CEST
International Conference on Environmental Science and Technology

Automatic detection of organic pollutants with characteristic time pattern in wastewater using computational approaches and chemometric tools on data acquired by LC-HRMS

Presenting author: Alygizakis Nikiforos

University of Athens
Department of Chemistry
Laboratory of Analytical Chemistry

TrAMS
Trace Analysis and Mass Spectrometry Group

Alygizakis N.A., Gago Ferrero P., Loos M., Singer H., Hollender J. and Thomaidis N.S.

Contents

- ✓ To demonstrate the motivation of finding analytes with high fluctuation between influent samples
- ✓ To describe a computational workflow capable to detect components with characteristic time pattern beginning from raw LC-HRMS data
- ✓ To describe the optimization of the crucial input parameters to the algorithms
- ✓ To demonstrate an interesting study case

Aim

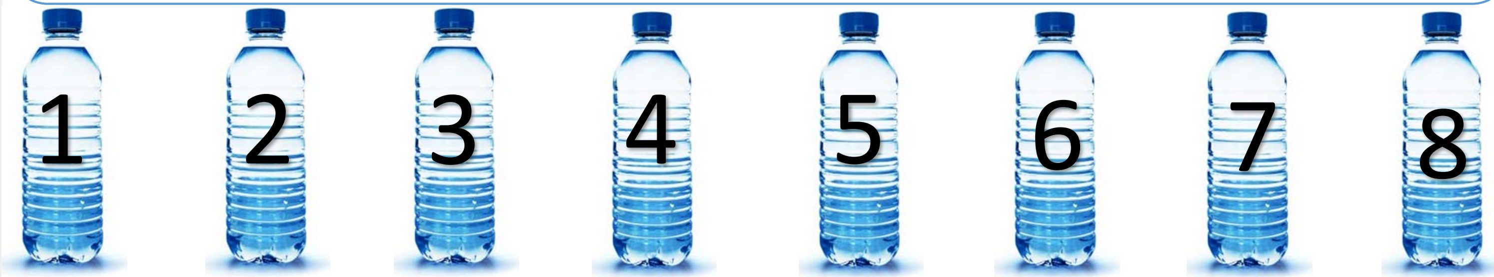
The aim of this study is to develop an automatic methodology which enables the screening of contaminants exhibiting characteristic time pattern in response, within daily influent samples.

Sampling

Athens Wastewater Treatment Plant (Psittalia)

Sampling Time period: Wednesday 4th of March–Wednesday 11th of March

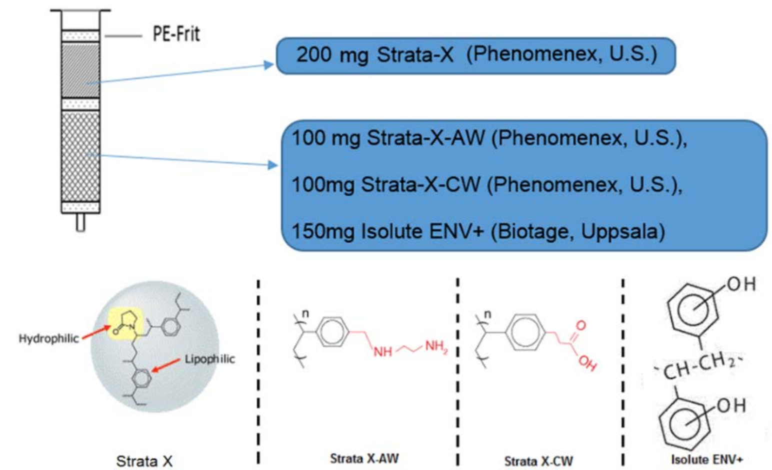
Representative samples following 24h flow proportional sampling



Cleanup and enrichments using Solid Phase Extraction

Conditioning: 5 mL Methanol, 10 mL Milli-Q Water

Elution: 4 mL MeOH:Ethyl Acetate (2% v/v NH_3) & 2 mL MeOH:Ethyl Acetate (1.7 % v/v Formic acid

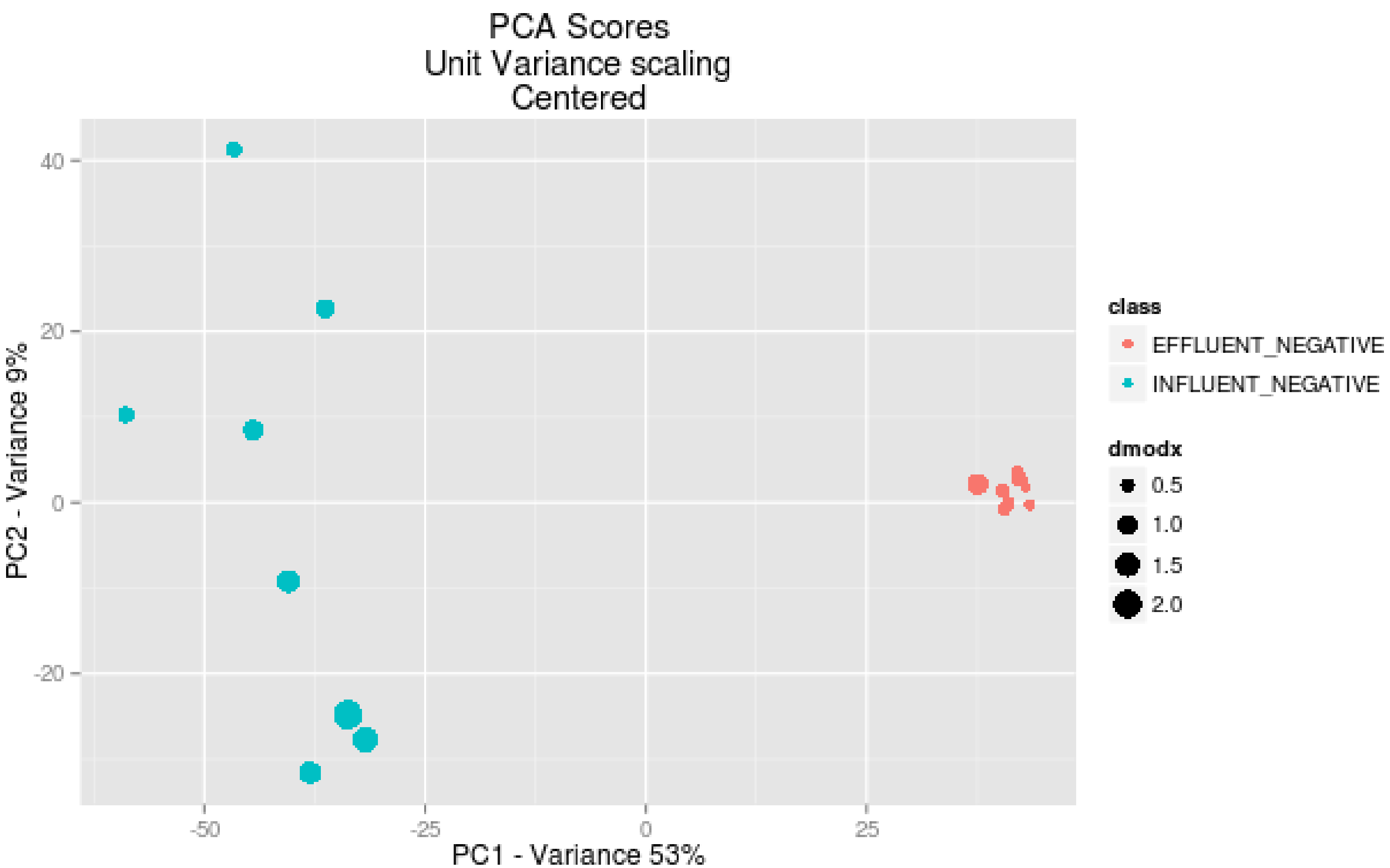


Data dependent AutoMS/MS acquisition using 5 precursor ions and then inclusion list of components

Kern et al. Environmental Science and Technology (2009) 43(18) p.p.7039–7046

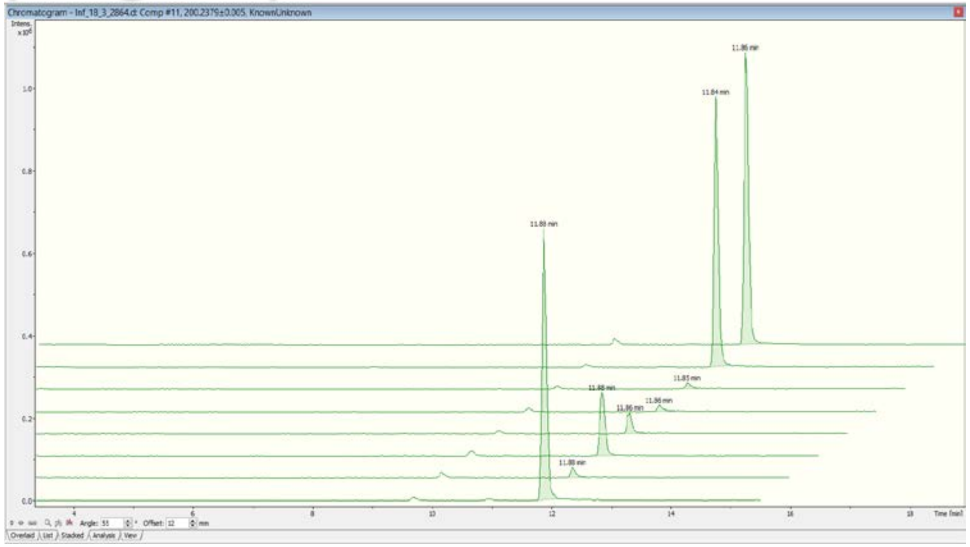
Trend Analysis

- Specific categories of emerging contaminants follow different consumption patterns and therefore concentration levels in influent wastewater can vary between different time sets.
- It is know that recreational drugs reach peak consumption during weekend
- X-ray contrast media and anticancer drugs have the opposite response.

