis (MARA): An Improved Method for Elemental Composition Assignment in Petroleomics.

Tibor Nagy, Ákos Kuki, Miklós Nagy, Miklos Zsuga, and Sándor Kéki

*Anal. Chem.*, **Just Accepted Manuscript** • Publication Date (Web): 26 Mar 2019

# *Mass-Remainder Analysis* (MARA): An Improved Method for Elemental Composition Assignment in Petroleomics.

Tibor Nagy, Ákos Kuki, Miklós Nagy, Miklós Zsuga, Sándor Kéki*

Department of Applied Chemistry, Faculty of Science and Technology, University of Debrecen, H-4032 Debrecen, Egyetem tér 1., Hungary

**ABSTRACT:** Data processing and visualization methods have an important role in the mass spectrometric study of crude oils and other natural samples. The recently invented data mining procedure: the *Mass-remainder* analysis (MARA), was further developed for the use in petroleomics. MARA is based on the calculation of the remainder after dividing by the exact mass of a base unit, in petroleomics by the mass of the $CH_2$ group. The two key steps in the MARA algorithm are the separation of the monoisotopic peaks from the other isotopic peaks and the subsequent intensity correction. The effectiveness of our MARA method was demonstrated on the analysis of lubricating mineral oil and crude oil samples by ultra-high resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) experiments. MARA is able to handle a huge portion of the overlapped peaks even in a moderate resolution mass spectrum. Using MARA, effective chemical composition assignment and visual representation were achieved for complex mass spectra recorded by a TOF analyzer with a limited resolution of 40 000 at *m/z* 400. In the absence of an ultra-high resolution mass analyzer, MARA can provide a closer look on the mass spectral peaks, like a digital zoom in a simple camera.

## Introduction

Crude oil is a very complex mixture of hydrocarbons and polar organic compounds. As the world's remaining oil reserves are becoming heavier and sourer, the knowledge of the molecular composition of crude oil is ever more important. The mass spectrometric analysis of this extremely complex mixture may require ultrahigh mass resolution and mass measurement accuracy to separate the tens of thousands of individual compounds present in crude oil. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) has been successfully applied for crude oil characterization.[1-10]. As alternatives, the performance of "zig-zag" multi-reflecting time-of-flight (TOF) analyzer with ultrahigh resolving power,[11] and Orbitrap Fourier transform mass spectrometer were also used for petroleomic studies.[12] Independently of analysers, fractionation of the sample is beneficial, especially for non-polar compounds.[13-15] The SARA (saturates, aromatics, resins, asphaltenes) fractionation is typical for crude oil, and the collection of subfractions can increase the number of assignations.[16-19] Beside the identification of the chemical composition, the structural analysis allows a better understanding of crude oil behavior.[20,21] Not only the high resolution and mass accuracy, but also the data processing and visualization methods, such as Kendrick mass defect[22] and van Krevelen analyses[23] have an important role in the study of crude oils. Kendrick suggested a new mass scale, which converts the mass of $CH_2$ from 14.01565 to an integer value of 14. The Kendrick mass defect (KMD) plot provides a tool for the sorting of compounds into homologous series. KMD plot simplifies the identification of the mass peaks[1,2,24] or enables the rapid characterization of the sample by filtering characteristic compound classes.[25] The compounds containing the same heteroatoms O, N and S (same class) and the number of rings plus double bonds (same type) but different numbers of $CH_2$ groups will have the same KMD values. Unfortunately, the Kendrick mass defect calculations can generate coincidences, namely identical KMD values for the compounds of different class and/or type (e.g. both the $[C_{30}H_{54}OS + H]^+$ and $[C_{30}H_{57}O_3N_3 + H]^+$ ions have almost the same KMD values 0.120619 and 0.120468, respectively). In order to differentiate between the above compositions, a better than 0.2 ppm mass accuracy is required at *m/z* 400. During the mass peak assignation, this issue can be eliminated by pre-sorting the compounds according to their nominal mass (the mass peaks are divided into 14 nominal mass series), as proposed by Hsu *et al.*[24], however these coincidences are still interfering during the visualization and in the filtering based on the KMD plots. Recently we have proposed a simple algorithm, the *Mass-remainder analysis* (*MARA*) for the processing of complex copolymer mass spectra.[26] MARA can not only handle the issues mentioned above, but does not require a multiple sorting algorithm. In this work, we demonstrate the ability of the *Mass-remainder analysis* for the elemental composition assignment and subsequent intensity correction by the analysis of lubricating mineral oil and crude oil samples recorded FT-ICR-MS experiments. Moreover, owing to its deisotoping procedure, MARA is expected to be able to analyze the complex crude oil mass spectra recorded by a TOF analyzer with moderate resolving power. With MARA, we report a new way for the data processing in petroleomics, and furthermore, to the best of our knowledge, there have been no studies on the characterization of crude oil by a TOF analyzer with limited resolution (40 000 at *m/z* 400).

## Experimental

**Chemicals.** The crude oil sample is originated from Russia, imported by the MOL group (Budapest, Hungary). LVO 100 lubricant oil (free of additives) was produced by Leybold (Vienna, Austria). Unused and used LVO 100 lubricants were measured. The used LVO 100 was applied in a rotary vacuum pump for 6 months. Methanol (HPLC grade) and acetic acid (100%, analytical reagent) were purchased from VWR International (Leuven, Belgium). Toluene was received from Sigma-Aldrich (Taufkirchen, Germany) and distilled before use. The samples

were dissolved in toluene and completed with methanol. Concentration of the samples were 1.5 mg/mL, the solvent was toluene methanol mixture (V/V 4:1). 1 mL of the samples were spiked by 3 µL acetic acid and prepared freshly before each measurement. For the comparison of FT-ICR and TOF measurements the concentration was 0.1 mg/mL.

