Cloud-based Analysis of Complex Sample Systems for Outlier and Trend Analysis in the Context of Water Monitoring

Poster Note 64791

Ralf Tautenhahn[1], Rahel Comte[2], Tim Stratton[1], Srividya Kailasam[1], Dipankar Ghosh[1], and Heinz Singer[2]
[1]Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, CA, USA, 95134
[2]Eawag, Überlandstrasse 133, 8600 Dübendorf, Switzerland

ABSTRACT

Many analytical challenges attempt to answer a common set of questions – what has or is changing in my system? Is this new sample like the previous ones? These are often asked in specific contexts such as authenticity of a food product or the change in the profile of contamination in water year over year due to new emerging contaminants. Here we introduce a cloud-based system which enables these kinds of problem solving analyses. It consists of a means of processing these potentially large datasets, creation of a representative profile database of components, and hypothesis testing tools to detect differences and trends within the data.

INTRODUCTION

Many complex analytical questions require techniques beyond simple targeted analysis. Determining the authenticity of claimed origin for a food product such as honey, orange juice, or whiskey requires understanding the entire profile of the sample in question compared to a known set of authentic samples. Identifying new emerging contaminants or trends in known contaminants in river or ground water requires the analysis and processing of samples from multiple sites across months or years. The comparison of a single unknown or challenge sample to an authentic is not sufficient given the complexity and natural changes that occur in the authentic or known material and as a result, "big data" analytical techniques are required to create a profile of what is "authentic", "good", or "known". Such an approach can assist with environmental monitoring, food origin authenticity, designer drug or sports doping agent detection, or many other applications. Here we present a cloud-based product profiling application to address such challenges and apply it to the analysis of multi-month water monitoring for trend detection.

MATERIALS AND METHODS

Sample collection: From November 2014 to January 2014, 87 daily time-proportional composite samples were collected at the outflow of a municipal waste water treatment plant (WWTP) located in a sub-catchment of the river Rhine. The sampled WWTP is equipped with a conventional tertiary treatment system (20M m^3/year; 150,000 inhabitants, no chlorination for disinfection). Samples were frozen for future analysis.

Sample preparation:
The samples were unfrozen and then about 5 ml were transferred into centrifugation vials. The centrifugation was needed to prevent any solid particles from entering the LC-HRMS system. The centrifuge (Megafuge 1.0 R, Heraeus Sepatech) was run for 30 minutes at 4000 rpm. 1 ml of the centrifuged sample was transferred to an HPLC vial and 10 µl of an internal standard mix containing 130 isotope labeled internal standards (ILIS) was added with a Hamilton syringe (resulting in a ILIS concentration of 5 µg/l). Thermo Scientific™ Nanopure™ water was used as a control and prepared through the same process.

LC-MS measurements:
The HPLC used consisted of a CTC PAL HTS-Xt (Zwingen/ Switzerland) autosampler, and a Thermo Scientific™ Dionex™ UltiMate™ 3000RS system containing a degasser, as well as a column oven. A Dionex XRS pump was used in the first measurement run but was replaced by the UltiMate pump later on. For the analysis of the WWTP samples acidified HPLC-grade water (NPW) and Methanol (MeOH) with 0.1% formic acid (FA) were used as a mobile phase. For the stationary phase the Atlantis® silica-based C18 column (T3 3µm 3.0x150mm, Waters Corporation) coupled to an Atlantis® pre-column (T3 3µm) were chosen. The gradient started at 95% NPW and 5% MeOH, increased the MeOH amount during 16 minutes up to 95%, stabilized on this level and then dropped to the starting ratio. The flow was during the whole gradient 300 µL/min. 100 µl of each sample was directly injected via the 100 µl loop. The separation on the column was approached with a constant temperature of 30° C degrees. For the analyte detection a Thermo Scientific™ Q Exactive™ Plus Orbitrap™ mass spectrometer (MS) was used. Full scan data was acquired at 280,000 resolution with AGC target 1e6 and max inject time 100msec, followed by data dependent MS2 at 17,500 resolution with AGC target 5e5 and max inject 100msec for the Top5 peaks. Fragmentation energy was set to 20, 50 and 80% NCE values. Dynamic exclusion was used with a value of 6 sec.

