The ROIMCR procedure as a powerful tool for sewage water sample characterization in environmental proteomics

Carlos Pérez López[a] , Damià Barceló[a,c] , Antoni Ginebreda[a] , Joaquin Abian[b], Montserrat Carrascal[b] and Romà Tauler[a]

a Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Department of Environmental Chemistry, Jordi Girona 18-26, 08034 Barcelona, Spain
b Proteomics Laboratory CSIC/UAB, Institute of Biomedical Research of Barcelona, Spanish National Research Council (IIBB-CSIC/IDIBAPS), Rosellón 161, E-08036 Barcelona, Spain
c Catalan Institute for Water Research (ICRA-CERCA), Emili Grahit 101, Parc Científic i Tecnològic de la Universitat de Girona, Edifici H2O, 17003 Girona, Spain

INTRODUCTION

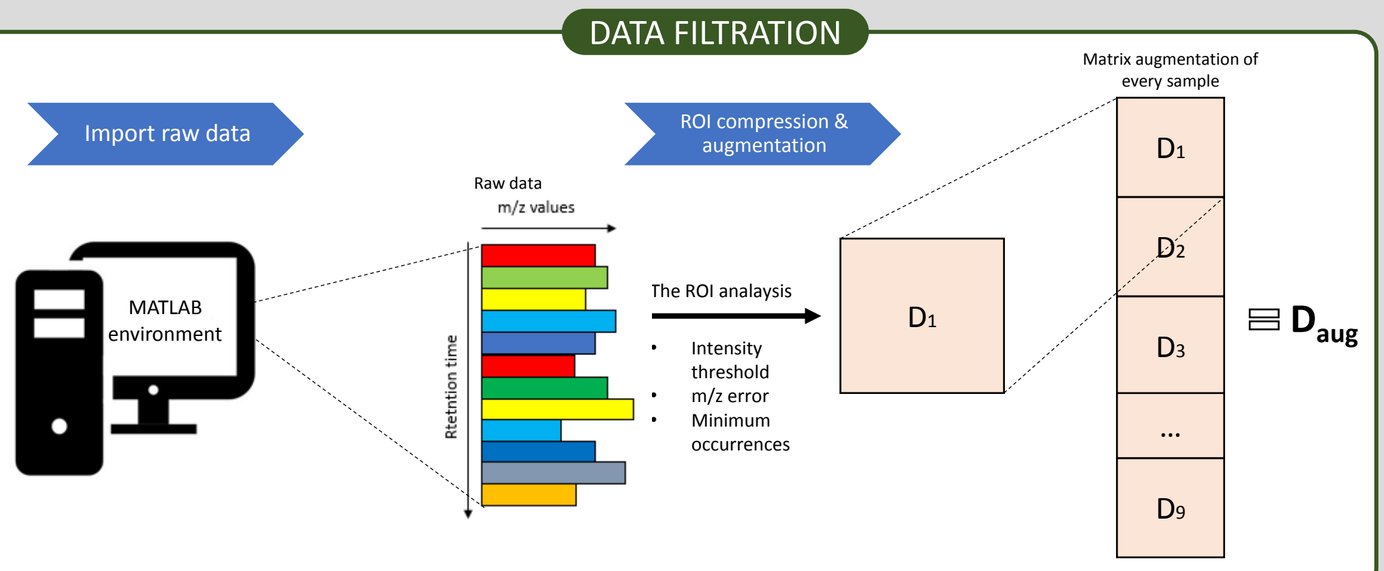
Sewage water provides a huge amount of information about the inhabitants of a region. Although sewage water has been used extensively in studies of small molecules e.g pharmaceuticals or illegal drugs, recently has also been used as a representative sampling source of protein biomarkers[1]. The large amount of information in these proteomics studies poses a difficulty in the characterization and quantification of biomarkers in sewage water samples. In this work, the Regions of Interest-Multivariate Curve Resolution (ROIMCR)[2] procedure, a recently proposed proteomic[3] tool, combined with Partial Least Squares-Discriminant Analysis (PLS-DA)[4] is proposed for the analysis of samples in environmental proteomic studies.