**Electrospray Quadrupole Time-of-Flight Mass Spectrometry (ESI-QTOF MS).** A Maxis II type Qq-TOF MS instrument (Bruker Daltonik, Bremen, Germany) was used equipped with an electrospray ion source where the spray voltage was 4.5 kV. The resolution of the instrument is 40 000 at *m/z* 400 (FWHM), the mass accuracy is <600 ppb (internal calibration). $N_2$ was utilized as drying (200°C, 4.0 L/min) and nebulizer gas (0.5 bar). The mass spectra were recorded by means of a digitizer at a sampling rate of 2 GHz. The spectra were calibrated in two steps, externally by ESI tune mix (first step), from Bruker, and internally using $[C_nH_{2n-13}N+H]^+$ (DBE = 8) well known series appeared with high intensity (second step). The spectra were evaluated with the Compass DataAnalysis 4.4 software from Bruker. The sample solutions were introduced directly into the ESI source with a syringe pump (Cole-Parmer Ins. Co., Vernon Hills, IL, USA) at a flow rate of 6 µL/min.

**Electrospray FT-ICR Mass Spectrometry (ESI-FT-ICR MS).** The measurements were carried out by SolarisX XR 15 T FT-ICR mass spectrometer (Bruker Daltonik, Bremen, Germany) equipped with an electrospray ion source (4.5 kV). The resolution of the instrument is 230 000 at *m/z* 400 (FWHM), the mass accuracy is <250 ppb (internal calibration). $N_2$ was utilized as drying (180°C, 4.0 L/min) and nebulizer gas (0.1 bar). The spectra were calibrated internally using $[C_nH_{2n-13}N+H]^+$ (DBE = 8) well known series and evaluated with the Compass DataAnalysis 4.4 software from Bruker. The sample solutions were introduced directly into the ESI source at a flow rate of 4 µL/min.

## Results and Discussion

**Mass-remainder analysis versus Kendrick mass defect analyses.** Kendrick has introduced a mass scale based on $CH_2 = 14$ for the handling of the large amount of mass data in the mass spectra of natural organic compounds,[22] and other base groups were also used, e.g. for the analysis of synthetic polymers.[27] The KMD analysis applies a conversion from the $^{12}C = 12$ scale to the Kendrick mass scale (equation 1 and 2 in Supporting Information).

The *Mass-remainder* analysis (MARA) does not require any transformation to a new mass scale, it is based on the calculation of the remainder after dividing by the exact mass of $CH_2 = 14.01565$.[26] *Mass-Remainder* (MR) values of the measured *m/z* mass peaks are calculated according to

$$\mathrm{MR} = \mathrm{m/z} \ \ \mathrm{MOD}\ 14.01565 \qquad (1)$$

where the modulo (MOD) operation finds the remainder after the division.

Figure 1a shows the ESI-FT-ICR mass spectrum of a mineral oil based lubricant (LVO 100), and Figure 1b depicts the MR *versus m/z* plot of this mass spectrum, after running our MARA algorithm, which will be detailed in the next chapter. As seen in Figure 1b, the homologous series, i.e., the compounds of the same class and type, differing only in the number of $CH_2$ groups have the same MR values and are plotted as horizontal lines. The ESI-TOF mass spectrum of the mineral oil based lubricant and the corresponding MR-*m/z* and KMD-NKM plots are shown in Figure S1 in the Supporting Information. One of the main aims of KMD method is the visual analysis of the complex mass spectra by creating the KMD *versus* nominal Kendrick mass (KM) plots,[1] as seen in Figure 1c. However, the complex mass spectra of natural samples probably contain mass peak pairs $m/z_1$ and $m/z_2$ with $m/z_2$ - $m/z_1$ = n × R/round(R) + ε differences, where R = 14.01565, round(R) = 14, *n* is an integer and ε (and ε' in Eq. 3) are small values.

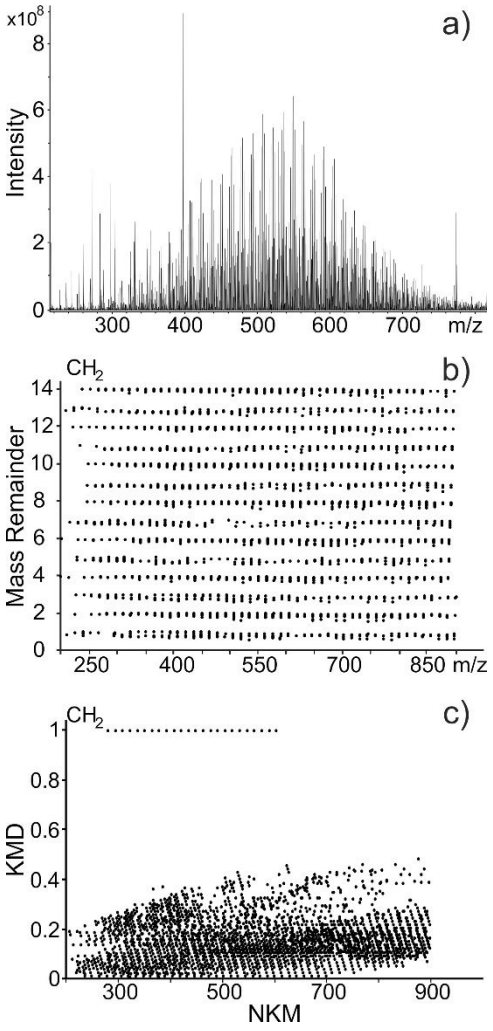


**Figure 1.** (a) ESI-FT-ICR mass spectrum of the mineral oil based lubricant Leybonol LVO 100, (b) *Mass-remainder* (MR) *versus m/z* plot, (c) Kendrick mass defect (KMD) *versus* nominal Kendrick mass (NKM) plot.

Lower mass accuracy means higher ε value. As Eq. 2 and 3 show, these pairs have KM values with an integer difference, and thereby they have identical KMD values.

$$\mathrm{KM}_1 = m/z \times \frac{\mathrm{round(R)}}{\mathrm{R}} \qquad (2)$$

$$\mathrm{KM}_2 = \left(\mathrm{m/z} + \mathrm{n} \times \frac{\mathrm{R}}{\mathrm{round(R)}} + \varepsilon\right) \times \frac{\mathrm{round(R)}}{\mathrm{R}} = \mathrm{KM}_1 + \mathrm{n} + \varepsilon' \qquad (3)$$

These overlaps, generated by the KMD calculation, are demonstrated in Figure S2b For instance, the compounds of type 4 and class OS have almost identical KMD values as the ones of type 4 and class $O_3N_3$ (namely, 0.120619 and 0.120468, respectively, as it was mentioned in the *Introduction*). The KMD difference is 0.000151 which requires higher mass accuracy than 0.2 ppm at *m/z* 400. These coincidences can be disturbing in the visual analysis of the complex mass spectra, and can cause false results when filtering for particular compounds. In contrast, when not the deviation from the nominal mass, but the one from the multiples of $CH_2$ is calculated (namely the *Mass-Remainder*

value), then the generated overlaps can be eliminated, as it is seen in Figure S2a, showing clearly separated lines for the compounds of different classes. The better visualization of the MR *versus m/z* plot is also obvious when it is compared to the KMD plot of the same spectrum (see Figure 1b and 1c, respectively).