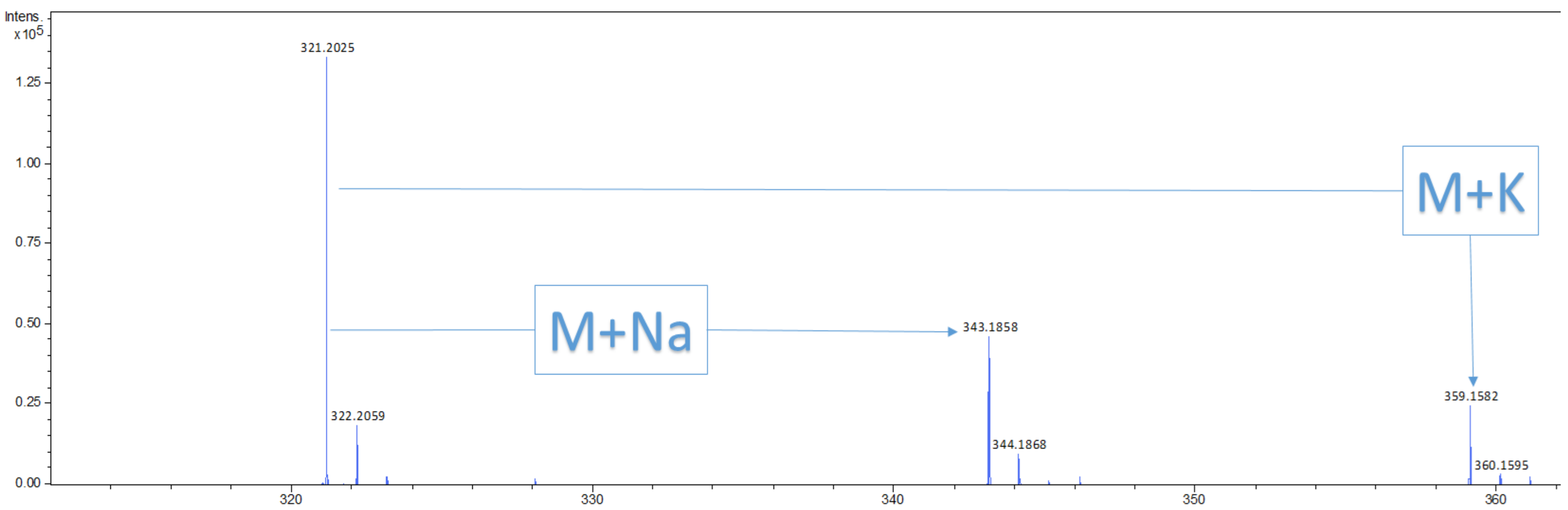


Explanation of the term "Grouping"

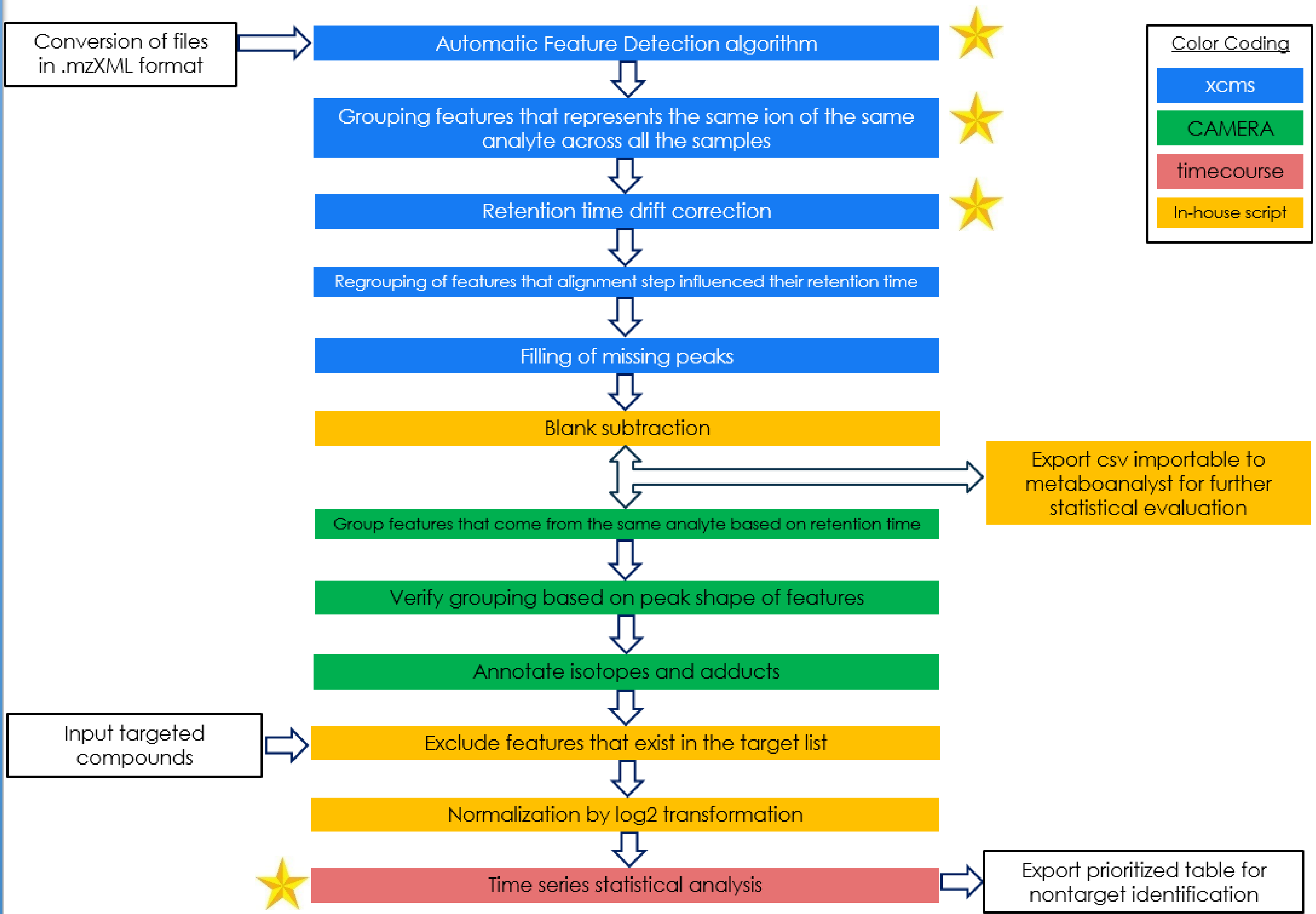
Grouping of peaks across the samples



Grouping of peaks that belongs to the same compound



Proposed computational Procedure



centWave approach

Conversion of files in .mzXML format → Automatic Feature Detection algorithm

↓

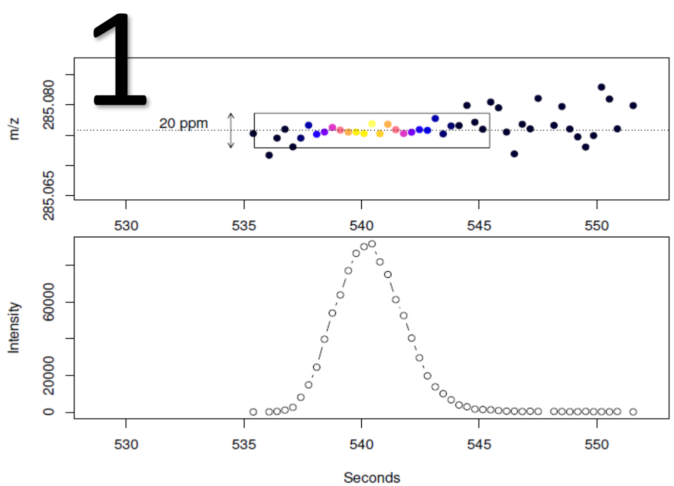
Grouping features that represents the same ion of the same analyte across all the samples