EXPERIMENTAL DESIGN

Polymeric devices were placed in the Influent water in the WWTP of Gavà-Viladecans (Barcelona, Spain) at 3 different times (3 devices at each sampling time) between April and May in 2020

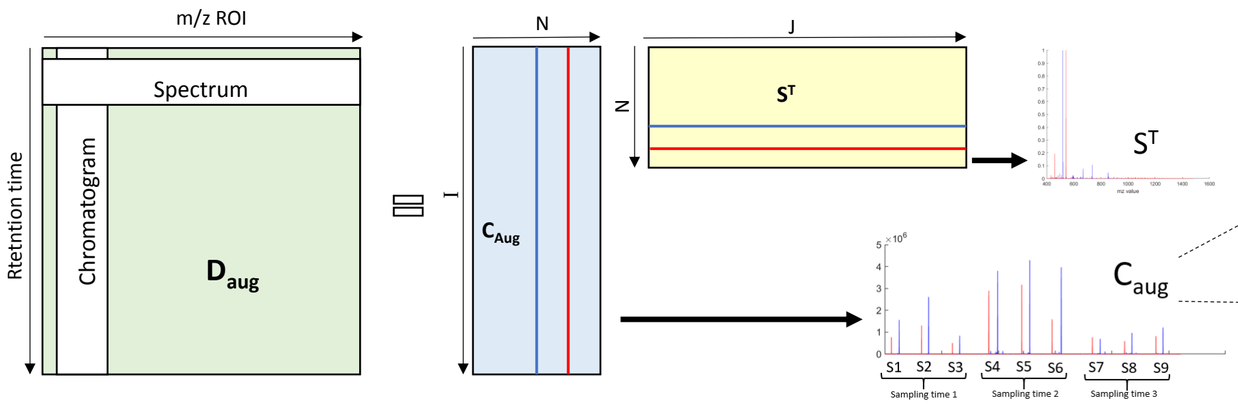
Polimeric devices were placed during 11 days → Surfaces from polimeric devices were cut and purified → Samples were digested with trypsin → The LC-HRMS/MS analysis

DATA FILTRATION



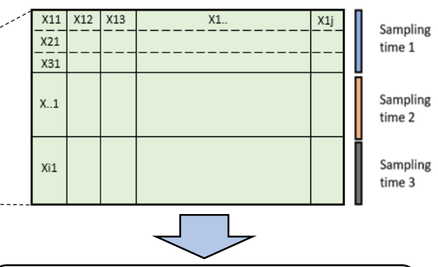
STATISTICS

MCR-ALS



PLS-DA

The heights of each MCR component from C_{aug} matrix are used as variables in PLS-Da analysis



MCR components stelected as the main responsible of the observed differences among samples collected at different times

PROTEIN IDENTIFICATION

The identification step focuses on the signals of those MCR components resultant from PLS-DA anlaisys as the responsibles of the variability among sampling times. The characteristics of the identification step are:

- **Software**: Proteome Discoverer
- **Methodology**: Bottom-up peptide identification approach
- **Limitations**: Only those signals previously fragmented have MS2 signals, necessary for protein identification
- **Procedure**: Only those matching signals between ROIMCR-PLSDA and previous results from MS2 fragmented signals were considered to be studied by Proteome Discoverer in Mascot format
- **Drawback**: There are signals from ROIMCR-PLSDA result that were not identified due to lack of MS2 signals.

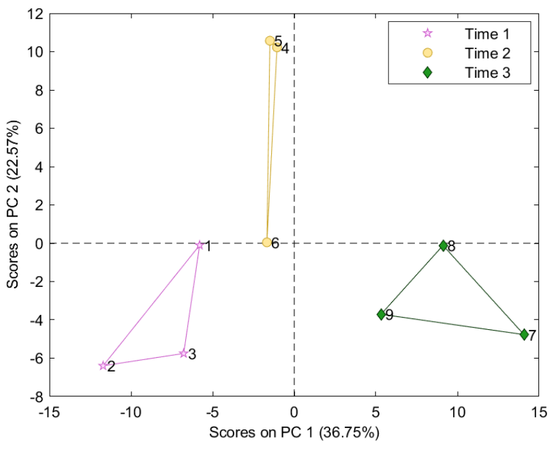
RESULTS

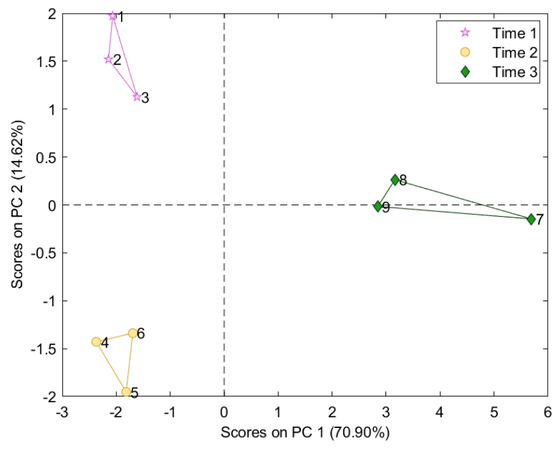
ROIMCR

As a result of the ROIMCR analysis, 181 'pure' components were obtained, achieving the explanation of 96.86% of the data variance . Most of these components can be associated with peptide signals. To assess the quality of these results the heights of the elution profiles of each components were analyzed.