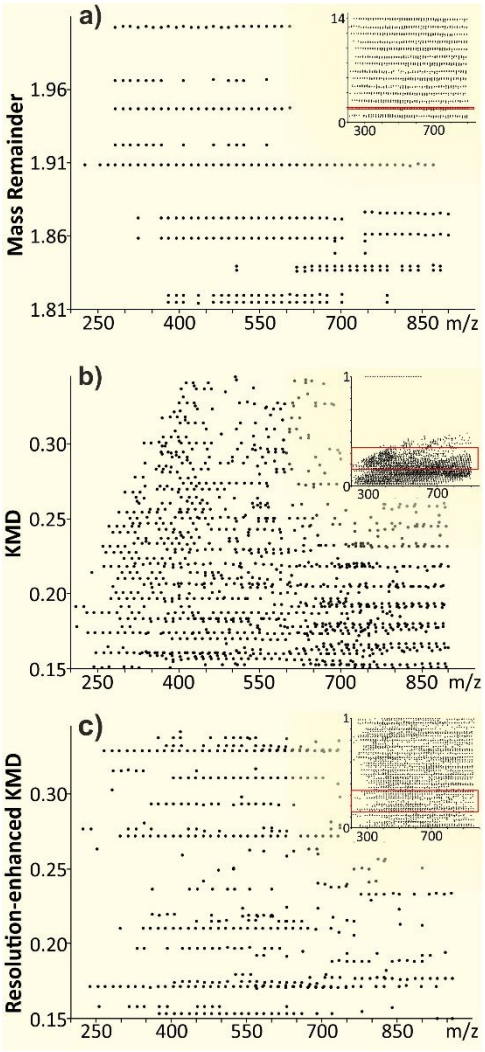


**Figure 2.** The zoomed MARA (a), KMD (b) and Resolution-enhanced KMD (c) plots with a spectral width of 0.2 of the mineral oil based lubricant Leybonol LVO 100. The insets show the full plots and the zoomed region.

Fouquet *et al.* suggested a modified KMD method called Resolution-enhanced KMD analysis, where the conversion to new mass scale is based on a fractional base unit R/X (where R is the base unit of KMD analysis and X is an integer)[28]. Figure 2 shows the zoomed KMD, Resolution-enhanced KMD[28] and MARA plots of the mineral oil based lubricant (LVO 100). As it is seen in Figure 2, the three different processing methods yield different plots. The KMD plot has high density in the range of 0.15-0.35 (Figure 2b). The resolution enhanced KMD plot (Figure 2c) contains fewer series in the same region, thereby decreasing the number of possible overlaps during the analysis. Comparing MARA to the previously mentioned methods allows the spectral width of 14.015650 for hydrocarbon analysis (Figure 2c) or even wider, depending on the applied base unit. The greatly increased scale enables better separation of the series and consequently our data mining process requires lower accuracy for effective identification and filtering.

**Mass peak assignation by Mass-remainder analysis (MARA).** As Eq. 1 shows, the compounds of the same class and type, differing only in the number of $CH_2$ groups, have identical MR values. A reference table was created in a spreadsheet containing the class/type – MR mapping. Our method is exemplified by the analysis of a mineral oil based lubricant (Leybonol LVO 100). The following restrictions were used for the reference table: $c$ unlimited, $h$ unlimited, $0 \le n \le 5$, $0 \le o \le 4$, $0 \le s \le 2$; $n+o+s \le 7$ in $C_cH_hN_nO_oS_s$, and $0 \le DBE \le 25$, where DBE stands for double bond equivalent corresponding to the number of rings plus double bonds (type). For example, the class $n = 1$, $o = 4$, $s = 0$ and type DBE = 1 (e.g. $[C_{25}H_{51}NO_4+H]^+$) is mapped to the MR value of 9.91959. The selection of the maximum numbers of the heteroatoms depends on the complexity of the sample. Sourer oil requires wider range of heteroatoms. In the literature the choice of the number of heteroatoms shows a great variety[1,3-5,29-31] (see Table S1 in the Supporting Information). Higher heteroatom content may result higher number of overlaps (depending on the resolution of the instrument), see Table S2 in the Supporting Information. Of course, if a different substance, e.g. an untreated crude oil is to be analyzed, the conditions and the reference table can be easily modified, for instance $0 \le o \le 5$, $0 \le s \le 5$ can be allowed. From the mass spectra, the *m/z* and intensity values for the ions in the *m/z* range 150 – 900 with at least 0.05% relative abundance and S/N > 3 were imported into the spreadsheet software. A slightly modified version of our original MARA algorithm[26] was developed using the built-in programming language of the spreadsheet software for the processing of the raw *m/z* – intensity lists with the following main steps:

(1) Calculation of the *Mass-remainder* values of the measured *m/z* peaks by Eq. 1 ($MR_{meas}$).

(2) Assignation of the reference table *Mass-remainder* value(s) ($MR_{ref}$) to the *m/z* peaks based on the mass accuracy with a mass tolerance of 1.0 and 1.5 ppm for FT-ICR and TOF measurements, respectively, and the subsequent calculation of the elemental composition(s). For TOF spectra, as a second step of the assignation the tolerance was increased based on the full width at half maximum (FWHM) of each peak, because the overlapping peaks result wider FWHM. Thereby the overlaps of the monoisotopic and isotopic peaks can be assigned and handled (see step 4). The unassigned peaks are removed from the mass list. Figure 3a shows a zoomed region of the ESI-FT-ICR spectrum of the mineral oil based lubricant. For example, the type/class 10 N (meaning DBE = 10 and $n = 1$, $o = 0$, $s = 0$) was assigned to the peak at *m/z* 404.33116, because the values of $MR_{meas}$ and $MR_{ref}$ are 11.89296 and 11.89297, respectively. The removed peaks are indicated in black in Figure 3.

