This item is the archived peer-reviewed author-version of:

InSourcerer : a high-throughput method to search for unknown metabolite modifications by mass spectrometry

Reference:

Mrzic Aida, Lermyte Frederik, Vu Trung N., Valkenborg Dirk, Laukens Kris.- InSourcerer : a high-throughput method to search for unknown metabolite modifications by mass spectrometry
Rapid communications in mass spectrometry - ISSN 0951-4198 - 31:17(2017), p. 1396-1404
Full text (Publisher's DOI): <https://doi.org/10.1002/RCM.7910>
To cite this reference: <http://hdl.handle.net/10067/1435090151162165141>

InSourcerer: a high-throughput method to search for unknown metabolite modifications by mass spectrometry.

Aida Mrzic^{5,6,+}, Frederik Lermyte^{1,2,4,7+}, Trung Nghia Vu^{5,6,8}, Dirk Valkenborg^{1,2,3,*} Kris Laukens^{5,6,*}

⁺ both authors share first authorship.

^{*} both authors share senior authorship and are corresponding author.

⁷ Currently with the School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom

⁸ Currently with the Department of Medical Epidemiology and Biostatistics, Karolinska Institute, 17177 Stockholm, Sweden.

To be submitted to RCM

Abstract

RATIONALE Using mass spectrometry, the analysis of known metabolite structures has become feasible in a systematic high-throughput fashion. Nevertheless, the identification of previously unknown structures remains challenging, partially because many unidentified variants originate from known molecules that underwent unexpected modifications. Here, we present a method for the discovery of unknown metabolite modifications and conjugate metabolite isoforms in a high-throughput fashion.

METHODS The method is based on user controlled in-source fragmentation which is used to induce loss of weakly bound modifications. This is followed by the comparison of product ions from in-source fragmentation and collision-induced dissociation. Diagonal MS²-MS³ matching allows the detection of unknown metabolite modifications, as well as substructure similarities. As the method relies heavily on the advantages of in-source fragmentation and its ability to ‘magically’ elucidate unknown modification, we have named it inSourcerer as a portmanteau of in-source and sorcerer.

RESULTS The method was evaluated using a set of 15 different cytokinin standards. Product ions from in-source fragmentation and collision-induced dissociation were compared. Hierarchical clustering revealed that good matches are due to the presence of common substructures. Plant leaf extract, spiked with a mix of all 15 standards, was used to demonstrate

the method's ability to detect these standards in a complex mixture, as well as confidently identify compounds already present in the plant material.

CONCLUSIONS Here we present a method that incorporates a classic LC-MS workflow with fragmentation models and computational algorithms. The assumptions upon which the concept of the method was built were shown to be valid and the method showed that in-source fragmentation can be used to pinpoint structural similarities and indicate the occurrence of a modification.

Keywords in-source fragmentation, metabolomics, metabolite identification, collision-induced dissociation, cytokinins

1. Introduction

Technological innovations have made the systematic profiling of a large number of small molecules in complex biological samples feasible, leading to a rapid growth of the field of metabolomics. Mass spectrometry (MS) has historically played a key role in the structural elucidation of small molecules.^[1, 2] Nevertheless, the metabolome remains extremely complex with an enormous diversity of small molecule structures. It is consequently still typical that besides the identification (and quantification) of known structures, a substantial fraction of the observed signals in MS-based metabolomics profiling experiments remains unidentified.

One factor that significantly contributes to this structural diversity is the fact that even well-characterized metabolites can undergo various modifications and interconversions *in vivo*, resulting in less known structural forms. Current bioinformatics tools perform very well with structures that are already known and present in metabolite databases, while unknown modifications and molecular structures are often overlooked and very hard to discover by software alone (*i.e.* without additional manual data interpretation). Although the analysis of known structures has become feasible in a systematic high-throughput fashion, the discovery and identification of previously unknown structures has not gained such throughput. There is a clear need for approaches to more systematically analyze and identify the unknown fraction of the metabolome.

Here we describe a comprehensive method which combines liquid chromatography (LC), with data-dependent acquisition tandem mass spectrometry (DDA-MSⁿ) measurements and subsequent data analysis and bioinformatics workflow. We propose to adjust the classical LC-MSⁿ workflow to screen for unknown metabolite structures and conjugate forms in the metabolome. The workflow consists of a unique but elegant integration of proven methodologies in LC-MS, fragmentation models and computational algorithms. The concept is based on user-controlled in-source fragmentation, which is used to induce loss of weakly bound modifications (*e.g.* glycosylation) and the subsequent analysis of spectral similarities across different fragmentation levels. The discovery of similarities across fragmentation levels can be assessed using computational techniques to pinpoint common substructures in the molecular ions.