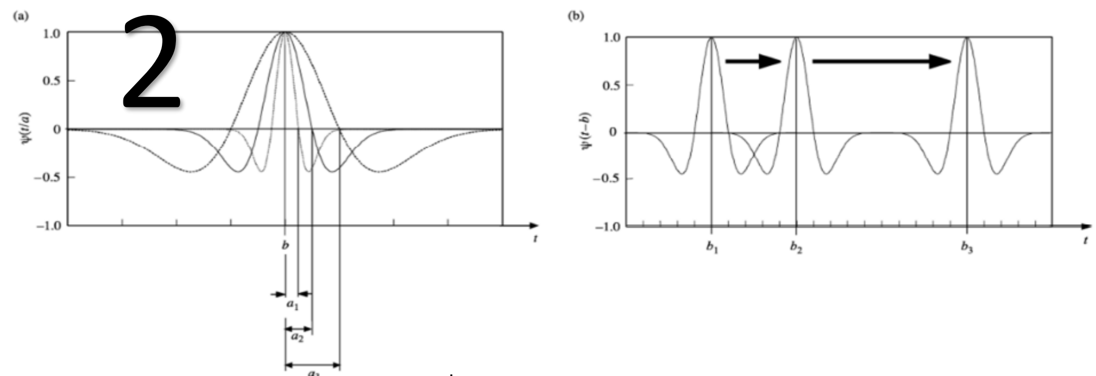
↓

Retention time drift correction

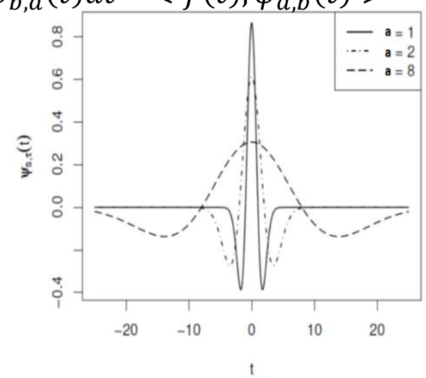
↓

Color Coding
xcms
CAMERA
timecourse
In-house script





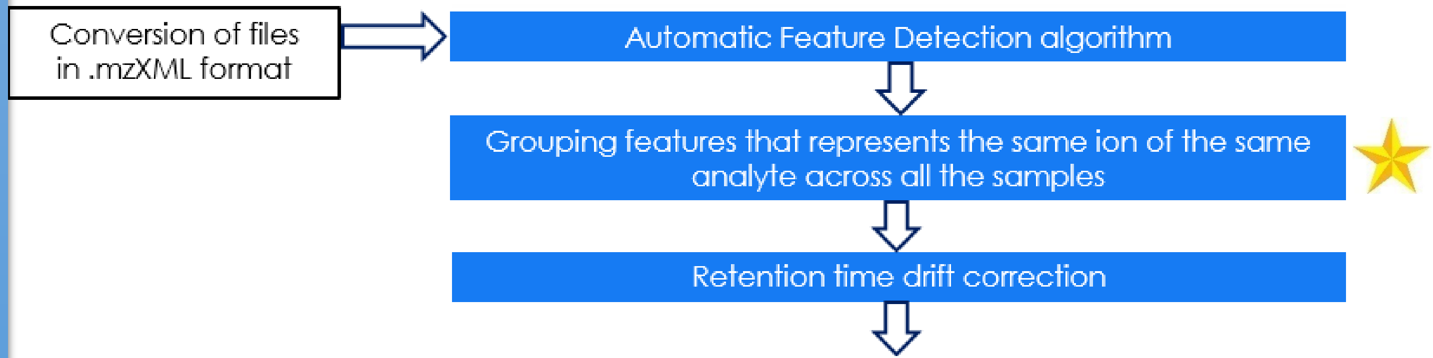
$$W_{\psi}f(a,b) = \int_{-\infty}^{+\infty} f(t)\overline{\psi_{b,a}(t)}dt =< f(t), \psi_{a,b}(t) >$$



Critical Parameters
ppm
Minimum and maximum peak width

Tautenhahn et al., 2008, BMC Bioinformatics

File organization and Grouping

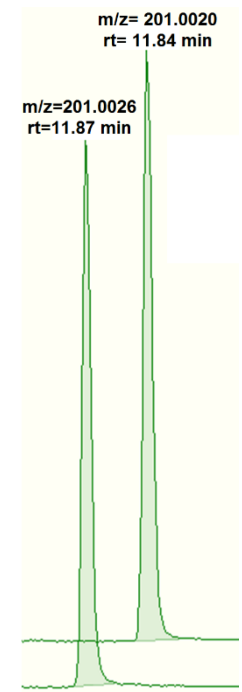


Color Coding

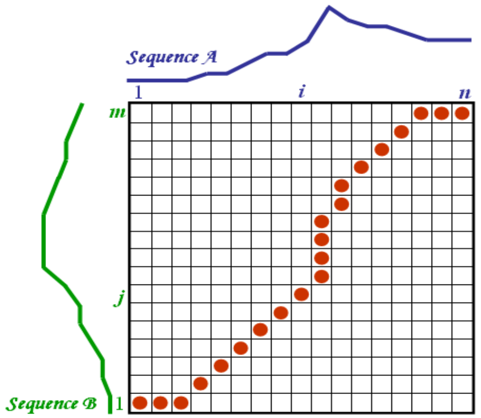
- xcms
- CAMERA
- timecourse
- In-house script

- Positive Ionization
  - 04.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 05.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 06.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 07.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 08.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 09.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 10.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - 11.03.2015
    - 4_3_2015_A1.mzXML
    - 4_3_2015_A2.mzXML
    - 4_3_2015_A3.mzXML
    - 4_3_2015_B1.mzXML
    - 4_3_2015_B2.mzXML
  - Procedural Blank
    - Blank_A1.mzXML
    - Blank_A2.mzXML

Critical Parameters	Explanation
Minsamp	Minimum number of samples
Minfrac	Minimum fraction of samples
bw	Bandwidth of kernel
mzwid	width of overlapping m/z slices

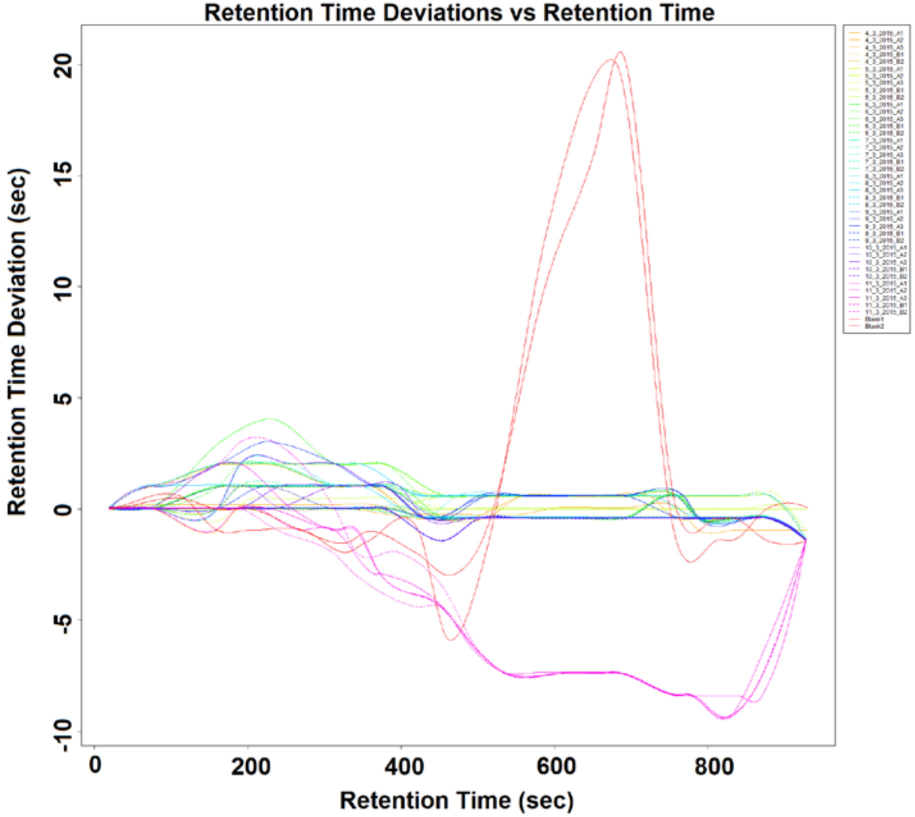


Retention time drift alignment

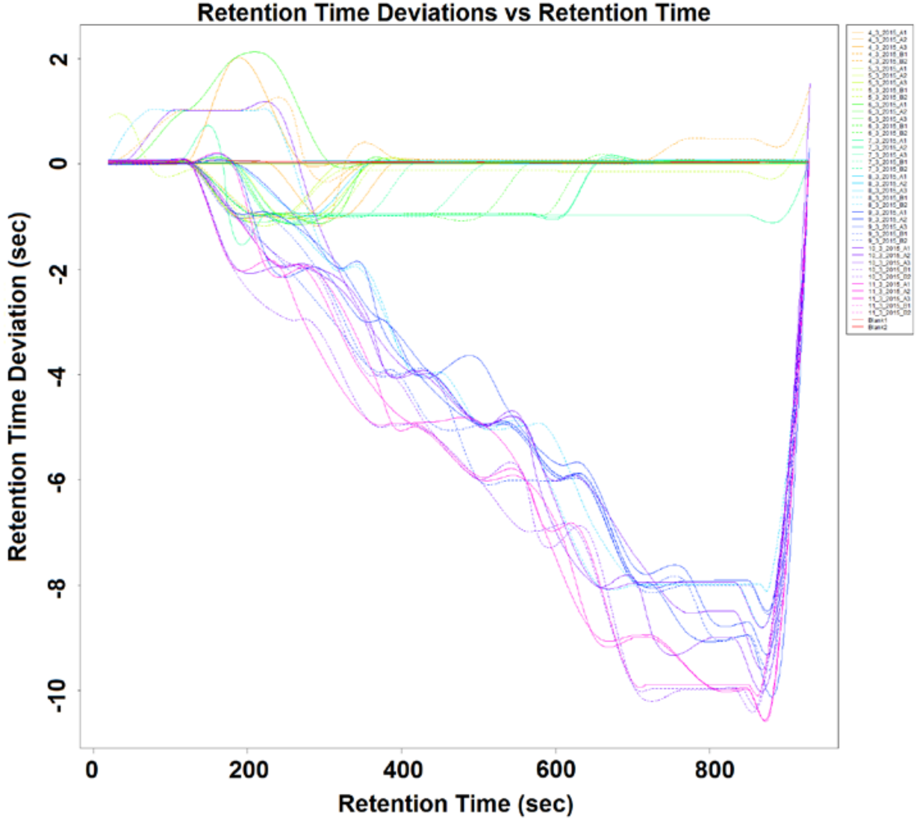


Critical Parameters	Explanation
gapInit	Penalty for Gap opening
gapExtend	Penalty for Gap enlargement

Negative ESI



Positive ESI



Parameters to be optimized

CentWave parameters		
ppm	?	?
Minimum peak width	?	?
Maximum peak width	?	?
Retention Time alignment based on OBI-Warp algorithm		
Distance function	cor_opt	cor_opt
gapInit	?	?
gapExtend	?	?
Grouping of features based on kernel density estimator		
bw	?	?
mzwid	?	?