(3) The number of multiple assignations can be reduced based on the numbers of heteroatoms.

(4) Deisotoping and intensity correction of overlapped peaks. One of the main strengths of MARA is the clear separation of the isotopic peaks, for example the $^{13}C_1$ isotope of the $[C_{29}H_{42}N+H]^+$ ion of the 10 N type/class, mentioned above, has the MR=12.89633 mass remainder value in contrast to the MR=11.89297 value of the $^{13}C_0$ peak. Therefore, in *step (2)*, the isotopic mass peaks are not assigned, instead they are removed from the mass list (as indicated in black color). The peak at *m/z* 404.21240 corresponds to the $^{13}C_1$ isotope of the $[C_{27}H_{31}OS+H]^+$ ion of 13 $OS^{13}C_1$ type/class. However, in the mass spectra recorded by a common TOF analyzer, a large number of the isotopic peaks are overlapped with potential monoisotopic peaks, which are therefore assigned in *step (2)*. It means: *i)* an incorrect assignation if the monoisotopic peak actually does not present in the oil sample (indicated in red in Figure 3a), or *ii)* an inaccurate

peak intensity of the monoisotopic peak if it is really present. MARA can handle this issue. The first ($^{13}C_1$) and second ($^{13}C_2$, $^{34}S$) isotope intensities are calculated for all the assigned mass peaks and in the case of coincidence, the measured intensities are decreased by these calculated isotopic intensities. It results *i)* the elimination of incorrect assignments because the measured intensity is decreased down to zero after the correction, which means, that the peak belongs to the isotope, the potential overlapping monoisotopic peak does not appear, or *ii)* the correction of overlapped peaks. An example for this correction can be seen in Figure 3b indicated in green. In this case the mass peak at *m/z* 431.2613 with intensity 698 may be an overlapped peak of the 8 $O_3S$ ion and the 13 $NS^{13}C_1$ ion (see 13 $NS^{13}C_0$ at *m/z* 430.2552). The measured intensity of the 13 $NS^{13}C_0$ ion is 125, which gives the calculated intensity of 41 for the 13 $NS^{13}C_1$ isotope, and this value is subtracted from the measured intensity of the *m/z* 431.2613 peak resulting the corrected intensity of 657 for the 8 $O_3S$ ion.

(5) Summing isotopic peak intensities. The intensity of every monoisotopic peak is increased by the summarized intensity of its $^{13}C_1$, $^{13}C_2$, and $^{34}S_1$ (if present) isotopic peaks.

Here we would like to emphasize, that the last two steps, namely the deisotoping and intensity corrections, enable the effective and exact evaluation of the highly complex mass spectra of crude oils recorded by TOF analyzers with moderate resolution, as will be shown in the next section.

The MARA process capable of resolving most of the overlaps comes from monoisotopic – isotopic and isotopic - isotopic isobar relations. Most of the overlaps are monoisotopic – isotopic, as it was shown by Hsu *et al*[7]. Table S2 shows the typical overlaps for monoisotopic – monoisotopic peaks, their mass differences, the number of hetero atoms acting in replacement and the required resolving power for their separation. Most of the overlaps replace numerous hetero atoms, which is not specific for the crude oil or their fractions and products. However, the replacement of 3 C to 4 H and 1 S result 0.003371 Da as difference. Mass spectrometers with medium resolution (40000) cannot separate these peaks, however the good mass accuracy allows the mono assignation if only one component is in the sample or their peak intensity ratios are large. Further problematic overlaps are the replacements of 5 C to 2 N and 2 O.

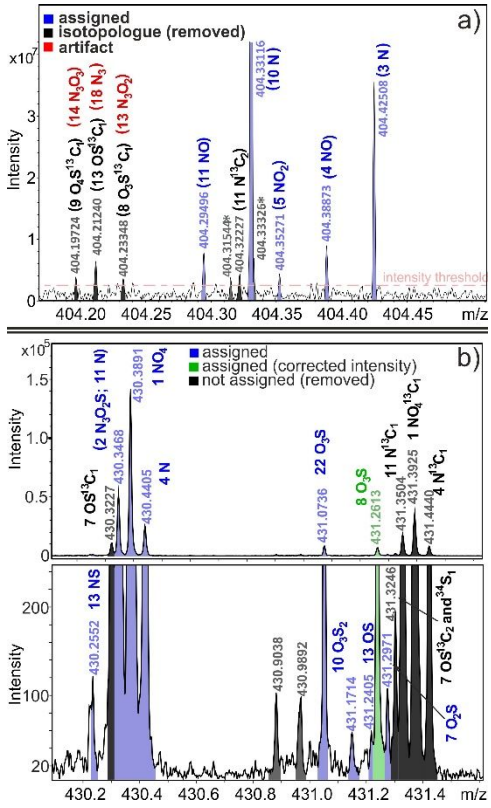


**Figure 3.** The zoomed *m/z* regions of the ESI-FT-ICR (a) and ESI-TOF (b) ((c) zoomed) spectrum of the mineral oil based lubricant Leybonol LVO 100.

**Applications of the Mass-remainder analysis (MARA).** In the previous section we showed how the Mass-remainder analysis assigns the type and class, and subsequently the elemental composition to the individual mass peaks. The mass spectrometric analysis of natural organic matters, such as crude oils, usually produces spectra with thousands or tens of thousands of mass peaks. As a novel challenge, MARA was applied for the mass spectrum of a mineral oil based lubricant (Leybonol LVO 100). The MARA procedure was able to assign a single elemental composition to approximately 2600 mass peaks, and performed multiple assignations to merely ca. 266 peaks in FT-ICR spectrum of mineral based lubricant LVO 100. This means, that using the filtering restrictions for the lubricant ($0 \le n \le 5$, $0 \le o \le 4$, $0 \le s \le 2$; $n+o+s \le 7$, $0 \le DBE \le 25$, as described above in more detail), 50% of the mass peaks were unambiguously identified. This high assignation ratio and the intensity correction of the overlapped and monoisotopic peaks, performed by MARA, enable the use of graphical-statistical methods for the presentation of the large number of individual elemental compositions. We have already shown a graphical analysis, namely the Mass-remainder *versus m/z* plot (see Figure 1b), and several advantages over the KMD plot were discussed. Once the correct elemental compositions were determined, the van Krevelen plots can be created, which are widely used for the visualization of molecular composition of complex mass spectra.[13,33,34] Figure 4a shows a van Krevelen plot of the mineral oil based lubricant LVO 100. Each dot corresponds to an identified mass spectrum peak, and the two coordinates are the atomic ratio of hydrogen to carbon (H/C) and the atomic ratio of nitrogen to carbon (N/C).