In prior studies, in-source fragmentation has been utilized to help with metabolite identification.^[3, 4] The comparison of MS² fragmentation levels has previously been used to distinguish common subgroups on a spectral network scale.^[5, 6] Comparison of all-ion in-source fragmentation against all-ion HCD fragmentation has been performed with the aim to evaluate the efficiency of both fragmentation techniques with respect to complete fragmentation results.^[7, 8] A recent approach paired in-source fragmentation with correlation analysis to detect molecules of interest.^[9] In contrast, the objective of our method is to utilize in-source fragmentation to get rid of weakly bound modifications for subsequent analysis via diagonal MS²-MS³ matching. Since the method largely depends on in-source fragmentation, we have named it inSourcerer as a portmanteau of in-source and sorcerer.

2. Materials and Methods

2.1. Concept of the method

A mass spectrometry-based method is devised to screen for unknown modifications and conjugate metabolite isoforms in a high-throughput fashion. The main principle is outlined in Figure 1. An analyte will elute from the liquid chromatography (LC) column prior to ionization by electrospray ionization (ESI) (Figure 1.a). The source conditions can be adjusted such that in-source fragmentation is induced. Often in-source fragmentation is regarded as a nuisance;^[10] however, we consider in-source fragmentation as beneficial, as it can be fine-tuned such that part of the metabolite ion population will lose a labile group, such as a metabolic modification, whilst the rest of the ions will maintain their intact form. This requires that the in-source transitions are reproducible and comparable to collision-induced dissociation (CID) in an ion trap, *i.e.*, labile modifications will be released first. As in-source fragmentation and CID in an ion trap both gradually increase the internal energy of an ion *via* many collisions, the assumption of distribution of this energy over all internal degrees of freedom and resulting thermodynamic control of the process seems reasonable.^[11] Consequently, the ionized metabolite population is split into two forms, an intact form and an in-source fragmented form, which has lost a labile modification.

This split results in two observed peaks for that analyte in the same MS^1 spectrum, since in-source fragmentation occurs after LC separation (Figure 1.b). An optimized LC separation will ensure that a DDA method will select both the intact and in-source fragmented ions for further interrogation in MS^2 , *i.e.* CID fragmentation in the ion trap. Since we assume that in-source and CID fragmentation will generate the same transitions; the intact form will release the same labile modification first. It is worth mentioning here that the CID fragmentation in the ion trap is a fragmentation technique that heats up a molecule based on its eigenfrequency. Since the intact molecule will easily detach a labile modification, the CID reaction is stopped after releasing the fragment (since the molecule's eigenfrequency is altered due to the mass loss), resulting in an abundant peak in the MS^2 spectrum that corresponds to the in-source fragmented form in the MS^1 spectrum. The most abundant MS^2 peak will be selected for further CID fragmentation and the result is captured in an MS^3 spectrum. Recall that the in-source fragmented form also has to be selected by DDA for CID fragmentation, which results in an MS^2 spectrum of the in-source fragmented form that corresponds to the MS^3 spectrum of the intact form. As before, the most abundant peak of the MS^2 spectrum is selected for further investigation in MS^3 . For the sake of clarity, we will introduce a new notation for different levels of MS of in-source fragmented ions at this point: CID fragmentation (in the ion trap) of the in-source fragmented ion peak in the MS^1 spectrum results in an MS^2 spectrum denoted by MS_{INS}^2 ; further fragmentation of the most abundant peak in spectrum MS_{INS}^2 results in an MS^3 spectrum denoted by MS_{INS}^3 .

This experimental set-up produces a rich source of information that can be used to discover correlations between the different fragmentation levels. Also, in the commonly used hybrid LTQ-Orbitrap setup, MS^2 is normally performed in the LTQ, meaning the mass accuracy and resolution are inferior to those of the MS^1 scan, which is usually performed in the Orbitrap. By

inducing in-source fragmentation, we already have fragment information at the MS^1 -level (with high resolution/accuracy) as opposed to the results in MS^2 . This allows one to obtain high-resolution and high-mass accuracy MS^2 information, which is normally inconvenient using LTQ-Orbitrap hybrid instruments, as the scan rate of the Orbitrap is considerably slower than that of the LTQ (especially for slightly older instruments still commonly in use). Furthermore, it should be mentioned that performing MS^3 of the in-source fragmented ions is equivalent to MS^4 information of the intact ion. So once a correlation between the multiple MS^n information layers has been established, analysis protocols like spectral trees can be used for de novo elucidation of the metabolite structure^[12, 13]. However, our main objective is diagonal MS^{n-1} matching, especially MS^2_{INS} - MS^3 matching. The principle that the MS^2_{INS} spectrum corresponds to the MS^3 spectrum can be further used to detect unknown metabolite modifications.

