Downloaded from orbit.dtu.dk on: Dec 12, 2024

DTU Library

Dynamic Cluster Analysis: An Unbiased Method for Identifying A+2 Element Containing Compounds in Liquid Chromatographic High-Resolution TOF Mass Spectrometric Data

Andersen, Aaron John Christian; Hansen, Per Juel; Jørgensen, Kevin; Nielsen, Kristian Fog

Published in:
Analytical Chemistry

Link to article, DOI:
10.1021/acs.analchem.6b03902

Publication date:
2016

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Andersen, A. J. C., Hansen, P. J., Jørgensen, K., & Nielsen, K. F. (2016). Dynamic Cluster Analysis: An Unbiased Method for Identifying A+2 Element Containing Compounds in Liquid Chromatographic High-Resolution TOF Mass Spectrometric Data. *Analytical Chemistry*, *88*(24), 12461–12469. https://doi.org/10.1021/acs.analchem.6b03902

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

analytical chemistry

DTU Library
Technical Information Center of Denmark

Subscriber access provided by DTU Library

Article

Dynamic Cluster Analysis: An Unbiased Method for Identifying A+2 Element Containing Compounds in Liquid Chromatographic High-Resolution TOF Mass Spectrometric Data

Aaron John Christian Andersen, Per Juel Hansen, Kevin Jørgensen, and Kristian Fog Nielsen

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.6b03902 • Publication Date (Web): 21 Nov 2016

Downloaded from http://pubs.acs.org on November 22, 2016

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.

ACS Publications

Analytical Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Published by American Chemical Society. Copyright © American Chemical Society. However, no copyright claim is made to original U.S. Government works, or works produced by employees of any Commonwealth realm Crown government in the course of their duties.

Dynamic Cluster Analysis: An Unbiased Method for Identifying A+2 Element Containing Compounds in Liquid Chromatographic High-Resolution TOF Mass Spectrometric Data

Author Names

Aaron John Christian Andersen,[1,2] Per Juel Hansen,[3] Kevin Jørgensen,[1] and Kristian Fog Nielsen[2]

Author Affiliations

[1] National Food Institute, Technical University of Denmark, Lyngby, Denmark

[2] Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark

[3] Marine Biology Section, University of Copenhagen, Helsingør, Denmark

Corresponding author

Kristian Fog Nielsen (kfn@bio.dtu.dk)

Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark

Keywords: mass spectrometry, dynamic cluster analysis, A+2, isotope spacing, *Prymnesium parvum*, prymnesin, harmful algal bloom

Abstract

Dynamic Cluster Analysis (DCA) is an automated, unbiased technique which can identify Cl, Br, S, and other A+2 element containing metabolites in liquid chromatographic high resolution mass spectrometric data. DCA is based on three features, primarily the previously unutilised A+1 to A+2 isotope cluster spacing which is a strong classifier in itself, but improved with the addition of the monoisotopic mass, and the well-known A:A+2 intensity ratio.

Utilizing only the A+1 to A+2 isotope cluster spacing and the monoisotopic mass it was possible to filter a chromatogram for metabolites which contain Cl, Br, and S. Screening simulated isotope patterns of the Antibase Natural Products Database it was determined that the A+1 to A+2 isotope cluster spacing can be used to correctly classify 97.4% of molecular formulas containing these elements, only misclassifying a few metabolites which were either over 2800 u or metabolites which contained other A+2 elements, such as Cu, Ni, Mg, and Zn.

It was determined that with an inter-isotopic mass accuracy of 1 ppm, in a fully automated process, using all three parameters, it is possible to specifically filter a chromatogram for S containing metabolites with monoisotopic masses less than 825 u. Furthermore, it was possible to specifically filter a chromatogram for Cl and Br containing metabolites with monoisotopic masses less than 1613 u.

Here DCA is applied on: i) simulated isotope patterns of the Antibase natural products databases; ii) LC-QTOF data of reference standards; and iii) LC-QTOF data of crude extracts of 10 strains of laboratory grown cultures of the microalga *Prymnesium parvum* where it identified known metabolites of the prymnesin series as well as over 20 previously undescribed prymnesin-like molecular features.

Introduction

It has been estimated that 15-20% of newly discovered natural products are halogenated, this is especially apparent with metabolites from marine ecosystems.[1–3] Every compound listed in the Stockholm Convention on Persistent Organic Pollutants contains Cl, Br, or S.[4] In comparison to the Antibase natural products database,[2] the molecular drugs on the 19th World Health Organisation Model List of Essential Medicines incorporate a proportionally higher number of Cl (2.4 times) and S atoms (2.8 times).[5]

Crude natural product extracts, environmental samples, tissue extracts, etc. are often very complex, often containing thousands of chemical features. These samples are often analyzed by high resolution mass spectrometry (HRMS) coupled to liquid chromatography. Given that halogenated and S containing compounds are often of biological interest, an automated approach to classify compounds containing S, Cl, and Br, especially within such complex chromatograms, would have applications in many areas.

By making use of the isotopic patterns and accurate mass, the incorporation of particular elements within a compound can be determined. Compounds containing Cl, Br, or S are notable for their comparatively high abundance of isotopologues which contribute to the A+2 centroid peak. These elements cause an increase in A+2 intensity relative to the A large enough for common mass analyzers (ion-trap, quadrupole, time-of-flight, and orbitrap) to detect, creating a distinctive isotope ratio. The ability to detect different isotopic ions has been present in vendor packages for many years and is used for centroid processing, as well as chemical formula calculations and chemical feature identification. Peak picking algorithms are also now common in open-source software packages, such as XCMS, MZmine 2, MetAlign, and mzMatch.[6–9]

While in principle it should be possible to manually screen data for metabolites containing these elements, in practice this increase in abundance can be difficult to identify. This difficulty is clearly apparent in molecules containing one S atom as the ^{34}S isotope abundancy is only 4.4%, similarly, screening data to identify molecules with one Cl atom can also become difficult with an increasing number of C atoms. As the number of C atoms increases in a molecular formula, the increase in abundancy of the A+2 due to the ^{37}Cl isotopic ion begins to become proportionally diluted into the abundances of the $^{13}C_2$ isotopomers in the centroid data. Identification can be further complicated by co-eluting dehydro isomers, as well as MS detector saturation, therefore an algorithm which relies solely

on an intensity ratio to classify metabolites as containing Cl, Br, and/or S can lead to both false positives and false negatives. Automated filtering algorithms, such as MeHaloCoA, can be used to analyze chromatograms for compounds containing Cl and Br, however, compounds with high molecular mass (> 800 u) and multiple charges may not be detected, and elements with lower A+2 abundance (S) cannot be distinguished.[10]

An important technique in metabolite screening is to filter for a particular fractional mass, known as mass defect filtering. This is often utilised in biomedical settings when unknown drug metabolites need to be identified.[11,12] Changes to a drug via metabolism are often slight, such as the addition or subtraction of a functional group, and therefore the change in the fractional mass will often be within a small mass defect range.[13] This has been proven to be a very useful technique, however it relies on a previously known molecular formula for a reference mass defect to search for.[13,14] Related techniques, referred to as *relative* mass defect, have been defined in a number of different ways and used for different purposes, such as a more selective way to filtering chromatograms,[15] a way a detect metabolites containing elements with large mass defects,[16] and as a tool for molecular formula calculations.[17]