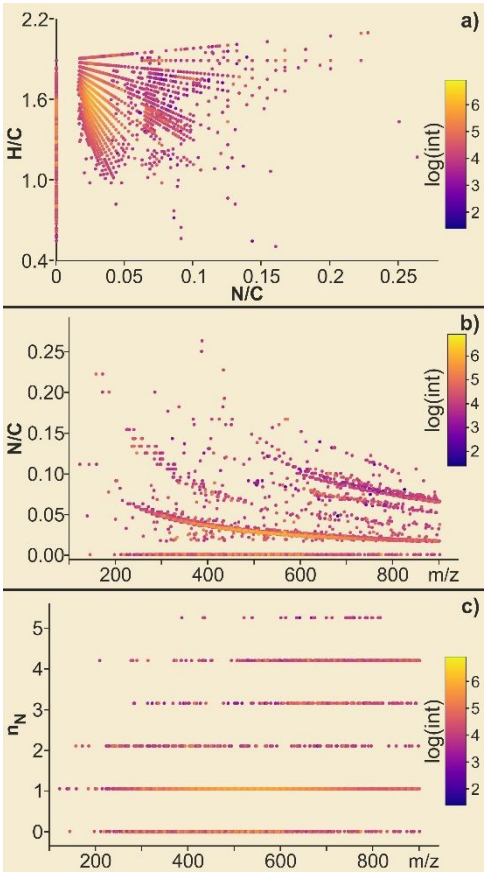


**Figure 4.** (a) The van Krevelen plot, (b) N/C atomic ratio *versus m/z* plot and (c) N class *versus m/z* plot calculated from the ESI-FT-ICR spectrum of the mineral oil based lubricant Leybonol LVO 100. $n_N$ stands for the number of nitrogen.

As seen in Figure 4a, most of the compounds fall into the H/C 1.4–2.0 region (lighter color), which reflects that the oil was treated. The appearance of H/C values around and above 2 was also expected, because the mineral based lubricants, contain saturated compounds. The van Krevelen plots can be used for assigning the identified compounds to different heteroatom classes (e.g. $N_1$, $N_2$ classes) or exploring the compositional differences between complex mixtures. However, the van Krevelen plots are not able to reveal the molecular mass distribution of the identified ions or classes. Very recently, Fedoros *et al.* proposed to complement the van Krevelen plots by H/C *versus* molecular mass plots for the visualization of mass spectra of lignin derivates.[34] A similar visualization, the N/C atomic ratio *versus m/z* plot of the mineral oil LVO 100 was created and is shown in Figure 4b. As seen in Figure 4b, the dots, representing the mass spectral peaks, are clearly aggregated according to the nitrogen classes ($N_0$ – $N_4$), contrary to the slightly messy and overlapping regions of Figure 4a. Even more definite classification is achieved by the direct construction of the N class versus *m/z* plot, as seen in Figure 4c. Figures 4b and 4c justify our initial restriction of $0 \le n \le 5$ for the nitrogen classes ($N_n$), since the $N_5$ class compounds are in minor abundance. The O/C atomic ratio *versus m/z* and O class *versus m/z* plots constructed from the LVO100 lubricant spectrum and the figures for ESI-TOF data are given in the Supporting Information in Figure S3.

It is important to emphasize, that the abundance of the various ions in the mass spectra, *ergo* the class distributions are greatly influenced by the ionization method. In our case, the use of electrospray ionization (ESI) puts the focus on the polar compounds (containing N, O, and S heteroatoms). However, this is not a disadvantage at all, because despite their small proportion (ca. 10 %), the detection of the polar compounds is crucial in the oil industry. In addition, the heteroatomic molecules are effective geochemical markers and are characteristic for degradation and maturation of oils[2,6,911]. Accordingly, we used the heteroatom class distributions (determined by MARA from the FT-ICR and ESI-TOF mass spectra) to follow the aging of the mineral oil based lubricant LVO 100. Figure S4 and Figure S5 show the relative abundances of the major heteroatom classes of the unused and used oils (used in a vacuum pump for half a year, at about 50°C).

As seen in Figure S4-S5, the relative abundance of class $N_1$ is overwhelming, probably due to the nitrogen atoms in the aromatic rings. We kept to the recommended oil lifetime and maximum temperature and indeed, Figure S4-S5 shows only minor changes between the unused and used oils in the heteroatom class distribution, no remarkable oxidation or polymerization can be observed.

The ESI-TOF data show similar distribution compared to a high-resolution FT-ICR mass spectrum despite the poor mass peak separation, owing to its deisotoping procedure, composition to approximately 1850 mass peaks, and performed multiple assignations to merely ca. 320 peaks.

As another application of the ESI-FT-ICR – MARA method, a Russian crude oil was also studied. MARA was able to assign ca. 4300 single elemental compositions from about 10900 detected peaks (the latter include the isotopic peaks). Figure 5a shows the relative abundances of the heteroatom classes calculated from the ESI-FT-ICR spectrum of the Russian crude oil. As seen in Figure 5a, the $N_1$ is the one major heteroatom class in the ESI(+)-MS study of the crude oil. The DBE distribution for this class is shown in Figure 5b, which is a usual visualization in petroleomics, as well.

As seen in Figure 5b, the DBE distribution of the $N_1$ class can be quite precisely approximated by the log-normal distribution (dashed line). The distribution of the heteroatom classes and the $N_1$ class is shown in Figure S6 in the Supporting Information.

