Accepted Manuscript

Saer Samanipour, Malcolm J. Reid, Kevin V. Thomas. 2017.
Statistical variable selection: an alternative prioritization strategy during the nontarget analysis of LC-HR-MS data.
Analytical Chemistry . 89(10): 5585-5591. ISSN 0003-2700.

Statistical variable selection: An alternative prioritization strategy during the non-target analysis of LC-HR-MS data

Saer Samanipour,*,† Malcolm J. Reid,† and Kevin V. Thomas†,‡

†*Norwegian Institute for Water Research (NIVA), 0349 Oslo, Norway*

‡*Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 39 Kessels Road, Coopers Plains QLD 4108, Australia*

E-mail: saer.samanipour@niva.no

Abstract

Liquid chromatography coupled to high resolution mass spectrometry (LC-HR-MS) has been one of the main analytical tools for the analysis of small polar organic pollutants in the environment. LC-HR-MS typically produces a large amount of data for a single chromatogram. The analyst is therefore required to perform prioritization prior to non-target structural elucidation. In the present study we have combined the F-ratio statistical variable selection and the apex detection algorithms in order to perform prioritization in data sets produced via LC-HR-MS. The approach was validated through the use of semi-synthetic data, which was a combination of real environmental data and the artificially added signal of 31 alkanes in that sample. We evaluated the performance of this method as a function of four false detection probabilities namely: 0.01, 0.02, 0.05, and 0.1%. We generated 100 different semi-synthetic data sets for each F-ratio and evaluated that data set using this method. This design of experiment created a population of 30,000 true positives and 32,000

true negatives for each F-ratio, which was considered sufficiently large enough in order to fully validate this method for analysis of LC-HR-MS data. The effect of both the F-ratio and signal to noise ratio (S/N) on the performance of the suggested approach were evaluated through normalized statistical tests. We also compared this method to the pixel-by-pixel as well as peak list approaches. More than 92% of features present in the final feature list via F-ratio method were also present in conventional peak list generated by MZmine. However, this method was the only approach successful in classification of samples, thus prioritization, when compared to the other evaluated approaches. The application potential and limitations of the suggested method discussed.

Introduction

A large number of small polar organic pollutants are considered as chemicals of emerging concern (CECs) due to their fate and behavior in the environment (as reviewed by Klečka et al.[1] and La Farre et al.[2]). Liquid chromatography coupled to the high resolution mass spectrometry (LC-HR-MS) has become the leading analytical instrumentation for analysis of these pollutants in different environmental compartments.[3–5] Measuring these pollutants in the environment takes place through three different and/or complementary approaches, namely target analysis, suspect analysis, and non-target analysis.[4,6–8] For target analysis the analyst has all the necessary information, including the retention time and the spectral information, for confident identification of a target analyte in complex environmental samples.[9] On the other hand for the suspect screening only limited information is available while during non-target analysis the analyst does not have any prior information regarding the identity of the analytes in the sample.[7,9] Even though non-target analysis is the most difficult and the least certain of the three mentioned approaches, this method is essential for the discovery of new CECs in the environment.[4,6,7]

Confident identification of pollutants based only on the data generated via non-target

analysis on LC-HR-MS, is a challenging task due to the volume and complexity of the data .[10,11] During the non-target analysis, each sample may produce thousands of features, where each includes a measured exact mass, intensity, and the retention time.[12,13] Therefore, the analyst may have to prioritize among the features for structural elucidation. For LC-HR-MS data, there have been different approaches used for prioritization during the non-target analysis. The simplest approach applies the absolute intesity and the detection frequency as the main criteria for prioritization.[9,12] However, high signal intensity and the detection frequency does not guaranty environmental relevance. Another approach utilizes either toxicity information (through effect-directed analysis[14]) or the elemental composition (through, for example, filters for halogenated compounds[12]). However, the mentioned approaches are complicated and may be biased towards a certain family of compounds, for example halogenated ones. A less used approach, particularly in the field of environmental analysis, has been the application of unsupervised and/or supervised statistical methods, such as principal component analysis (PCA) and partial least square discrimination analysis (PLS-DA) for prioritization of the relevant features.[13,15] These statistical methods perform well when used in metabolomics due to a more clear change in the sample composition. However, these same methods may suffer when there is a high level of redundancy/similarity in the analyzed samples.[16,17] Recent studies have shown the superior performance of the supervised F-ratio method combined with PCA for analysis of the data via gas chromatography coupled to low resolution mass spectrometry (GC-MS) of complex samples.[16–20] However, the F-ratio method has never been used/optimized for the non-target analysis of the data recorded via LC-HR-MS and particularly for complex environmental samples.

