Aquasearch: a new software for fast proteomic characterization and classification of wastewater samples analyzed using MALDI-TOF.

Carlos Pérez-López (1), Antoni Ginebreda (1), Joaquín Abian (2), Damiá Barcelò (1), Montserrat Carrascal (2)

(1) Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Department of Environmental Chemistry, Jordi Girona 18–26, 08034 Barcelona, Spain
(2) Biological and Environmental Proteomics, Institute of Biomedical Research of Barcelona, Spanish National Research Council (IIBB-CSIC/IDIBAPS), Rosellón 161, E-08036 Barcelona, Spain

Email: carlos.perezlopez@cid.csic.es

INTRODUCTION

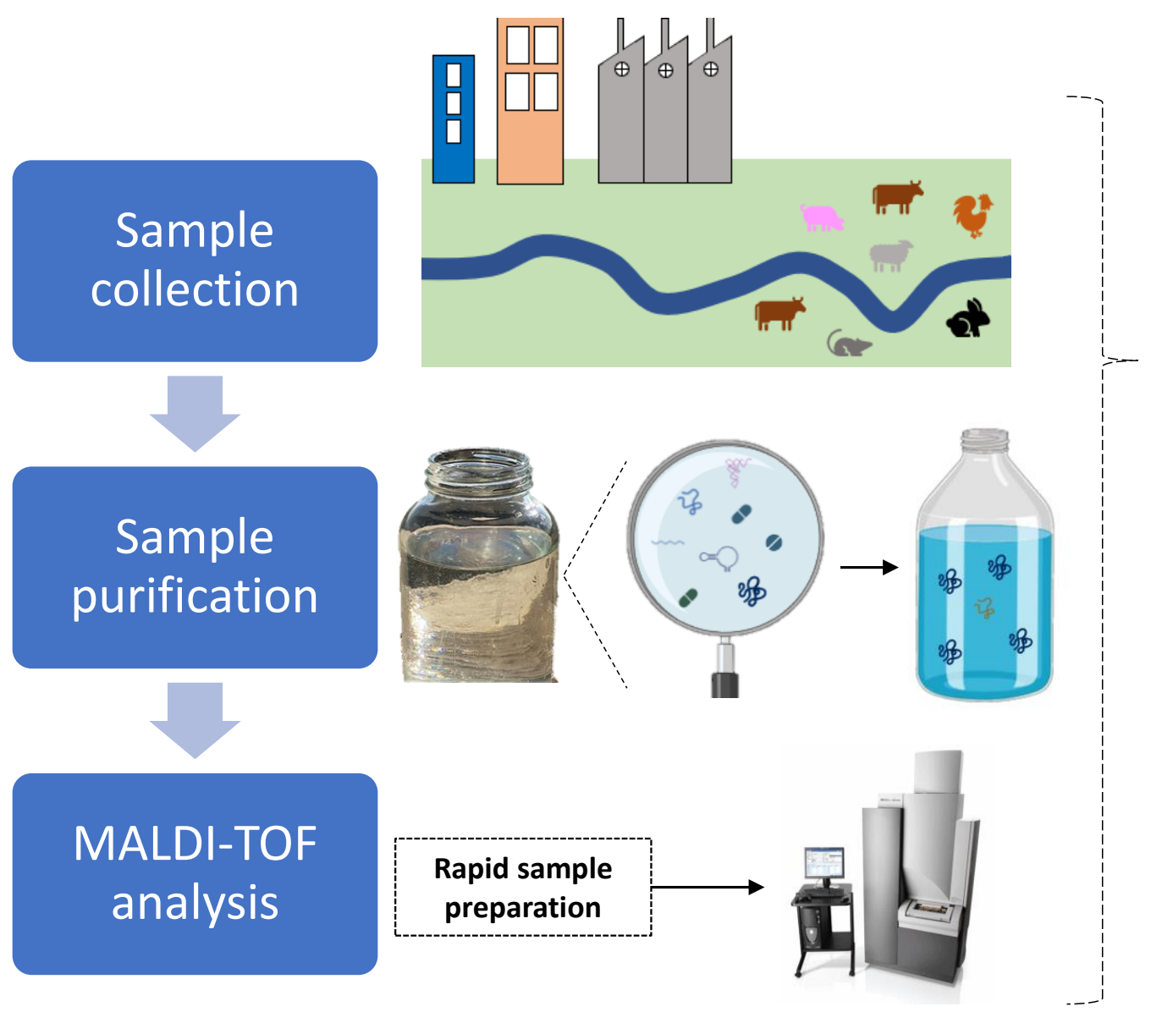
Traditionally, the study of wastewater has been focused on small molecules such as pharmaceuticals, illegal drugs or pesticides among others. However, recent studies have highlighted the valuable information provided by large molecules (proteins) present in wastewater [1, 2, 3], regarding the health and lifestyle of the population served by the system. Chromatographic techniques usually employed in shotgun proteomics obtains comprehensive information can be expensive and time-consuming. Therefore, Matrix-Assisted Laser Desorption/Ionization coupled with Time of Flight (**MALDI-TOF**) is proposed as a high-throughput instrumental approach for faster and more cost-effective sample characterization. In this work, we present **Aquasearch**, a newly developed software for a rapid characterization and classification of **proteomics in wastewater samples** analyzed with MALDI-TOF.