SOFTWARE

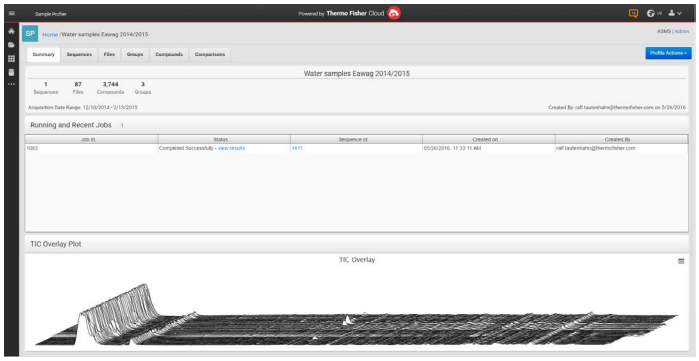
The Sample Profiler™ solution is a newly developed cloud-based software platform for comparative analysis. The software allows the creation of extensive databases (*profiles*) based on large numbers of samples, for example based on long term monitoring of river, ground, or run-off water or to create profiles of authentic products. New samples can easily be added to an existing profile without the need to reprocess the entire dataset. The system automatically aligns all samples, performs component detection and matches fragmentation spectra against the mzCloud™ spectral database for compound identification. For each sample and each compound the relative intensities are stored together with the chromatograms. The result is a representative profile database of all compounds present in the samples. The data stored in a profile can be partitioned into logical groups based on the experimental design. For example a group can be created that contains all the samples from a certain time period or a group the includes samples from a certain region, type or producer in the context of product authenticity. The groups are updated automatically, when samples are being added or removed that match these criteria.

Challenge samples can be compared either against the entire profile or against specific groups, hypothesis testing is performed and differences are reported. Different visualizations are available for reviewing chromatograms and intensity distribution for all detected compounds, trends over time as well as box-and-whiskers plots to highlight the detected differences. Finally, marker compounds can be selected within the profile to give quick access to compounds of particular interest.

DATA PROCESSING

Water samples for a period of 3 months (November 6th 2014 – January 31st 2015) of routine monitoring were uploaded into the Thermo Fisher Cloud environment, imported into the Sample Profiler application and assigned to a new profile. After finishing the initial sample processing, the application reported the number of compounds that were detected and rendered a TIC overlay of all samples (Fig.1). A total of 3,744 compounds were detected and out of those 250 were automatically identified using mzCloud, a curated online high-resolution spectral library, as well as a local database containing accurate mass and retention time (Fig.2). To allow for trend analysis within different timeframes (i.e. long term and short term trends), multiple sample groups were created within the profile, consisting of groups that contained water samples from the entire 3 month period, from the last month, and from only the final two weeks, respectively. These sample groups were then used to discover specific and conspicuous patterns (e.g. spikes) for certain compounds, to find correlations between compounds and to compare different timeframes.

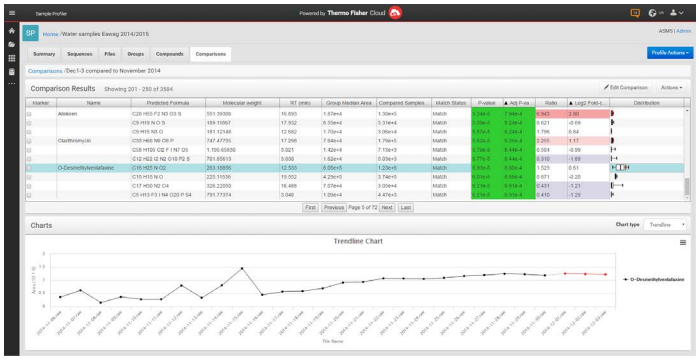
Figure 1. Summary page



Screenshot of the summary page in Sample Profiler for the profile that was created based on 87 water samples. Data is represented in tabs and hyperlinks can be used to directly switch between views of the same data in different tabs. The table in the middle displays the status of recent data processing jobs. At the bottom a TIC overlay of all samples currently stored in the profile is shown.

Thermo SCIENTIFIC

Figure 2. Trends and differential analysis



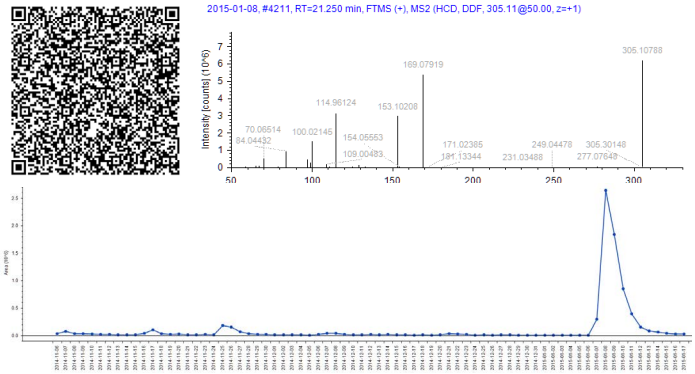
Differential analysis in Sample Profiler. Comparisons between data stored in the profile and challenge samples can be calculated on-the-fly without the need to reprocess all the samples in the profile. The distribution plot in the rightmost column now displays overlaid box-and-whiskers plot for both profile and challenge samples to visualize differences. Hypothesis testing is performed and p-values, adjusted p-values, ratio and fold-change are reported for all compounds in the table. Significant differences are highlighted. Bottom: Trendline chart. Trend analysis can be performed for the entire profile, within a sample group or as part of a comparison.