The mentioned statistical variable selection approaches can be applied to either a peak list[11,13,15] or the whole chromatogram.[16–24] Even though processing of the peak list is faster than the whole chromatogram due to its smaller size compared to the whole chromatogram, the raw chromatogram must go through preprocessing steps such as signal deconvolution,

peak finding, peak picking, and peak integration in order to generate a final peak list useful to a prioritization method. All these preprocessing steps are prone to error when dealing with highly complex samples.[22–25] The application of the statistical variable selection approaches to the whole chromatogram has been shown to result in reliable models and therefore, reliable prioritization.[16–25]

The aim of this study is to adapt, comprehensively validate, and test the applicability of the F-ratio method for the non-target analysis of LC-HR-MS chromatograms of complex environmental samples. The F-ratio was applied to the whole chromatogram in order to minimize the data manipulation and produce a reliable statistical model. We combined the F-ratio method with the apex detection as well as adduct and isotope removal algorithms, in order to adapt this method to be used for non-target analysis of LC-HR-MS data. We comprehensively validated this method using a semi-synthetic data set, which consisted of the background signal generated from the real environmental samples with the addition of the signal of 31 alkanes randomly distributed as true positives and true negatives, and noise. This data set was evaluated 400 times where the random selection of the alkanes and the background signal caused generation of a completely different sample for each evaluation. Finally, the chromatograms of 15 sludge extracts from three different locations in Norway and three blanks were analyzed using the F-ratio method as well as conventional peak picking algorithms. We also applied the F-ratio method to the peak list and compared this feature list to the one produced via using the whole chromatogram. The feature lists via F-ratio were compared to the peak lists generated by a conventional peak pick method, in order to further evaluate and/or validate the applicability of this method for non-target analysis of LC-HR-MS data.

Experimental Methods

Environmental Sampling and Sample Preparation

15 sludge samples from three different wastewater treatment plants (WWTP) in Norway were collected (i.e. five replicates for each WWTP), during the spring of 2015. These WWTPs were located at Oslo, Hamar, and Gjøvik. The plants in Oslo and Hamar were equipped with a three stage treatment process, including physical, chemical, and biological treatment whereas the plant in Gjøvik has only the physical and chemical treatments. More details on the chemicals, the suppliers, and the sample preparation steps are provided in section S1 of Supporting Information.

Instrumental Conditions and Analysis

All the extracts were analyzed employing, Waters Acquity UPLC system (Waters Milford, MA, USA). An Acquity UPLC HSS C18 column (2.1 × 150 mm, particle size 1.8 mm) (Waters, Milford, MA, USA) was used for all the separations. A mixture of solvent A, 5 mM ammonium formate at pH 3.0 and solvent B, acetonitrile with 0.1% formic acid at a constant flow rate of 0.4 ml min^{-1} was used for the chromatographic separations. The gradient varied from 87% of solvent A to 5% of solvent A. More details regarding this method are provided elsewhere.[26] Both the analytical column and the column guard were kept as 50 °C during the separations.

Xevo G2-S Q-TOF-MS (Waters Milford, MA, US) was used for analysis of all 18 samples, including the 15 sludge extracts and 3 blanks. The MS^1, with a collision energy of 6 eV, and the MS^2, with a collision energy ramp between 15 to 45 eV were simultaneously recorded during the whole chromatogram. We employed a mass range of between 90 Da and 700 Da with a sampling frequency of 1.8 Hz. More information regarding the mass spectrometer conditions is available elsewhere.[26]

Data Processing and Workflow

All the chromatograms were recorded in profile mode employing MassLynx (Waters Milford, MA, US). The chromatograms were then exported as netCDF files, using DataBridge package (Waters Milford, MA, US) incorporated in the MassLynx software. The MS^1 channel (i.e. low collision energy channel 6 eV) chromatograms were used for the data analysis. All the exported chromatograms were then imported into Matlab.[27] These files were processed using the following sequnece of steps in the stated order including the binning, retention alignment, data matrix generation, F-ratio calculation, null-distribution validation, zero mask application, chromatogram folding, apex detection, and finally adducts and isotope removal. All the steps taken in this workflow are explained in detail below and section S2 of the Supporting Information.

We binned the exported chromatograms using a bin thickness of 10 mDa, which was based on the observed mass accuracy of $\pm$ 5 mDa in our data set (section S2.1). The mass accuracy was defined based on the shift observed in the measured mass of the calibrant injected every 20 s into the source. The binned chromatograms were then retention aligned with a home-developed algorithm inspired by the piecewise method previously developed and validated by Synovec group.[25,28] We added an additional mass spectral correlation control in order to increase the accuracy of the retention alignment. More details regarding both binning and retention alignment processes are provided in Supporting Information section S2. The retention aligned chromatograms were then unfolded to create a long vector of intensities for every single measured m/z value. These vectors were then stacked on top of each other in order to produce a large matrix which was used for the statistical prioritization. Every row in this matrix was a sample while every column was an independent variable. The F-ratio was calculated for each variable,[16] or column of the matrix, based on *a priori* knowledge of the sample classification (section S2.3). An F-ratio threshold was calculated using the probability distribution generated via null-distribution analysis.[19] This procedure

aims to minimize the number of false positive detections as well as the method validation during the analysis (see section S2.4). The variables that had an F-ratio smaller than the defined threshold, based on the null-distribution, were set to zero in the data matrix. This process was referred to as the zero mask application. Each zero mask applied chromatogram then was folded back into a matrix where a row was one scan and a column was the signal for a m/z value. We performed apex detection in the folded chromatograms (see section S2.5 of Supporting Information). The apex detection groups the non-zero and statistically meaningful variables which can be represented as a feature in the chromatogram. For example all the non-zero variables in a chromatographic peak can be grouped and represented via only one pair of retention time and m/z value, thus a feature. Therefore, the apex detection generates a list of unique retention time and m/z value pairs for each sample. This differs from conventional peak picking algorithms in that apex detection does not perform signal modeling and/or integration therefore minimizes the signal manipulation. Finally, the adducts and the isotopes were removed from this list in order to create the final unique feature list for each chromatogram. This workflow provides the necessary initial information for discovery-based non-target analysis of complex samples analyzed via LC-HR-MS.

