HaloSeeker v1.0, a user-friendly software application for screening halogenated chemicals from untargeted high resolution mass spectrometry data

Cariou R[1], Léon A[1,2], Hutinet S[1], Guitton Y[1], Hurel J[1], Pourchet-Gellez M[1], Antignac JP[1], Munschy C[2], Tixier C[2], Dervilly-Pinel G[1], Le Bizec B[1]

[1]LABERCA, Oniris, INRA, Université Bretagne-Loire, F-44307, Nantes, France, contact.haloseeker@oniris-nantes.fr; [2]Ifremer, Laboratoire Biogéochimie des Contaminants Organiques, F-44311, Nantes, France

Introduction

Monitoring chemical contaminants in environmental abiotic compartments and along trophic webs, including in human, is a major needed input for supporting risk assessment. Highly specific and sensitive targeted analytical workflows involving chromatography coupled to mass spectrometry are available to monitor a wide range of known chemical hazards, including Persistent Organic Pollutants (POPs) listed in the Stockholm Convention and alternatives emerging substances. However, such strategies only allow characterizing a limited number of compounds of interest, consequently describing a truncated contamination picture. Indeed, many other chemicals, not yet (fully) described may participate to the global contamination profile and be considered as potential emerging hazards. Thus, these substances may deserve more consideration in order to depict an exhaustive overview of the overall chemical exposome.

Recent advances in high resolution mass spectrometry (HRMS) today opens the way to full scan untargeted profiling approaches, which are powerful for screening substances which are not yet known or not yet included in existing targeted methods. However, the challenge shifted to post-acquisition data interpretation since dedicated advanced bioinformatics tools are required to extract rapidly and efficiently the signals of interest from such huge datasets. Among the possible ways to perform this extraction, a chemistry based approach may exploit the fact that a common feature to all POPs is that they are polyhalogenated (F, Cl, Br). In particular, Cl- and Br-containing substances may be screened on the basis of two properties:

(i) the discriminating mass defect engendered by heteroatoms. H/Cl-scale mass defect (MD) plots[1], derived from Kendrick plots[2], are an elegant way to visualize halogenated ions in complex HRMS fingerprints.

(ii) the highly specific isotopic patterns arising from the occurrence of naturally abundant and stable isotopes of Cl and Br. Automated filtering algorithms, such as MeHaloCoa[3] or DCA[4] were developed along with user interfaces to ensure the accessibility of such tools.

In the present work, we describe HaloSeeker v1.0, an ergonomic interface aiming at gathering all necessary post-acquisition processing and interpretation tools in a single user interface, ensuring efficient accessibility, interactivity and traceability. Through these innovative integrated capacities, HaloSeeker assists the user in the semi-automatic screening of Cl- and Br-containing substances from full scan HRMS datasets, without any interaction with the underlying coding part. In this proceeding, the capabilities and related outputs offered by this tool are described using two marine samples (sediment and mussel), which are typically used as integrative matrices to monitor contaminations in marine environments.

Materials and methods

HaloSeeker was developed under the open access R environment and aims at processing full scan HRMS datasets in order to filter halogenated signals and assign chemical formula. Three requirements were considered:

(i) Developing a user-friendly interface that does not require any manual installations.

(ii) Securing the data and users' actions into a dedicated and traceable database.

(iii) Drastically decreasing the time spent for data interpretation via graphical interactivity.

HaloSeeker v1.0 is available on request at contact.haloseeker@oniris-nantes.fr. The zip file (~1 Go) includes all required packages. The run.vbs file launches the interface as a Firefox browser page. Main processes and key features integrated into HaloSeeker are described below.

Conversion. The open access software MSConvert (proteowizard) is embedded in HaloSeeker to convert raw data into the open format *mzXML* prior to post-processing. Although MSConvert manages most manufacturer formats, users have also the possibility to input directly *mzXML* files.

