A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry

Saer Samanipour[a,*], Katherine Langford[a], Malcolm J. Reid[a], Kevin V. Thomas[a]

[a] *Norwegian Institute for Water Research (NIVA), 0349 Oslo, Norway*

Abstract

Gas chromatography coupled with high resolution time of flight mass spectrometry (GC-HR-TOFMS) has gained popularity for the target and suspect analysis of complex samples. However, confident detection of target/suspect analytes in complex samples, such as produced water, remains a challenging task. Here we report on the development and validation of a two stage algorithm for the confident target and suspect analysis of produced water extracts. We performed both target and suspect analysis for 48 standards, which were a mixture of 28 aliphatic hydrocarbons and 20 alkylated phenols, in 3 produced water extracts. The two stage algorithm produces a chemical standard database of spectra, in the first stage, which is used for target and suspect analysis during the second stage. The first stage is carried out through five steps via an algorithm here referred to as unique ion extractor (UIE). During the first step the m/z values in the spectrum of a standard that do not belong to

[*] Saer Samanipour
Email address: saer.samanipour@niva.no (Saer Samanipour)
[1] NIVA, Gaustadalléen 21, 0349 Oslo, Norway
Tel: +47 98222087

that standard are removed in order to produce a clean spectrum and then during the last step the cleaned spectrum is calibrated. The Dot-product algorithm, during the second stage, uses the cleaned and calibrated spectra of the standards for both target and suspect analysis. We performed the target analysis of 48 standards in all 3 samples via conventional methods, in order to validate the two stage algorithm. The two stage algorithm was demonstrated to be more robust, reliable, and less sensitive to the signal-to-noise ratio (S/N), when compared to the conventional method. The Dot-product algorithm showed lower potential in producing false positives compared to the conventional methods, when dealing with complex samples. We also evaluated the effect of the mass accuracy on the performances of Dot-product algorithm. Our results indicated the crucial importance of HR-MS data and the mass accuracy for confident suspect analysis in complex samples.

Keywords: Produced water, GC-HR-TOFMS, Dot product, Matching algorithm, Unique ion extractor, Reverse match, Suspect analysis, Target analysis

1. Introduction

Gas chromatography coupled with mass spectrometry (GC-MS) is one of the common analytical techniques for analysis of complex samples for volatile and semi volatile compounds [1–5]. The three main approaches to perform this type of analysis are: target analysis, where the analytical standard of the analyte is available; suspect analysis, where the analytical standard is not available however information, such as exact mass and the fragmentation pattern is available for that analyte; and finally non-target analysis, where no prior information is available for that analyte [6]. Confident detection of an analyte in a complex sample is a challenging task, particularly during suspect and non-target analysis [6, 7]. The introduction of high resolution and/or high accuracy mass spectrometers improved drastically the levels of confidence in the suspect analysis, however difficulties still persist [6, 8, 9].

For target analysis, depending on the target analyte and the data processing tools used for analysis, few m/z values and the absolute retention time are used for identity confirmation of a target analyte in the sample [10–13]. Regarding suspect analysis, the identity confirmation is carried out employing either the direct analysis or reverse analysis [9, 14, 15]. Direct analysis consists of first performing mass spectral deconvolution of the suspect peak in the sample, and then comparing the deconvoluted spectra to a standard database [16–18] (e.g. Mass spectral library of National Institute of Standards and Technology, NIST [19]). As a result of the spectral comparison the chemical structures with the highest similarity score are reported as a hit list. Lu et al. demonstrated that the conventional deconvolution algorithm may cause introduction of artifacts into the final deconvoluted spectrum, depending on the complexity of the sample [20], which translates into errata library matching and scoring. In case of reverse analysis, the spectra of a chemical standard is compared to the whole chromatogram of the sample and where the analyte is present in the sample a higher level of similarity score is observed [21]. A large number of scoring systems have been developed and tested on different datasets (as reviewed by Scheubert et al. 2013 [9]). Amongst the tested scoring algorithm the dot product has been recognized as one of the most reliable matching methods, for both direct and reverse analysis [16, 21, 22]. The direct matching algorithms appear to be highly sensitive to the quality of deconvolution, spectral weighting function, binning step, and Signal-to-Noise ratio (S/N) [9, 20, 23]. Also the mentioned scoring systems often do not produce high enough levels of confidence in the detection [23] . The reverse matching method shown to be less sensitive to levels of S/N [9, 14, 24]. For example, in the study by Sinha et al. the authors were able to detect trimethylsilyl in urine samples by employing a unit mass spectra of trimethylsilyl and reverse dot product methodology [21]. The confidence in the detection for the reverse matching algorithms, is highly dependent to the quality and the levels of mass accuracy of the standard spectra [16, 23]. Limited studies have focused on the matching algorithms for the GC-HR-MS data [22, 24], particularly the reverse matching methodology, due to the lack of GC-HR-MS spectral database of standards.

Herein we report on a two stage algorithm for target and suspect analysis in complex samples using GC-HR-MS data. In the first stage the unique ions of a standard spectra are extracted from the raw data (via unique ion extractor algorithm, UIE) in order to produce a chemical standard database of HR spectra. In the second step the clean spectra of a target/suspect analyte is compared to the whole GC-HR-MS chromatogram of the sample employing reverse dot product methodology (via Dot-product algorithm). The comparison between the standard spectra and the sample spectra results in a similarity matrix with higher levels of similarity for the analytes which are present in the sample compared to the background signal. This approach was validated by comparing the results of the two stage algorithm to the conventional target and suspect analysis method. Higher levels of reliability and robustness were observed for the two stage algorithm when compared to the conventional methods. The validation was carried out through the analysis of 48 analytes in 3 produced water extracts. The produced water samples consisted of a total extract of produced water, the non-polar fraction of produced water, and the polar fraction of produced water. The produced water extracts provided a high level of complexity for the validation study, due to the commonalities in the fragmentation pattern of the target/suspect analytes and the background signal. The two stage algorithm proved to be able to distinguish the signal of target/suspect analytes from the background signal successfully. The two stage algorithm produced 0 cases of false positive compared to 1 via the conventional method. Moreover, this algorithm showed to be less sensitive to the levels of S/N.

2. Experimental

2.1. Chemicals

A mixture of 28 aliphatic hydrocarbons and 20 alkylated phenols were purchased from Sigma-Aldrich, Norway. A complete list of the standards is provided in the Supporting Information, Table S1. ACS grade ethanol, dichloromethane, methanol, hydrochloric acid, sodium hydroxide, and sodium sulphate were also obtained from Sigma-Aldrich. We obtained technical grade glass fiber filter (GF/C) from VWR, Norway.

For our analysis we used an extract of produced water. Produced water is a petrogenic by-product of offshore petroleum extraction. Produced water is a complex mixture containing thousands of compounds including heavy metals, hydrocarbons, phenols, organic acids, and oil production chemicals [11]. An extract of produced water at pH 2, using dichloromethane was provided by Stiftelsen for Industriell og Teknisk Forskning, Trondheim, Norway (SINTEF). Herein we refer to this sample as total extract. The extraction was performed according to the guidelines of Norwegian Environmental Protection Agency for the sampling and analysis of oil and gas [2]. In short 2.5 L of produced water was extracted employing 60 mL of dichloromethane, via liquid-liquid extraction, for three constitutive times. The final extract was dried using sodium sulphate.