We also performed F-ratio analysis on the peak list produced by conventional peak picking algorithm, MZmine 2[29](explained in detail below). The peak list was retention-aligned using a home-developed method using a mass window of 2 mDa and a retention window of 2 S. The retention aligned peak tables were used for F-ratio and null-distribution calculations. The peaks in the peak list with an F-ratio larger than the threshold were kept in order to produce the feature list. The feature list, finally, was processed for adduct and isotope removal in order to generate the final feature list.

Data pretreatment

During the validation process of the F-ratio method (i.e. analysis of the semi-synthetic data), we did not employ any data pre-treatment methods such as mean-centering, standardization, and normalization. This choice enabled us to comprehensively evaluate the effect of introduced noise on the performance of the F-ratio method. For the environmental sample analysis, we tested different data pre-treatment methods such as mean-centering, standardization, and normalization before processing the data set with F-ratio method. However, these pre-treatments did not affect the final unique feature list for the analyzed data set. Therefore, we decided to work with the raw data and avoid performing any type of pre-treatment.

Computations

All the mentioned data processing steps were performed via Matlab, employing a Windows 7 Professional version (Microsoft Inc, USA) workstation computer with 12 CPUs and 128 GB of memory.

MZmine Peak Picking

The conventional peak list for each chromatogram was generated using MZmine 2.[29] The peak picking was performed by mass detection followed by GridMass 2D peak detection. A five scan window was selected for the smoothing of the chromatogram in the time dimension and a 10 mD window was used in the mass dimension. A minimum signal of 300 counts was required for a peak to be considered as a meaningful peak. These parameters were optimized based on the observed mass accuracy and the peak widths in both time and mass domains. These parameter settings resulted in feature numbers varying between 7,500 for blanks and 12,500 for the samples from Oslo WWTP.

Principal Component Analysis (PCA)

We employed principal component analysis (PCA)[30] for classification/separation of the sample groups. We performed PCA on the peak list generated via MZmine, variable selected peak list (i.e. the F-ratio applied to the peak list generated via MZmine) with F-ratio method, the whole chromatogram (i.e. pixel-by-pixel), and the variable selected chromatogram employing F-ratio method. The PCA was performed on the mean centered data utilizing the singular value decomposition algorithm.[30]

Results and Discussion

We validated the F-ratio method for data generated via LC-HR-MS, employing both semi-synthetic data and the real environmental data. The use of semi-synthetic data enabled us to perform a large number of evaluations (i.e. total number of detection cases 62,000×4) knowing exactly the added signal, noise, and relative intensity of the added signal which, translated into comprehensive validation of the proposed method. This would not have been possible using spiked samples due to the limitation on the number of standards and injections as well as the potential interference between the sample and the standard mixture. This study is the first implication of this method for the data generated via LC-HR-MS as well as adaptation of this method in order for it to be included in non-target identification workflows. This method enables the direct prioritization of the unique features, which are the main cause of the separation of different sample groups. Therefore, the identification efforts can be focused on the prioritized unique features.

Validation via Semi-synthetic Data

We employed a semi-synthetic data set, which consisted of a combination of real environmental data and synthetic signal, for comprehensive validation of the F-ratio method. The signal of 31 alkanes (i.e. the neutral monoisotopic masses, Table S1) was added at different

concentrations to a background signal, which came from the real environmental data. During each analysis, these 31 alkanes were divided in two randomly selected groups where the first group of 15 alkanes was added to the background signal at concentration levels that were statistically meaningful. Therefore, for these 15 alkanes the resulting F-ratios were larger than the threshold. For the second group, 16 alkanes were added to the background at a statistically constant concentration. Four different F-ratios of 208, 30, 28, and 13 having false positive detection probabilities of 0.01, 0.02, 0.05, and 0.1 %, respectively, were evaluated. Each F-ratio value was evaluated 100 times with different: background signal, combination of alkanes, concentration levels, and retention times of true positives (i.e. 15 alkanes) and true negatives, thus a total of 400 evaluations. The generation of the these semi-synthetic data is described in detail in Supporting Information, section S3. Alkanes were selected for our analysis because these compounds are not ionized by ESI source therefore we were sure that these compounds were not present in the real background signal. This design of experiment created a set of 15 true positives, 16 true negatives, and a different background signal during each evaluation, which enabled us to comprehensively examine the capabilities and limitations of the F-ratio method. The number of evaluation (i.e. 100 for each F-ratio) was selected based on our preliminary assessment, that showed that 100 analysis for each F-ratio would generate a large enough population of true positives (TP) 30,000 and true negatives (TN) 32,000 for that F-ratio, in order to fully validate this method. To compare the effect of different F-ratio probability value on the final results, we employed normalized statistical parameters such as rate of false positive, rate of false negative, sensitivity, specificity, and accuracy.[31]

