Phytochemistry Letters 21 (2017) 297–305

Contents lists available at ScienceDirect

Phytochemistry Letters

journal homepage: www.elsevier.com/locate/phytol

Short communication

Successes and pitfalls in automated dereplication strategy using liquid chromatography coupled to mass spectrometry data: A CASMI 2016 experience

Samuel Bertrand[a,b,*], Yann Guitton[c], Catherine Roullier[a,b]

[a] Groupe Mer, Molécules, Santé-EA 2160, UFR des Sciences Pharmaceutiques et Biologiques, Université de Nantes, France
[b] ThalassOMICS metabolomics facility, Plateforme Corsaire, Biogenouest, 44035 Nantes, France
[c] Laboratoire d'Etude des Re'sidus et Contaminants dans les Aliments (LABERCA), LUNAM Universite', Oniris, Nantes 44307, France

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 6 December 2016
Accepted 19 December 2016
Available online 15 January 2017

Keywords:
Annotation
Automated method
CASMI contest
Dereplication
LC–HRMS
Natural products

ABSTRACT

Automated annotation of data, originating from liquid chromatography coupled to high-resolution mass spectrometry profiles (LC–HRMS), remains a highly challenging task. Therefore, the Critical Assessment of Small Molecule Identification (CASMI) Contest (http://casmi-contest.org/) represents a unique opportunity to blindly evaluate annotation workflows. The 2016 CASMI contest consisted of 16 LC–HRMS/MS profiles with 18 detected peaks to annotate. Those peaks corresponded to compounds from natural origin. An R script based on the XCMS, IPO, RMassBank, CAMERA and MeHaloCoA packages was devised. Two other external tools: SIRIUS3 and CFM-ID were also integrated for molecular formulae and *in silico* fragmentation calculation, respectively. This script was used to perform peak picking, spectral interpretation, molecular formula determination and database search for structural determination. Finally, the structures were further discriminated based on *in silico* fragmentation. After the release of the CASMI contest solutions, successes and failures of the proposed script were investigated. In most cases, no differences were observed in the rank of the correct structure when using raw LC–HRMS data or manually obtained MS and MS/MS spectra. However, the study of the few cases where differences were detected tends to show that automatic detection of MS^2 data within the raw LC–MS data yielded more accurate identification. The failures in proposing the correct structure within the submission list were related to the absence of the right structure in the interrogated databases. However, very close structure were proposed in first rank indicating that such approaches are able to rapidly determine the carbon skeleton of the structure; the medium rank of the correct structure in the proposed list for each peak of interest being 2^{nd}.

1. Introduction

In the natural product (NP) field, early identification of the crude extract composition remains a very challenging task (Wolfender et al., 2015). To achieve this task high performance liquid chromatography (HPLC) coupled to high-resolution mass spectrometry (HRMS) is considered as a very promising approach due to its high dynamic range and versatility. However, even if a good separation of the compounds is achieved and high quality HRMS and HRMS/MS spectra are acquired, the correct identification of the molecules is still difficult to accomplish. This dereplication is generally performed manually by interpretation of the acquired spectra combined with a search within databases (DBs) (Wolfender et al., 2015). However, this process is very time-consuming, therefore automated approaches are necessary.

With the recent development of LC–HRMS metabolomics approaches, automated dereplication strategies are currently developed (Bertrand et al., 2017; Brown et al., 2011; Creek et al., 2012; Wang et al., 2016) which should benefit the NP research area. To challenge automated dereplication workflows, the Critical Assessment of Small Molecule Identification (CASMI) Contest (http://casmi-contest.org/) represents a unique chance to blindly evaluate annotation scripts (Nikolic et al., 2017; Nishioka et al., 2014; Schymanski and Neumann, 2013). In 2016, the NP side of this contest (Category 1–Best Structure Identification on Natural

* Corresponding author at: Groupe Mer, Molécules, Santé–EA 2160, UFR des Sciences Pharmaceutiques et Biologiques, Université de Nantes, 9 Rue Bias, Nantes, France.
E-mail address: Samuel.bertrand@univ-nantes.fr (S. Bertrand).

http://dx.doi.org/10.1016/j.phytol.2016.12.025
1874-3900/

Products) consisted of 18 MS and MS/MS spectra from compounds reported in the literature.

This 2016 submission represents our second attempt at this challenge (Bertrand et al., 2017) and was used to challenge an in-house automatic dereplication workflow dedicated to LC–HRMS peak annotation. Based on our personal results in the 2014 CASMI challenge, the script was corrected. In addition, widely used packages and programs were implemented in the workflow: CAMERA (Kuhl et al., 2012) (for spectral interpretation) and SIRIUS 3 (Böcker et al., 2009) (for molecular formula determination). Finally, the MS^2 comparison was improved to be more discriminant (Allen et al., 2017). While the script has not been made publically available yet, its successes and failures in the 2016 CASMI challenge are discussed in the present paper and reveals major key points to consider for automatic dereplication. It raises relevant questions and issues to tackle in this growing field combining bioinformatics and NP chemists.

2. Materials and methods

The proposed automated procedure consisted of a complete script. It was written in R 3.0 (R Core Team, 2015) with the XCMS (Tautenhahn et al., 2008), CAMERA (Kuhl et al., 2012), IPO (Libiseller et al., 2015), MeHaloCoA (Roullier et al., 2016) and RMassBank (Stravs et al., 2013) packages. Two sets of data were treated in parallel, namely the raw LC–HRMS data and their corresponding manually detected MS and MS/MS spectra (raw MS data). The results of each step were stored in a MYSQL database using RODBC for future retrieval of the results.