An aliquot of the total extract was fractioned into polar and non-polar portions. For this fractionation, we dissolved 1 mL of the total extract into 1 L of water at pH 11, which was carried out by shaking the solution for 24 h at 150 rpm. This solution was extracted using liquid-liquid extraction with 60 mL of dichloromethane for three consecutive times. The final extract was dried on a bed of sodium sulphate. The volume of the final extract was reduced to 1 mL of dichloromethane employing a turbovap system under a gentile flow of N_2. For the non-polar fraction, the pH of the water was reduced to 1 from 11. The same liquid-liquid extraction procedure was carried out for the acidified sample. The final extract of the acidified sample was considered the non-polar fraction of the total extract.

All the extracts were stored immediately at -20 °C until analysis.

2.2. GC-HR-TOFMS analysis

We analyzed mixtures of standards at three concentration levels (2, 10, and 20 ng/mL), the total extract (i.e. the total extract of produced water received from SINTEF), and the polar and non-polar fractions of the total extract with a GC-HR-TOFMS (GCT Premier, Waters, USA) equipped with electron impact ion source (EI). The separations were carried out on a BD-5 column (30 m 0.25 m 0.25 mm, Agilent). All the injections were performed in splitless mode having an injection volume of 1 μL. Helium was used as the carrier gas. The TOFMS collected 2 spectra every second between 50 Da and 600 Da. The detector exhibited a resolution of $\sim$ 8000 at half width full range (i.e. 50 Da to 600 Da). The detector was operated at 2850 V and a filament current of $\sim$ 1 mA. More information about the instrumental setup is provided in section S2 of Supporting Information.

2.3. Data analysis

The raw chromatograms were exported as netCDF files employing MassLynx (Waters, Manchester, UK). The raw chromatograms then were imported into matlab (R2015b) [25] for further processing. All the scripts for both the UIE and Dot-product algorithms were developed in matlab. As a validation tool for UIE algorithm as well as the target analysis, we used the software package TargetLynx (Waters, Manchester, UK) within the Masslynx. A target analyte was considered detected in TargetLynx if we observed a positive match between the retention times ± 5 s and the exact mass ± 10 mDa of the standard and the target peak in the sample. Both the retention window and the exact mass window were selected based on the observed variabilities in our dataset for these parameters. The minimum S/N required for a positive detection was set to 10.

The S/N calculations were performed via MassLynx. The signal was defined as the 50% of the peak hight whereas the noise was defined as the root mean square error of the 10 scans in one side of the peak. The ratio of these two values resulted in the S/N.

All the calculations were performed on a personal computer with an Intel i7, 2.8 GHz processor, and 16 GB of memory. The operating system was Windows 7 enterprise version.

3. Theory

The chromatograms of the standards were further processed with the UIE algorithm. We obtained clean and calibrated spectra of all 48 standards by processing their raw data via UIE algorithm. All the steps taken during the UIE are explained in detail in Section 3.1. These clean and calibrated spectra (i.e. the standard spectra) were used for both suspect and target screening via Dot-product algorithm (see Section 3.2 for more explanations regarding the Dot-product algorithm).

3.1. Unique ion extractor (UIE)

The unique ion extractor (UIE) is applied to the HR mass spectra of each standard before its storage in the personal library. The UIE algorithm produces the pure spectra that belongs to the chromatographic peak of a standard. This process takes place in total of 5 steps. During the data processing the user can decide the number of necessary steps to take in order to produce a final clean spectra of the target analyte.

1. Peak detection was performed using a lab-developed algorithm. In order to perform the peak detection, we generated the Savitsky-Golay smoothing vectors of first and second derivatives of the total ion chromatogram (TIC) [26, 27]. The apex of a peak was defined as the scan number, which has its first derivative equal to zero, and in the second derivative it has a negative minimum, and surrounded by two positive maximums. In order to optimize the smoothing functions (i.e. both the first and second derivatives), we tested different polynomial functions from first to fourth orders with smoothing window varying between 3 to 15 scans. For both the first and second derivatives, the best results were observed when employing a third order polynomial as the smoothing function and a smoothing window of 7 scans. We also recorded the location of the two positive maximums in the Savitsky-Golay second derivative vector (Figure 1, step 1). These locations, for a completely resolved peak (i.e. chromatographic resolution larger than 3), were considered a conservative estimate of the starting and the end points of a peak. However, these points could be fed manually to the UIE algorithm. Therefore, any other peak detection algorithm could be employed for this task, as long as these three parameters are recorded for each peak (i.e. peak apex, starting point, and the end point of the peak).
2. The spectral averaging step is an optional step, which follows the peak detection step. The peak apex, start, and end information recorded during the peak detection are used during this step. For the spectral averaging, the MS spectra

of 3 to 5 scans are averaged, where the peak apex is the central point in the averaging window (Figure 1, step 2). With an averaging window of 3 scans we were able to find the best conditions. The 3 scans averaging window enabled us to avoid the MS signal, which belongs to the background signal independently from the peak intensity. Throughout this article we refer to the apex averaged spectra as the ”apex spectra”.

3. The background signal subtraction is also an optional step, where the background signal is subtracted from the apex spectra of the peak. The background signal is defined as the average spectra of 40 neighboring scans of the peak. In other words, the spectra of 20 scans before the peak start point and 20 scans after the peak end point are averaged and then subtracted from the apex spectra (Figure 1, step 3). The dimension of the background window is defined by the user and depends on the chromatographic resolution of the peak. In our case a window of 20 scans guaranteed the removal of background signal and also enabled a faster unique ion selection.

4. The unique ion selection is carried out by comparing the retention time of the extracted ion chromatogram (XIC) for every single m/z value, which has an intensity larger than zero. An m/z peak is excluded from the apex spectra if it produces a retention time larger or smaller than the peak retention time ± 2 scans (Figure 1, step 4). This retention window may be modified based on the TOF-MS sampling rate. In other words, this window may be larger than 2 scans for instruments with a sampling rate larger than 2 Hz.

5. The final step is the calibration of the clean apex spectra. This step also is

optional depending on the instrumentation. We calibrated the clean spectra employing the calibrant signal (heptacosa), which was injected into the source during each scan. We generated two vectors consisting of the exact masses of the calibrant fragments and the measured masses for those fragments. We fitted a third order polynomial with four fitting parameters to the measured mass vector and the mass residuals (i.e. the difference between the exact mass and measured mass). The fitted function enabled us to calculate the shift for each m/z value during each scan, thus calibration.

Finally, the cleaned and calibrated spectra is stored in a database including some chemical specific information, such as CAS number, retention time, boiling point, and log K_{ow}. Both boiling point and log K_{ow} were estimated employing EPISuite [28].