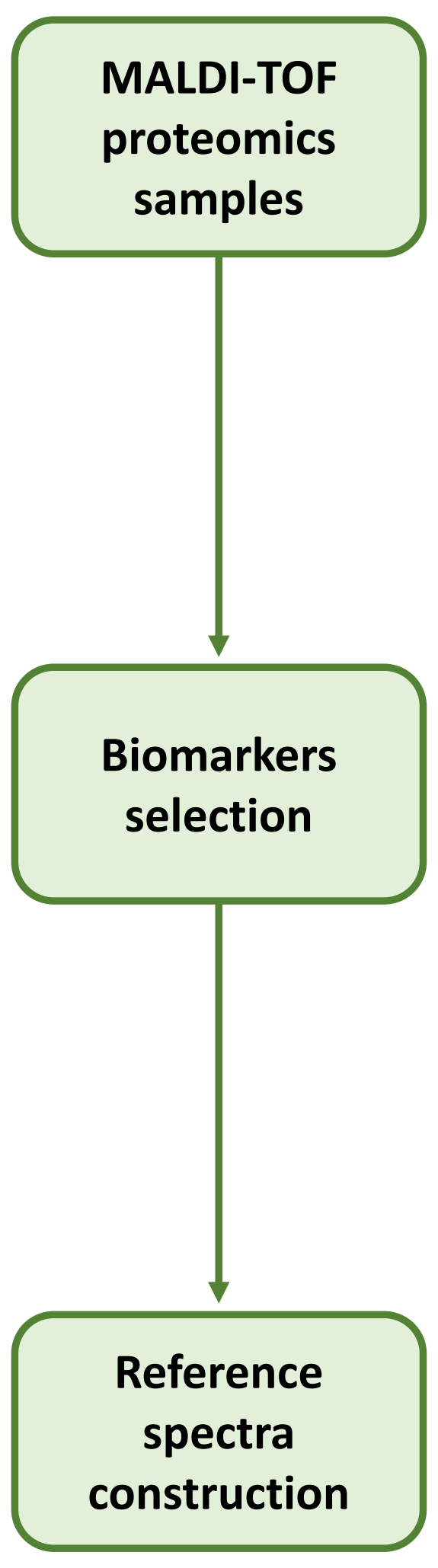
OBJECTIVES

- Construction of an in-house database with the peptides characteristics of each biomarker
- Development of matching and scoring systems to assess the presence or absence of a biomarker and the classification of the samples
- Complete an accurate pipeline for rapid characterization of proteomics in wastewater samples

METHODOLOGY

Database construction


Sample collection
Sample purification
MALDI-TOF analysis
Rapid sample preparation


MALDI-TOF proteomics samples
Biomarkers selection
Reference spectra construction

Spectral signals filtering and management — PYTHON

Database management — SQLite

Table 2: example of in-silico reference spectrum

Protein	Sequence	m/z	Relative intensity
P49065	211	977.4443	100
ʺ	212	1288.6256	1.96
ʺ	213	1292.6565	1.3
ʺ	156	1311.6927	8.38
ʺ	187	1410.6372	100
ʺ	157	1467.7164	100
ʺ	188	1472.6852	40.45
ʺ	214	1525.7769	4.19
ʺ	215	1528.7169	8.27
ʺ	189	1568.6667	38.26
ʺ	216	1902.8682	13.63
ʺ	190	2058.9517	67.61
ʺ	82	2113.8494	13.78
ʺ	83	2247.9231	19.76
ʺ	191	2315.0488	33.33
ʺ	62	2612.0861	97.27