PLS-DA

The 181 pure components from the ROIMCR analysis were studied using PLS-DA to evaluate which were the components responsibles of the differenciation among sampling times. Finally 41 of these components were selected.





This table shows the different proteins represented by the peptides selected as the main responsibles of the observed differences among samples collected at different times. The last 3 columns represent the number of peptides at each sampling time. This proteins are representatives of a large variety of species e.g human, mouse or bacterial species among others.

Proteins	Protein name	Organism name	Time 1	Time 2	Time 5
P35908	Keratin, type II cytoskeletal 2 epidermal	Homo sapiens	7	10	5
Q6IFZ6	Keratin, type II cytoskeletal 1b	Mus musculus	7	9	5
P04264	Keratin, type II cytoskeletal 1	Homo sapiens	12	17	6
Q981J9	60 kDa chaperonin 5	Mesorhizobium japonicum	1	1	1
B5YJN3	60 kDa chaperonin	Thermodesulfovibrio yellowstonii	1	1	1
A1K436	60 kDa chaperonin 1	Azoarcus sp.	1	1	1
Q5P7G2	60 kDa chaperonin	Aromatoleum aromaticum	1	1	1
A4G837	60 kDa chaperonin	Herminiimonas arsenicoxydans	1	1	1
Q1H4F2	60 kDa chaperonin	Methylobacillus flagellatus	1	1	1
A4SZV4	60 kDa chaperonin	Polynucleobacter asymbioticus	1	1	1
Q3A0V2	60 kDa chaperonin	Pelobacter carbinolicus	1	1	1
P13645	Keratin, type I cytoskeletal 10	Homo sapiens	1	9	2
P04259	Keratin, type II cytoskeletal 6B	Homo sapiens	2	4	0
A8EV70	ATP synthase subunit beta	Arcobacter butzleri	0	1	0
Q99895	Chymotrypsin-C	Homo sapiens	0	6	0
P01876	Immunoglobulin heavy constant alpha 1	Homo sapiens	0	7	2
Q6FF97	Elongation factor Tu	Acinetobacter baylyi	0	4	0
A3M1F6	Elongation factor Tu	Acinetobacter baumannii	0	4	0
P09093	Chymotrypsin-like elastase family member 3A	Homo sapiens	0	14	0
A5FLS1	ATP synthase subunit beta	Flavobacterium johnsoniae	0	1	0
A1ALL7	ATP synthase subunit beta 1	Pelobacter propionicus	0	1	0
Q82XP8	ATP synthase subunit beta	Nitrosomonas europaea	0	1	0
Q9Y6R7	IgGFc-binding protein	Homo sapiens	0	1	0
Q14533	Keratin, type II cuticular Hb1	Homo sapiens	0	0	1
P78385	Keratin, type II cuticular Hb3	Homo sapiens	0	0	1
P78386	Keratin, type II cuticular Hb5	Homo sapiens	0	0	1
O43790	Keratin, type II cuticular Hb6	Homo sapiens	0	0	1
A5A6M5	Keratin, type I cuticular Ha1	Pan troglodytes	0	0	1
O76009	Keratin, type I cuticular Ha3-I	Homo sapiens	0	0	1

CONCLUSIONS

- ✓ Data compression and analysis performed by ROIMCR allows the reduction of the amount of data without losing the accuracy of the proteomic signal of each sample
- ✓ The PLS-DA analysis allows the selection of those pure components with the peptide signals which represent the variability among sampling times
- ✓ Further work is pursued at present using LC-HRMS/MS analysis of the target MS signals selected by the combination of the ROIMCR and PLSDA procedures.

ACKNOWLEDGEMENTS

This work has been financially supported by the Generalitat de Catalunya (Consolidated Research Group "2017 SGR 1404 - Water and Soil Quality Unit"), by the Spanish Ministry of Science and Innovation (Project CEX2018-000794-S, PID2019-105732GB and PID2020-114065RB-C22)

REFERENCES

1- M. Carrascal et al. (2020) Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes. Sci. Total Environ., 747 Article 141145.

2- C. Perez-Lopez et al. (2021) Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method. J. Environ. Chem. Eng. Volume 9, Issue 4, August, 105752.

3- E. Gorrochategui et al. (2019) ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinforma., 20 (1), pp. 1-17.

4- S. Wold et al. (2001) PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab Syst., 58 (2) (2001), pp. 109-130