MS^2_{INS} and MS^3 spectra can also be used in a clustering approach within and between the different MS layers to mine for substructures. The shared substructures can be identified since molecules with the same basic structure but different modifications (*e.g.* a glycosylated and phosphorylated variant) generate the same product ions on different fragmentation levels. This of course implicitly assumes not only that different forms of collision-induced dissociation will result in the same transition, but also that fragmentation of the same common basic structure carrying different modifications (see Figure 1) will typically result in loss of the modification, while keeping the basic structure intact. However, as these modifications are typically linked *via* bonds to a heteroatom, and often need to be easily added or removed from a biological perspective, the assumption that this link will be (one of) the most energetically labile in the molecule is also a reasonable one.

It is worth providing a list of assumptions upon which the concepts of inSourcerer are built. These should be kept in mind when evaluating the result of the experiments:

- 1) Source conditions can be optimized such that both intact form and in-source fragmented form will co-exist.
- 2) LC-MS can be optimized such that a DDA will select both forms for further interrogation on MS^2/MS^3 .
- 3) In-source CID and CID in the ion trap will generate reproducible transitions.
 - 3.1) Fragmentation of the same common basic structure carrying different modifications will typically result in loss of the modification while preserving the intact basic structure.

2.2. Experimental setup

Chemicals

Typical examples of a group of important small molecules that naturally occur in a diversity of forms due to such interconversions are cytokinins. Upon the discovery of their role in the promotion of plant cell division in the fifties, cytokinins have always been the subject of intense study.^[14] Natural cytokinins are N⁶-substituted adenine derivatives.^[15] As a consequence of a variety of cytokinin biosynthetic and metabolic enzymes, a number of different cytokinin species and conjugates are present in plants (e.g. trans-zeatin, cis-zeatin, dihydrozeatin, etc.) that are recognized by downstream proteins with different affinities and that generate distinct downstream responses.^[15-17] Cytokinins are thus diverse, this diversity is biologically important and different cytokinins share several common structural features. Together with the fact that several different cytokinins can be purchased in pure form, these aspects make cytokinins an interesting class of small biomolecules to study new workflows for the systematic elucidation of the uncharacterized fraction of the metabolome.

Sample origins and preparation

H₂O and acetonitrile (ULC/MS grade) were obtained from BioSolve (Valkenswaard, the Netherlands). Formic acid (99%, for analysis) was obtained from Acros Organics (Pittsburgh, PA). Cytokinin standards were acquired from Olchemim Ltd (Olomouc, Czech Republic). Preparation of *Psychotria kirkii* samples was performed according to the protocol described in Prinsen et al. (1998)^[18].

LC-MS setup

All experiments were performed using a Thermo Finnigan LTQ (ThermoFisher Scientific, Bremen, Germany) mass spectrometer, using an ESI voltage of 3.8 kV, a capillary voltage of 50 V and a tube lens voltage of 110 V. External calibration was performed using a standard calibration mixture containing caffeine, MRFA, and Ultramark 1621 (Pierce LTQ ESI Positive Ion Calibration Solution, catalog #88322). In the data-dependent experiments, the 15 most intense peaks in the mass spectrum were subjected to CID fragmentation. The single most intense peak in each MS² spectrum was isolated and also subjected to CID fragmentation. Fragmentation in both MS² and MS³ scans was performed using normalized collision energy of 25 V and an activation time of 30 ms.

For chromatographic separation, a Waters Acquity UPLC system was used in combination with an Acquity UPLC BEH C18 1.7µm 2.1 mm × 100 mm Column (Waters, Milford USA) and an Acquity UPLC VanGuard TM Pre-Column 1.7 µm 2.1 mm × 5 mm (Waters, Milford USA). As performing a “top 15 MS² – top 1 MS³” DDA experiment increases the cycle time of the mass spectrometer considerably, chromatographic peak width needed to be increased, while retaining separation. For this, a custom gradient was developed as follows (where A = 0.1% formic acid in H₂O and B = 0.1% formic acid in CH₃CN), with the flow rate set to 0.03 mL/min: t=0-30 min, 95% A (5% B); t=30.1-35 min, 40% A (60% B); t=35.1-36 min, 1% A (99% B); t=36.1-41 min, 95% A (5% B). This resulted in chromatographic peaks approximately 42 seconds wide.