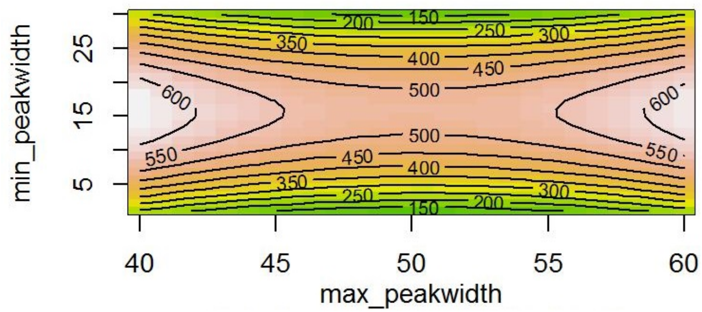
Optimization of parameters of peak picking

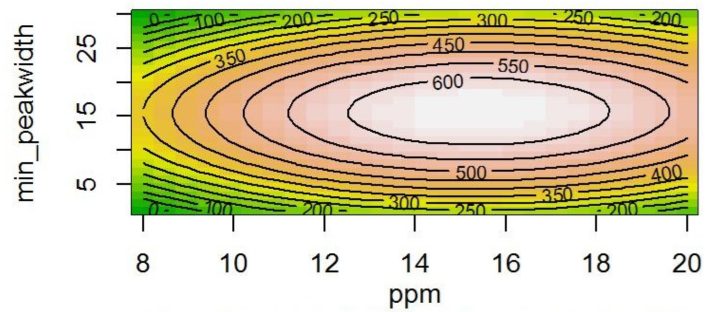
- Optimization was based on Box-Behnken (BBD) experimental design three step:

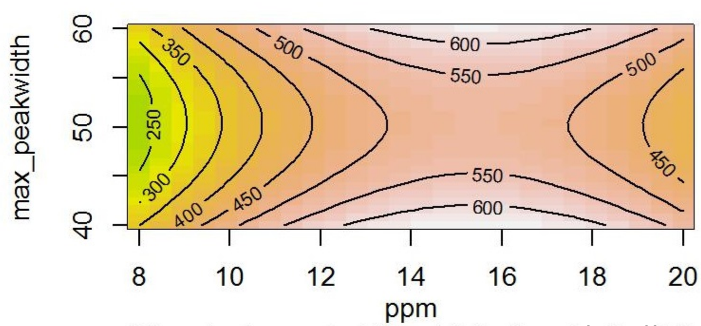
$$PPS = \frac{RP^2}{All\ peaks - LIP}$$

Where;

- PPS=Peak picking score (Response)
- RP=Reliable Peaks (M+H successfully identified)
- LIP=Low intensity peaks







Input Parameters	POSITIVE ESI	negative ESI
CentWave parameters		
ppm	17.6	17.6
Minimum peak width	14.34	15.5
Maximum peak width	50	50

Libiseller et al. BMC Bioinformatics (2015) 16(118)

Optimization of grouping of features and retention time alignment

Response function for retention

- $RCS(x) = \left(\frac{sum\left(\frac{\sum_{n=1}^{k} |median(x) - x_n|}{k}\right)}{k} \right)^{-1}$

RCS=Retention time score,

x are symbolized the retention times of features within a group

k is the number of retention times

Response for grouping of features across samples:

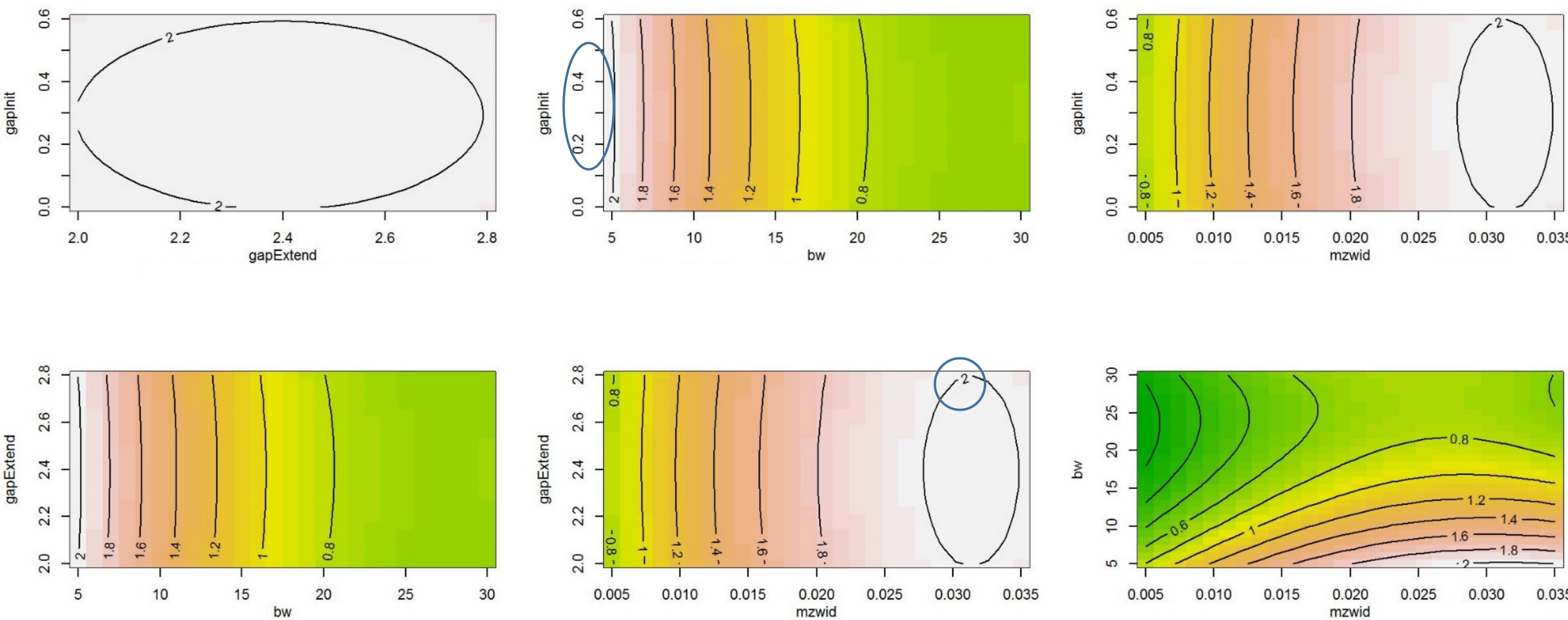
- $GS = \frac{reliable\ groups^2}{\text{non reliable groups}}$

GS= Grouping score,

reliable groups

Total score is a weighted combination of responses GS and RCS.