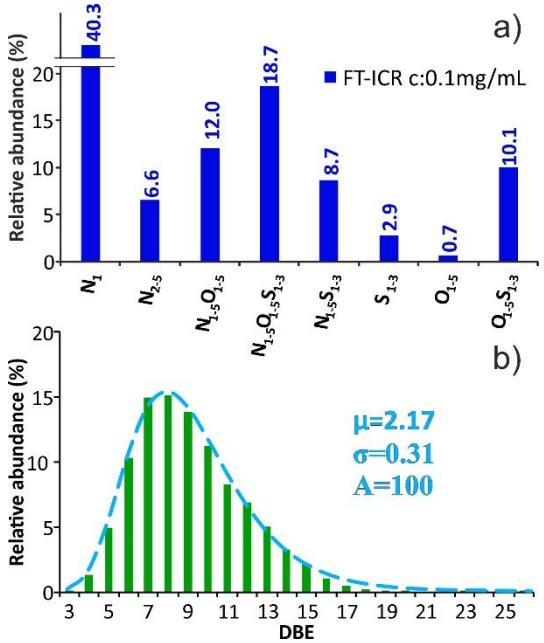


**Figure 5.** (a) The relative abundances of the heteroatom classes, (b) the DBE distribution for class $N_1$ calculated from the FT-ICR spectrum of the Russian crude oil. The dashed line is the fitted curve of the log-normal probability density

function with the parameters: $\mu$=2.2, $\sigma$=0.31 and $A$=100 ($A$ is a scaling factor).

**Comparison of ESI-TOF and ESI-FT-ICR measurements.** Figure 6 shows the zoomed mass spectra of the crude oil measured by ESI-TOF (at two different sample concentrations) and ESI-FT-ICR instrument. The resolutions of the mass spectra ***a***, ***b***, and ***c*** are 40 000, 40 000 and, 253 000 respectively, calculated from the mass peak of the 11 N compound.

As seen in Figure 6, the main peaks are similar, 5 major compounds were assigned in this range of each mass spectra. Of course, the FT-ICR spectrum contains more peaks (8), but the additional peaks have low intensity.

The number of the peaks identified by our MARA method and their percentage to the total peaks are compared. For example, 10800 (40.5 %) (FT-ICR, c:0.1 mg/mL), 8000 (49 %) (TOF, c:1.5 mg/mL) and 6000 (50 %) (TOF, c:0.1 mg/mL) peaks were identified in the mass spectra of the Russian crude oil. (The percentage values show the ratio of the monoassigned peaks to all the detected ones. In the case of the FT-ICR lower values are due to the large number of detected isotopic peaks compared to the TOF measurements.) However, it must be highlighted that only a small portion of the compounds in the unfractionated crude oil samples can be characterized due to selective ionization. Therefore, the analysis of whole petroleum samples provides only limited information.

Figure S7 in the Supporting Information shows the comparison of the compound class distributions calculated for the Russian crude oil.

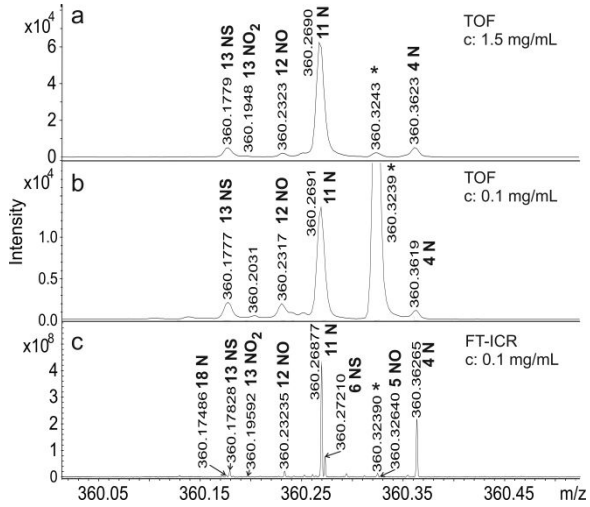


**Figure 6.** The mass spectra of Russian crude oil measured by (a) ESI-TOF (c: 1.5 mg/mL) (b), ESI-TOF (c: 0.1mg/mL) (c) and FT-ICR (c: 0.1 mg/mL). The peaks marked with asterisk are assigned as background peaks.

As seen in Figure S7, $N_1$ is the main class, however, the relative total ion abundances of this class are different: 47.9 %, 71.5 %, 40.3 % for TOF (c:0.1 mg/mL), TOF (c:1.5 mg/mL), and FT-ICR respectively. The deviation due to the concentration difference was described by Ruddy *et al.*[35]. They have shown, that higher concentration is preferable to detect N containing compounds in positive ion mode, but it can result peak loss due to the suppression of the other peaks. Figure 7 shows the DBE distributions of class $N_1$ in the Russian crude oil measured by FT-ICR and TOF.

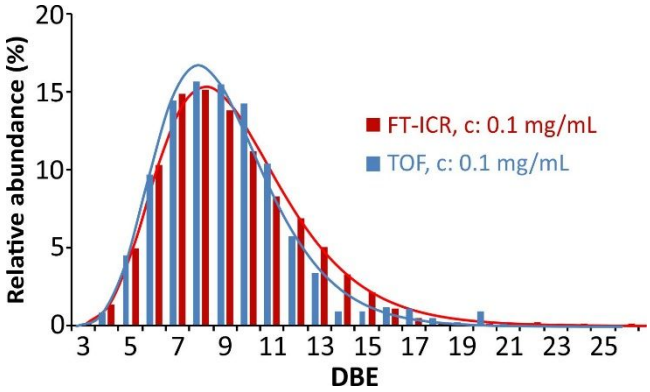


**Figure 7.** The DBE distribution of class $N_1$ in the Russian crude oil measured by different instruments.

As seen in Figure 7, the distributions of this main class are very similar, namely the $N_1$ class appears from the DBE value of 4, its relative abundance tops at the DBE value of 8 with the maximum of ca. 15 % and finally disappears at the values around 20.