Datasets

dataset 1: This dataset served to validate the assumption that in-source fragmentation would yield comparable transitions to conventional CID fragmentation in the ion trap, as well as the assumption that instrument parameters could be optimized to generate a combination of intact and in-source fragmented ions for a wide variety of analytes. For this, we used a set of 15 cytokinin standards, ranging between 203 and 403 Da, *i.e.* zeatin, dihydrozeatin, zeatin riboside, zeatin-N-glucoside, zeatin-O-glucoside, 2-methylthiozeatin, 2-methylthiozeatin riboside, isopentenyladenine, isopentenyladenosine, isopentenyladenine-N-glucoside, methylthio-isopentenyladenine, methylthio-isopentenyladenosine, ortho-topolin, ortho-topolin-9-glucoside, ortho-topolin riboside. The structures of these compounds can be found in Table 1. These experiments were performed through direct infusion of these compounds (1 μ M in 50/50/0.1 H₂O/acetonitrile/formic acid) at a flow rate of 10 μ l/min. The compounds were infused one by one in order to avoid potential ion suppression as well as ambiguity as to which precursor was isolated (as some of the standards have the same mass). Rather than using data dependent acquisition, the $[M+H]^+$ ion generated from each of the 15 standards was manually selected in these experiments and used to generate MS² and MS³ (of the most abundant fragment in MS²) spectra. Subsequently, the same was done under harsher conditions leading to in-source fragmentation. This allowed us to both optimize instrument conditions to induce efficient in-source and conventional CID fragmentation for the range of standards used here, as well as verify the assumption that in-source fragmentation and conventional CID lead to formation of the same (most abundant) fragment.

dataset 2: In this experiment, we sought to validate the assumption that a data-dependent LC-MSⁿ method could be optimized such that both the intact and in-source fragmented ions belonging to the same precursor would be selected in a data-dependent acquisition set-up. For this purpose, a sample with limited complexity (*i.e.* a solution containing all of the previously mentioned 15 cytokine standards; dissolved at concentrations of 10, 1, and 0.1 μ M) was evaluated by LC-MSⁿ. This dataset was also used to test the use of MassBank for automated identification.^[19]

dataset 3: As the next step, a metabolite extract from *Psychotria kirkii* was spiked with 0.5 μ M of a mix of all 15 standards and analyzed using LC-MSⁿ. This aim of this experiment was to verify that the standard compounds would be detected in a complex sample matrix, and to investigate whether the instrument method optimized for dataset 2 would be capable of detecting compounds already present in the plant extract.

Bioinformatic implementation

To compare two spectra, the open source spectral library search implementation SpectraST was used.^[20] The software runs on Windows XP embedded within the TPP framework [(TPP v4.4 VUVUZELA rev 1, Build 201010121551 (MinGW))].^[21] This software performs spectrum-spectrum matching (SSM) using dot product as a method for computing the similarity between spectra. The precursor masses were removed to avoid their impact on spectral similarity score calculation (which would lead to a similarity overestimation).

SpectraST uses a so-called binning or bucketing technique to divide a certain m/z range into equidistant intervals to transform the observed mass spectra into equal-size vectors, which are then further used for matching. Typically, rounding is involved in this binning procedure and it can occur that peaks that do match between two spectra according to the instrument precision, are placed in distinct bins. In this implementation, to avoid the case of corresponding peaks of two spectra being put in wrong bins by SpectraST, each MS_{INS}^2 spectrum is matched to a MS^3 spectrum with an m/z error tolerance of 0.5 Da. We removed noise peaks by eliminating peaks with intensity less than 10.00 and 0.01 (arbitrary units provided by the instrument software) for MS_{INS}^2 and MS^3 spectra, respectively.

The results of the spectral comparison are used for consecutive database searching and are filtered as follows: (1) The mass difference between the parents in MS_{INS}^2 and MS^3 is greater than 60 Da (*i.e.* we expect the absolute value of the modification mass to be greater than 60 Da); (2) the similarity score (dot product value) is larger than 0.5 (which is considered a good cutoff value).^[22] A higher score indicates a better match. It should be noted that the cutoff value of 60 Da is user-defined and serves our application to screen for glycosylated conjugates. A different value can be set by the user depending on the application.