Isotope cluster analysis, a method that utilises both the A to A+2 isotope cluster spacing and A:A+2 intensity ratio, is another useful tool. This approach has been successful in filtering various halogenated compounds, such as drug metabolites and highly chlorinated flame retardants.[18,19] By setting a value for each of these parameters, as well as an acceptable margin of error, it is possible to quickly screen a complex chromatogram for compounds containing a particular proportion of Cl and/or Br.[19] Due to the need for pre-set parameters this technique is typically used when targets within particular mass ranges are being investigated, as over a typical scan range the isotope cluster spacing and intensity ratios can vary significantly depending on the number of C atoms. Other algorithms designed to identify Cl and Br containing metabolites in a more automated process also do not compensate for the changes in isotope pattern characteristics with increasing monoisotopic mass and therefore can operate within a comparatively restricted mass range.[10]

Many of the limitations associated with these filtering techniques can be overcome with sufficient resolution.[20] The two Fourier transform (FT) mass analyzers: i) ion cyclotron resonance (ICR), ii) orbitrap mass analyzers, can resolve most isotopic ions, such as $^{13}C_2{}^{35}Cl$ from $^{12}C^{37}Cl$, over a wide mass range, however the resolution is inversely proportional to acquisition time, therefore considerations of scan rate in relation to chromatographic

separation need to made.[21,23] Finally, the FT-ICR mass analyzers are significantly more expensive, in both purchase price and especially running costs, in comparison to TOF and orbitrap mass analyzers.

Presented here is the method of Dynamic Cluster Analysis (DCA), an automated technique that is directly applicable to target metabolites across a wide mass range, effective with samples of high complexity, and without the need for compound specific, pre-defined parameters. DCA utilises intra-isotopic pattern characteristics, most significantly the A+1 to A+2 isotope cluster spacing, in conjunction with monoisotopic mass and A:A+2 intensity ratio to classify metabolites of particular elemental compositions. The DCA algorithms have the ability to screen for metabolites containing Cl and Br, as well as S, on instruments with mass accuracies up to or better than 5 ppm. We present here the results of applying DCA to simulated isotope patterns from metabolite databases, the results of applying DCA to reference standards analyzed on two different TOF instruments, and finally, we demonstrate the effectiveness of this method when used to screen complex matrixes of marine microalgae, specifically crude extracts from 10 strains of the haptophyte *Prymnesium parvum*, known to produce the halogenated series of ichthyotoxins, the prymnesins.

Materials and Methods

Chemicals and Standards

LCMS grade H_2O and acetonitrile for UHPLC-HRMS analysis were purchased from Sigma-Aldrich (St Louis, Missouri, MO). Formic acid (≥ 96%) was purchased from Thermo Fisher Scientific (Waltham, MA). HPLC grade acetone and MeOH for culture extraction and standard preparation were purchased from Sigma-Aldrich.

Reference standards amphotericin B, folic acid, malformin A, neosolaniol, penicillin G, penitrem A, roridin A, cephalosporin C, phalloidin, and vitamin B1 were purchased from Sigma-Aldrich. BE52440 A and bis-sclerotioramin were purchased from Analyticon Discovery (Potsdam, Germany). Enniatin A and phomopsin A were purchased from BioAustralis (Smithfield, NSW, Australia). Ochratoxin alpha, dichlorodiaportin, nidulin, and citreo-isocoumarin were available from previous studies.[24] Reference standard structures and molecular formula are listed in **Supporting Table S-1**.

Prymnesium parvum *Samples*

Biomass samples from 10 strains of *P. parvum* from 100 mL cultures were acquired from a previous study.[25] Growth conditions and strain information listed in **Supporting Table S-2**.

Liquid Chromatography and Mass Spectrometry

Two UHPLC-QTOF-HRMS instruments were utilised to compare two different instruments under different standard conditions. Both used a linear reversed phase separation system with water-acetonitrile containing 20 mM formic acid.

UHPLC-QTOF-1. An Ultimate 3000 UHPLC (Dionex; Sunnyvale, CA) equipped with a 2.6 µm, 100 Å, 100 × 2.1 mm Phenomenex Kinetex C_{18} column (Torrance, CA) was coupled to a Bruker Maxis HD QTOF (Bruker Daltonics, Bremen, Germany). The gradient started at 10% acetonitrile, increased to 100% over 10 min, held at 100% for 3 min, at a constant flow rate of 0.4 mL/min and column temperature of 40 °C. For the analysis of reference standards the nebuliser gas pressure was 1.8 bar, drying gas temperature 200 °C, drying gas flowrate 12 L/min, and spectra recorded at *m/z* 75-1250.[24] For the analysis of *P. parvum* extracts the drying gas flowrate was set to 10 L/min, and the scan range was *m/z* 300-2500, the ion path optimized for ions in the range.

UHPLC-QTOF-2. An Agilent 1290 Infinity UHPLC (Agilent Technologies; Santa Clara, CA) equipped with an Agilent Poroshell 120 Phenyl Hexyl column (2.7 µm, 250 × 2.1 mm) was coupled to an Agilent 6545 QTOF. The gradient started at 10% acetonitrile, increased to 100% over 15 min, held at 100% for 2 min, at a constant flow rate of 0.35 mL/min and column temperature of 60 °C.[24] For the analysis of reference standards the scan range was *m/z* 100-1600.[24]

LCMS Data Extraction

LCMS data was exported from vender software (Agilent: Qualitative Analysis B.06.00; Bruker: Compass DataAnalysis 4.2) as XML data files. These exported files were extracted using the R package XCMS, with the following settings: ppm 10, peakwidth c(4,15), snthresh 4, bw 2, and mzwid 0.010.[6,26,27] Isotopomers which were observed in both sample and blank chromatograms were removed. The compiling of isotopomers into chemical features, molecular features, and subsequent DCA classification was achieved using a program developed for this project called DCAnalysis v1.07 (**Supporting Code S-1**).[28,29] All programs developed in python for this project were constructed using the Tkinter GUI.[29] All programs

developed for this project are available from the authors and the source code for all programs is published in supporting information. All programs created for this project were developed in python 2.7.6.[28] All data was processed on a Windows 7 PC, i7-4800MQ 2.70 GHz, 8 Gb RAM.

Database Isotope Pattern Simulation and Extraction

All molecular formulas from the 2012 Marinlit[1] and 2012 Antibase[2] natural products databases that contain more than one C atom, and have a monoisotopic mass greater than or equal to 50 u (9121 and 16155 unique molecular formulas respectively) were compiled. The isotope patterns of the compiled molecular formula lists for their $[M+H]^+$ adduct were simulated and a centroid process was applied to all isotopologues within an isotopic interval using weighted averages. From the centroid data, monoisotopic ion masses, the A:A+2 intensity ratios, and the isotope cluster spacing for the A+1 to A+2 and A to A+1 isotope clusters were calculated using FormExtract v1.01 (**Supporting Code S-2**).