3.2. Dot-product algorithm for HRMS data

The Dot-product algorithm is based on the similarity between the spectra of a standard and the sample, which is a modified version of the reverse match originally developed by Stein [16]. A recent report showed the applicability of this algorithm for comprehensive two-dimensional gas chromatography coupled to a low resolution TOF-MS dataset [21]. Herein we report on the combination of UIE and an adaptation of DotMap algorithm for GC-HR-TOFMS data analysis. The Dot-product algorithm computes the vectorial product of scaled, normalized, and weighted clean mass spectra of the standard and the sample mass spectra, for each scan. More detail information about the algorithm is provided elsewhere [21]. Additionally, we

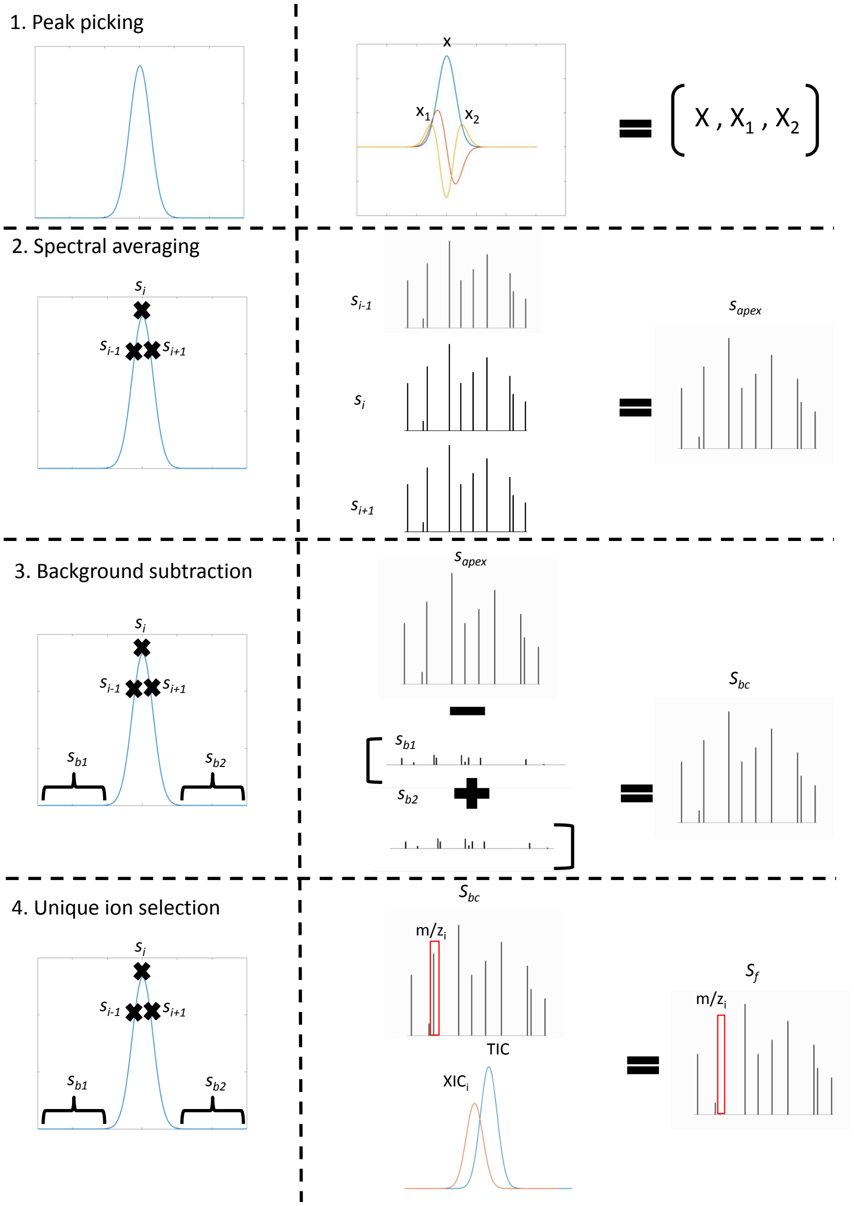

Figure 1: Conceptual schematics of the steps in the UIE algorithm with synthetic data. In this figure: x, x_1, and x_2 depict the the scan number of the peak apex, peak start, and peak end; s_i, s_{i-1}, s_{i+1}, and s_{apex} represent the spectra for the scan numbers i, $i-1$, $i+1$, and the average spectra of the three scans; s_{b1} and s_{b2} illustrate the average spectra of noise before and after the peak, whereas s_{bc} shows the background corrected spectra; m/z_i depicts an m/z value with a non-zero intensity, XIC_i and TIC illustrate the extracted ion chromatogram for the m/z_i and total ion chromatogram; and finally s_f is the clean spectra.

combined the results of the Dot-product algorithm with the exact or estimated retention time, and 4 to 5 XICs for the m/z values with the highest relative intensity and the exact mass of the chemical. The combination of this information provided an identification confidence level of 1 for target screening and level 2 for suspect screening [6]. The identification confidence level 1 refers to an ideal situation where there are positive matches of both the retention time and the mass spectra between the reference standard and the considered peak in the sample (i.e. target analysis) whereas the confidence level 2 refers to a case where there is a positive match between the library spectrum and the spectrum of the peak in the sample (i.e. suspect analysis) [6].

4. Results and discussions

We processed the MS spectra of all 48 standards with the UIE algorithm. A chemical standard database was created based on the results of UIE algorithm. We performed both target and suspect analysis for 48 compounds in three complex samples. These samples consisted of a total extract, an extract of polar fraction, and an extract of the non-polar fraction of produced water. The target analysis were performed employing both the Dot-product algorithm and the commercially available TargetLynx software package. The results of the two mentioned approaches enabled an objective validation of the Dot-product algorithm. For the suspect screening, we tested the Dot-product algorithm by analyzing the 3 complex samples for all 48 standards. In this case, the retention time of each suspect analyte was estimated by taking advantage of its boiling point.

4.1. Unique Ion Extractor (UIE)

The UIE algorithm is a fully automized approach for the extraction of the unique ions, which belong to a chemical, and creation of a chemical standard database. This algorithm removes the m/z values which caused the background. The background signal is defined as the signal produced by noise, carryover due to the previous analysis, and overlapping peaks. The UIE proved effective for all the peaks where the chromatographic resolution was larger than 0.5.

The UIE successfully removed the m/z values introduced into the spectra by noise, background and other interfering signals for all 48 standards. As an example we selected the peak of octadecane with chromatographic resolution of 0.8 and scan number of 592, Figure 2. This peak was partially overlapped with a neighboring peak therefore its pure spectra was buried in the background signal. The m/z value, which theoretically should have had the highest intensity, i.e. 71.084 ± 10 mDa [19], appeared to have an intensity roughly one order of magnitude lower than the m/z value with the highest intensity (i.e. 218.985) in the octadecane raw spectra, Figure 2. Before the UIE treatment the m/z value with the highest intensity in the spectra of the apex, excluding the m/z of the calibrant (i.e. 218.985), was 130.990 whereas after treatment the m/z value with the highest intensity in that peak was 71.084 ± 10 mDa, which was in agreement with the literature spectra published for octadecane [19]. Major part of the m/z values larger than 254.297, such as m/z values 363.978, 413.976, 436.977, and 501.972 were removed during the spectral subtraction. These m/z values showed to have similar intensities in the surrounding scans of the peak (i.e. the octadecane peak). The m/z values 163.992, 168.987, 213.988, and 219.989 were removed during the unique ion selection process. These m/z values did not have an apex within the retention window of octadecane (see section 3.1 for more details regarding unique ion selection process). We also processed the spectra of the same peak (i.e. octadecane) without spectral subtraction. We observed 100% agreement between the final spectra of octadecane processed with and without spectral subtraction. We observed an increase in the time necessary for the UIE algorithm for processing the spectra of octadecane when the spectral subtraction was skipped. The observed increase in the analysis time was caused by the step 4 of the UIE, due to larger number of non-zero intensity m/z values compared to the case where the spectral subtraction was not skipped. It is worth noting that the analyzed standard mixture was a particularly difficult one due to the similarity in the fragmentation pattern of different standards in the mixture. For example m/z values 57.068 and 85.100 were observed in the spectra of almost all of the analyzed alkanes. Therefore, we observed traces of these m/z values in the spectra of the standards which theoretically should not have had these m/z values (e.g. 2,4,6-trimethylphenol).