Input Parameters	POSITIVE ESI	negative ESI
Retention Time alignment based on OBI-Warp algorithm		
gapInit	0.3	0.27
gapExtend	2.4	2.36
Grouping of features based on kernel density estimator		
bw	5	5
mzwid	0.032	0.0305

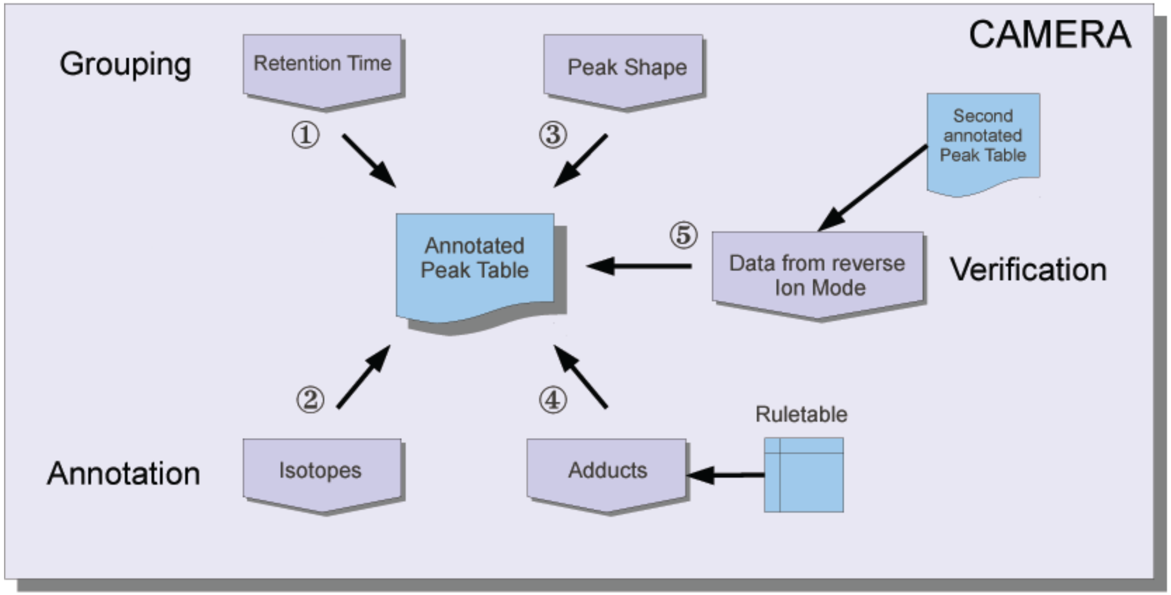


Libiseller et al. BMC Bioinformatics (2015) 16(118)

Optimum parameters

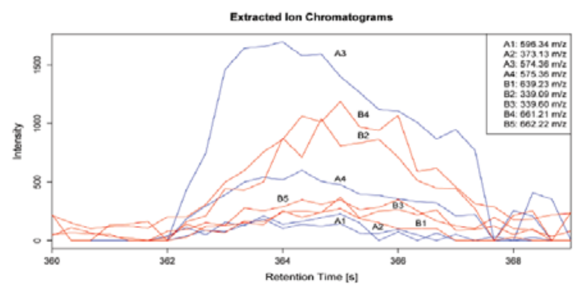
Input Parameters	POSITIVE ESI	negative ESI
CentWave parameters		
ppm	17.6	17.6
Minimum peak width	14.34	15.5
Maximum peak width	50	50
Retention Time alignment based on OBI-Warp algorithm		
Distance function	cor_opt	cor_opt
gapInit	0.3	0.27
gapExtend	2.4	2.36
Grouping of features based on kernel density estimator		
bw	5	5
mzwid	0.032	0.0305
minfrac	0.5	0.5
minsamp	2	2
max	50	50

CAMERA



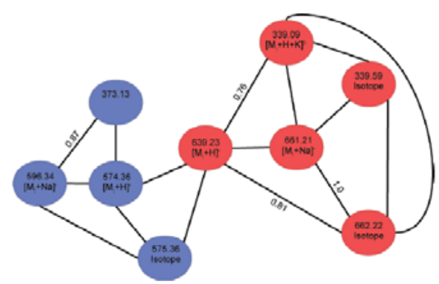
id	m/z	rt [s]
1	339.09	369.1
2	339.59	369.1
3	373.13	368.6
4	574.36	368.7
5	575.36	368.7
6	596.34	368.4
7	639.23	368.7
8	661.21	369.1
9	662.22	369.1

Initial compound spectrum



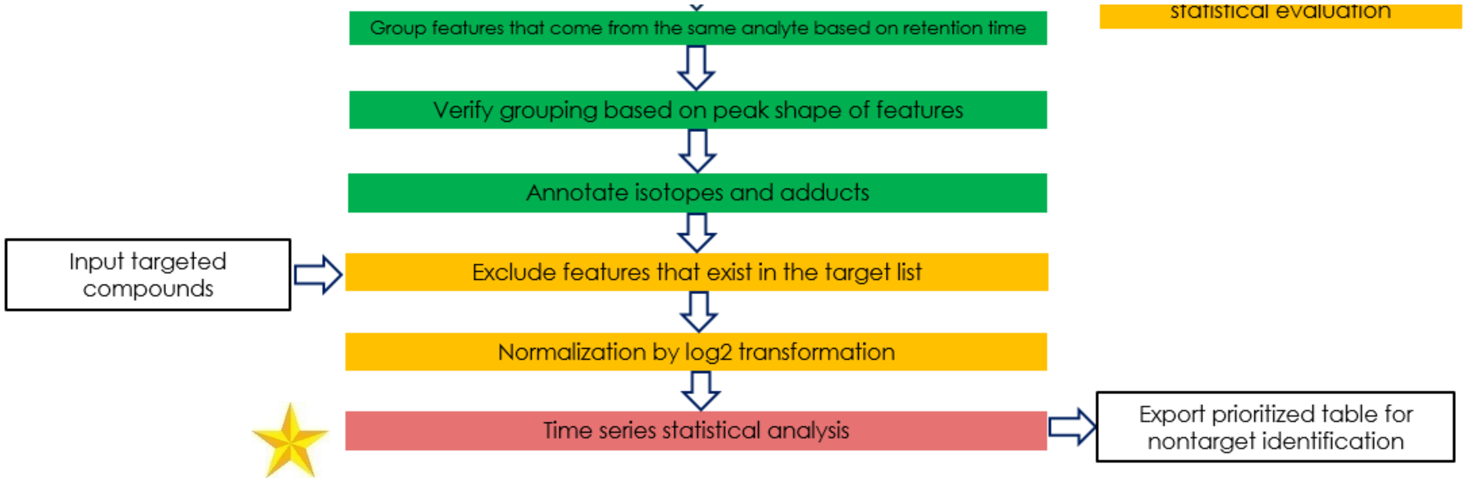
	1	2	3	4	5	6	7	8
2	1.0							
3	0.34	0.36						
4	0.72	0.67	0.90					
5	0.68	0.47	0.68	1				
6	0.38	0.38	0.87	0.86	0.75			
7	0.76	0.72	0.73	0.82	0.88	0.70		
8	0.94	0.79	0.35	0.70	0.67	0.51	0.79	
9	0.88	0.84	0.46	0.62	0.70	0.62	0.81	1.0

Relationship scoring table



Kuhl et al., Analytical Chemistry (2012) 84(1), p.p. 283-289

Prioritization methods-Review



- Intensity-based (Schymanski et al., 2014)
- Cl, Br, S compounds
  - Characteristic isotope pattern (like Hug et al., 2014)
  - Characteristic mass defect (like Chiaia-Hernandez et al., 2014)
- Venn diagrams (operators of union, intersect and complement) (Muller et al., 2011)
- Effect-directed analysis (Weiss et al., 2011)

Time-series Analysis

- There are two kinds of time course experiments
  - **Periodic time courses (**specific pattern)

  Typically concern natural biological processes such as circadian rhythms

  - **Developmental time courses** (less expectation for specific patterns)

  Example: concentration levels at a series of times in a developmental process

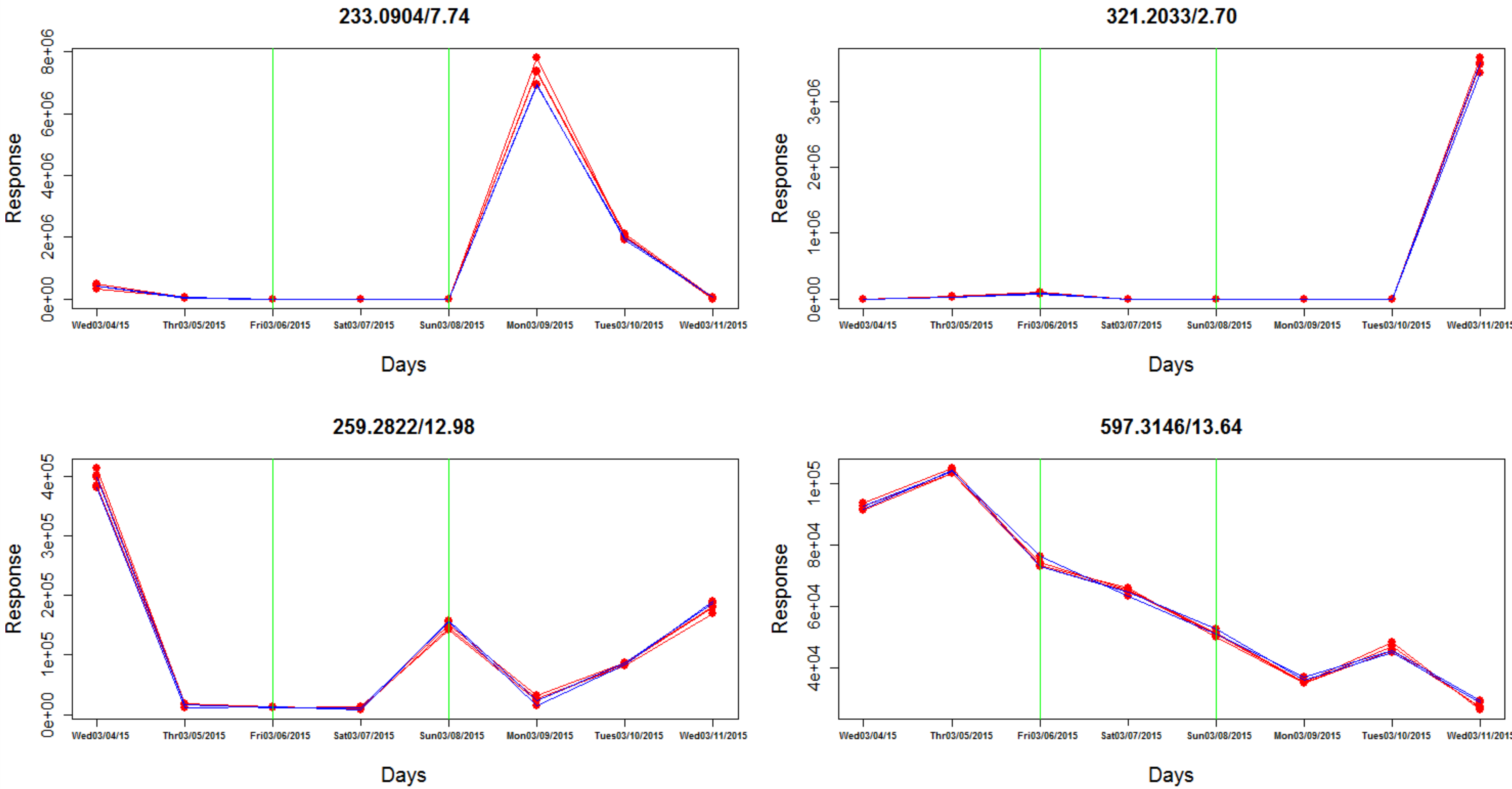
  Features are ranked with **one-sample Multivariate empirical Bayes approach**, which is suitable for REPLICATED, SHORT developmental time courses.