2.1. Automatic peak detection

The peak detection from each LC–HRMS raw files were achieved using XCMS (Tautenhahn et al., 2008) with appropriate parameters (Table S1) (no MS/MS data were used during the peak picking process). In most cases, optimum parameters were not defined manually but selected using the Isotopologue Parameter Optimisation (IPO) (Libiseller et al., 2015) separately on each raw data file of interest. This optimisation was not achieved targetedly on the peak of interest of each challenge but on the global LC–HRMS data. The two parameters *ppm* and *mzdiff* used by XCMS were systematically optimized except for challenges 17 and 18. In the case of those two challenges, the peaks of interest were not automatically detected after the use of IPO due to a very low intensity, therefore the parameters were manually selected. However, to speed-up peak detection the peak duration parameters were manually selected in the case of challenges 10–13 and the *S/N* (signal to noise ratio) was selected to 1 for challenges 4, 10–13. In the case of raw LC–HRMS data stored in continuous mode, they were converted to centroid mode using ProteoWizard (Kessner et al., 2008) prior to any peak picking procedure. After peak detection, MS/MS data were automatically retrieved from the raw data using RMassBank (Stravs et al., 2013).

2.2. Spectral interpretation

The ions detected by XCMS (Tautenhahn et al., 2008) were grouped using CAMERA (Kuhl et al., 2012) based on retention time similarities yielding MS pseudospectra (pcgroups). Each pseudospectrum was interpreted using CAMERA (with an in-house extended adduct list) yielding to the detection of isotopes, adducts and neutral losses. The used parameters are reported in Table S1 and S2. In the case of raw MS data, for MS spectra of challenges 1, 2, 6 and 7, the intensity [M + 1] and [M + 3] were added manually to allow data interpretation by CAMERA.

2.3. Molecular formula determination

For each ion (except isotopes), a list of possible molecular formulae (MFs) was deduced by SIRIUS3 (Böcker et al., 2009) based on MS and MS/MS spectra, using $C_xH_yO_{27}N_{25}P_9F_{34}$ as maximum possible atoms based on existing MFs in the Dictionary of Natural Products (DNP) (Chapman and Hall, 2014) as reported by Kind and Fiehn (2007). The maximum number of carbon (x) and hydrogen (y) was estimated based on the detected m/z; x was set to $(m/z)/12$ and y to $(m/z)/2$. Only MFs in a specific *ppm* mass accuracy range were kept for future steps (ppm value are listed in Table S1 and S2). The potassium and sodium atoms were also added only in the case of $[M+K]^+$ and $[M+Na]^+$ adducts occurrence (predicted by CAMERA), respectively. When no adduct information were detected during the spectral interpretation step, only $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M]^+$ were considered in positive ionisation and only $[M-H]^-$, $[M+HCOOH-H]^-$ in negative ionisation. In addition, S, Cl and Br were automatically detected from the isotopic patterns using a script adapted from MeHaloCoA (Roullier et al., 2016) to reduce calculation time when those atoms may be present (Meusel et al., 2016). Finally, the SIRIUS3 score (S_{SIRIUS}) was used to discriminate between possible MF. Compound MF was then deduced by adduct correction.

When multiple adducts were detected in an MS^1 spectrum, yielding some MF to be detected for multiple times in one spectrum (after adduct correction), the S_{SIRIUS} was corrected based on MF redundancy score (S_{red}) between all adducts (Bertrand et al., 2017). This correction was based on the addition to the S_{SIRIUS} score of the S_{red} calculated as follows: 10 times the number of occurrences of the MF among all adducts over the maximum number of occurrences of a MF among all proposed MF of a given challenge.

2.4. Structure determination by database search

Each of the compound possible MFs was searched within various online and local DBs for matches. The DBs used were: AntiBase (Laatsch, 2008), ChEBI (Degtyarenko et al., 2008); Dictionary of Natural Products (DNP) (Chapman and Hall, 2014), Dictionary of Marine Natural Products (DMNP) (Blunt and Munro, 2013), HMDB (Wishart et al., 2013), KeGG (Kanehisa and Goto, 2000), KNApSAcK (Shinbo et al., 2006), Lipid Maps (Fahy et al., 2007), Universal Natural Product Database (UNPD) (Gu et al., 2013). From all these DBs, as much information as possible was retrieved (such as CAS number, InChI, InChIKey, SMILES and MOL file). During this process missing structural information were possibly obtained from conversion tools: OpenBabel (O'Boyle et al., 2011), Chemical Identifier Resolver (CACTUS, http://cactus.nci.nih.gov/chemical/structure), Chemical Translation Service (CTS) (Wohlgemuth et al., 2010) and ChemSpider (Pence and Williams, 2010).

To further discriminate between all proposed structures for a given peak, a strategy based on *in silico* fragmentation was achieved. This step was undertaken by competitive fragmentation modelling using CFM-ID (Allen et al., 2014) using an MS/MS similarity score ($S_{MS/MS}$ see Eq. (1)) adapted from Allen et al. (2017).

$$\begin{aligned} S_{MS/MS} = {} & 4 \times Jaccard_{10} + 4 \times Jaccard_{20} + 2 \times DotProduct_{10} \\ & + 2 \times DotProduct_{20} + 2 \times ExplPeaks_{20} + 2 \times ExplPeaks_{10} \\ & + 2 \times ExplIntensities_{20} + 2 \times ExplIntensities_{10} - 5 \end{aligned} \quad (1)$$

Where, for x (*10* or *20*) the energy level provided by CFM-ID, $Jaccard_x$ corresponds to the Jaccard distance between the simulated and the real spectra; $DotProduct_x$ corresponds to the dot product between the simulated and the real spectra (Jaccard