RESULTS

The database created in the previous steps was used to answer two questions – What is different in an acute time scale and what major trends in the profile could be identified over time. For the first, a new dataset for a recent time point was added to the platform and a comparison performed to the constructed database. Compounds were highlighted that were either new to this sample, missing in the new sample, or were present at levels statistically significantly different than the historical data. For the second question, an analysis of the data for trends was performed considering both seasonal cycle and trend over time. Components which were identified as potentially trending upward were subjected to specific scrutiny for identification.

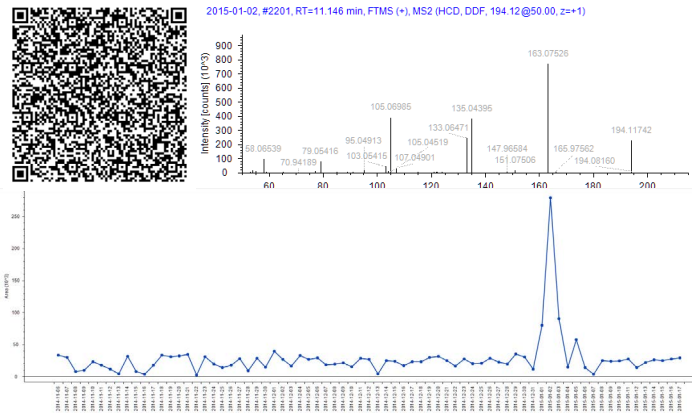
For the first question, samples collected in a more recent two-week time range (January 1st 2015 to January 14th 2015) were compared to a database which contains samples from November 6th 2014 to December 31st 2014. 98 compounds were found to be present at much higher levels (ratio ≥ 5, p-value ≤ 0.001) than the historical data. Two examples are Diazinon, an insecticide, which is present at elevated levels from January 7h to January 12th (Fig.3). The presence of Diazinon outside of the agricultural growing season most probably stems from use in private households. MDMA, a synthetic drug, is elevated on January 1st and 2nd (Fig.4) and indicates the increased use over the new year holiday break.

Figure 3. Diazinon



Trendline chart, MS² spectrum and QR code (see section "QR codes") for Diazinon.

Figure 4. MDMA

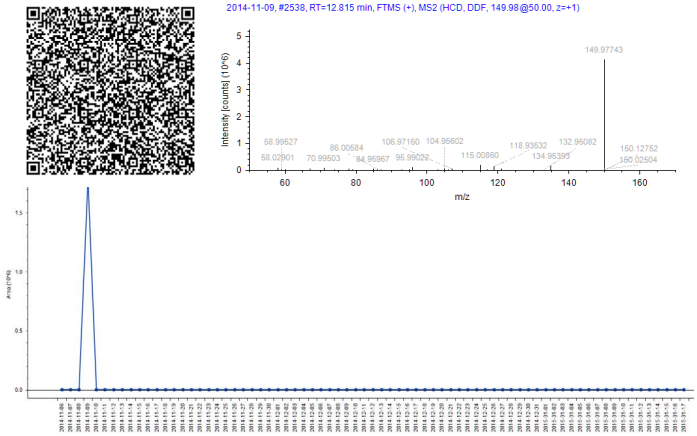


Trendline chart, MS² spectrum and QR code (see section "QR codes") for MDMA.

For the second question, long and short term trends were investigated by defining samples groups over different time periods and applying statistical tests as well as manual interpretation using the trendline visualization tool. The following compounds are selected examples from the several hundred compounds that showed significant trends over time:

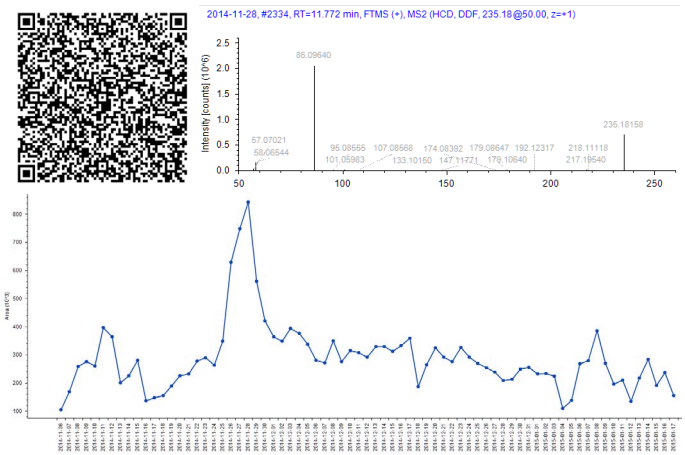
- 5-Chloro-2-methyl-4-isothiazolin-3-one, an antimicrobial agent for industrial applications, was detected in a single sample collected on November 11, 2014 (Fig. 5)
- Lidocaine, a local anesthetic, elevated between November 26 and November 29, 2014 (Fig. 6)
- Propiconazol, a fungicide, detected at elevated levels on several days in December 2014 and January 2015 (Fig. 7)