Database Modelling and Theoretical Discrimination

Modelling and formulation of decision boundaries for discriminatory analysis was preformed using the Marinlit database (n = 22588, 9129 unique molecular formulas) as a training dataset and the larger Antibase database (n = 40065, 16134 unique molecular formulas) was later used as the test dataset.

Simulated isotope patterns of the Marinlit database were categorised into three groups based on molecular formula: Group-Cl, molecular formulas containing one or more Cl and/or Br; Group-S, molecular formulas containing one or more S (excluding those which also contain Cl and/or Br); and Group-C, all remaining molecular formulas. Decision boundaries for the A+1 to A+2 isotope cluster spacing and the A:A+2 intensity ratios between these groups were then determined by assessing realistic hypothetical molecular formulas based on the characteristics of metabolites within the Marinlit database (**Supporting Text S-1**). This resulted in the series of equations which defined the boundaries between the three groups of molecular formulas.

Simulated isotopomers of the molecular formulas from Antibase were calculated using a program called PatExtract v1.01 (**Supporting Code S-3**). The decision boundary equations derived from the Marinlit database were then applied to the simulated isotopomers using DCAnalysis v1.07 to determine the theoretical discriminatory robustness of the equations.

UHPLC-QTOF-MS Analyses of Reference Standards

18 reference standards, 6 for each of the classification groups (Group-Cl, -S, and -C), were chosen with varying degrees of Cl and S incorporation over a wide mass range (256-923 u). All standards were analyzed on both UHPLC-QTOF instruments at various concentrations to produce spectra with peak height signal-to-noise ratios greater than 140 but below the detectors point of saturation.

P. parvum *Extraction and Analyses*

P. parvum biomass samples were first extracted with acetone to remove comparatively lipophilic constituents followed by extraction with MeOH to produce samples for UHPLC-QTOF analysis. The full extraction protocol can be found in **Supporting Text S-2**.

Results and Discussion

A+1 to A+2 Isotope Cluster Spacing

When investigating the potential of the A+1 to A+2 isotope cluster spacing for chromatogram filtering analysis of the simulated isotope patterns of the Marinlit database revealed an enhanced resolution of discrimination when using the A+1 to A+2 isotope cluster spacing in comparison to the A to A+2. This enhancement was particularly noticeable at the boundary of Group-S and Group-C, between 1000 and 2000 u (**Fig. 1**).

This effect was due to the presence of heteroatoms, particularly N. ^{15}N has a natural abundance of 0.365% and contributes to the A+1 isotope cluster, and to a lesser extent the A+2, the presence of this isotope causes a decrease in the A to A+2 isotope cluster spacing in comparison to the A+1 to A+2. For metabolites containing no heteroatoms, the A to A+2 isotope cluster spacing will be approximately 2.006 u. The addition of N will decrease this value, increasing the A to A+2 isotope cluster spacing spread within the group. A similar effect occurs with the A+1 to A+2 isotope cluster spacing, though to a lesser extent as the A+1 isotope cluster incorporates ^{15}N isotopomers. This has the consequence of increasing the isotope cluster spacing, resulting in increased resolution of discrimination.

In general, a A+1 to A+2 isotope cluster spacing lower than 1.001 u is indicative of metabolites containing S, Cl and/or Br. A dynamic boundary between metabolites containing these elements and those which do not, can be made by adjusting this 1.001 u limit in regards to the *m/z* of the monoisotopic ion of a metabolite, and can decrease the inter-isotopic mass accuracy (**Supporting Figure S-1**) needed to differentiate these metabolites.

In order to determine the dynamic boundary between groups of metabolites containing S, Cl and/or Br and those which do not, realistic hypothetical molecular formulas were produced based on characteristics of the molecular formulas within the Marinlit database. The maximum O:C and N:C ratios and the minimum H:C ratio that included 99.8% of the Marinlit database were found to be 0.80, 0.65, and 0.49, respectively. Using these ratios, it was possible to produce realistic hypothetical molecular formulas across a mass range exceeding the databases. The monoisotopic ion masses and A+1 to A+2 isotope cluster spacing for these formulas were simulated and modelled using polynomial equations ($R^2 > 0.981$) (**Fig. 2**).

The midpoints between the polynomial equations for Group-Cl and Group-C and for Group-S and Group-C were calculated. This resulted a set of midpoint equations expressed using polynomial equations, each having R^2 values greater than 0.996. These midpoint equations were then used as the decision boundaries between Group-Cl and Group-C (**Eq. 1**) and between Group-S and Group-C (**Eq. 2**) (**Supporting Figure S-2**).

$$V = 1.5644 \times 10^{-23}(mz)^6 - 2.46 \times 10^{-19}(mz)^5 + 1.5135 \times 10^{-15}(mz)^4 - 4.485 \times 10^{-12}(mz)^3 + 5.954 \times 10^{-9}(mz)^2 - 9.019 \times 10^{-7}(mz) + 0.99832 \qquad (eq.1)$$

$$V = 4.288 \times 10^{-23}(mz)^6 - 4.975 \times 10^{-19}(mz)^5 + 2.0086 \times 10^{-15}(mz)^4 - 2.735 \times 10^{-12}(mz)^3 - 2.07 \times 10^{-9}(mz)^2 + 8.454 \times 10^{-6}(mz) + 0.99684 \qquad (eq.2)$$

Where V is the A+1 to A+2 isotope cluster spacing, m is the monoisotopic ion mass, and z is the ion charge.

The lower decision boundary for Group-Cl and Group-S was determined to be the lowest value of the A+1 to A+2 isotope cluster spacing expected for a saturated molecular formula containing one Cl or S atom, 0.9936, minus an acceptable error of 5 ppm (**Eq. 3**).

$$V = 0.9936 - \left(\frac{(mz)}{1000000} \times 5\right) \qquad (eq.3)$$

Using the polynomial equations derived from the Group-Cl and Group-C hypothetical molecular formulas it was also possible to calculate the mass limit of discrimination, the first intercept of Group-Cl equations with Group-C equations. This was determined to be 2793 u (**Fig. 2**). The same calculation was performed on the Group-S and Group-C equations and was found to be 1081 u (**Fig. 2**). These values indicated that, for centroid data, it may not be

possible to differentiate Cl/Br or S containing compounds with monoisotopic ion masses greater than 2793 and 1081 u, respectively. These values were then considered the theoretical limits of DCA.

When considering instrument mass accuracy, it was revealed that an instrument with an inter-isotopic mass accuracy of 5 ppm would be able to classify Group-Cl metabolites with monoisotopic masses up to 900 u, and 533 u for Group-S metabolites. For higher mass accuracy instruments 1 ppm limits were also derived, which demonstrated classification limits of 1612 u for Group-Cl and 824 u for Group-S. Between classification limits and the theoretical limits it was expected that there would be an increase in misclassified metabolites proportional to increasing monoisotopic mass at a constant inter-isotopic mass accuracy.