The parameters of the log-normal distribution can be used as good measures in the comparison of oil samples (with various types, origin and degradation). Accordingly, the $N_1$ class abundance *versus* DBE plots reported by Pakarinen *et al.* (Fig. 3 in ref. 6) were digitalized, and the probability density function of the log-normal distribution was fitted, as it was also done on our DBE distribution plots (Figure 7). The parameter triplets [μ, σ, $A$] ($A$ is a scaling factor) of the log-normal distribution were obtained to be [2.2, 0.31, 100] [2.2, 0.28, 100], [2.2, 0.28, 54], and [2.2, 0.34, 86] for our crude oil sample measured by ESI-FT-ICR and ESI-TOF instruments, and for the Russian and North Sea crude oil samples studied by Pakarinen *et al.*, respectively. This agreement may justify the effectiveness of our MARA method.

The DBE distributions of the minor classes show more significant differences, which can be explained by the followings: i) Interestingly, many compounds are detected in the ESI-TOF spectra which couldn't be observed in the FT-ICR spectrum, and of course *vice-versa*. Typically, the intensity of these peaks is low, but they can affect the DBE distributions of minor series. ii) The number of detected peaks strongly depends on the sample concentration as well. iii) The different instrumental conditions can result different relative mass peak intensity ratios, for example the intensity ratio of the peaks 11 N to 13 NS is 6:1 and 16:1 calculated from the TOF (c: 0.1 mg/mL) and FT-ICR spectra, respectively (see Figure 6).

The similar class distribution and the DBE distribution of the main class suggests, that in the absence of an expensive FT-ICR instrument, MARA is capable to effectively process and statistically analyze the complex mass spectra recorded by a TOF analyzer with moderate resolving power. It is due to, particularly, the deisotoping and subsequent intensity correction steps. It must be emphasized, that by these steps, MARA improves the validity of the data processing of high resolution (e.g. FT-ICR) mass spectra, as well.

The error distribution of the assignations for the FT-ICR-MARA and ESI-TOF-MARA procedures are shown in Figure S8 in the Supporting Information. As seen in Figure S8 the TOF-MARA has higher error tolerance due to the handling of the overlaps as it was discussed in *step (2)* in the *Mass peak assignation by Mass-remainder analysis (MARA)* section. The limit for the FT-ICR instrument was set to 1 ppm. Contrary, the error for the TOF was defined as 1.5 ppm, based on the FWHM, in the first and second round of the assignation (*step 2*), respectively.

## Conclusions

The *Mass-remainder* analysis (MARA), a recently invented data mining procedure, was adopted for the analysis of mineral based lubricant and crude oil samples recorded by FT-ICR-MS experiments. Moreover, owing to its deisotoping and intensity correction operations, MARA was able to characterize these complex mass spectra with about 10.000 ion signals recorded by a common TOF analyzer. Despite the imperfect mass peak separation of the TOF mass spectra, MARA provided reliable results, for example a single elemental composition was assigned to 85% of the assigned mass peaks of the mineral oil sample. We propose the TOF–MARA study as a cheaper, more accessible alternative of the usual FT-ICR-MS for the analysis of complex natural samples. However, we have to stress that the lubricant and crude oil samples lack the spectral complexity compared to the ultra-complex petroleum fractions. Furthermore, positive-ion atmospheric pressure photoionization (+APPI), which is routinely used in Petroleomics yields even more complex mass spectra. Therefore, further experiments are needed to test the TOF–MARA method for the analysis of samples with extreme complexity. Nevertheless, we believe, that MARA can be especially effective for the filtering of special compound classes in the complex mass spectra.

## ASSOCIATED CONTENT

Supporting Information

Supporting_Information.doc

Figure S1. (a) ESI-FT-ICR mass spectrum of the mineral oil based lubricant Leybonol LVO 100, (b) Mass-remainder (MR) versus m/z plot, (c) Kendrick mass defect (KMD) versus nominal Kendrick mass (NKM) plot.

Figure S2 The 2-7 OS and 1-6 N3O3 (type class) series depicted in the (a) Kendrick mass defect (KMD) versus nominal Kendrick mass (NKM) plot and (b) Mass-remainder (MR) versus m/z plot (theoretical values).

Figure S3. The van Krevelen diagrams and the distributions of N/C, nN, O/C and nO values as a function of m/z for the LVO 100 unused lubricant oil measured by ESI-FT-ICR and ESI-TOF instruments. nO and nN stand for the number of oxygen and nitrogen, respectively.

Figure S4 The relative abundances of the major heteroatom classes calculated from the ESI-TOF spectrum of the unused and used mineral oil based lubricant Leybonol LVO 100.

Figure S5 The relative abundances of the major heteroatom classes calculated from the ESI-FT-ICR spectrum of the unused and used mineral oil based lubricant Leybonol LVO 100.

Table S1 Maximum number of heteroatoms selected for crude oil analysis.

Table S2 Possible overlaps within 10 ppm error at m/z 800 and the resolving power requirements for their separation at different m/z. n(O)=0-5, n(N)=0-5, n(S)=0-5, N(heteroatom)=0-10.

Figure S6. (a) The relative abundances of the heteroatom classes, (b) the DBE distribution for class N1 calculated from the ESI-TOF spectrum of the Russian crude oil. The red line is the fitted curve of the log-normal probability density function with the parameters: $\mu$=2.2, $\sigma$=0.28 and A=100 (A is a scaling factor).

Figure S7. Class distributions for the Russian crude oil measured by ESI-TOF (lower concentration) (blue), ESI-FT-ICR (red), and ESI-TOF (higher concentration) (green)

Figure S8. The error distribution of the assignation as function of m/z applying FT-ICR (a) and TOF (b) instruments.

## Author Information

Corresponding Author

*E-mail: keki.sandor@science.unideb.hu. Fax: +36 52 518662

ORCID

Sándor Kéki: 0000-0002-5274-6117

Notes:

The authors declare no competing financial interest.

## Acknowledgment

The work was supported by the GINOP-2.3.2-15-2016-00041, GINOP-2.3.3-15-2016-00004, and GINOP-2.3.3-15-2016-00021 projects. The projects were co-financed by the European Union and the European Regional Development Fund. Furthermore, this paper was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (Miklós Nagy). Thanks for the sample and the financial support provided by the Mol Group, Hungary. We thank Arpad Somogyi (Associate Director of Mass Spectrometry and Proteomics Facility, Campus Chemical Instrument Center, Ohio State University) for providing us the ultrahigh resolution FT-ICR data. The 15 T Bruker SolariXR FT-ICR instrument was supported by NIH Award Number Grant S10 OD018507.