The UIE algorithm showed high levels of robustness with respect to the variation in the S/N ratio. We evaluated the effect of the S/N ratio on the performances of the UIE algorithm by decreasing the concentration of the standard mixture, roughly, to the instrument limit of detection (i.e 2 ng/mL). The S/N for the analyzed standards varied from 32 for undecane at 2 ng/mL to 2640 for heneicosane at 20 ng/mL, Table S1. The algorithm was able to produce the clean spectra for all 48 standards at all

3 analyzed concentration levels or S/N.

Despite the difficulties posed by the analyzed sample complexity and the levels of S/N, the UIE algorithm showed its ability to remove the irrelevant m/z values from the spectra of a peak and produce a clean calibrated spectra for all 48 analyzed standards. Finally, the UIE algorithm takes around 20 s for processing the spectra of a peak including all 5 steps, i.e. peak detection, spectral averaging, spectral subtraction, unique ion selection, and the mass calibration.

4.2. Target analysis of produced water extracts

We analyzed all 3 produced water extracts for 48 target analytes. For the target analysis we took advantage of the retention information recorded in the standard database during UIE spectral processing. We defined a retention window of 21 scans (i.e. 10.5 s) with the absolute retention time of the target analyte in the center of this window. We used the Dot-product algorithm to calculated the similarity matrix, Eq. 1.

$$SIM_{i,j} = (\frac{m_j(\sqrt{S_{sample}})_i}{\sum_{j=1}^{k}(m_j(\sqrt{S_{sample}})_i)}) \cdot (\frac{m_j(\sqrt{S_f})}{\sum_{j=1}^{k}(m_j(\sqrt{S_f}))}) \qquad (1)$$

where $SIM_{i,j}$ represents the similarity matrix, m represents an m/z value in both the sample spectra (i.e. S_{sample}) and the standard spectra (i.e. the clean and calibrated spectra produced via UIE, S_f), i is the index for the number of spectra recored in the retention window (e.g. for a retention window of 21 scans i is a number $1 \leq i \leq 21$), and j is the index for the number of m/z values recored in spectra with the maximum value of k. The $SIM_{i,j}$ computed for each scan number and m/z values

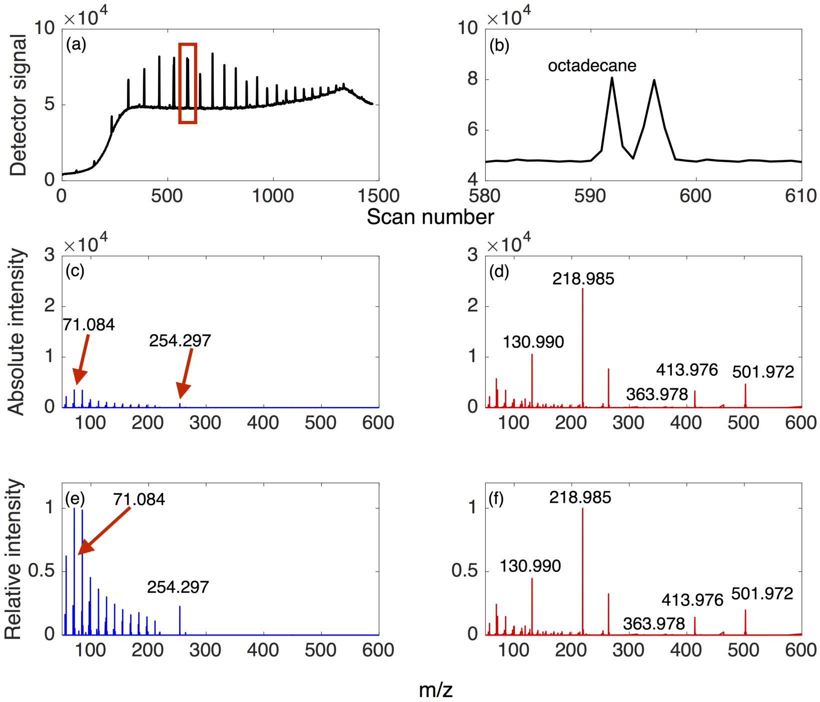

Figure 2: Figure showing (a) the TIC of the chemical standards at lowest concentration level (i.e. 2 ng/mL); (b) the zoomed in region of the TIC where the peak of octadecane is located; (c) the clean and calibrated spectra of octadecane with two m/z assigned; (d) the raw spectra of the octadecane peak with few m/z values assigned; (e) the normalized clean spectra relative to the m/z value with the highest intensity (i.e. 71.084); and (f) the normalized raw spectra of the octadecane peak relative to the calibrane m/z (i.e. 218.985).

within the retention window of a target analyte produces a similarity matrix. If a target analyte is present in the analyzed sample, the scan numbers where the target analyte is located in the sample show higher level of spectral similarity compared to the other scan numbers in that retention window (Figure 3). A perfect match between the sample spectra and the standard spectra produces a similarity value of 1 whereas a perfect orthogonality between the two spectra produces a similarity value of 0. In addition to the similarity matrix, we increased the confidence level in the positive (i.e. confirmed presence) and/or negative (i.e. confirmed absence) detections by extracting the XIC of 3 m/z values with the highest relative intensities and the XIC of the exact mass of the target analyte (Figure 3). The presence of the signal for the 4 XICs within the accepted retention window indicates that those ions belong to the target analyte and not to the background signal. Therefore, a target analyte detected in the sample must show an apex in the similarity matrix at scan number of the absolute retention time (i.e. the retention time of standard) ± 1 scans, and show apexes at the same location for at least 3 out 4 XICs (i.e. the 3 m/z values with the highest intensity and the exact mass). This implies a five-point criterion (i.e. similarity peak, 3 out 4 XICs, and the retention time match between these signals) for both positive and negative detections, which guaranties a high level of confidence in detections [6, 29].

For both the total extract and non-polar fraction of produced water, we successfully detected 37 out of 48 target analytes whereas for the polar fraction, we detected 35 out 48 target analytes, using the Dot-product algorithm (Table S2). As a validation tool we performed the same target analysis of the 3 produced water extracts, employing TargetLynx (section 2.3). Except two cases, we did not observed any discrepancies between the two approaches. Target analyte undecane was detected in the non-polar fraction of produced water via Dot-product algorithm whereas it was reported as non detected in the same sample by TargetLynx (Table S2). Within the retention window of undecane, we observed a clear peak in the similarity matrix. We also observed 3 peaks with correct retention time in the XIC of the 3 m/z values with the highest intensity. However, we did not observe any peak in the XIC based on the exact mass of undecane. Further inspections into the data showed that due to low levels of S/N of this target analyte, the m/z value of the exact mass of the undecane had recorded an intensity of zero. Therefore this target analyte was considered absent in the sample by TargetLynx. On the other hand, with the Dot-product algorithm 5 out of 6 criteria for positive detection were met and therefore it was considered a positive detection. For the target analyte 4-n-penthylphenol the Dot-product algorithm resulted in the negative detection (Figure 4) whereas the TargetLynx appeared to have detected this target analyte in the polar fraction of produced water (Table S2). In the retention window of 4-n-penthylphenol we did not observe a clear peak in the similarity matrix (Figure 4). However, a small peak appeared in the XIC of the exact mass near the absolute retention time of 4-n-penthylphenol. Also we only observed a peak for the m/z value of 150.09 but not for the other two m/z values (i.e. 135.06 and 117.06). All these evidences combined strongly suggested the negative detection (i.e. the absence) of 4-n-penthylphenol in the analyzed sample. Further inspection of the MS spectra of the peak located at the location of 4-n-penthylphenol in the polar fraction of produced water, demonstrated that several important m/z values (e.g. 135.06, 117.06, 105.06) were not present in the spectra (Figure S1), which confirmed the lack of detection of this target analyte in that sample. These results again indicate the importance of the application of the whole spectra rather than few selected ions in order to avoid results containing false positive and/or false negatives.