When applying **Eq. 2** and **Eq. 3** as decision boundaries on the simulated isotope patterns of Antibase it was found that it was possible to correctly classify 97.4% of molecular formulas containing S, Cl, or Br from those which contained none of these elements. The vast majority of misclassifications were due to the incorporation of atypical elements, specifically Se, Cu, Te, Fe, Ni, Mg, or Zn. Using the same method to derive decision boundaries for the A to A+2 isotope cluster spacing (**Supporting Equation S-1**, **Supporting Equation S-2**, and **Supporting Equation S-3**), it was found that this classification rate dropped to 96.7%. The increase in misclassified molecular formulas was due, in equal parts, to an increase in the misclassification of S containing metabolites above 1200 u and the misclassification of metabolites containing B and V. This demonstrated the effectiveness of the A+1 to A+2 isotope cluster spacing as a filtering tool in itself, and its benefit over the A to A+2 isotope cluster spacing.

A:A+2 Intensity Ratio

The A:A2 intensity ratio provided another dimension which refined classification. The A:A2 intensity ratio of the molecular formula equations that corresponded to complete saturation and no heteroatoms diverged greatly from the Marinlit dataset. This indicated that complete saturation of metabolites with the absence of heteroatoms was unlikely, particularly at high molecular masses, and therefore the molecular formula equations used to model the A:A+2 intensity ratio decision boundaries were $C_nH_{2n+0.65n+1}N_{0.65n}Cl$ for Group-Cl, and $C_nH_{2n+0.65n+2}N_{0.65n}S$ for Group-S, illustrated in **Fig. 3**.

To allow for instrument error a buffer of 0.05 was subtracted from each of the A:A+2 intensity ratios of the hypothetical molecular formulas, this is equivalent to a ±5% isotope

ratio error, a value suggested previously for molecular formula calculations.[30] This buffer adjusted data was then expressed using polynomial equations, each having R^2 values greater than 0.999, resulting in two equations for the decision boundaries determining the lower limit for Group-Cl (**Eq. 4**), and the lower limit for Group-S (**Eq. 5**).

$$I = 1.611 \times 10^{-7}(mz)^2 - 1.319 \times 10^{-5}(mz) + 0.2702 \qquad (eq.4)$$

$$I = 1.611 \times 10^{-7}(mz)^2 - 7.982 \times 10^{-6}(mz) + 0.00471 \qquad (eq.5)$$

Where *I* is the A:A+2 intensity ratio, *m* is the monoisotopic ion mass, and *z* is the charge of the ion. A molecular ion with an intensity ratio greater than that calculated by **Eq. 4** was assigned to Group-Cl. A molecular ion with an intensity ratio less than the ratio calculated by **Eq. 4** and greater than that calculated by **Eq. 5** was assigned to either Group-S or Group-C.

An upper decision boundary for Group-Cl was used (**Eq. 6**) to capture a maximum number of 12 Br atoms in a molecular formula (the most Br atoms in a molecular formula within the Marinlit database was 8). The purpose of this decision boundary was to exclude ratios greater than this equation to prevent atypical elements, such as Fe, from being misidentified as halogens.

$$I = 1.611 \times 10^{-7}(mz)^2 - 1.319 \times 10^{-5}(mz) + 12.05 \qquad (eq.6)$$

Eq. 6 can be adjusted to encompass molecular formulas with more than 12 Br atoms by replacing the value of 12.05 in the equation with the value of the maximum number Br atoms to filter for plus the isotope abundancy error of 0.05. This is a general rule for this equation as the relative abundancy of ^{79}Br to ^{81}Br is close to 1 (0.97).

Appling **Eq. 4**, **Eq. 5**, and **Eq. 6** to the Antibase database demonstrated that Group-Cl molecular formulas could be separated from those in Group-S and Group-C, however Group-S could not be fully distinguished from Group-C.

Dynamic Cluster Analysis

By combining the decision boundaries of the A+1 to A+2 isotope cluster spacing and the A:A+2 intensity ratio the major classification limitations of the individual approaches were eliminated. By applying the A+1 to A+2 isotope cluster spacing equations compounds not containing A+2 elements are removed from the data, then by applying the A:A+2 intensity ratio equations it is possible to distinguish which A+2 elements are present and allows for the filtering of A+2 elements with lower A+2 abundances such as S. Molecular formulas which fall below **Eq. 1**, above **Eq. 3**, above **Eq. 4**, and below **Eq. 6** are assigned to Group-Cl, with

this combination of formulas referred to as DCA-Halogen (DCA-Hal). Molecular formulas which fall below **Eq. 2**, above **Eq. 3**, below **Eq. 4**, and above **Eq. 5** are assigned to Group-S, with this combination of formulas referred to as DCA-Sulfur (DCA-Sul). The two methods, DCA-Hal and DCA-Sul, were applied to simulated isotope patterns of the molecular formulas from Antibase. This analysis resulted in an average correct classification rate of 98.2% (**Supporting Table S-3**). The majority of misclassified molecular formulas contained atypical A+2 elements with unusual isotope patterns, such as Zn, Mg and Ni.

The misclassification of metabolites with atypical elements suggested that DCA may be useful in identifying metabolites which contain other A+2 elements, such as Mg in DCA-Sul and Zn, and Ni in DCA-Hal. As these other A+2 elements are comparatively rare, manual inspection of isotope patterns can be made to determine elemental composition (**Supporting Table S-4**).

UHPLC-QTOF-MS Analysis and Classification of Standards

The two methods, DCA-Hal and DCA-Sul, were incorporated into the program DCAnalysis v1.07 for the filtering of LCMS data. DCAnalysis v1.07 deconvolutes isotope clusters into chemical features, such as $[M+H]^+$, $[M+Na+H]^{2+}$, or $[2M–H_2O+H]^+$. These chemical features are deconvoluted into molecular features, a collection of adducts and fragmentation ions from the same chemical compound. The algorithms then assess each of the chemical features within a molecular feature to determine the classification for the molecular feature. The classification of a molecular feature was determined correct if a majority of its chemical features fall within the decision boundaries. A majority was used to minimise false positives/negatives due to the possibility that some chemical features can lose their distinctive elements (e.g. $[M–SO_3+H]^+$, and $[M–HCl+H]^+$) and chemical features with significant signal-to-noise interferences can augment their isotope pattern.

Based on these processes 100% of standards on both instruments were categorised correctly using DCA-Hal and DCA-Sul. Standards of Group-C were considered correctly classified if their molecular features were not present in the DCA-Hal and DCA-Sul analysis. The average correct classification for chemical features within a molecular feature of Group-Cl and Group-S standards was 87.5% for standards analyzed on the Bruker QTOF instrument, and 96.1% for standards analyzed on the Agilent QTOF instrument (**Supporting Table S-5**).