3. Results and discussion

3.1. Assessment of the assumptions using molecular standards

In the first experiment, we evaluated the experimental application of the method using a set of 15 different cytokinin standards. These were infused and subjected to in-source fragmentation using the instrument settings described in the Experimental section, which were found to be optimal for the detection of both intact and in-source fragmented standards simultaneously. The similarity between the MS_{INS}^2 spectrum of the in-source generated fragment and the corresponding MS^3 spectrum of the intact precursor was determined. Figure 2 plots the distribution of the resulting similarity scores, each based on 20 independent MS_{INS}^2 - MS^3 comparisons. For 10 out of 15 comparisons, this similarity score was consistently very high, *i.e.* above 0.9. The average score for 2-methylthiozeatin was also greater than 0.9, but a higher inter-spectra variability was observed. For the 4 cytokinins consisting of adenine and a labile modification, *i.e.* zeatin, dihydrozeatin, ortho-topolin, and isopentenyladenine, the similarity score was lower, though on average still above 0.7. A possible explanation for this lower similarity is the high stability of the resulting adenine ‘core’ that leads to limited information in the MS_{INS}^2 and MS^3 spectra (see Supporting Information S1) which are compared using SpectraST (recall that the precursor signal is not included in calculation of the similarity score). The high similarity scores observed conclusively indicate that that CID fragmentation in the ion trap leads to the same transition as in-source fragmentation.

We can extend the diagonal tandem MS matching to compare all MS^3 spectra to all MS_{INS}^2 spectra. In this case, for each tandem MS pair between the MS layers, the algorithm will return a similarity score that can be visually depicted in a matrix. The result, presented in the heatmap in Figure 3(a), shows that the majority of inter-spectral comparisons exhibit low similarity as indicated by the light colors. Matches from the same precursor are present on the diagonal and consistently score high or very high as indicated by the dark blue colors (see Supporting Information S1 for spectral matches). Some other molecular classes yield high inter-spectral matches. To reveal whether any common substructure underlies the high matches for these cases, we performed hierarchical clustering using the inter-spectral match score as a distance score. Indeed, some highly similar analytes could be found based on spectral correlations displayed in Figure 3(b); the same product ions are generated from precursor ions that share the same basic structure, but contain different labile modifications (see scenario (c) in Figure 1). This outcome could have been expected as our experimental design contains cytokines that indeed share some common substructures. Specifically, this is shown by the high similarity scores found for (1) isopentenyladenosine and isopentenyladenine-N-glucoside, (2) ortho-topolin riboside and ortho-topolin glucoside, (3) zeatin-N-glucoside and zeatin-riboside, (4) methylthio-isopentenyladenine and 2-methylthiozeatin, (5) zeatin-O-glucoside and zeatin riboside, and (6) zeatin-O-glucoside and zeatin-N-glucoside (see Supporting Information S2).

3.2. Application to LC-MS with spiked samples

A mix of all 15 cytokinin standards was subjected to data-dependent LC-MS analysis using in-source fragmentation as described in the Experimental section. This experiment was performed to test whether both the intact and in-source fragmented ions would be selected for tandem MS in the ion trap in data dependent acquisition. As we also wished to estimate the sensitivity of this method, the experiment was performed with analyte concentrations of 10^{-7} , 10^{-6} and 10^{-5} M (Table 2). At 10^{-5} M, spectral matches for all 15 compounds except for 2-methylthiozeatin were detected. At 10^{-6} M, matches for dihydrozeatin and ortho-topolin were also no longer detected. This observation is in agreement with the results from Section 3.1, in which the three components which seem most difficult to detect in the standard mix, also did not consistently result in very high similarity scores, due to limited information in the MS^2_{INS} and MS^3 spectra. Interestingly, isopentenyladenine and zeatin – which also showed fairly low pairwise matching scores in Section 3.1 – were detected at a concentration of 10^{-6} M in the standard mix. At 10^{-7} M, none of the 15 standards were detected. Retention times, parent masses, fragment masses and intensities for this dataset at each concentration were provided as Supporting Information S3. We also found that 11 of the 14 detected standards (at 10^{-5} M) were correctly identified using MassBank. Upon further analysis of the remaining 3 (ortho-topolin, ortho-topolin-9-glucoside and ortho-topolin-riboside), it was found that these molecules were not present in the MassBank database.

With this experiment, we have illustrated that instrument setup can be optimized in such a way that in data-dependent acquisition mode, the in-source fragmented ions and intact ions are selected for MS^2 and MS^3 interrogation. The detection limit depends on sample complexity, chromatographic separation and the speed of MS acquisition. LC and MS parameters can be customized to optimize the detection of low-abundance metabolites. In the next experiment, we performed the same analysis in biological samples that were spiked with molecular standards.

3.3. Implementation in a complex sample screening workflow

Leaf extract from *Psychotria kirkii* was prepared according to Prinsen et al (1998).^[18] This extract was then diluted in a 1:1 ratio with the 1 μ M standard mix described in the previous section, yielding a final concentration of 500 nM for the standards. The resulting sample was then analyzed using the inSourcerer method, *i.e.* data-dependent LC-MS with controlled in-source fragmentation. This experiment was performed for two reasons; first of all, to test whether the standards would still be detected in a fairly low concentration while in a complex matrix, and second of all, to test whether this method would be able to confidently identify any biomolecules already present in the plant material.