The Dot-product algorithm was able to detect and confidently confirm the presence of a target analyte in complex samples. In cases with low levels of S/N the Dot-product algorithm showed more effective in target analysis than conventional approach (i.e. TargetLynx with an m/z value as qualifier). Moreover, when we tried to include more than one m/z qualifier in the TargetLynx detection setup, the automated target analysis algorithm failed to detect the target analyte in the analyzed samples. As a consequence of these failures we had to manually add the mentioned peaks into the detected target analyte list. Finally, performing target analysis via Dot-product algorithm takes around 40 s and it produces detection confidence level of 1 for both positive and negative detections.

4.3. Suspect analysis of produced water extracts

For the suspect analysis, we used the same 3 produced water extract chromatograms and the standard database of 48 chemicals. However, for the suspect analytes we did not use the retention time information during the analysis. The retention times of the suspect analytes were estimated using a linear model with 2 fitting parameters between the retention time of target analytes and their boiling points. The linear model showed to have a $R^2 \approx 0.98$, assuming a 95% confidence

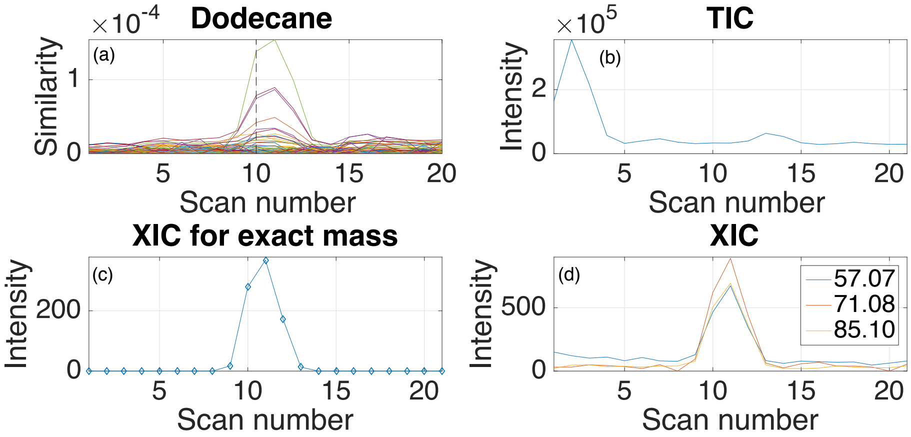

Figure 3: Figure depicting (a) the similarity matrix for dodecane with a mass window of ± 10 mDa in the polar fraction of produced water and the absolute retention time of the standard showed by the dotted line, (b) the TIC of the retention window for dodecane in the polar fraction of produced water, (c) the XIC of the exact mass (170.203 ± 10 mDa) of dodecane in the polar fraction of produced water chromatogram, and (d) the XIC for 3 m/z values (mass window of ± 10 mDa) with the highest intensity, based on the standard spectra, in the polar fraction of produced water.

interval. We divided the 48 standards in target analytes, which were a random pool of 18 chemicals selected from the 48 standard, and suspect analytes, which were the remainder 30 compounds. Every time this process repeated a new set of target and suspect analytes were created. Thus, we repeated this process 10 times in order to make sure that every single standard was considered as a suspect analyte at least once. We defined the retention window as the estimated retention time ± 60 scans, with the estimated retention time in the center of the window (Figure 5). The width of the window (i.e. 121 scans or 60.5 s) was defined based on the 95% confidence interval of the estimated retention time. The width of the retention window is defined by the user, therefore the operator can choose this parameter based on the instrumental setup and also the uncertainty in the estimated retention time. The larger

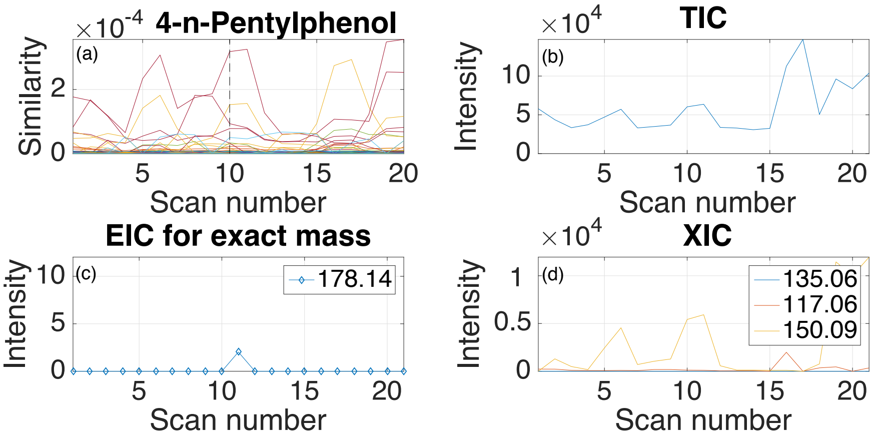

Figure 4: Figure depicting (a) the similarity matrix for 4-n-penthylphenol with a mass window of ± 10 mDa in the polar fraction of produced water and the absolute retention time of the standard showed by the dotted line, (b) the TIC of the retention window for 4-n-penthylphenol in the polar fraction of produced water, (c) the XIC of the exact mass (178.14 ± 10 mDa) of 4-n-penthylphenol in the polar fraction of produced water chromatogram, and (d) the XIC for 3 m/z values (mass window of ± 10 mDa) with the highest intensity, based on the standard spectra, in the polar fraction of produced water.

is the retention window the longer is the time needed for the analysis. Additionally, for the suspect screening we used 5 XICs, consisting of the exact mass and 4 m/z values with the highest intensities. Also for the suspect analysis the presence of a suspect was confirmed in the sample if and only if it met at least 6 out of 7 criteria.

We observed 100% agreement between the results of suspect and target analysis of the 3 samples. The Dot-product algorithm also in this case successfully detected 35 out of 48 target analytes in the polar fraction of produced water, and 37 out of 48 target analytes in both the total extract and the non-polar fraction of produced water. The Dot-product algorithm takes less than 2 min for confident detection of a suspect analyte in a complex sample. Differently from the conventional method (i.e. application of one or two m/z values as qualifiers) where the analyst must further inspect the data in order to increase the level of confidence in the positive and/or negative detections, the Dot-product algorithm does not require further inspection in the data. However, the analyst must make sure that the provided retention window to the algorithm is relevant to the analyzed suspect. For example if due to the high levels of uncertainty in the estimated retention time and an inappropriate selection of the width of the retention window the signal of suspect analyte happens to be outside of the provided retention window, the Dot-product algorithm may produce a false negative. All considered, the Dot-product algorithm provides the tools for an objective, fast, and confident suspect screening.