For comparative purposes the standards were also assessed using the MeHaloCoA method for halogen identification.[10] As MeHaloCoA is not able to identify compounds

containing sulfur nor is it able to identify multiply charged species, the direct comparison was made between singly charged chemical features identified by DCA-Hal with those found by the MeHaloCoA method. Of a combined total of 113 chemical features found by both methods, DCA-Hal resulted in 18 misclassifications, whereas MeHaloCoA resulted in 38 misclassifications. 6 of the misclassifications by DCA-Hal were false positives. Of the remaining 12 misclassifications, 5 were chemical features misidentified as containing sulfur, and the 7 remaining were errors in chemical feature compiling, such as missing the A+1 isotopomer (**Supporting Table S-6a-r**). All comparisons between DCA-Hal and MeHaloCoA were performed on the same XML data files and extracted using the same XCMS centwave settings. Default MeHaloCoA filtering settings were used in the comparisons.

The extraction of LCMS data using R was considered a potential source of error that could contribute to the misclassification of chemical features. Analysis of the most abundant adduct from each of the standards found that the error introduced by the extraction of data using XCMS was only slight (**Supporting Table S-7** and **S-8**). The largest errors were found in the comparison of the acquired data to that of theoretically calculated isotope patterns for the standards, these were ±0.9 ppm in the A+1 to A+2 inter-isotopic mass accuracy and ±2.0% in the A:A+2 intensity ratio. The average A+1 to A+2 inter-isotopic mass accuracy was ±0.9 ppm for the Bruker QTOF, and ±0.8 ppm for the Agilent QTOF. The average error in the A:A+2 intensity ratio was ±3.3% for Bruker QTOF and ±0.8% for the Agilent QTOF (**Supporting Tables S-7** and **S-8**).

Simulated isotope patterns indicated greater resolution of separation using the A+1 to A+2 isotope cluster spacing compared to A to A+2. Although theoretically this improvement was shown, it was considered possible that the inter-isotopic mass accuracy of A+1 to A+2 would be less than that of A to A+2 due to the centroid process of averaging multiple isotopologues. The data from standard analysis showed no consistent difference between the two isotope cluster spacings. The A to A+2 inter-isotopic mass accuracy was found to be ±1.2 ppm on the Bruker QTOF and ±1.1 ppm on the and Agilent QTOF, whereas the A+1 to A+2 inter-isotopic mass accuracy was found to be ±0.9 ppm on the Bruker QTOF and ±0.8 ppm on the Agilent QTOF, indicating no loss in mass accuracy by using the A+1 to A+2 isotope cluster spacing (**Supporting Table S-7**). Although there was no loss, this illustrates the importance of accurate centroid processing and that the inter-isotopic mass error might not necessarily be completely dependent on the accurate mass error.

The decision boundary equations of DCA accommodate multiple charged ions by multiplying the mass of each isotope cluster by the charge state, however the reduced spacing between isotope clusters of multiply charged ions may result in a decrease in inter-isotopic mass accuracy. Of the 115 identified ions produced by reference standards the average error for signally charged species was 1.1 ppm and 1.2 ppm for multiply charged species (**Supporting Table S-9**).

DCAnalysis v1.07 compiles chemical features from centroid isotope data, and based on accurate mass differences from a matrix of common positively charged adducts (H^+, NH_4^+, Na^+, K^+), fragmentations ($-H_2O$, –HCl, –HBr, $-SO_3$), multiple charges (allowing for the detection of singly, doubly and triply charged ions), and dimers, it compiles these chemical features into molecular features. The authors acknowledge that applications exist which identify chemical and molecular features, such as those available in vendor software, as well as in open source packages, such as CAMERA in R,[31] however for direct comparison between different LCMS instruments and due to advantages of combining molecular feature compiling and filtering into one graphical user interface, these features were combined into the one program (GUI illustrated in **Supporting Figure S-3**).

Screening of Algae Extracts

All 10 extracts of *P. parvum* were analyzed for known prymnesins, and prymnesin-like molecular features using DCA-Hal filtering.[25,32,33] For the purposes of this study, a molecular feature was considered prymnesin-like if it contained a $[M+2H]^{+2}$ ion with a *m/z* between 800 and 1200,[25] was identified by DCA-Hal, and eluted within 4-7 min on the Dionex LC or between 7-10 min on the Agilent LC. The identification of known prymnesins was determined by the identification of the exact mass of their $[M+2H]^{+2}$ ion, and the neutral loss fragmentation ion of their corresponding carbohydrate moieties.

Applying DCA-Hal to the *P. parvum* extracts resulted in the identification of all 16 previously identified prymnesins.[25] In total 51 prymnesin-like molecular features were identified across the 10 strains analyzed on the Bruker QTOF (**Supporting Table S-10**), an example of which is illustrated in **Fig. 4**. 22 of the identified molecular features had *m/z* for $[M+2H]^{2+}$ which corresponded to previously identified prymnesins,[25,32,33] and 29 of the molecular features had monoisotopic molecular masses that were previously unknown in relation to prymnesins. A total of 39 prymnesin-like molecular features were identified across the 10 strains analyzed on the Agilent QTOF (**Supporting Table S-11**). 21 of these molecular

features had *m/z* for $[M+2H]^{2+}$ which corresponded to previously identified prymnesins,[25,32,33] and 18 of the molecular features had monoisotopic molecular masses that were previously unknown in relation to prymnesins. All compounds had different elution patterns with expected peak widths, showing no indication of false positives due to matrix interference, or co-elution of compound adducts and isotopologues from different compounds. **Fig. 4** also shows that even though the prymnesins are small peaks compared to the matrix of chlorophylls, lipids, and other cellular components, the algorithm could still differentiate these compounds. However future studies needs to show how the algorithm can work in a dirty matrix, but it will be very dependent on the peak picking and spectral clean up algorithms.

During analysis another metabolite was observed in the DCA-Hal data which was present in 6 strains. This metabolite however was not considered prymnesin-like due to its low monoisotopic ion mass (*m/z* 644.2413, $[M+H]^+$; $C_{28}H_{47}Cl_2NO_9S$). It had a characteristic isotope pattern of a halogenated metabolite and, due to its presence across many strains, it was considered an interesting addition to the prymnesin and prymnesin-like molecular features. The observation of this additional, lower molecular weight molecular feature, illustrates the ability of DCA to classify metabolites over wide molecular mass ranges in one automated process.

The MeHaloCoA method was also tested on a representative algae extract to compare the results to those of DCA-Hal (**Supporting Figure S-4**). Although MeHaloCoA did identify prymnesin-like chemical features (30 identified, compared to 76 identified by DCA-Hal, **Supporting Table S-12**), MeHaloCoA also resulted in significantly more false positives. These were attributed to peak saturation, the misclassification of multiply charged ions, and of ions greater than 800 u (**Supporting Table S-12** and **S-13**).

Analysis of algae extracts with DCAnalysis v1.07 revealed that the program in most cases allows for the processing of one complex algae extract (the formation of chemical features, compiling into molecular features, and DCA filtering) in less than 1 min, with an average processing time for an algae extract data file of 26 sec (**Supporting Tables S-10** and **S-11**), and an average processing time for a reference standard data file of 3 sec (**Supporting Table S-5**).

