AutoSuspect: an R package to Perform Automatic Suspect Screening based on Regulatory Databases

Reza Aalizadeh[1], Emma L. Schymanski [2] and Nikolaos S. Thomaidis[*, 1]

[1] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zographou, 15771, Athens, Greece, *e-mail: ntho@chem.uoa.gr

[2] Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland.

EXTENDED ABSTRACT

Recent advances on Liquid Chromatography-High Resolution Mass Spectrometry (LC-HRMS) have revolutionized the identification of compounds in the environment. The continuous growing of application of LC-HRMS workflows increased the “peak inventories” reported in environmental samples. This was achieved by using three general workflows, namely “target”, “suspect” and “non-target” screening [1]. Although targeted analysis remains the best way to confirm the identification of a compound, it is sometime not widely applicable due to limited access to a high number of reference standards. The vast majority of the substances or peaks detected in samples typically remain unidentified and supportive information such as retention time prediction, MS/MS evaluation and ionization behavior could help increasing the identification confidence.

As “peak inventories” increase and the number of regulatory databases grows, an automatic approach is greatly needed to screen for new compounds in environmental samples using “suspect screening”. A systematic approach is required to check the mass accuracy of the precursor ion and the experimental and theoretical isotopic fit, prior to chemical structural elucidation based on their retention time, MS/MS and ionization behaviour.

The aim of this study is to propose an automatic workflow to screen environmental samples, such as influent wastewater, with a wide-scope regulatory database of chemicals. All the influent wastewater samples considered in this study were collected from the WWTP of Athens (Greece) during 7 days in 2016. Analyses were carried out by reversed-phase liquid chromatography quadrupole-time-of-flight mass spectrometry (RPLC-QToF-MS) with electrospray ionization (ESI), operating in positive mode. More details about the analytical method used can be found in [2].

An MS-ready database of environmentally-relevant substances was compiled by Schymanski *et al.* [3] (it can be found online at: *http://www.norman-network.com/?q=node/236*) and it was used to screen the samples. This database offers specific format such as an optimized and stereochemistry-free chemical identifier, molecular formula and name, Retention Time Indices (RTI) [4] in both ESI modes and monoisotopic mass.

The proposed workflow starts with an optimized peak picking algorithm, using XCMS with IPO package behind the optimization task, for a set of samples (here n=16) in which the MS information were recorded in data independent acquisition mode. Retention time correction and alignment (optimized with IPO) was performed using the “obiwarp” algorithm (Ordered Bijective Interpolated Warping) [5]. Componentization and annotation of peaks list were achieved using “CAMERA” [6]. This step is needed to focus on those m/z that are found to be potential molecular ions. Then, all the retention time of the detected m/z are converted to RTI to calibrate the elution information and support identification. The next step was to use the compiled MS-ready database to screen it against the peaks list generated by XCMS with a mass deviation of 0.01 and retention time indices tolerance of ±300 RTI units. Then, the theoretical isotopic pattern was calculated for a given molecular formula that was found to be matched to peaks. The theoretical isotopic pattern is derived for each case by “enviPat” and then compared with extracted experimental isotopic pattern [7]. The fitting measurement was carried out by Support Vector Machine Regression (SVR) with an inherited cross-validated approach to optimize the SVR internal parameters in each case. The final score, which shows the match of the isotopic pattern for each molecular formula, is derived based on true presence of theoretical isotopic pattern and its match to the experimental ones. Finally, the results including the isotopic match, RTI and its uncertainty, mass error and chemical identifiers are exported to further evaluate the experimental MS/MS information.

“AutoSuspect” offers an automatic screening tool using an environmentally related regulatory database in the background which could accelerate the identification task in given samples. After performing “AutSuspect”, 130 from overall 390 potential masses (molecular ions) detected by CAMERA were matched to the compounds in the compiled regulatory database with high mass accuracy and acceptable RTI tolerance (experimental and predicted RTI). Among these matched compounds, 22 compounds were confirmed at level of identification ***3*** [1] as their MS/MS information were available and could explain the majority of fragments created by MetFrag (*in silico* fragmentation approach) [8]. Two compounds (Metformin and Theophylline) identified by AutoSuspect were confirmed at identification level of ***2a*** as their experimental MS/MS matched reference spectra in MassBank (www.massbank.eu), used also to “validate” the procedure. Among the detected compounds, some homologue series were identified such as 2-[2-(3-aminopropoxy)ethoxy]ethanol and tetraethylene glycol, which are active ingredients in laundry, dishwashing and automotive care products. 32 structures were remained at identification level ***4*** (the predicted and experimental RTI was matching), since MS/MS information was not available. 33 matched compounds were also found to be false positives, as their characteristic ions were not present. The rest of compounds matched to the peak list were also rejected either due to low intensity, unclear MS spectra or high RTI deviation. “AutoSuspect” greatly facilitates a higher confidence, rapid screening of samples for suspects, providing an overview of tentative identifications and likely false positive matches for subsequent follow-up.

Keywords: Suspect Screening, Regulatory Database, Isotopic fit, Retention Time Indices, Structure Elucidation

References

[1] E. L. Schymanski, *et al*. “Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence” Environmental Science & Technology, 48 (4), pp. 2097–2098, 2014.
[2] P. Gago-Ferrero *et al*. "Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS" Environmental Science & Technology, 49(20), pp. 12333-12341, 2015.
[3] E. L. Schymanski, *et al*. "International Suspect Screening: NORMAN Suspect Exchange meets the US EPA CompTox Chemistry Dashboard" in 16th International Conference on Chemistry and the Environment (ICCE2017), Oslo, Norway, 2017.
[4] R. Aalizadeh, *et al*. "Development and Prediction of Liquid Chromatographic Retention Time Indices (RTI) to facilitate non-target identification" In preparation, 2017.
[5] J. T. Prince, and Edward M. Marcotte "Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping" Analytical Chemistry, 78(17), pp. 6140-6152, 2006.
[6] C. Kuhl *et al*. "CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets" Analytical Chemistry, 84(1), pp. 283-289, 2012.
[7] M. Loos *et al*., "Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees" Analytical Chemistry, 87(11), pp. 5738-5744, 2015.
[8] C. Ruttkies *et al*. "MetFrag relaunched: incorporating strategies beyond in silico fragmentation" Journal of Cheminformatics, 8(1): pp. 3, 2016.