All tandem MS data were also queried against the MassBank database via spectrum-to-spectrum matching, similarly as for datasets 1 and 2. Spectra with the same parent mass as the query spectrum were extracted from the database and compared to the observed spectrum,

again using the dot product as a similarity metric with 0.7 cutoff value for automated identification.

Eight of the fifteen standard compounds (specifically, zeatin, zeatin riboside, zeatin-N-glucoside, zeatin-O-glucoside, 2-methylthiozeatin riboside, isopentenyladenine, isopentenyladenosine, and methylthio-isopentenyladenosine) were identified by diagonal matching MS_{INS}^2 - MS^3 and automated identification (Figure 4). It was found upon manual analysis of the data that spectral matches for ortho-topolin and ortho-topolin riboside had also been detected, but, as previously mentioned, these are not present in the MassBank database and therefore could not be identified. These findings illustrate that the inSourcerer method is able to find diagonal matches between MS_{INS}^2 and MS^3 signal in a complex matrix (see Supporting Information S4 for retention times, parent masses, fragment masses and intensities for this dataset).

In addition to this, three common flavonoids already present in the leaf extract, quercetin-3- β -O-galactoside, isorhamnetin-3-O-glucoside and naringenin-7-O-glucoside, were also confidently identified by this method. In order to be considered a confident identification, we required that the matches resulting from both the in-source fragmented precursor (MS^1) - MS^2 and MS_{INS}^2 - MS^3 spectra were consistent. To further verify these matches, manual interpretation of the MS_{INS}^2 and MS^3 spectra was performed and the identity of these compounds was confirmed. Other spectral matches were found in the sample, but these could not be confidently identified according to our criteria.

4. Conclusions

We have introduced a novel method, inSourcerer, for the detection of unknown metabolite modifications based on in-source fragmentation and the subsequent comparison of different fragmentation levels. The results show that the assumptions upon which the concept of inSourcerer is built, are valid: (1) the same optimized instrument settings could be used for simultaneous detection of intact and in-source fragmented ions for a range of standards used in the experiments; (2) LC-MS setup can be optimized in such a way that in data-dependent acquisition mode the in-source fragmented ion and intact ion are both selected for MS^2 and MS^3 interrogation; (3) conventional CID and in-source CID will generate reproducible transitions. As the timescale of the CID process implies that the thermodynamically most unstable group is lost first, this can allow us to easily assign analytes to a certain class of compounds, while the different species can be distinguished by the different mass of the intact (*i.e.* not the product of in-source fragmentation) precursor (e.g. isopentenyladenosine 336 Da, isopentenyladenine-N-glucoside 366 Da). Isomers such as zeatin-N-glucoside/zeatin-O-glucoside, consisting of the same basic structure carrying the same labile modification in a different position, however, cannot be easily distinguished by this method, but can be separated by liquid chromatography methods.

There are several extensions and applications of the presented method. For example, the search could be extended to look for the biological conjugates over the entire LC run, whereas for now the comparison of the data-dependent acquisition cascade has to be connected within a single MS¹ scan. This could then be used for substructure mining, in a similar way as indicated in Figure 3. Next, the in-source fragmentation could be quantified by screening for MS² base peaks present in the corresponding MS¹ scans. The use of a faster mass spectrometer with higher mass accuracy and resolution, such as the LTQ Orbitrap Velos or the Orbitrap Fusion Tribrid in this workflow would lead to a greater number of confident identifications in a high-throughput fashion. The advantage of higher resolution and mass accuracy is obvious, while increased speed of tandem MS experiments would allow the current method to be used with narrower LC-peaks, so that a more conventional LC-method (with higher resolution and peak capacity) could be used. This would improve the sensitivity of the method by reducing undersampling, which is especially important as splitting the precursor population between an intact and in-source fragmented form, implies a tradeoff between sensitivity and the additional information provided by the inSourcerer method.

In this study, we used one optimal setting for the production of in-source fragments. In case of a heterogeneous sample, where multiple molecular classes are present (*e.g.* metabolites and peptides), it is opportune to use different in-source settings to optimally produce fragments for both cases. In principle, this can be achieved by continuously switching between the different regimes in a single run; however, the data structure would then become more complex by convoluting different MS² and MS³ spectra with source conditions. Furthermore, for DDA, selecting the same compound in its in-source fragmented and intact form under the same source conditions would potentially become less likely. As a result, it is better to perform this type of experiment *via* two separate runs with different (but constant within each run) in-source settings.