  Has advantages over other statistical approaches, since it does not cluster but ranks features.

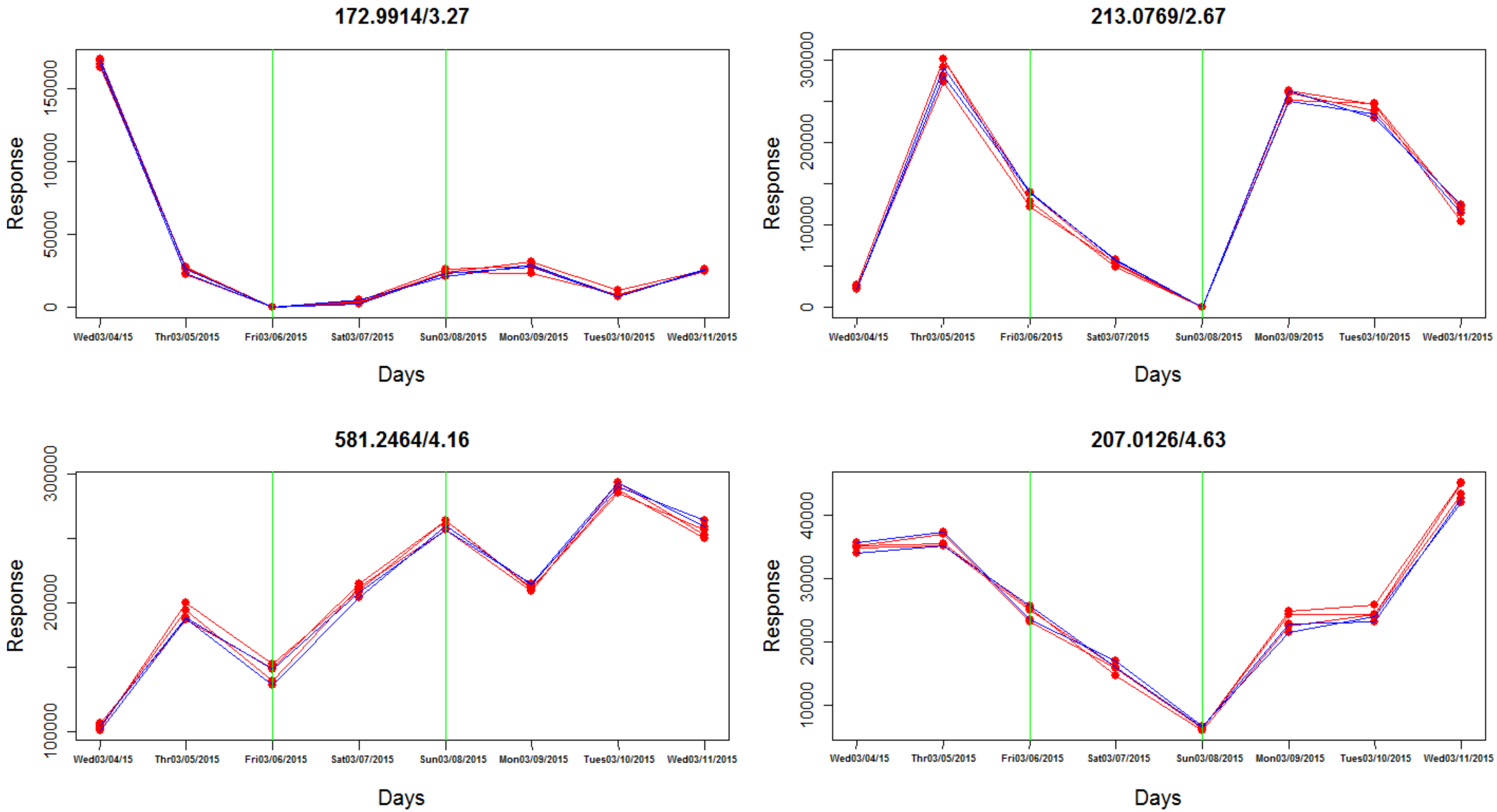
H_0: The expected temporal profile of an analyte is constant

Tai and Speed, The Annals of statistics, 2006

Top ranked components in Positive ESI

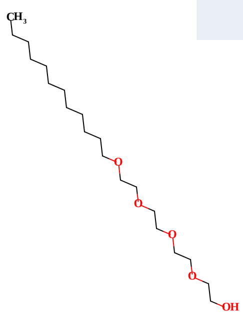
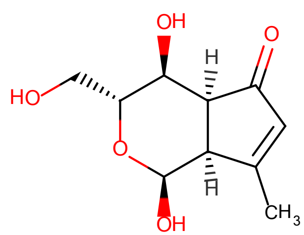
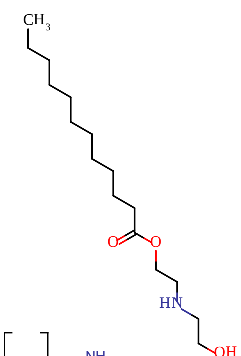
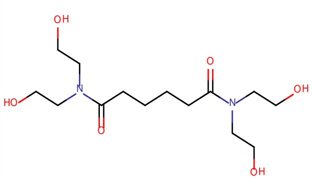
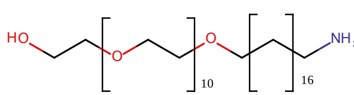
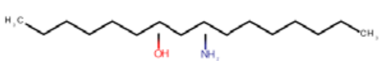
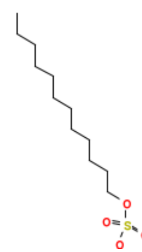
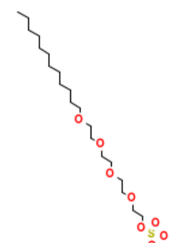
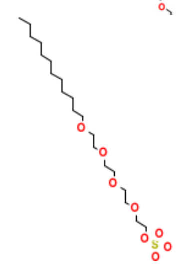
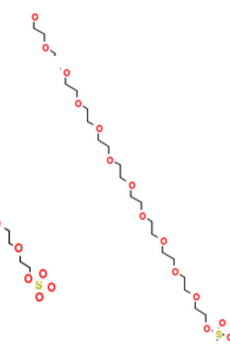
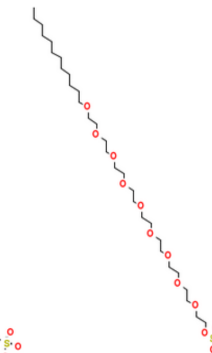
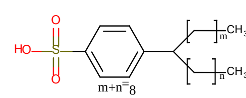
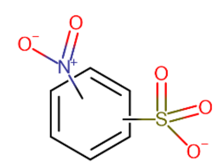
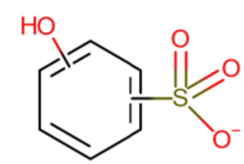


Top ranked components in Negative ESI



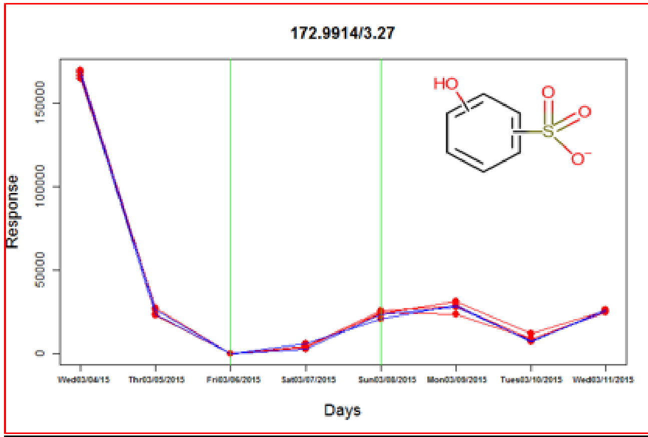
Results

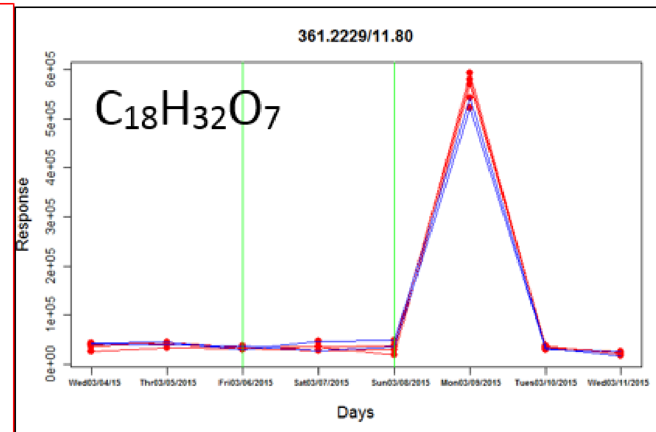
Identification level	Positive ESI	Negative ESI
LEVEL 2A	0	1
LEVEL 2B	4	5
LEVEL 3	2	2
LEVEL 4	13	12
LEVEL 5	1	10
Sum	20/30	30/30