Peak picking. The *centWave* algorithm from the *xcms* package is embedded for chromatographic peaks integration[5]. By default, parameters are set at optimized values for our specific instrumental conditions as follows: *ppm*=3, *peak width*=5-60, *snthresh*=10, *prefilter step*=3, *prefilter level*=10,000, *mzdiff*=0.001. Output tables contain integrated features characterized by *m/z*, retention times (Rt) and intensities.
Pairing. We previously developed a VBA script for pairing halogenated isotopologue ions (clusters), based on the exact mass difference between Cl and Br isotopes[6]. The script was translated into R language and incorporated in HaloSeeker, drastically decreasing the processing time. Also, the script was slightly modified, mass difference being now calculated from the base peak (A) for all isotopologues. The script allows to discriminate "paired clusters" (at least an A+2 paired feature) from "non-paired clusters", thus acting as a first filter. Default parameters were set as follows: *retention time tolerance*=1 s, *mass deviation tolerance*=0.5 Da.
Interactive H/Cl-scale MD plots are drawn by HaloSeeker. Display options are available to narrow down searches. They include (i) a retention time window, (ii) an intensity threshold and/or a number of most intense clusters, (iii) cluster assignation status filters (all, all but discarded, assigned or non-assigned), (iv) *m/z*-pairing and ion ratio filters including "F0", "F1", "F2" and "F2+" and (v) cluster slope filters allowing to specifically display clusters according to the slope of the linear regression curve of isotopologues (A±n, n being even). Indeed, in H/Cl MD plots, the theoretical slope is -3.3×10^{-4} and $+1.2\times10^{-4}$ with correlation coefficient of 1 for (poly)chlorinated and (poly)brominated clusters, respectively, whereas mixed clusters exhibit intermediary slopes with lower correlation coefficients. More precisely, "F0" filter displays all picked peak features and "F1" filter the A+2–paired clusters. Within "F1" clusters, additional rules related to ion ratios aim at removing false positive halogenated clusters. Based on theoretical considerations, isotopologue areas relative to base peak (A) rules are, for "F2" filter:

$$[A\text{-}2 = 0\% \text{ AND } A+2 \geq 25\%] \text{ OR } [A\text{-}2 \geq 60\% \text{ AND } A+2 \geq 20\%] \text{ OR } [A\text{-}2 \geq 27\% \text{ AND } A+2 \geq 36\%].$$

For "F2+" filter, "F2" clusters suspected to relate to monohalogenated ions are removed if:

$$[A\text{-}2 = 0\%] \text{ AND } [A+4 = 0\%] \text{ AND } [A+2 \in (25\text{-}39) \cup (77\text{-}117)\%].$$

Raw chemical formula assignment. Part of the ergonomics capabilities of HaloSeeker lies in the interactive nature of MD plots. A click on a cluster displays a pop-up window dedicated to formula assignment. It includes, the mass spectrum, the extracted ion chromatograms and information about the cluster slope. We adapted Rdisop[7], a script dedicated to formula ranking using high precision mass spectrometry, to allow calculation of the formula from either the monoisotopic or the base peak isotopologue. Mass deviation and a developed pattern scoring (using *EnviPat*[8] theoretical spectra calculation) are provided for each possible hit.
Dereplication. The dereplication module allows to search within the database populated along the projects (assigned chemical formula or manually recorded).

Illustration datasets. Sediment and mussel samples were collected in 2002 and 2017, respectively, from the river Seine estuary (France). Five grams lyophilized aliquotes were extracted by pressurized liquid extraction with dichloromethane, and transferred in 6 mL *n*-hexane. The mussel extract was treated by liquid-liquid partitioning with concentrated sulfuric acid (4×3 mL) and water for neutralization (2×6 mL). The sediment extract was treated by hydrochloric acid-activated copper. Procedural blanks were processed similarly. The purified extracts were spiked with $^{2}H_{18}$-γ-hexabromocyclododecane (HBCDD) as external standard and reconstituted in a mixture of water/acetonitrile 1:4 (*v/v*, 100 µL). Extracts (10 µL) were analyzed by LC-ESI-HRMS (Q-Exactive Orbitrap, Thermo Fischer Scientific, San José, CA, USA) in the negative mode. Chromatographic separation was performed on a Hypersil Gold column (100 mm × 2.1 mm, 1.9 µm) kept at 45 °C and a gradient of acetonitrile against water, both containing 10 mM ammonium acetate. HRMS data were acquired in full scan mode within the *m/z* range 120-1000 at a resolving power of 140,000 FWHM@200, using *m/z* 305.02307 ($[C_2H_3O_2.(NaC_2H_3O_2)_3]^-$) as lock mass. Data were processed on a Windows 7 64 bits PC with 8 GB RAM and 3.2 GHz i5 CPU.

Results and discussion

Overview. Total ion current chromatograms appeared informative with satisfying procedural blanks. The retention time range between front solvent and column reconditioning was selected for the investigations. Table 1 reports the filtering results according to *m/z* pairing and ion ratio filters for both samples. HaloSeeker processed thousands of features within few minutes for both datasets (mean value of 3.7 min for peak picking and 1.5 min for pairing).

Table 1: Number of features, paired clusters and cumulated intensities ($\times 10^9$ AU) according to *m/z* pairing and ion ratio filters for both samples with a 2×10^6 intensity threshold.