## References

(1) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. *Anal. Chem.* **2001**, *73*, 4676-4681.

(2) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53-59.

(3) Krajewski, L. C.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2017**, *89*, 11318-11324.

(4) Sim, A.; Cho, Y.; Kim, D.; Witt, M.; Birdwell, J. E.; Kim, B. J.; Kim, S. *Fuel* **2015**, *140*, 717-723.

(5) Hur, M.; Ware, R. L.; Park, J.; McKenna, A. M.; Rodgers, R. P.; Nikolau, B. J.; Wurtele, E. S.; Marshall, A. G. *Energy & Fuels* **2018**, *32*, 1206-1212.

(6) Pakarinen, J. M. H.; Teräväinen, M. J.; Pirskanen, A.; Wickström, K.; Vainiotalo, P. *Energy & Fuels* **2007**, *21*, 3369-3374.

(7) Hsu, C. S. *Energy & Fuels* **2012**, *26*, 1169-1177.

(8) Smith, D. F.; Podgorski, D. C.; Rodgers, R. P.; Blakney, G. T.; Hendrickson, C. L. *Anal. Chem.* **2018**, *90*, 2041-2047.

(9) Chacón-Patiño, M. L.; Blanco-Tirado, C.; Orrego-Ruiz, J. A.; Gómez-Escudero, A.; Combariza, M. Y. *Energy & Fuels* **2015**, *29*, 1323-1331.

(10) Jarvis, J. M.; Robbins, W. K.; Corilo, Y. E.; Rodgers, R. P. *Energy & Fuels* **2015**, *29*, 7058-7064.

(11) Klitzke, C. F.; Corilo, Y. E.; Siek, K.; Binkley, J.; Patrick, J.; Eberlin, M. N. *Energy & Fuels* **2012**, *26*, 5787-5794.

(12) Schmidt, E. M.; Pudenzi, M. A.; Santos, J. M.; Angolini, C. F. F.; Pereira, R. C. L.; Rocha, Y. S.; Denisov, E.; Damoc, E.; Makarov, A.; Eberlin, M. N. *RSC Advances* **2018**, *8*, 6183-6191.

(13) Klein, G. C.; Angström, A.; Rodgers, R. P.; Marshall, A. G. *Energy & Fuels* **2006**, *20*, 668-672.

(14) Cho, Y.; Na, J.-G.; Nho, N.-S.; Kim, S.; Kim, S. *Energy & Fuels* **2012**, *26*, 2558-2565.

(15) Gaspar, A.; Zellermann, E.; Lababidi, S.; Reece, J.; Schrader, W. *Energy & Fuels* **2012**, *26*, 3481-3487.

(16) Podgorski, D. C.; Corilo, Y. E.; Nyadong, L.; Lobodin, V. V.; Bythell, B. J.; Robbins, W. K.; McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. *Energy & Fuels* **2013**, *27*, 1268-1276.

(17) Rowland, S. M.; Robbins, W. K.; Corilo, Y. E.; Marshall, A. G.; Rodgers, R. P. *Energy & Fuels* **2014**, *28*, 5043-5048.

(18) Giraldo-Dávila, D.; Chacón-Patiño, M. L.; Orrego-Ruiz, J. A.; Blanco-Tirado, C.; Combariza, M. Y. *Fuel* **2016**, *185*, 45-58.

(19) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. *Energy & Fuels* **2018**, *32*, 314-328.

(20) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. *Energy & Fuels* **2017**, *31*, 13509-13518.

(21) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. *Energy & Fuels* **2018**, *32*, 9106-9120.
(22) Kendrick, E. *Anal. Chem.* **1963**, *35*, 2146-2154.
(23) van Krevelen, D. *Fuel* **1950**, *29*, 269-284.
(24) Hsu, C. S.; Qian, K.; Chen, Y. C. *Anal. Chim. Acta* **1992**, *264*, 79-89.
(25) Dier, T. K. F.; Egele, K.; Fossog, V.; Hempelmann, R.; Volmer, D. A. *Anal. Chem.* **2016**, *88*, 1328-1335.
(26) Nagy, T.; Kuki, Á.; Zsuga, M.; Kéki, S. *Anal. Chem.* **2018**, *90*, 3892-3897.
(27) Sato, H.; Nakamura, S.; Teramoto, K.; Sato, T. *J. Am. Soc. Mass. Spectrom.* **2014**, *25*, 1346-1355.
(28) Fouquet, T.; Sato, H. *Anal. Chem.* **2017**, *89*, 2682-2686.
(29) Teräväinen, M. J.; Pakarinen, J. M. H.; Wickström, K.; Vainiotalo, P. *Energy & Fuels* **2007**, *21*, 266-273.
(30) Huba, A. K.; Gardinali, P. R. *Sci. Total Environ.* **2016**, *563-564*, 600-610.
(31) Roach, P. J.; Laskin, J.; Laskin, A. *Anal. Chem.* **2011**, *83*, 4924-4929.
(32) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336-5344.
(33) Wu, Z.; Rodgers, R. P.; Marshall, A. G. *Anal. Chem.* **2004**, *76*, 2511-2516.
(34) Fedoros, E. I.; Orlov, A. A.; Zherebker, A.; Gubareva, E. A.; Maydin, M. A.; Konstantinov, A. I.; Krasnov, K. A.; Karapetian, R. N.; Izotova, E. I.; Pigarev, S. E.; Panchenko, A. V.; Tyndyk, M. L.; Osolodkin, D. I.; Nikolaev, E. N.; Perminova, I. V.; Anisimov, V. N. *Oncotarget* **2018**, *9*, 18578-18593.
(35) Ruddy, B. M.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. *Energy & Fuels* **2018**, *32*, 2901-2907.

For Table of Contents only

# *Mass-Remainder Analysis* (MARA): An Improved Method for Elemental Composition Assignment in Petroleomics.

Tibor Nagy, Ákos Kuki, Miklós Nagy, Miklós Zsuga, Sándor Kéki*

*Department of Applied Chemistry, Faculty of Science and Technology, University of Debrecen, H-4032 Debrecen, Egyetem tér 1., Hungary*