Switching between positive and negative ion mode would allow interrogation of modifications which are simply observed as neutral losses in positive mode, *e.g.* a sugar moiety. Combining data from both ion modes could potentially reveal more structural information about the analytes.

References

- [1] K. Scheubert, F. Hufsky, S. Bocker. Computational mass spectrometry for small molecules. *Journal of Cheminformatics* **2013**, 5.
- [2] F. Hufsky, K. Scheubert, S. Bocker. Computational mass spectrometry for small-molecule fragmentation. *Trac-Trends in Analytical Chemistry* **2014**, 53, 41.
- [3] J. C. Dominguez-Romero, J. F. Garcia-Reyes, R. Martinez-Romero, P. Berton, E. Martinez-Lara, M. L. Del Moral-Leal, A. Molina-Diaz. Combined data mining strategy for the systematic identification of sport drug metabolites in urine by liquid chromatography time-of-flight mass spectrometry. *Analytica Chimica Acta* **2013**, 761, 1.
- [4] S. Heiling, S. Khanal, A. Barsch, G. Zurek, I. T. Baldwin, E. Gaquerel. Using the knowns to discover the unknowns: MS-based dereplication uncovers structural diversity in 17-hydroxygeranyllinalool diterpene glycoside production in the Solanaceae. *Plant Journal* **2016**, 85, 561.
- [5] J. Y. Yang, L. M. Sanchez, C. M. Rath, X. T. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, W. R. Wong, R. G. Linington, L. X. Zhang, H. M. Debonisi, W. H. Gerwick, P. C. Dorrestein. Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* **2013**, 76, 1686.
- [6] L. Sanchez, J. Y. Yang, C. Rath, P. Boudreau, X. Liu, N. Bruns, A. Wodtke, R. d. Felicio, A. Fenner, E. Glukhov, W. Gerwick, P. Dorrestein. in *Proceedings of the 61st ASMS Conference on Mass Spectrometry and Allied Topics*, Minneapolis, Minnesota, **2013**.
- [7] C. D. Broeckling, F. A. Afsar, S. Neumann, A. Ben-Hur, J. E. Prenni. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Analytical Chemistry* **2014**, 86, 6812.
- [8] C. Broeckling, J. Prenni. in *the 61st ASMS Conference on Mass Spectrometry and Allied Topics*, Minneapolis, Minnesota, **2013**.
- [9] J. Godzien, E. G. Armitage, S. Angulo, M. P. Martinez-Alcazar, V. Alonso-Herranz, A. Otero, A. Lopez-Gonzalvez, C. Barbas. In-source fragmentation and correlation analysis as tools for metabolite identification exemplified with CE-TOF untargeted metabolomics. *Electrophoresis* **2015**.
- [10] Y. F. Xu, W. Lu, J. D. Rabinowitz. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Anal Chem* **2015**, 87, 2273.
- [11] S. A. McLuckey, D. E. Goeringer. Slow heating methods in tandem mass spectrometry. *Journal of Mass Spectrometry* **1997**, 32, 461.
- [12] J. E. Peironcelly, M. Rojas-Cherto, A. Tas, R. Vreeken, T. Reijmers, L. Coulier, T. Hankemeier. Automated Pipeline for De Novo Metabolite Identification Using Mass-Spectrometry-Based Metabolomics. *Analytical Chemistry* **2013**, 85, 3576.
- [13] M. Rojas-Cherto, M. van Vliet, J. E. Peironcelly, R. van Doorn, M. Kooyman, T. te Beek, M. A. van Driel, T. Hankemeier, T. Reijmers. MetiTree: a web application to organize and process high-resolution multi-stage mass spectrometry metabolomics data. *Bioinformatics* **2012**, 28, 2707.
- [14] C. O. Miller, F. Skoog, M. H. Vonsaltza, F. M. Strong. Kinetin, a Cell Division Factor from Deoxyribonucleic Acid. *Journal of the American Chemical Society* **1955**, 77, 1392.
- [15] J. J. Kieber, G. E. Schaller. The perception of cytokinin: a story 50 years in the making. *Plant Physiol* **2010**, 154, 487.
- [16] K. Yonekura-Sakakibara, M. Kojima, T. Yamaya, H. Sakakibara. Molecular characterization of cytokinin-responsive histidine kinases in maize. Differential ligand preferences and response to cis-zeatin. *Plant Physiol* **2004**, 134, 1654.
- [17] G. A. Romanov, S. N. Lomin, T. Schmulling. Biochemical characteristics and ligand-binding properties of Arabidopsis cytokinin receptor AHK3 compared to CRE1/AHK4 as revealed by a direct binding assay. *J Exp Bot* **2006**, 57, 4051.