Increased F-ratios resulted in a smaller number of false positives and a larger number of false negatives. The number of false positive detection ranged from 2,518 cases for the F-ratio of 208 having a probability of 0.01% to 9,525 cases for the F-ratio of 13 with a probability of 0.1%, Table 1 and Figure S7. The largest number of false negative detections of 2204 was

observed for an F-ratio of 208 whereas the smallest number of false negative detections of 1,404 was caused by an F-ratio of 13, Table 1. These trends were due to the fact that the selection of a large F-ratio value (i.e. more strict selection criterion) lowers the probability of false positive detection while increasing the probability of false negative detections. The observed changes in F-ratio method performance were better projected through normalized statistical parameters such as rate of false positive detection, rate of false negative detection, sensitivity, specificity, and accuracy, Table 1. For example, the drop in the specificity and accuracy observed for F-ratio of 13 (probability of 0.1%) showed the inadequacy of this F-ratio for the analyzed data set (Figure S7). This drop also indicated that this F-ratio may cause a large number of false positive detections when analyzing this data set. Therefore, the analyst is required to find an optimized F-ratio value in order to minimize the number of potential false positive detection while limiting the number of false negatives. Among the four F-ratios evaluated, the value of 28 (probability of 0.05%) showed to be the optimized one, considering that this value provided the largest accuracy level, second largest sensitivity level while maintaining a high level of specificity (Figure S7).

Further evaluation of our data set, showed that for F-ratios ≥ 28 more than 70% of the false positive cases were coming from the background signal rather than the true negatives (i.e. added signal of alkanes at constant concentrations). The observed trend was caused by the high level of variability artificially introduced into the background signal during the background generation. For F-ratio of 13, around 50% of the false positives were true negatives. In this case even though the signal of true negatives did not have a large level of variability between sample groups, once added to the background, the variability in that signal increased due to the inherent large variance in the background signal. Therefore, these true negatives produced a large enough F-ratio, which met the F-ratio threshold and were selected as positive detections. When looking at the false negative cases, for all four F-ratios, the main causes of false negative detection were the large variability in the back-

ground signal and the S/N threshold setting during the apex detection. Also in this case, the random variability introduced into the true positives signal was not large enough to overcome the variability present in the background signal. We tested these hypothesis by increasing the initial concentration of the added signal of the true positives from 5% to 15% and also increasing the concentration factor from 2-8 to 2-20 (see section S3 of SI for more details). With an F-ratio of 28, increasing these parameters reduced drastically the number of false positive detection from 2,864 to 253 cases as well as the number of false negatives from 1,570 to 35 cases after 100 simulations. These results indicated that the combination of low level concentration of added alkanes, their low between sample group variability, and finally the large level of variability introduced into the background signal have an important effect on the performance of this algorithm.

Mean centering and standardization (i.e. division by the square root of standard deviation of each variable) with an F-ratio of 28, added signal of 5%, and the concentration factor of 2-8 reduced the number of both false positive detection and false negative detection from 2,864 to 350 cases and from 1,570 to 97 cases, respectively after 100 simulations. These pre-treatments' approaches both decrease the noise levels in the data set while emphasizing the underlying trend.[30] This implies that these data pre-treatments reduced the effect of artificially introduced noise in the data set while emphasizing the between group variability, thus a decrease in the number of false positive and false negative detection. The type of data pre-treatments employed prior to the F-ratio analysis is data set and objective dependent.[30] Therefore, the analyst is required to optimize these data pre-treatments approaches in advance in order to be able to produce reliable results. Further investigation on the effect of these parameters on the F-ratio method are needed and will be subject of our future studies.

Table 1: The number of false positive detections, number of false negative detections, rate of false positive, rate of false negative, sensitivity, specificity, and accuracy parameters calculated for four different F-ratio values based on 100 evaluations for each F-ratio probability value.

	F-ratio values (probability of false positive detection %)			
Parameter	208 (0.01)	30 (0.02)	28 (0.05)	13 (0.10)
False positive detection[a] (FP)	2,518	3,172	2,864	9,525
False negative detection[a] (FN)	2,204	2,220	1,570	1,404
Rate of false positive detection[b] (%)	7.3	9.0	8.0	23.0
Rate of false negative detection[c] (%)	6.8	6.9	5.0	4.5
Sensitivity[d] (%)	93.2	93.1	95.0	95.5
Specificity[e] (%)	92.7	91.0	91.8	77.1
Accuracy[f] (%)	92.9	92.0	93.3	85.0

[a]This parameter represents the number false positive detection out of total number of detections of 62,000, including 30,000 true positives (TP) and 32,000 true negatives (TN); [b]The rate of false positive[31] was calculated as FP/(FP+TN); [c]The rate of false negative[31] was calculated as: FN/(TP+FN); [d]The sensitivity[31] values were calculated using: TP/(TP+FN); [e]The specificity[31] values were calculated with: TN/(TN+FP); [f]The accuracy[31] values were calculated employing: (TP+TN)/(TP+FP+FN+TN).

The effect of S/N on F-ratio algorithm

The S/N is an important parameter, which affects the performance of the F-ratio algorithm particularly during the apex detection. The apex detection step aims to reduce the level of redundancy in the data set by grouping variables, that can be represented by a unique one (see section S2.5 for more detailed information). We evaluated the effect of S/N on the results of the algorithm with an F-ratio of 28. This evaluation was performed by varying the S/N from 1 to 10 (i.e. 1, 3, and 10) and performing 20 analysis for each S/N value. The F-ratio of 28 was selected based on the fact that it appeared to be the optimized F-ratio for the evaluated data set.