Conclusion

The two algorithms developed in this project, DCA-Sul and DCA-Hal, are effective tools for filtering complex chromatograms, and can help in identifying unknown metabolites of particular elemental compositions. By using the A+1 to A+2 isotope cluster spacing the algorithms are able to eliminate potential false positives from detector saturation and co-eluting dihydro/dehydro isomers, a common problem in complex chromatograms. The effectiveness of these tools may be enhanced by combining them with other techniques such as mass defect filtering or isotope cluster analysis. Applying mass defect filtering or an isotope cluster analysis to a DCA chromatogram has the potential to be even more selective than either of the techniques on their own.

DCA is highly dependent on the inter-isotopic mass accuracy of the MS. This study has shown that this method of filtering can be effective on instruments with inter-isotopic mass accuracies of approximately 1 ppm. Theoretical calculations suggest that for data acquired on an instrument with an inter-isotopic mass accuracy of approximately 1 ppm, DCA-Sul and DCA-Hal would be effective in filtering metabolites with masses up to 824 u and 1612 u, respectively. Metabolites with monoisotopic ion masses above these values can still be classified, as shown here with the prymnesins, however with an increasing degree of uncertainty proportional to increasing mass, up to the point of the theoretical limits of the decision boundaries, 1081 u for DCA-Sul and 2793 u for DCA-Hal.

Of the total number of unique molecular formulas compiled from the Antibase database, 12.6% of them were assigned to Group-Cl, for Marinlit this value was 23.8%. The greater proportion of Cl and Br containing unique molecular formulas within the Marinlit database may be expected due to greater availability of these elements in the marine environment, however this may suggest that DCA would be of particular interest to natural product chemists investigating organisms of marine origin, as demonstrated here with microalgae.

Acknowledgements

This study was funded by the Danish Council for Strategic Research through the project "HABFISH" (Project No. 0603-00449B). Agilent Technologies is acknowledged for the *Thought Leader Donation* of the 6545 UHPLC-QTOF.

We would also like to thank Dr Christopher Phippen at the Technical University of Denmark, Department of Biotechnology and Biomedicine, for his help during the project.

References

(1) Marinlit http://pubs.rsc.org/marinlit.
(2) Laatsch, H. Antibase 2012 - The Natural Compound Identifier http://www.wiley-vch.de/stmdata/antibase.php.
(3) Gordon W. Gribble. *Naturally Occurring Organohalogen Compounds - A Comprehensive Update*; Springer: Wien/New York, 2010.
(4) United Nations. *Stockholm Convention on Persistent Organic Pollutants*; 2001.
(5) World Health Organisation. **2015**, No. April.
(6) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779–787.
(7) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.
(8) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8* (4), 719–726.
(9) Scheltema, R. A.; Jankevics, A.; Jansen, R. C.; Swertz, M. A.; Breitling, R. *Anal. Chem.* **2011**, *83* (7), 2786–2793.
(10) Roullier, C.; Guitton, Y.; Valery, M.; Amand, S.; Prado, S.; Robiou du Pont, T.; Grovel, O.; Pouchus, Y. F. *Anal. Chem.* **2016**, acs.analchem.6b02128.
(11) Zhu, M.; Ma, L.; Zhang, D.; Ray, K.; Zhao, W.; Humphreys, W. G.; Skiles, G.; Sanders, M.; Zhang, H. *Pharmacology* **2006**, *34* (10), 1722–1733.
(12) Bateman, K. P.; Castro-Perez, J.; Wrona, M.; Shockcor, J. P.; Yu, K.; Oballa, R.; Nicoll-Griffith, D. A. *Rapid Commun. Mass Spectrom.* **2007**, *21* (9), 1485–1496.
(13) Zhang, H.; Zhang, D.; Ray, K.; Zhu, M. *J. Mass Spectrom.* **2009**, *44* (7), 999–1016.
(14) Rousu, T.; Pelkonen, O.; Tolonen, A. *Rapid Commun. Mass Spectrom.* **2009**, *23* (6), 843–855.
(15) Ekanayaka, E. A. P.; Celiz, M. D.; Jones, A. D. *Plant Physiol* **2015**, *167* (4), 1221–1232.
(16) Shah, M.; Meija, J.; Caruso, J. A. *Anal. Chem.* **2007**, *79* (3), 846–853.
(17) Thurman, E. M.; Ferrer, I. *Anal. Bioanal. Chem.* **2010**, *397* (7), 2807–2816.
(18) Ancleregg, R. J. *Anal. Chem.* **1981**, *53* (14), 2169–2171.
(19) Ionas, A. C.; Ballesteros Gómez, A.; Leonards, P. E. G.; Covaci, A. *J. Mass Spectrom.* **2015**, *50* (8), 1031–1038.
(20) Nagy, K.; Sandoz, L.; Craft, B. D.; Destaillats, F. *Food Addit. Contam.* **2011**, *28* (11), 1492–1500.
(21) Makarov, A.; Denisov, E.; Lange, O. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (8), 1391–1396.
(22) Radionova, A.; Filippov, I.; Derrick, P. J. *Mass Spectrom. Rev.* **2015**, *47* (12).
(23) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. *Anal. Chem.* **2008**, *80* (11), 3985–3990.
(24) Klitgaard, A.; Iversen, A.; Andersen, M. R.; Larsen, T. O.; Frisvad, J. C.; Nielsen, K. F. *Anal. Bioanal. Chem.* **2014**, *406* (7), 1933–1943.
(25) Rasmussen, S. A.; Meier, S.; Andersen, N. G.; Blossom, H. E.; Duus, J. Ø.; Nielsen, K. F.; Hansen, P. J.; Larsen, T. O. *J. Nat. Prod.* **2016**, *79* (9), 2250–2256.
(26) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9* (1), 504.
(27) Benton, H. P.; Want, E. J.; Ebbels, T. M. D. *Bioinformatics* **2010**, *26* (19), 2488–2489.
(28) Python Software Foundation. .
(29) Shipman, J. W. *Computer (Long. Beach. Calif).* **2013**, 1–118.
(30) Kind, T.; Fiehn, O. *BMC Bioinformatics* **2007**, *8*, 105.

(31) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84* (1), 283–289.
(32) Igarashi, T.; Satake, M.; Yasumoto, T. *J. Am. Chem. Soc.* **1996**, *118* (2), 479–480.
(33) Murata, M.; Yasumoto, T. *Nat. Prod. Rep.* **2000**, *17* (3), 293–314.