- [18] E. Prinsen, W. Van Dongen, E. L. Esmans, H. A. Van Onckelen. Micro and capillary liquid chromatography tandem mass spectrometry: a new dimension in phytohormone research. *Journal of Chromatography A* **1998**, 826, 25.
- [19] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **2010**, 45, 703.
- [20] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, R. Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, 7, 655.
- [21] A. Keller, J. Eng, N. Zhang, X. J. Li, R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **2005**, 1, 2005 0017.
- [22] T. N. Vu, W. Bittremieux, D. Valkenborg, B. Goethals, F. Lemiére, K. Laukens. Efficient Reduction of Candidate Matches in Peptide Spectrum Library Searching Using the Top k Most Intense Peaks. *Journal of Proteome Research* **2014**, 13, 4175.

Acknowledgements:

This work was supported by SBO grant “InSPECtor” (120025) of the Flemish agency for Innovation by Science and Technology (IWT). F.L. thanks the Research Foundation – Flanders (FWO) for funding a PhD fellowship.

All authors wish to thank the reviewers for their insightful comments.

Table 1. Chemical structures of the 15 cytokinin standards used in this study.

Dihydrozeatin (DHZ)	Isopentenyladenine (iP)	Isopentenyladenine-N-glucoside (iPG)

Isopentenyladenosine (iPA)	methylthio-isopentenyladenine (ms-iP)	methylthio-isopentenyladenosine (ms-iPA)

2-methylthiozeatin (ms-zeatin)	2-methylthiozeatin riboside (ms-ZR)	ortho-topolin (ortho-top)

ortho-topolin-9-glucoside (o-top-G)	ortho-topolin-riboside (o-top-R)	Zeatin (zeatin)

zeatin-N-glucoside (zng)	zeatin-O-glucoside (t-ZO-glu)	zeatin riboside (tZR)

Table 2. Detection of compounds present in datasets 2 and 3 at different concentrations.

Compound	dataset 2			dataset 3
	10 μ M	1 μ M	0.1 μ M	0.5 μ M
Dihydrozeatin	+	-	-	-
Isopentenyladenine	+	+	-	+
Isopentenyladenine-N-glucoside	+	+	-	-
Isopentenyladenosine	+	+	-	+
Methylthio-isopentenyladenine	+	+	-	-
Methylthio-isopentenyladenosine	+	+	-	+
2-methylthiozeatin	-	-	-	-
2-methylthiozeatin riboside	+	+	-	+
Ortho-topolin	+	-	-	+
Ortho-topolin-9-glucoside	+	+	-	-
Ortho-topolin riboside	+	+	-	+
Zeatin	+	+	-	+
Zeatin-N-glucoside	+	+	-	+
Zeatin-O-glucoside	+	+	-	+
Zeatin riboside	+	+	-	+

+, the compound is detected; -, the compound is not detected

Figures

Figure 1: Outline of the concept. (a) Part of the ion population will lose a labile modification in the ESI source due to in-source fragmentation. (b) This controlled in-source fragmentation will produce two observed peaks for that analyte in the same MS¹ spectrum. After selection of both observed peaks by data-dependent acquisition for tandem MS, the most abundant peak in the MS² spectrum of the intact form corresponds to the in-source fragmented form in the MS¹ spectrum; this relation is visualized by the thin diagonal arrow. The MS² spectrum of the in-source fragmented form is denoted by MS_{INS}^2 and corresponds to the MS³ spectrum of the most abundant MS² peak from the intact form. The latter relation is illustrated by the wide diagonal arrow. The MS³ spectrum of the most abundant peak in MS_{INS}^2 is denoted by MS_{INS}^3 and represents pseudo-MS⁴ level information. (c) Metabolite consisting of the same basic structure and a different labile modification. The modification will be cleaved off resulting in a MS² - MS³ match between MS² and MS³ level information, or could be clustered within MS² or MS³ level information indicated by the thin and wide horizontal arrows, respectively.

Figure 2: Pairwise matching score between MS^2_{INS} and MS^3 runs: box plot for pairwise matches. Each box is composed out of 20x20 SpectraST runs of the MS^2_{INS} run against the equivalent MS^3 run for the same compound.

a)

b)

Figure 3: Pairwise matching score between MS_{INS}^2 and MS^3 runs: (a) Heatmap for all (unpaired) runs. (b) Heatmap for all runs including hierarchical clustering. Each box is the average value of the 20x20 similarity scores of SpectraST between the 20 spectra of MS_{INS}^2 (columns) and the 20 spectra of MS^3 (row).

Figure 4: Example of the MS_{INS}^2 - MS^3 match for zeatin riboside (similarity score: 0.99).