We observed a slight decrease in the number of false positive detection as a function of increase in the S/N while the increase in the S/N had a positive effect on the number of detected false negatives, Table 2. However, the changes in the S/N did not appear to cause a large variation in the normalized statistical parameters such as rate of false positive, rate

of false negative, sensitivity, specificity, and accuracy.[31] This suggested that the S/N ratio has a less relevant effect on the performance of this method compared to the F-ratio value. However, these results may be case dependent, therefore optimization of this parameter based on the data set should be considered by the analyst.

Table 2: The number of false positive detections, number of false negative detections, rate of false positive, rate of false negative, sensitivity, specificity, and accuracy parameters calculated for four different S/N values, having an F-ratio of 28, based on 20 simulations for each S/N.

	S/N values		
Parameter	1	3	10
False positive detection[a] (FP)	654	629	583
False negative detection[a] (FN)	122	151	166
Rate of false positive detection[b] (%)	9.0	9.0	8.0
Rate of false negative detection[c] (%)	2.0	2.5	2.7
Sensitivity[e] (%)	98.0	97.5	97.3
Specificity[f] (%)	90.7	91.1	91.7
Accuracy[g] (%)	94.1	94.1	94.3

[a]This parameter represents the number false positive detection out of total number of detections of 12,400, including 6,000 true positives (TP) and 6,400 true negatives (TN); [b]The rate of false positive[31] was calculated as FP/(FP+TN); [c]The rate of false negative[31] was calculated as: FN/(TP+FN); [e]The sensitivity[31] values were calculated using: TP/(TP+FN); [f]The specificity[31] values were calculated with: TN/(TN+FP); [g]The accuracy[31] values were calculated employing: (TP+TN)/(TP+FP+FN+TN).

Comparison between the unique feature list and the conventional peak list

Once the F-ratio method was validated via semi-synthetic data, we processed the chromatograms of the 15 sludge samples plus 3 method blanks using this algorithm. The same data set was also processed via MZmine, employing previously optimized parameters. The F-ratio method produced a list of unique features for each sample whereas MZmine created a conventional peak list for the same samples. We compared the unique feature lists produced via F-ratio method to the conventional peak lists by MZmine as well as the unique feature

lists produced via application of F-ratio method to both the whole chromatogram and the peak list by MZmine. These comparisons enabled us to further evaluate/validate the F-ratio method for analysis of the data generated via LC-HR-MS.

More than 92% of the unique features via F-ratio method were also present in the conventional peak list via MZmine. For example, for one of the Oslo samples after the adducts and isotopes removal 109 out of total 112 (i.e. 97%) unique features were also present in the peak list of the same sample generated by MZmine. The number of features, via F-ratio method, before adducts and isotope removal ranged from 403 features for one of the blank samples to 127 for the Oslo sample whereas after the adducts and isotope removal the unique features numbers ranged between 302 for the blank sample and 112 for the Oslo sample. For the conventional peak list, we observed around 7500 peaks for the blank whereas this number was around 12500 for the sludge samples. When comparing the unique feature list to the conventional peak list, the number of discrepancy cases varied between 3 cases for Oslo sample and 23 cases for the blank samples. A discrepancy case is defined as a unique feature detected via F-ratio that is not present in the conventional peak list. All the discrepancy cases were classified in two categories, for ease of explanation. The first category and the most dominant one, particularly in the blank samples belonged to unique features, which appeared to be noise rather than analytical signal. Considering the large number of variables evaluated occurrence of a certain number of false positives was likely. The second category was mainly caused by the fact that MZmine performs peak modeling during the peak picking and uses the modeled apex for estimation of both m/z value and the retention time. Using this approach, this algorithm may group shoulders of a peak with the main peak. The F-ratio method however treats the shoulders as independent variables and thus potential unique features. This category of discrepancy cases may also be caused by the resolution of our instrument of 35,000. All considered, the F-ratio combined with the apex detection algorithm method showed to have a large number (i.e. $\leq$ 92%) of common

unique features with the conventional peak picking approach which is an indication of its robustness. Furthermore, these results imply that this method can be implemented in the non-target workflows for structural elucidation.

The F-ratio applied to the whole chromatograms of the environmental samples resulted in 250 unique feature in average while producing 3 unique features in average when applied to the peak list generated via MZmine. A large number of the observed discrepancy cases were due to the signal deconvolution, which was caused by the complexity of the analyzed samples. The unsuccessful signal deconvolution was directly translated into the large within group variability in the area of the integrated peaks, thus their lack of detection. The second group of discrepancies was due to the peak modeling algorithm in MZmine, which failed to detect the shoulder of a peak in the m/z domain as a separate peak, therefore their absence from the unique feature list. It should be noted that the mentioned sources of failure in the F-ratio applied to the peak list may be case dependent and may vary from dataset to dataset. Further investigation of the potential sources of discrepancy between the F-ratio variable selection applied to the whole chromatogram vs the peak list are needed.