Figure Legend

Figure 1: Group separation comparison of the A+1 to A+2 and the A to A+2 isotope cluster spacings for the Marinlit natural products database

Comparison of group separation between the A+1 to A+2 and the A to A+2 isotope cluster spacings for the Marinlit database for *m/z* values for $[M+H]^+$ between 1000 and 2000. The dotted lines illustrate the separation of 0.00049 u between Group-S and Group-C in the A+1 to A+2 dataset. The group separation is reduced to 0.00007 u in the A to A+2 dataset, a 86% reduction, or equivalent to a reduction in inter-isotopic mass accuracy for separation from 0.33 ppm to 0.05 ppm at 1500 u. The y-axis' are aligned to an increase of 1.00342 u in the A to A+2 plot, this value is the expected difference for the two values for a saturated hydrocarbon with a monoisotopic mass of 1500 u.

Figure 2: A+1 to A+2 isotope cluster spacings of unique molecular formulas of the Marinlit natural products database

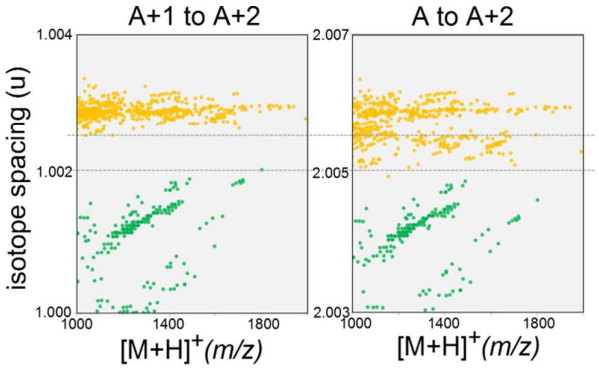
The molecular formulas of Marinlit plotted as their isotopic properties *m/z* of $[M+H]^+$ on the x-axis, and A+1 to A+2 isotope cluster spacing on the y-axis, split into three groups: Group-Cl (blue), metabolites containing Cl and/or Br; Group-S (green), metabolites containing S and no Cl/Br; and Group-C (red), the remaining metabolites of the Marinlit database. Molecular formula equations which defined the lower limit of Group-C: $C_nH_{0.49n}O_{0.8n}$ (solid yellow line), and $C_nH_{2n+0.65n}N_{0.65n}O_{0.8n}$ (dashed yellow line). Molecular formula equations which defined the upper limit of Group-S: $C_nH_{2n+2}S$ (solid green line), and $C_nH_{2n+0.65n}N_{0.65n}O_{0.8n}S$ (dashed green line). Molecular formula equations which defined the upper limit of Group-Cl: $C_nH_{2n+1}Cl$ (solid blue line), and $C_nH_{2n+0.65n+1}N_{0.65n}Br$ (dashed blue line). The heteroatom and H ratios for these hypothetical molecular formulas, for $n > 4$, fall within the limits suggested for molecular formula filtering.[30] Three outliers of Group-C (large black ringed points) are due to unusually high proportion of heteroatoms relative to C, molecular formulas: $C_2H_3IO_2$, $C_2H_2I_2O_2$, $C_3H_6As_4O_3$, from left to right in plot. Other outliers of Group-C are due to the incorporation of the less frequently occurring elements Mg, Ni, Zn and Cu.

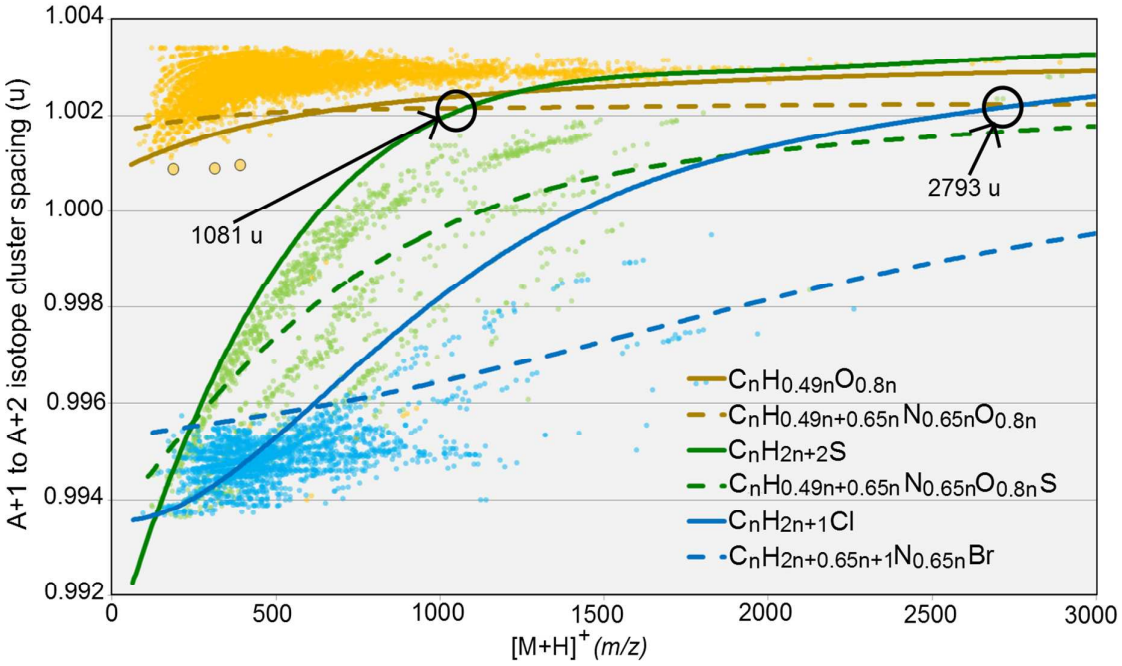
Figure 3: A:A+2 intensity ratios of unique molecular formulas of the Marinlit natural products database

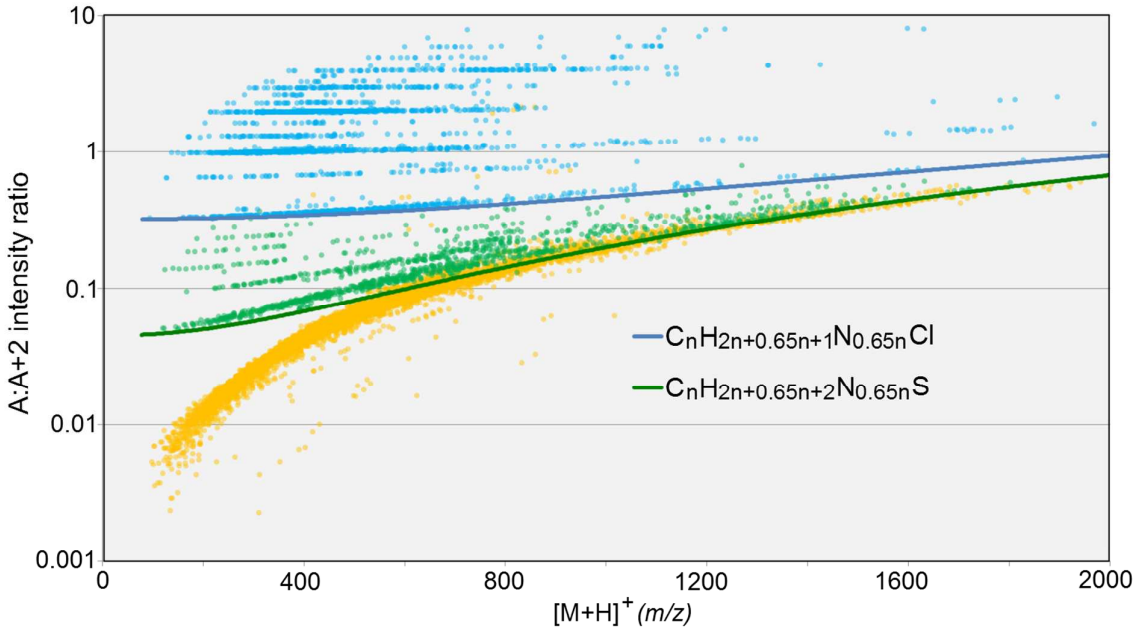
The molecular formulas of Marinlit plotted as their isotopic properties *m/z* of $[M+H]^+$ on the x-axis, and A:A+2 intensity ratio on the y-axis, split into three groups: Group-Cl (blue), metabolites containing Cl and/or Br; Group-S (green), metabolites containing S and no Cl/Br; and Group-C (yellow), the remaining metabolites of Marinlit database. The molecular formula equation which defined the lower boundary of Group-Cl: $C_nH_{2n+0.65n+1}N_{0.65n}Cl$ (solid blue line). The molecular formula equation which defined the lower boundary of Group-S: $C_nH_{2n+0.65n+2}N_{0.65n}S$ (solid green line).