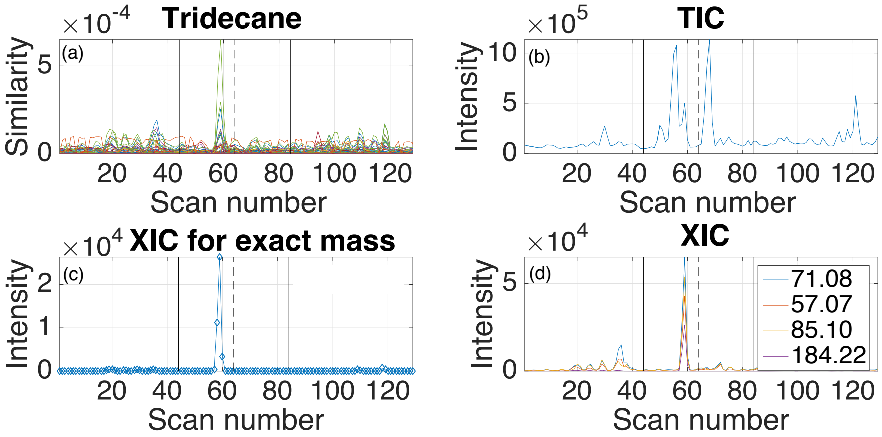

Figure 5: Figure depicting (a) the similarity matrix for tridecane with a mass window of ± 10 mDa in the non-polar fraction of produced water and the estimated retention time of the standard showed by the dotted line, (b) the TIC of the retention window for tridecane in the non-polar fraction of produced water, (c) the XIC of the exact mass (184.219 ± 10 mDa) of tridecane in the non-polar fraction of produced water chromatogram, and (d) the XIC for 4 m/z values (mass window of ± 10 mDa) with the highest intensity, based on the standard spectra, in the non-polar fraction of produced water.

4.4. Evaluation of the odds of false positive detections

We examined the odds of false positive results for both the Dot-product algorithm and TargetLynx, based on the complexity of the background signal. We generated two types of background signals and calculated the similarity values between all 48 analytes and these background signals. The background signals consisted of 5 randomly selected scans of the total extract chromatogram and 5 randomly selected scans of an analytical blank sample. Both background signals were considered analyte free (see section S4 in the SI). We also estimated the minimum and maximum similarity thresholds for all 48 analytes included in this study. The calculated similarity value of the full spectral comparison between the analyte spectrum and background signal was considered the minimum similarity threshold whereas the calculated similarity value of the analyte spectra with itself was assumed the maximum similarity threshold. The minimum similarity threshold was considered the minimum similarity signal necessary for a positive detection whereas the maximum similarity threshold was considered the effective similarity value achieved by a perfect match. We considered an algorithm to, potentially, results in a false positive if and only if the similarity value for the analyte and background (i.e. negative detection) was larger than maximum similarity threshold, Figure 6. For example, the similarity values between tetracosane and the noisy background signal (i.e. produced water background signal), when less than 10 ions were used for similarity calculation, were larger than the maximum threshold of similarity. This implied that, in that case, if an algorithm uses less than 10 ions for identification of tetracosane, this algorithm may result in a false positive.

The minimum threshold of similarity appeared to be dependent on the complexity of the background signal. The averaged minimum similarity threshold for the Dot-product algorithm varied from 1×10^{-5}, for the analytical blank background signal, to 1×10^{-4} for the produced water background signal, based on 960 evaluated cases. In other words, for the less noisy background (blank) the Dot-product algorithm needed less signal in order to confidently confirm the presence of chemical in the sample, whereas for the more noisy sample (produced water background) more signal was necessary in order to identify the target/suspect analyte in the sample. For the maximum similarity threshold, we observed a similar value of 3×10^{-3} for all 48 analytes.

The Dot-product algorithm resulted in a rate of false positive (RF) of zero for the produced water sample whereas the TargetLynx produced an RF of 0.34 (i.e. 25 analytes out of 48) for the same sample. Both evaluated methods resulted in RF values of zero for the analytical blank background. For the blank background, independently from the number of ions included in the similarity calculations, the similarity value for the background signal (i.e. the negative detection) was always smaller than the similarity value observed for the analyte signal (i.e. positive detection), Figure 6. This implied that confident identification was possible employing only one ion, thus $RF = 0$ for both algorithms. However, for a more complex background signal for 25 out of 48 analytes the application of the whole spectrum appeared to be necessary in order to ovoid false positive results (e.g. tetracosane Figure 6). These results may indicate the higher odds of the conventional methods to produce a false positive result for highly complex samples compared to the two stage algorithm. Our data also demonstrate that the full spectral comparison is necessary for a confident identification in the complex samples. It should be noted that the RFs and the similarity thresholds are only indicative values and their absolute values may change according to the analyzed sample and/or the analytes. Also further investigations regarding this subject are needed.

4.5. The effect of mass accuracy on the Dot-product algorithm

We evaluated the effect of mass accuracy on the Dot-product algorithm. Our instrument after mass calibration showed to have a mass accuracy of $\leq$ 10 mDa for the whole measured mass range (i.e from 50 Da to 600 Da). We modified the mass accuracy of our dataset by changing the thickness of the bins alongside the m/z vector. For example, with a mass accuracy of 10 mDa the thickness of each bin is 0.01 which implies that the distance between two m/z values is 0.01. This produces a sequence of m/z values such as 100.01, 100.02, 100.03, and so on for the whole measured mass range. Therefore, the signal for all the m/z values between 100.015 and 100.025 were stored as one single intensity in the 100.02 bin. As a consequence, by changing the thickness of the bins we were able to modify the level of mass accuracy in our data set. We computed the similarity matrix of 5 target analytes detected by both Dot-product algorithm and the MassLynx (i.e. dodecane, heneicosane, hexacosane, 4-ethylphenol, and 2,4,6-trimethylphenol) in the total extract of produced water at 4 different levels of mass accuracy, i.e. unit mass, 100 mDa, 10 mDa, and 1 mDa (Figure 7). It is worth remembering that our instrument is not capable of producing

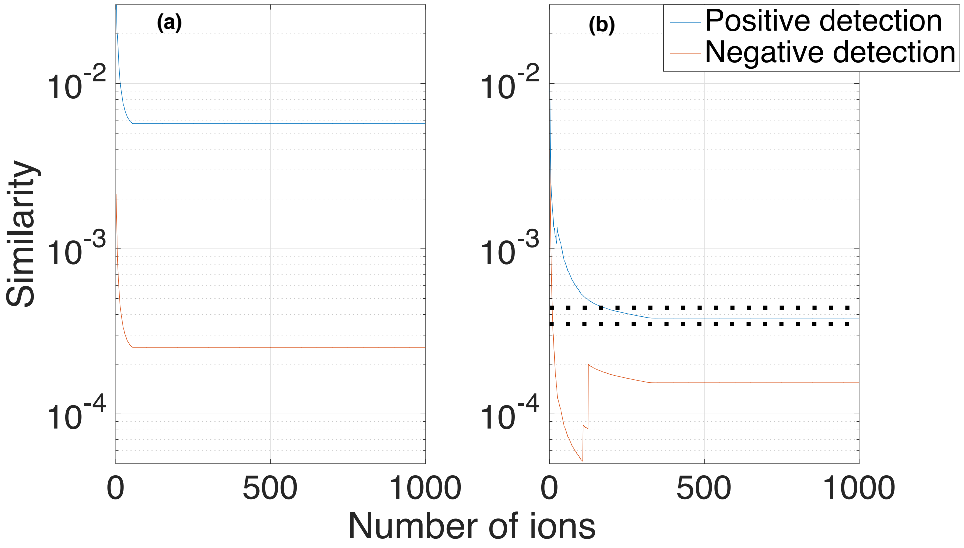

Figure 6: The similarity values of tetracosane as a function of the number of ions included for the similarity calculation in (a) analytical blank sample, and (b) in produced water sample. The negative detection depicts the background signal, the positive detection depicts the analyte signal, and the dotted lines indicate the similarity values for < 11 ions which are larger than the maximum threshold of similarity, thus potential false positive detections.

a mass accuracy of 1 mDa.