The F-ratio method appeared to be able to successfully separate the sample groups while both peak list and pixel-by-pixel methods failed in carrying out this task, Figure 1. Multivariate statistical methods such as principal component analysis (PCA) when dealing with large, complex datasets with a large level of noise and redundancy may fail to classify the samples in logical groups. Consequently, univariate methods such as F-ratio are used prior to these tests in order to reduce the redundancy in the data set. Therefore, a clear and logical separation of the samples in the score plots is a crucial indication of a successful prioritization/variable selection. We performed PCA on zero mask applied chromatograms following the variable selection, the retention aligned peak list via MZmine, and the whole chromatogram (i.e. pixel-by-pixel analysis). In the case of the sludge samples the inherent

complexity of the background signal was translated into inability of both peak list based and pixel-by-pixel based methods to separate these sample groups from each other properly. The F-ratio method, on the other hand, was able to perform separation of the sample groups because this method retains the variables that are causing the clustering of samples within a particular group. We also performed the F-ratio variable selection on the peak list generated via MZmine. In this case also the PCA was not able to separate the sample groups from each other, Figure 1. Therefore, it was not possible to perform a prioritization based on the peak list using the F-ratio method. Despite the mentioned complexity, the F-ratio method was able to separate the sample groups from each other, thus performing successful prioritization. These results also indicate the applicability of F-ratio method within the structure elucidation workflows during non-target analysis of complex samples analyzed via LC-HR-MS.

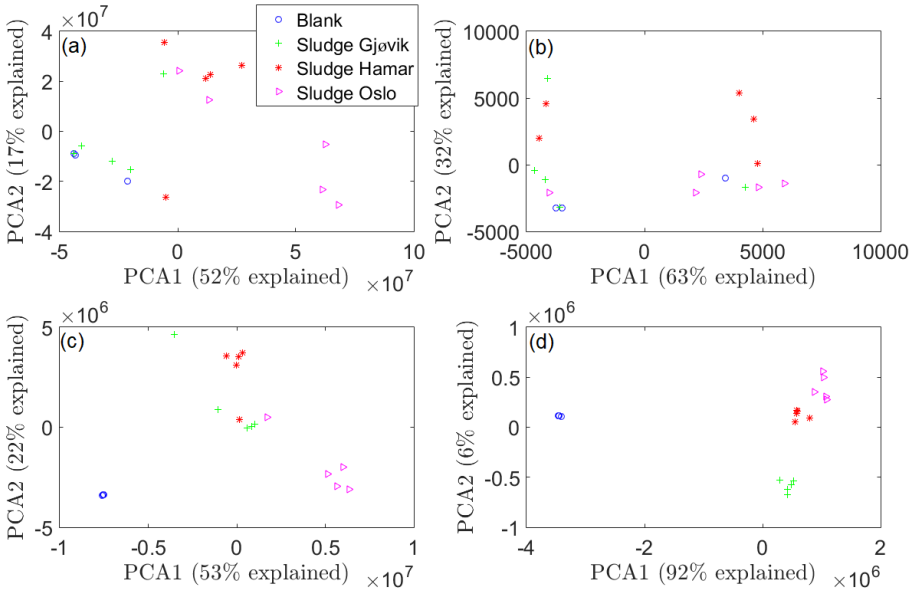


Figure 1: Figure depicting the PCA score plots of (a) peak list based classification, (b) peak list based after F-ratio variable selection (c) pixel-by-pixel analysis, and (d) the F-ratio method during the prioritization for non-target analysis of the 15 sludge samples plus 3 blanks.

Potential and Limitations

The F-ratio method combined with the apex detection showed to be a robust and reliable approach for prioritization of the unique features that are relevant to the sample classification. This method was effective at prioritization even for cases where the other conventional methods may fail due to the complexity of the analyzed data set, Figures 2 and 1. This method minimizes the data manipulations such as peak picking and/or modeling and at the same time results in a list of unique features which can be used for structure elucidation. The F-ratio method reduces the redundancy in the data set and detects the relevant variables in the data set enabling the analyst to focus on the identification of only the unique and relevant features. This method also has the advantage of being less dependent on the absolute intensity of the each chemical signal in the sample compared to the conventional prioritization methods. In other words, as long as a chemical signal causes large enough variability between the sample groups, independently from its absolute intensity, it will be detected as a unique relevant feature (Figure 2). Additionally, this method can be used for a battery of discovery-based non-targeted applications as long as there are replicates present. Furthermore, by changing the initial hypothesis, one can interrogate the data set in a completely different way. For example in case of the sludge samples in this study, by assuming that all the sludge samples belonged to one group and the blanks to another group, we could have selected the unique features which are in common in all the sludge samples and simultaneously subtracted the blanks from our samples.

There are also some limitations to application of F-ratio method for non-target analysis of LC-HR-MS data. This method is computationally expensive due to the large data sets produced when employing LC-HR-MS. For example, each sludge sample chromatogram in this study produced around 180 million variables (Figure 2), which requires a large computational power in order to be done in a timely manner. Moreover, this method has to be complemented with target and suspect analysis using the conventional methods for ubiq-

uitous chemicals or pollutants where measurable concentrations are more uniform. These pollutants would not be detected as unique features with statistically significant differences between sample groups. Also the use of data pre-treatment should be evaluated by the analyst on a case study base.