	Sediment				Mussel			
Filter	**F0**	**F1**	**F2**	**F2+**	**F0**	**F1**	**F2**	**F2+**
Features	4532	1066	961	827	4035	843	729	457
Clusters	-	245	211	165	-	210	163	79
Intensity	131	6.45	4.21	3.77	374	10.8	9.12	3.32

An intensity threshold of 2×10^6 UA was applied to remove low intense clusters. The *m/z* pairing script (F1) considerably reduced the number of potential signals of interest, which was then further refined by adding complementary filtering rules based on isotopic ratio (F2 and F2+). Both filters permit to reduce potential "false positives" such as silicon and sulfur-containing compounds, as well as irrelevantly picked peaks and/or paired features. Still, few hundreds of clusters remained after filtering, which made manual investigation a realistic option. The difference between F2 and F2+ suggests that the two samples may contain a significant proportion of monohalogenated substances, accounting for 10% and 64% of the cumulated intensities, respectively. Part of the monohalogenated clusters could arise from chlorine adducts, as observed for HBCDD, thus increasing "false positive" hits. The observed mass deviation (Δppm) for a selection of analytical standards covering most of the *m/z* range of interest (HBCDDs, tetrabromobisphenol A, halogenophenols, mixed halogenated hydroxydiphenyl ethers), ranged from 0.06 to 0.92 ppm (mean=0.33±0.23 ppm). Such sub-ppm accuracy and precision of the instrument appears decisive for final chemical formula assignment. Indeed, along with the isotopic pattern, the exact mass stands as key information for chemical formula assignment. Consequently, a tolerated value of ±1.5 ppm was set for the formula generator tool.

In the following example, focus on polyhalogenated related signals only (F2+ filter) was made. The interactivity of the H/Cl-scale MD plot combined with the fact that the application gives access to all useful information in one single pop-up window allowed a manual investigation of hundreds of clusters in both samples within a realistic amount of time (about two working days) to be performed. For chemical formula assignment, the following elements were considered: ^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{31}P, ^{35}Cl, ^{37}Cl, ^{79}Br and ^{81}Br.

<u>*Mussel*</u>. Chemical formulas with a relatively high confidence level could be assigned to 59 out of the 79 clusters within the F2+ filter (60% of cumulated intensities). Among these signals, four could be attributed to halogenophenols (some of which confirmed by standards), tribromophenol being the most intense cluster. The second most intense cluster was assigned to tribromoanisole, is a transformation product of tribromophenol.
A series of 10 horizontally aligned clusters (7 distinct *m/z*) particularly drew our attention (cumulated intensities: 7.1×10^7). Figure 1a shows a zoom-in of the area of interest. This series exhibited horizontal vectors characteristic of -H/+Br, -H/+Cl and -Cl/+Br substitutions on H/Cl-scale MD plots. Cluster slopes are also useful to support data interpretation, acting as an additional filter. Indeed, in the H/Cl-scale, there is a slightly positive mass defect between the fractional part of ^{81}Br (9.14×10^{-3}) and ^{79}Br (8.90×10^{-3}), which corresponds to the Y-axis difference between two consecutive dots of a brominated cluster on the MD plot. Thus, brominated clusters are characterized by a positive slope (purple arrow in Figure 1). Conversely, the mass defect between ^{37}Cl (8.32×10^{-3}) and ^{35}Cl (8.98×10^{-3}) leads to a negative slope for chlorinated clusters (orange arrow). Consequently, considering the spectrometric resolution, mixed halogenated clusters present slopes in between the aforementioned ones, which leads to nearly horizontal slopes (light blue arrow). In the case of our series of interest, all clusters slopes were attributed to mixed Cl/Br formulas. Combining precise mass, isotopic pattern and slope information, the general chemical formula $C_8H_xOCl_yBr_z$ was found to be the most relevant. To our knowledge, none of the 10 chemical formulas were related to previously-reported compounds.

<u>*Sediment*</u>. Prior to graphical interpretation, the sediment dataset was subjected to the dereplication module populated with the mussel dataset, which allowed to quickly assign formulas to 32 clusters. Formulas with relatively high confidence level could be assigned to 87 out of the 165 clusters within the F2+ filter (69% of cumulated intensities). Among these signals, similar halogenophenol related clusters (n=7) were observed. A cluster was assigned to halogenated carbazoles ($C_{12}H_5NCl_2Br_2$), substances of growing concern due to their presence in the environment, persistence and potential dioxin-like activity.

A series of 22 horizontally aligned clusters (13 distinct *m/z*) particularly drew our attention (relatively high cumulated intensities: 6.7×10^8). Figure 1b shows a zoom-in of the area of interest. The general formula $C_{12}H_xO_2Br_yCl_z$ was found to be the most relevant for this series of mixed halogenated (n=3, orange circles on Figure 1), fully chlorinated (n=16, Cl_2 to Cl_9) and fully brominated (n=3) clusters corresponding to mixed halogenated hydroxylated diphenyl ethers (HO-XDEs). Triclosan ($C_{12}H_6O_2Cl_3$, the most intense signal, blue circle) and 6-hydroxy-2,2',3,4,4'-pentabromodiphenyl ether ($C_{12}H_4O_2Br_5$, green circle) were confirmed against pure analytical standards.