Figure 4: DCA-Hal results of *P. parvum* strain K-0374

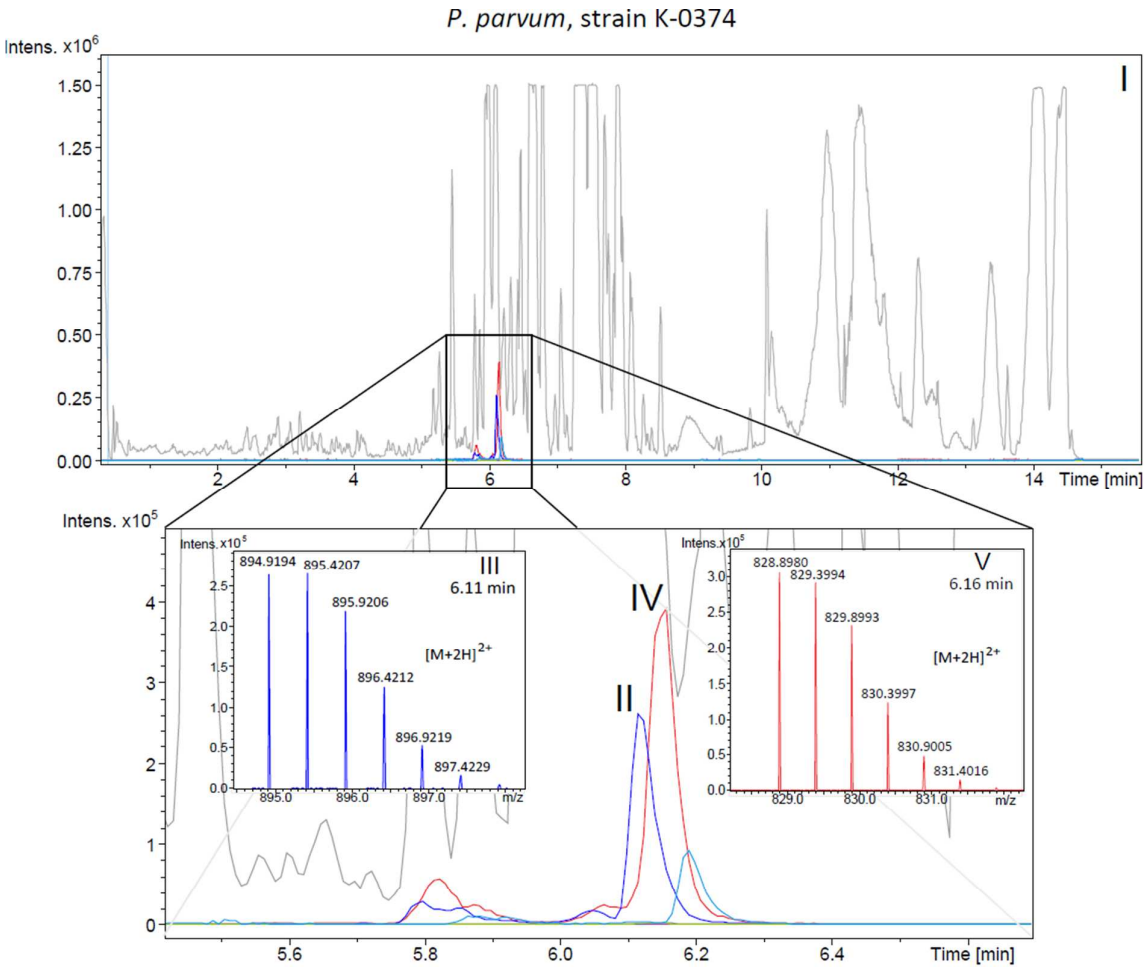
EIC of detected $[M+2H]^{2+}$ ions from DCA-Hal analysis on the crude extract of *P. parvum* strain K-0374 from the UHPLC-QTOF-1 system. I: The BPC of the crude extract in grey and the EIC of the $[M+2H]^{2+}$ masses of the molecular features detected by DCA-Hal in red and blue. II: The EIC of a mass and retention time identified by DCA-Hal. III: The mass spectra of the $[M+2H]^{2+}$ ion of the metabolite identified in II. IV: The EIC of a mass and retention time identified by DCA-Hal. V: The mass spectra of the $[M+2H]^{2+}$ ion of the metabolite identified in IV.


A+1 to A+2
A to A+2
isotope spacing (u)
1.004
1.002
1.000
2.007
2.005
2.003
1000
1400
1800
[M+H]+(m/z)


1.004
1.002
1.000
0.998
0.996
0.994
0.992
A+1 to A+2 isotope cluster spacing (u)
0
500
1000
1500
2000
2500
3000
[M+H]+ (m/z)
1081 u
2793 u
CnH0.49nO0.8n
CnH0.49n+0.65nN0.65nO0.8n
CnH2n+2S
CnH0.49n+0.65nN0.65nO0.8nS
CnH2n+1Cl
CnH2n+0.65n+1N0.65nBr


10
1
0.1
0.01
0.001
A:A+2 intensity ratio
0
400
800
1200
1600
2000
[M+H]+ (m/z)
CnH2n+0.65n+1N0.65nCl
CnH2n+0.65n+2N0.65nS

22


P. parvum, strain K-0374
Intens. x10^6
I
Time [min]
Intens. x10^5
III
6.11 min
894.9194
895.4207
895.9206
896.4212
896.9219
897.4229
[M+2H]2+
IV
II
V
6.16 min
828.8980
829.3994
829.8993
830.3997
830.9005
831.4016
m/z
Time [min]

For TOC Only