Considering capabilities and the limitations of the F-ratio method, this approach has a great potential to be applied to the LC-HR-MS non-target discovery-based analysis. The application of this method as well as its combination with the structural elucidation workflows are going to be subject of our future studies.

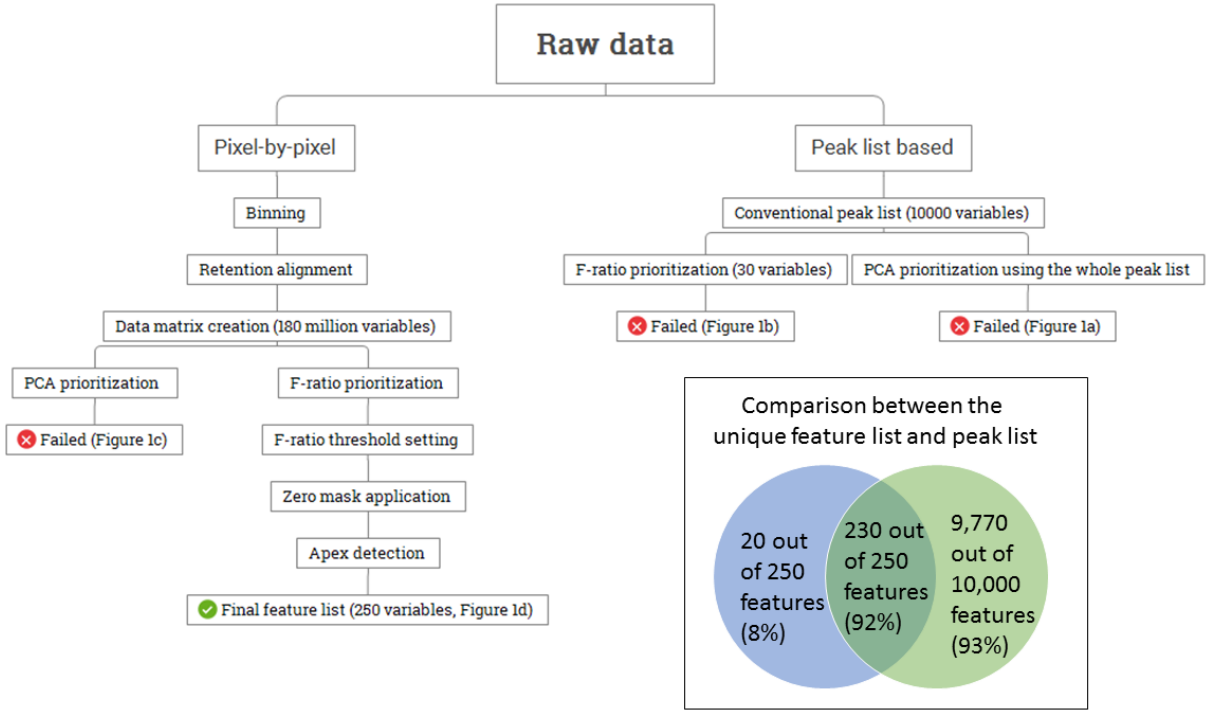


Figure 2: Figure depicting an overview of the F-ratio method vs the conventional methods as well as the venn diagrams of the comparison between the unique feature list and the conventional peak list generated via MZmine.

Associated Content

Acknowledgement

The authors are thankful to Prof. Bert van Bavel, Dr. Merete Grung and Dr. Jose A. Baz-Lomba for their editorial input. We are also grateful to the Research Council of Norway for the financial support of this project (RESOLVE, 243720).

Supporting Information

The Supporting Information including details regarding the sample preparation, analysis, sateps taken during the data processing, and semi-synthetic data generation is available free of charge on the ACS Publications website.

Author Information

Corresponding Author:

Saer Samanipour

E-mail: saer.samanipour@niva.no

Phone: +47 98 222 087

Address: Norwegian Institute for Water Research (NIVA)

0349 Oslo, Norway

References

(1) Klečka, G.; Persoon, C.; Currie, R. *Reviews of Environmental Contamination and Toxicology Volume 207*; Springer, 2010; pp 1–93.

(2) La Farre, M.; Pérez, S.; Kantiani, L.; Barceló, D. *TrAC Trends Anal. Chem.* **2008**, *27*, 991–1007.

(3) Giger, W. *Anal. Bioanal. Chem.* **2009**, *393*, 37.

(4) Krauss, M.; Singer, H.; Hollender, J. *Anal. Bioanal. Chem.* **2010**, *397*, 943–951.

(5) Richardson, S. D. *Anal. Chem.* **2009**, *81*, 4645–4677.

(6) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48*, 1811–1818.

(7) Gago-Ferrero, P.; Schymanski, E. L.; Bletsou, A. A.; Aalizadeh, R.; Hollender, J.; Thomaidis, N. S. *Environ. Sci. Technol.* **2015**, *49*, 12333–12341.

(8) Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V. *J. Chromatogra. A* **2016**, *1463*, 153–161.

(9) others,, et al. *Anal. Bioanal. Chem.* **2015**, *407*, 6237–6255.

(10) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *Trends Anal. Chem.* **2016**, *82*, 425–442.

(11) Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. *Anal. Chem. acta* **2016**, *914*, 17–34.

(12) Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J. *Anal. Bioanal. Chem.* **2014**, *406*, 7323–7335.