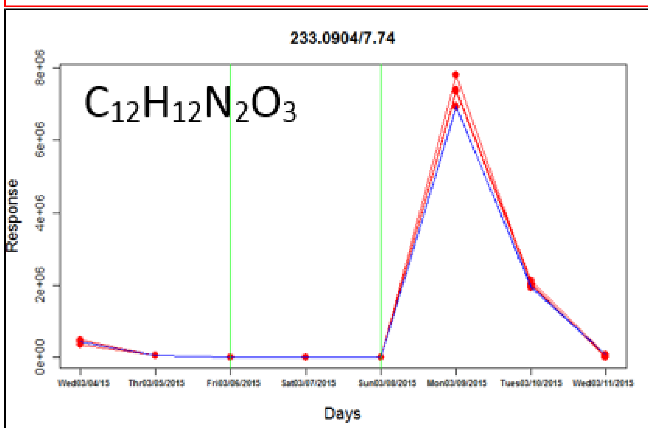















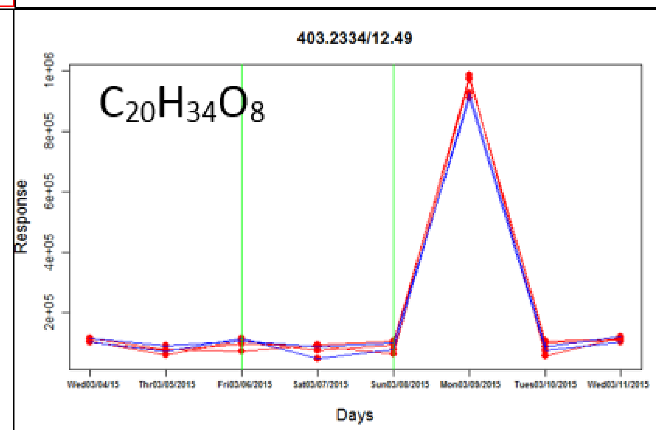
Non-target Workflow based on Gago-Ferrero et al., 2015, EST, Submitted

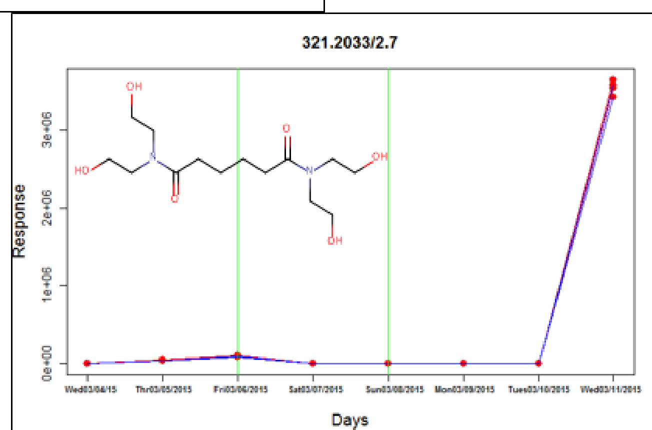
Events of direct disposal



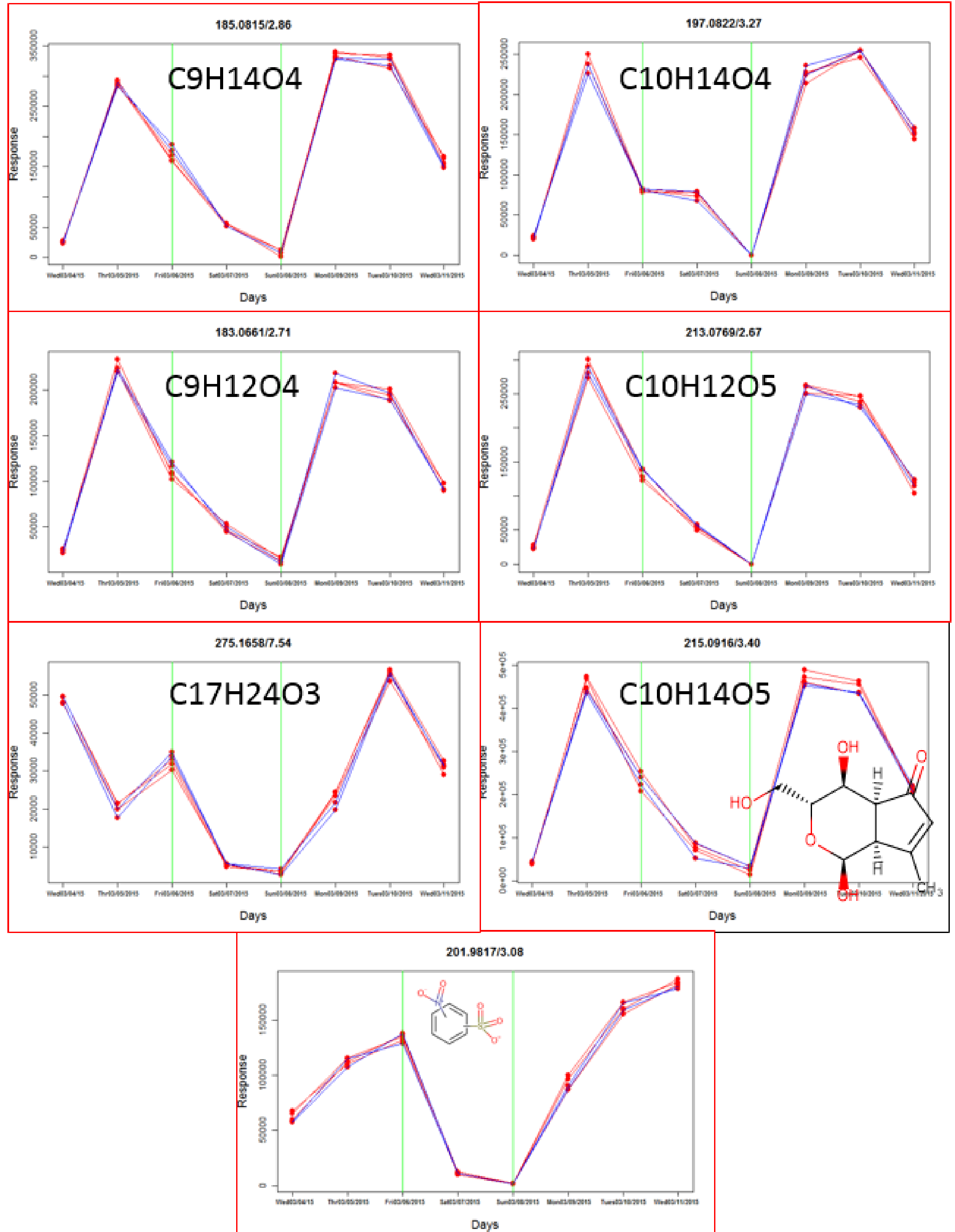






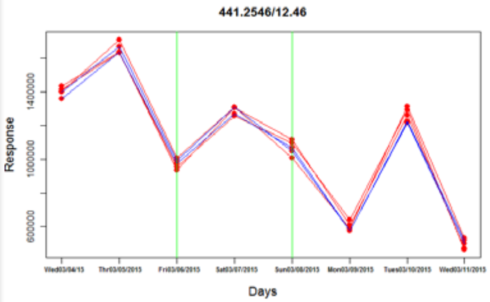


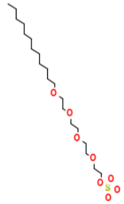
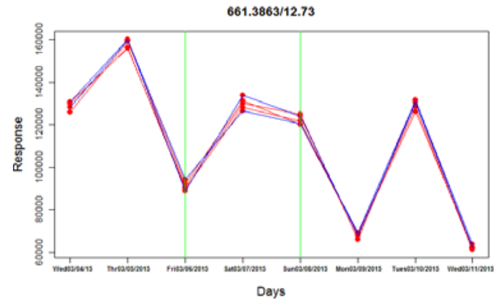
Compounds with low concentrations during the weekend

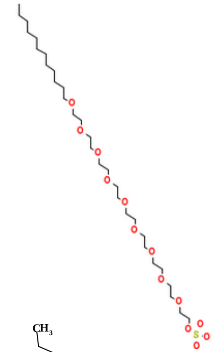
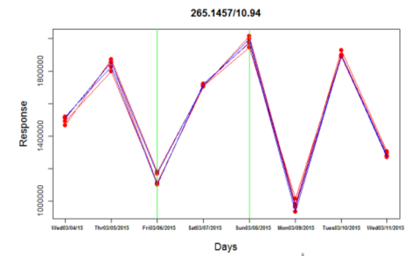