We observed the highest level of distinction between target/suspect analyte signal and the background at 10 mDa mass accuracy (Figure 7). This trend was observed for all 5 standards. As an example, we focus on standard heneicosane, which appeared to be representative for all 5 analyzed standards. At the unit mass accuracy the signal of heneicosane in the similarity matrix was covered by the background signal. Based on the similarity matrix at unit mass accuracy this standard was not detected in the sample, even though we previously confirmed its presence by both Dot-product algorithm and MassLynx. This was attributed to the complexity of the sample, high level of noise, and the abundance of the commune fragments between the heneicosane and the background. Therefore, unit mass accuracy appeared to be not enough for separating the signal of heneicosane from the background. Increasing the mass accuracy from unit mass to 100 mDa and further to 10 mDa, as expected, caused a clear distinction between the signal of heneicosane and background. The signal of heneicosane with a mass accuracy of 10 mDa was 6 times larger than the background signal whereas with the mass accuracy of 100 mDa it was only a factor of 2. In case of mass accuracy of 1 mDa due to the instrumental limitations the signal of both heneicosane and background were suppressed, which suggested zero similarity between the standard spectra and the sample spectra. Our data indicated that the Dot-product algorithm performs the best with the highest level of mass accuracy permitted by the instrumental limitations. Our data also may explain the difficulties observed by analysts while using unit mass libraries, such as NIST library. However, the Dot-product algorithm with an appropriate level of mass accuracy showed to be a powerful tool for both target and suspect analysis.

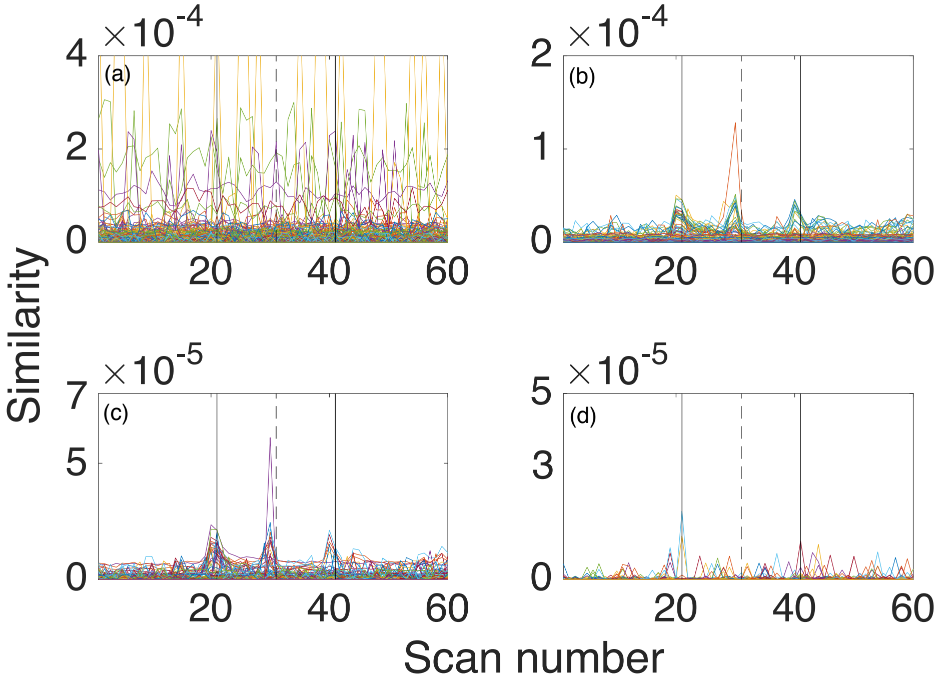

Figure 7: Computed similarity matrix of heneicosane in the total extract of produced water (a) with a unit mass accuracy, (b) with 100 mDa mass accuracy, (c) with 10 mDa mass accuracy, and (d) with 1 mDa mass accuracy.

5. Conclusions

Suspect and target screening of volatile and semi volatile organic compound in complex samples is challenging task. Here we report on the development and validation of a two stage method which enables the confident target and suspect analysis. A chemical spectra database was created by processing the raw spectra of the standards using UIE. The database of the clean spectra was used for both target and suspect analysis of complex samples, via Dot-product algorithm. The results of the two stage algorithm were cross validated employing conventional method (via MassLynx).

The UIE algorithm showed to be able to extract the unique ions of a chromatographic peak, even under difficult circumstances, such as low levels of S/N and sample complexity. The UIE successfully produced the clean and calibrated spectra of 48 standards at concentration levels near limit of detection. This algorithm removed the ions introduced by the background signal, even when the analyte signal was shadowed by the background. However, further investigation into the effect of concentration on the UIE and commercially available algorithms are needed. The necessary time for processing the spectra of a standard varied between 15 to 35 s, based on the number of steps included in the algorithm as well as the user defined parameters. This method demonstrated to be a fast, reliable, and robust algorithm for creation of personal databases of HR spectra.

The Dot-product algorithm can be used for both target and suspect analysis of complex samples. The comparison between the Dot-product algorithm and the conventional method (via TargetLynx) indicated that the Dot-product algorithm has lower probability of false positives. However, particular care should be taken in selection of the algorithm parameters, e.g. the retention window and the mass accuracy. The Dot-product algorithm enabled the detection of a target/suspect analyte in a complex sample with confidence levels of 1 for target analysis and 2 for suspect analysis. Differently, from the conventional methods of target and suspect analysis, the Dot-product minimizes the post inspection of the positive and negative detection, by providing the clear evidence for both positive and negative detections. Also, this method showed to be more robust and effective than the conventional target and suspect analysis methods for particularly difficult samples (e.g. produced water). This method demonstrated to be less affected by the sample complexity caused by high levels of noise and fragmentation pattern similarities between the target/suspect analytes and the background. Considering that the similarity score follows the chromatographic peak shape in the Dot-product algorithm, the analyst can verify the presence of an actual chromatographic peak and not only a match factor. Moreover, Dot-product algorithm does not require deconvolution of the sample chromatogram, which has been shown to be a challenging task [20]. Our analysis showed that the Dot-product algorithm is a powerful method for confident identification of target/suspect analytes in complex samples. The target analysis via Dot-product took less than a min whereas the suspect analysis in average took roughly 2 min. The time necessary for the analysis was highly dependent on the width of the retention window, particularly for suspect analysis.