(13) Schollee, J. E.; Schymanski, E. L.; Avak, S. E.; Loos, M.; Hollender, J. *Anal. Chem.* **2015**, *87*, 12121–12129.

(14) Thomas, K. V.; Langford, K.; Petersen, K.; Smith, A. J.; Tollefsen, K. E. *Environ. Sci. Technol.* **2009**, *43*, 8066–8071.

(15) Kalogiouri, N. P.; Alygizakis, N. A.; Aalizadeh, R.; Thomaidis, N. S. *Anal. and Bioanal. Chem.* **2016**, *408*, 7955–7970.

(16) Pierce, K. M.; Hoggard, J. C.; Hope, J. L.; Rainey, P. M.; Hoofnagle, A. N.; Jack, R. M.; Wright, B. W.; Synovec, R. E. *Anal. Chem.* **2006**, *78*, 5068–5075.

(17) Beckstrom, A. C.; Humston, E. M.; Snyder, L. R.; Synovec, R. E.; Juul, S. E. *J. Chromatogra. A* **2011**, *1218*, 1899–1906.

(18) Christensen, J. H.; Tomasi, G. *J. Chromatogr. A* **2007**, *1169*, 1–22.

(19) Parsons, B. A.; Marney, L. C.; Siegler, W. C.; Hoggard, J. C.; Wright, B. W.; Synovec, R. E. *Analytical chemistry* **2015**, *87*, 3812–3819.

(20) Parsons, B. A.; Pinkerton, D. K.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2016**, *1440*, 179–190.

(21) Sinkov, N. A.; Sandercock, P. M. L.; Harynuk, J. J. *Forensic Sci. Int.* **2014**, *235*, 24–31.

(22) Sinkov, N. A.; Harynuk, J. J. *Talanta* **2011**, *83*, 1079–1087.

(23) Sinkov, N. A.; Harynuk, J. J. *Talanta* **2013**, *103*, 252–259.

(24) Adutwum, L.; Harynuk, J. *Anal. Chem.* **2014**, *86*, 7726–7733.

(25) Watson, N. E.; VanWingerden, M. M.; Pierce, K. M.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2006**, *1129*, 111–118.

(26) Baz-Lomba, J. A.; Reid, M. J.; Thomas, K. V. *Anal. Chem. acta* **2016**, *914*, 81–90.

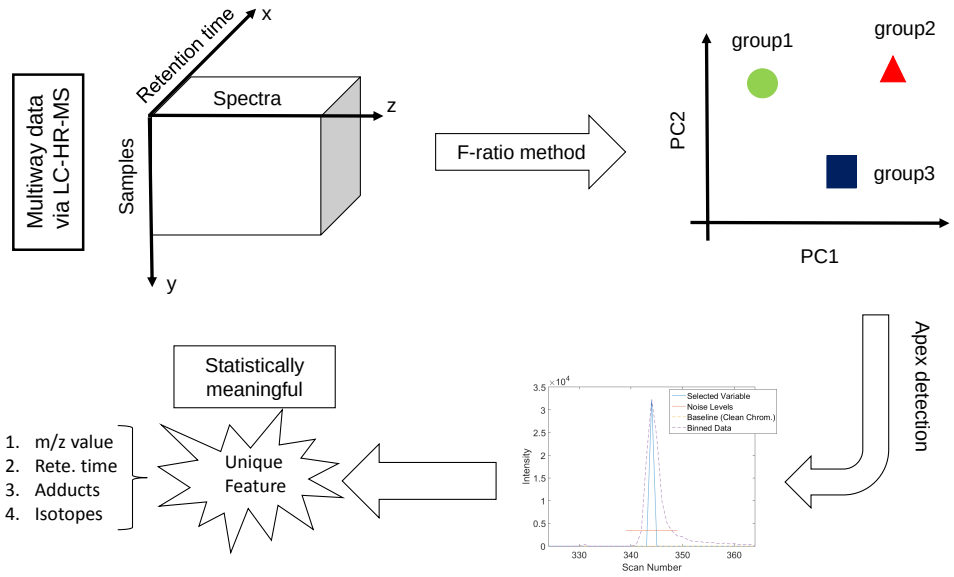
(27) MATLAB version 9.1 Natick, Massachusetts: The MathWorks Inc.,

(28) Nadeau, J. S.; Wright, B. W.; Synovec, R. E. *Talanta* **2010**, *81*, 120–128.

(29) Katajamaa, M.; Miettinen, J.; Orešič, M. *Bioinformatics* **2006**, *22*, 634–636.

(30) Brereton, R. G. *Applied chemometrics for scientists*; John Wiley & Sons, 2007.

(31) Burke, D. S.; Brundage, J. F.; Redfield, R. R.; Damato, J. J.; Schable, C. A.; Putman, P.; Visintine, R.; Kim, H. I. *N. Engl. J. Med.* **1988**, *319*, 961–964.


Multiway data via LC-HR-MS
Retention time
x
Spectra
z
Samples
y
F-ratio method
group1
group2
group3
PC2
PC1
Apex detection
Statistically meaningful
1. m/z value
2. Rete. time
3. Adducts
4. Isotopes
Unique Feature
Selected Variable
Noise Levels
Baseline (Clean Chrom.)
Binned Data
Intensity
Scan Number

for TOC Only