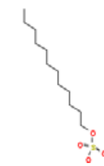
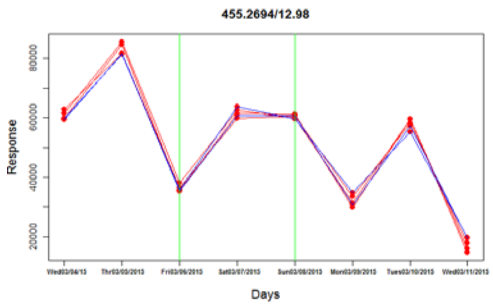
→Compounds with similar elemental composition exhibit similar trend

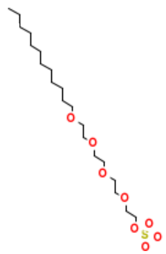
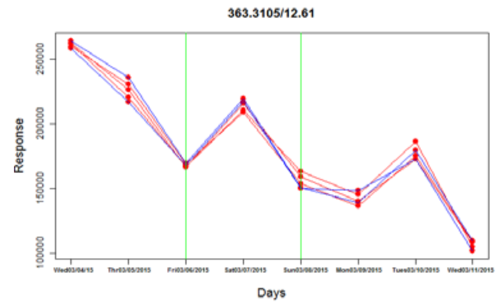
Surfactants and related substances share similar trend

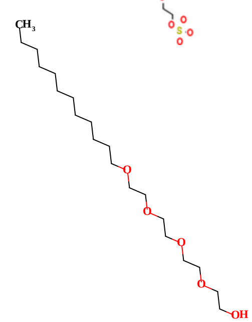


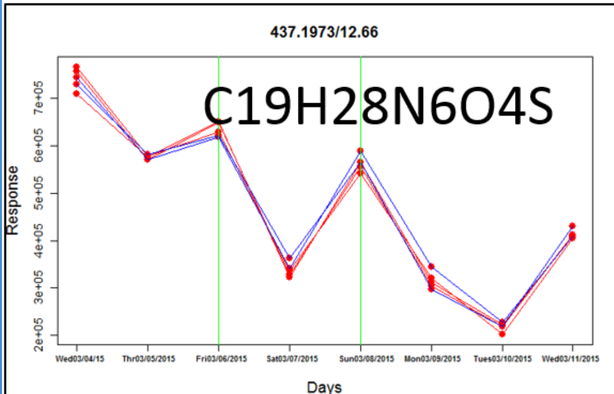



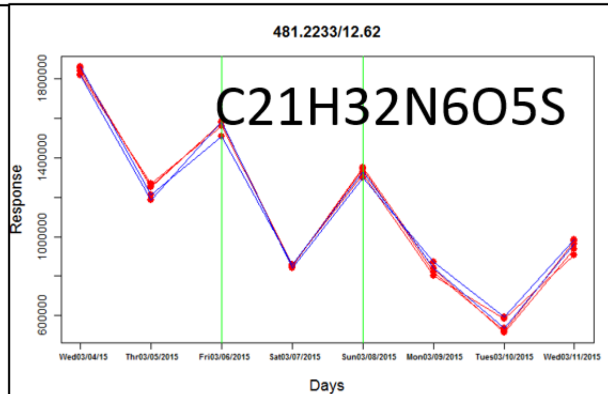



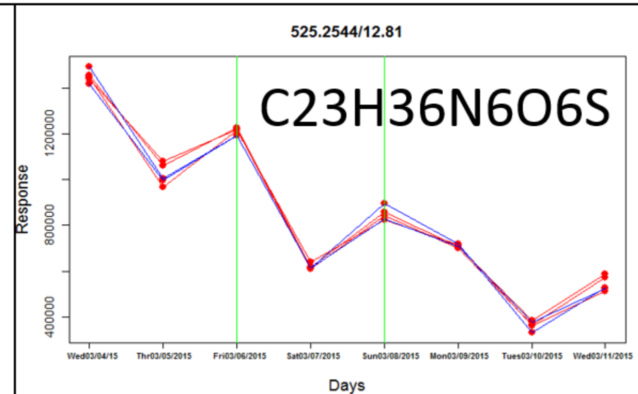





→Compounds with same origin share common trend graphs



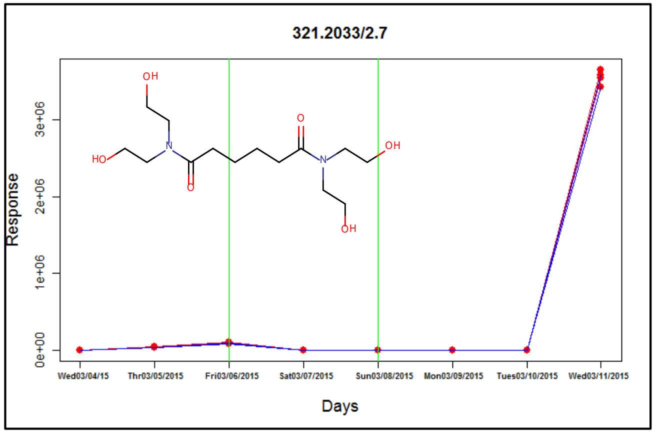


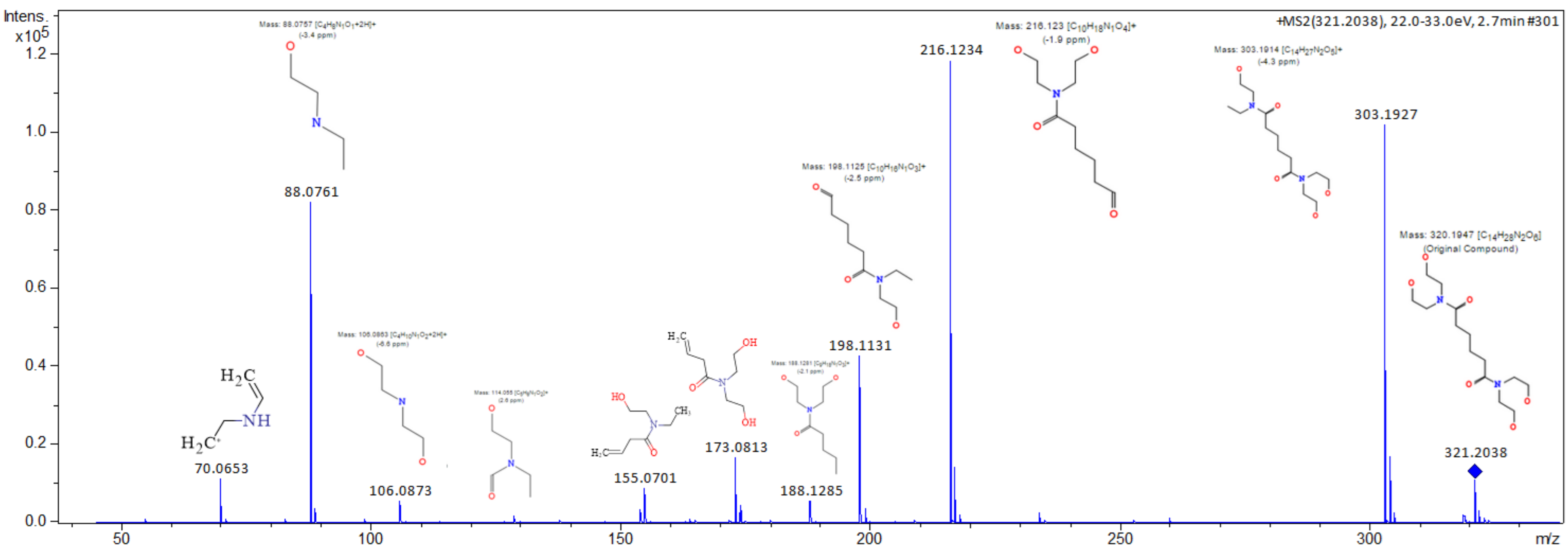


→Compounds in homologue series share common trend graphs

Interesting case study

Identified in 3 out of 8 days (Tuesday 5th of March 2015 Intensity: $(3.34 \pm 0.62) \times 10^4$; Thursday 6th of March 2015 Intensity: $(9.39 \pm 1.19) \times 10^4$ and Wednesday 11th of March 2015 $(3.56 \pm 0.01) \times 10^7$).





Conclusions

- A computational workflow with a novel prioritization method was implemented successfully on real samples.
- Crucial input parameters to the algorithms were optimized.
- Non-target identification of the top 30 components per ionization was conducted and the identity of many compounds was revealed.
- We demonstrated that relevant compounds with common origin share common time-trend. This information can be used to assist detection and identification of relevant compounds.

Thank you for your attention

Time for questions and discussion

Acknowledgements: This research has been co-financed by the European Union and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – ARISTEIA 624 (TREMEPOL project).

European Union
European Social Fund
OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS, CULTURE & SPORTS
MANAGING AUTHORITY
Co-financed by Greece and the European Union
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

TrAMS
Trace Analysis and Mass Spectrometry Group

John S. Latsis
Public Benefit Foundation

CEST
International Conference on Environmental Science and Technology