We also evaluated the effect of the mass accuracy on the performances of the Dot-product algorithm. We observed a clear improvement in the performances of Dot-product algorithm with respect to the mass accuracy. The Dot-product algorithm was not able to detected the target and suspect analytes in the total extract of produced water at unit mass accuracy. This failure in the performances of Dot-product algorithm was attributed to the complexity of the analyzed sample and low levels of S/N. However, the Dot-product algorithm demonstrated capable of processing the same complex sample (i.e. the total extract of produced water) with mass accuracies of 100 and 10 mDa. Our results indicated the crucial importance of HR-MS data for confident target and suspect analysis.

In overall, the two stage algorithm demonstrated to be a fast and robust method for confident target and suspect analysis of complex samples via GC-HR-TOFMS. The evaluation of the two stage algorithm for LC-MS data will be the subject of future studies.

6. Acknowledgments

The authors are thankful to the Research Council of Norway for the financial support of the project (RESOLVE, 243720). We also would like to thank the NIVA staff for their technical support during the project and SINTEF for providing the produced water extract.

7. References

[1] Weiguang Xu, Xian Wang, and Zongwei Cai. Analytical chemistry of the persistent organic pollutants identified in the stockholm convention: A review. *Anal. Chim. Acta*, 790:1–13, 2013.

[2] Norog. *Norwegian Oil and Gas recommended guidelines for sampling and analysis of produced water, translated version*, 2012 edition, 2003.

[3] Derek Muir and Ed Sverko. Analytical methods for pcbs and organochlorine pesticides in environmental monitoring and surveillance: a critical appraisal. *Anal. Bioanal. Chem.*, 386(4):769–789, 2006.

[4] Douglas A Skoog and Donald M West. *Principles of instrumental analysis*, volume 158. Saunders College Philadelphia, 1980.

[5] Edmond Hoffmann. *Mass spectrometry.* Wiley Online Library, 1996.

[6] Emma L Schymanski, Heinz P Singer, Jaroslav Slobodnik, Ildiko M Ipolyi, Peter Oswald, Martin Krauss, Tobias Schulze, Peter Haglund, Thomas Letzel, Sylvia Grosse, et al. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.*, 407(21):6237–6255, 2015.

[7] Fang Zhang, Haoyang Wang, Li Zhang, Jing Zhang, Ruojing Fan, Chongtian Yu, Wenwen Wang, and Yinlong Guo. Suspected-target pesticide screening using gas chromatography–quadrupole time-of-flight mass spectrometry with high

resolution deconvolution and retention index/mass spectrum library. *Talanta*, 128:156–163, 2014.

[8] Emma L Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P Singer, and Juliane Hollender. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.*, 48(4):2097–2098, 2014.

[9] Kerstin Scheubert, Franziska Hufsky, and Sebastian Böcker. Computational mass spectrometry for small molecules. *J. Cheminf.*, 5:12, 2013.

[10] Saer Samanipour, Petros Dimitriou-Christidis, Jonas Gros, Aureline Grange, and J Samuel Arey. Analyte quantification with comprehensive two-dimensional gas chromatography: Assessment of methods for baseline correction, peak delineation, and matrix effect elimination for real samples. *J. Chromatogr. A*, 1375:123–139, 2015.

[11] Kevin V. Thomas, Katherine Langford, Karina Petersen, Andy J. Smith, and Knut E Tollefsen. Effect-directed identification of naphthenic acids as important in vitro xeno-estrogens and anti-androgens in north sea offshore produced water discharges. *Environ. Sci. Technol.*, 43(21):8066–8071, 2009.

[12] Matthias Onghena, Els Van Hoeck, Joris Van Loco, María Ibáñez, Laura Cherta, Tania Portolés, Elena Pitarch, Félix Hernandéz, Filip Lemière, and Adrian Covaci. Identification of substances migrating from plastic baby bottles using a combination of low-resolution and high-resolution mass spectrometric analysers

coupled to gas and liquid chromatography. *J. Mass Spectrom.*, 50(11):1234–1244, 2015.

[13] Ricardo JN Bettencourt da Silva. Evaluation of trace analyte identification in complex matrices by low-resolution gas chromatography–mass spectrometry through signal simulation. *Talanta*, 150:553–567, 2016.

[14] Arvind Visvanathan. *Information-theoretic mass spectral library search for comprehensive two-dimensional gas chromatography with mass spectrometry.* PhD thesis, 2008.

[15] L Cherta, T Portolés, E Pitarch, J Beltran, FJ López, C Calatayud, B Company, and Felix Hernández. Analytical strategy based on the combination of gas chromatography coupled to time-of-flight and hybrid quadrupole time-of-flight mass analyzers for non-target analysis in food packaging. *Food chem.*, 188:301–308, 2015.

[16] Stephen E Stein. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, 10(8):770–781, 1999.

[17] Imhoi Koo, Seongho Kim, and Xiang Zhang. Comparative analysis of mass spectral matching-based compound identification in gas chromatography–mass spectrometry. *J. Chromatogr. A*, 1298:132–138, 2013.

[18] Imhoi Koo, Xiang Zhang, and Seongho Kim. Wavelet-and fourier-transform-

based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Anal. Chem.*, 83(14):5631–5638, 2011.

[19] National Institue of Standards and Technology (NIST). Nist chemistry webbook.

[20] Hongmei Lu, Yizeng Liang, Warwick B Dunn, Hailin Shen, and Douglas B Kell. Comparative evaluation of software for deconvolution of metabolomics data based on gc-tof-ms. *TRAC-Trends Anal. Chem.*, 27(3):215–227, 2008.

[21] Amanda E Sinha, Janiece L Hope, Bryan J Prazen, Erik J Nilsson, Rhona M Jack, and Robert E Synovec. Algorithm for locating analytes of interest based on mass spectral similarity in GC× GC–TOF-MS data: analysis of metabolites in human infant urine. *J. Chromatogr. A*, 1058(1):209–215, 2004.

[22] Michael Edberg Hansen and Jørn Smedsgaard. A new matching algorithm for high resolution mass spectra. *J. Am. Soc. Mass Spectrom.*, 15(8):1173–1180, 2004.

[23] Cuiping Li, Jiuqiang Han, Qibin Huang, Baoqiang Li, Zhongyao Zhang, and Chuntao Guo. An effective two-stage spectral library search approach based on lifting wavelet decomposition for complicated mass spectra. *Chemometr.Intell. Lab.*, 132:75–81, 2014.

[24] Rasmus Bro. Review on multiway analysis in chemistry2000–2005. *Crit. Rev. Anal. Chem.*, 36(3-4):279–293, 2006.

[25] MATLAB User's Guide. The mathworks. *Inc., Natick, MA*, 5, 1998.

[26] Jean Steinier, Yves Termonia, and Jules Deltour. Smoothing and differentiation of data by simplified least square procedure. *Anal. Chem.*, 44(11):1906–1909, 1972.

[27] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[28] U.S. Environmental Protection Agency, Office of Pollution Prevention and Toxics: Washington, DC. Exposure assessment tools and models, estimation program interface (epi) suite, version 3.12;. http://www.epa.gov/oppt/ exposure/pubs/episuitedl.htm., 2005.

[29] EU-Commission. Commission decision ec 2002/657 of 12 August 2002 implementing council directive 96/23/ec concerning the performance of analytical methods and the interpretation of results. *Off. J. Eur. Communities L*, 221, 2002.