Figure 5. 5-Chloro-2-methyl-4-isothiazolin-3-one



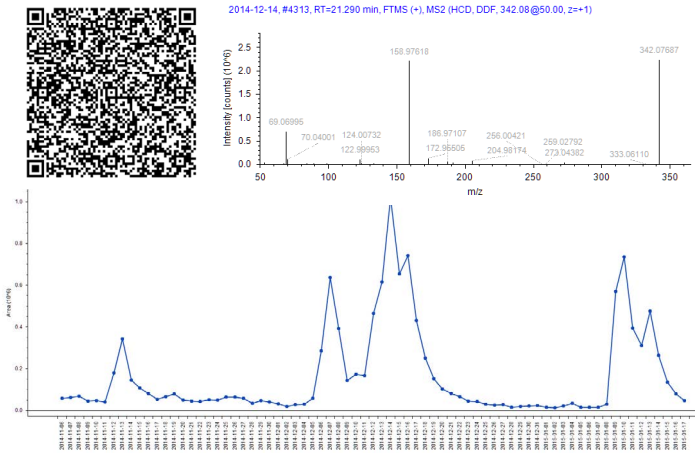
Trendline chart, MS² spectrum and QR code (see section "QR codes") for 5-Chloro-2-methyl-4-isothiazolin-3-one.

Figure 6. Lidocain



Trendline chart, MS² spectrum and QR code (see section "QR codes") for Lidocain.

Figure 7. Propiconazol



Trendline chart, MS² spectrum and QR code (see section "QR codes") for Propiconazol.

CONCLUSIONS

- Large datasets can be analyzed with a cloud-based data processing workflow to create broad profiles of sample matrices
- These profiles can be used to determine both specific localized changes, in our examples the emergence of a contaminant at specific times, or to identify trends in samples such as the change in measured amount of compound over time in our effluent profiles
- The cloud-based Sample Profiler solution was also able to provide identifications for several detected components using a continuous search against mzCloud, an online high resolution accurate mass fragmentation spectral database.

We have demonstrated the application of the Sample Profiler solution to analyze large datasets which can continuously grow over time, to detect new emerging trends within the data. This approach can be applied to other situations where such large dataset analysis is useful ,such as the analysis of suspect food samples to confirm or refute claims of product origin, or potentially to assist in detection of new emerging drugs or doping agents, as well as our demonstrated application of emerging environmental contaminant and contaminant trend analysis.

QR codes

The QR codes shown on this poster contain fragment spectrum information. The QR code can be used to scan, display and search the spectrum with the mzCloud app (available for iPhone® and Android™ platforms). Spectra Teleporter, which creates QRSpectrum™, is available at **www.mzCloud.org**.

ACKNOWLEDGEMENTS

We would like to acknowledge Robert Mistrik and his team for giving us early access to the QRSpectrum application and the mzCloud app. Furthermore, we thank the personnel of the WWTP for their support with sampling.

www.thermofisher.com

Africa +43 1 333 50 34 0
Australia +61 3 9757 4300
Austria +43 810 282 206
Belgium +32 53 73 42 41
Brazil +55 11 2730 3006
Canada +1 800 530 8447
China 800 810 5118 (free call domestic)
400 650 5118
Denmark +45 70 23 62 60
Europe-Other +43 1 333 50 34 0
Finland +358 10 3292 200
France +33 1 60 92 48 00
Germany +49 6103 408 1014
India +91 22 6742 9494
Italy +39 02 950 591
Japan +81 6 6885 1213
Korea +82 2 3420 8600
Latin America +1 561 688 8700
Middle East +43 1 333 50 34 0
Netherlands +31 76 579 55 55
New Zealand +64 9 980 6700
Norway +46 8 556 468 00
Russia/CIS +43 1 333 50 34 0
Singapore +65 6289 1190
Sweden +46 8 556 468 00
Switzerland +41 61 716 77 00
Taiwan +886 2 8751 6655
UK/Ireland +44 1442 233555
USA +1 800 532 4752

PN64791-EN 0816S

Thermo SCIENTIFIC
A Thermo Fisher Scientific Brand