Conclusion

HaloSeeker v1.0 allowed to data mine hundreds of polyhalogenated clusters among HRMS datasets using advanced bioinformatics tools without requiring extensive knowledge in informatics. Beyond formula assignment relying on mass deviation and isotopic pattern scorings, unambiguous structural elucidation is still challenging and requires complementary assays, such as comparison with reference pure standards and/or MS and MS/MS fragmentation.

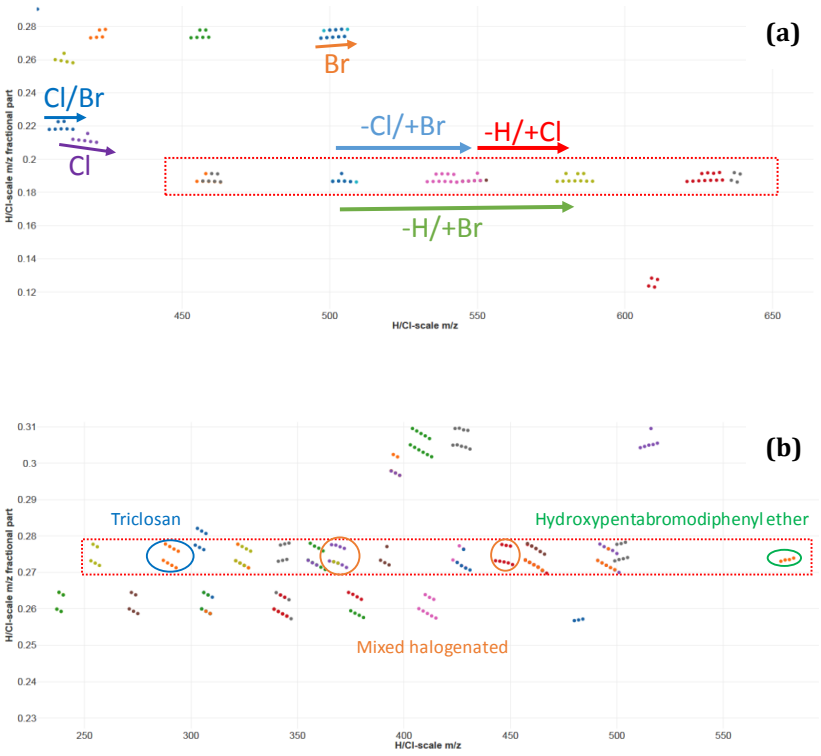

Figure 1: PNG HaloSeeker export of H/Cl-scale MD plot zoom-in for the mussel (a) and the sediment (b) samples (Rt range 1-20 min, filter F2+), highlighting the $C_8H_xOCl_yBr_z$ and $C_{12}H_xO_2Br_yCl_z$ series (dashed rectangles).

It is now expected that sharing the developed tool with the scientific community will allow defining/broadening the scope of its applicability (environmental sciences, halogenated active compounds screening, metabolism studies…) and provide feedback for potential improvements. So far, we identified that HaloSeeker v1.0 could be further enhanced to speed up the interpretation step. Indeed, ongoing developments relate to the implementation of (i) an alignment module from the *xcms* package allowing for processing several datasets at once and (ii) an adducts/fragments pairing based on the *CAMERA* script.

Acknowledgements

AL received a PhD grant from both Ifremer and the "Région Pays de la Loire", France" to conduct this research. The authors want to express their acknowledgments to the French General Directorate for Food and the HBM4EU project (EU H2020 SC1-2016-RTD call, Cofund project nb. 733032) for their financial support.

References

1. Taguchi VY, Nieckarz RJ, Clement RE, Krolik S, et al. (2010); *J Am Soc Mass Spectrom*. 21: 1918-1921.
2. Marshall AG, Rodgers RP (2004); *Acc Chem Res*. 37: 53-59.
3. Roullier C, Guitton Y, Valery M, Amand S, Prado S, et al. (2016); *Anal Chem*. 88: 9143-9150.
4. Andersen AJ, Hansen PJ, Jorgensen K, Nielsen KF (2016); *Anal Chem*. 88: 12461-12469.
5. Tautenhahn R, Bottcher C, Neumann S (2008); *BMC Bioinformatics*. 9, 504.
6. Cariou R, Omer E, Léon A, Dervilly-Pinel G, Le Bizec B (2016); *Anal Chim Acta*. 936, 130-138.
7. Bocker S, Liptak Z, Martin M, Pervukhin A, Sudek H (2008); *Bioinformatics*. 24, 591-593.
8. Loos M, Gerber C, Corona F, Hollender J, Singer H (2015); *Anal Chem*. 87, 5738-5744.