Table 1: Chosen biomarkers for sample characterization

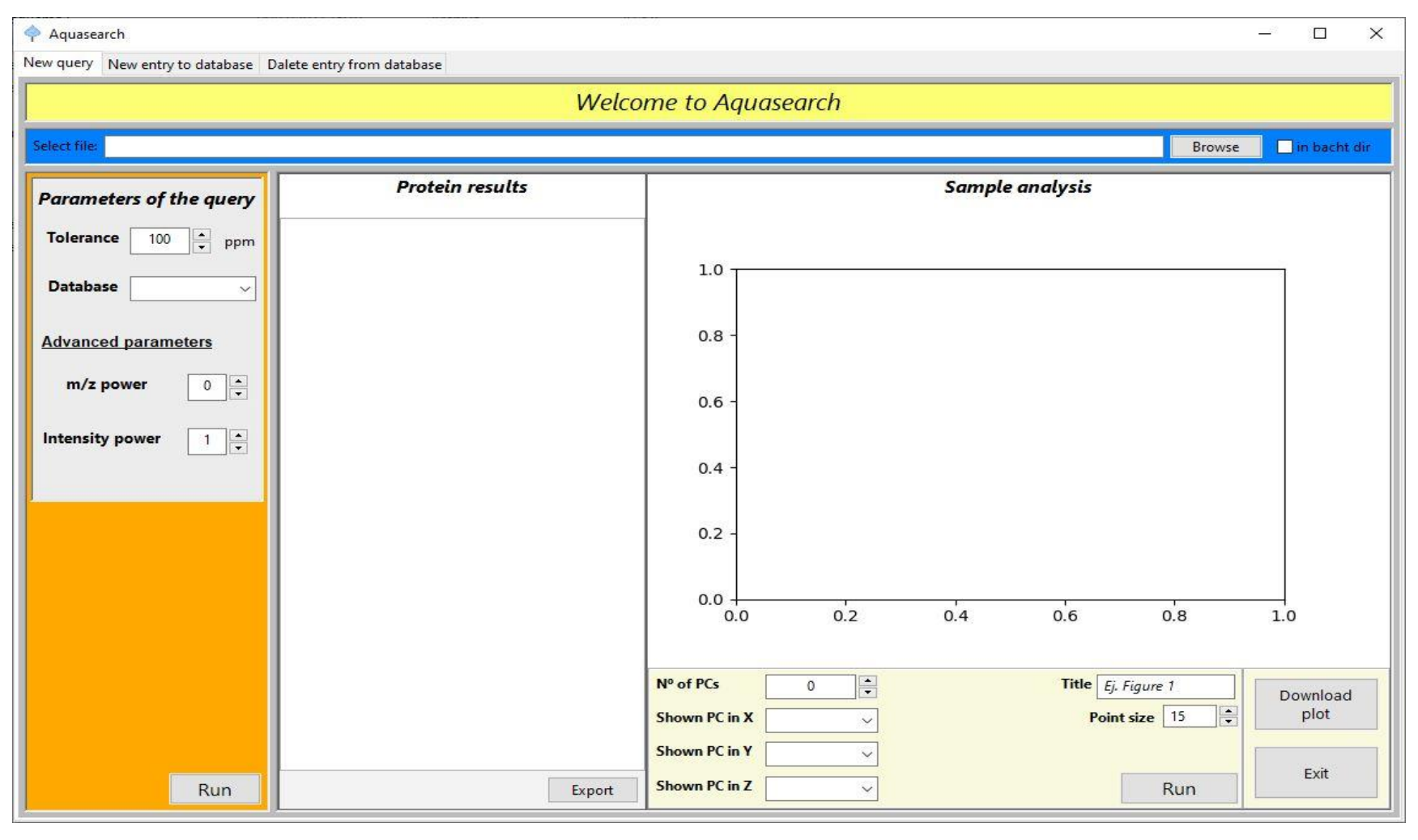
Protein code	Protein name	Organism
P04746	Pancreatic alpha-amylase	*Homo sapiens*
P08835	Albumin	*Sus scrofa*
P0DUB6	Alpha-amylase 1A	*Homo sapiens*
P19961	Alpha amylase 2B	*Homo sapiens*
P02769	Albumin	*Bos taurus*
P19121	Albumin	*Gallus gallus*
P02768	Albumin	*Homo sapiens*
P14639	Albumin	*Ovis aries*
P35747	Albumin	*Equus caballus*
P49065	Albumin	*Oryctolagus cuniculus*
P07724[a]	Albumin	*Mus musculus*
P02770[a]	Albumin	*Rattus norvegicus*
P00687[b]	Alpha-amylase 1	*Mus musculus*
P00689[b]	Pancreatic alpha-amylase	*Rattus norvegicus*

[a] peptides enclosed in 'Murid albumin (P07724;P02770)'
[b] peptides enclosed in 'Murid pancreatic (P00687;P00689)'

Application development

Graphic user interface (GUI) development tools

wxGlade — PYTHON


Aquasearch
New query | New entry to database | Delete entry from database
Welcome to Aquasearch
Select file: Browse ☐ In bacht dir
Parameters of the query: Tolerance 100 ppm; Database; Advanced parameters: m/z power 0; Intensity power 1; Run
Protein results: Export
Sample analysis: Nº of PCs 0; Shown PC in X; Shown PC in Y; Shown PC in Z; Title Ej. Figure 1; Point size 15; Download plot; Run; Exit

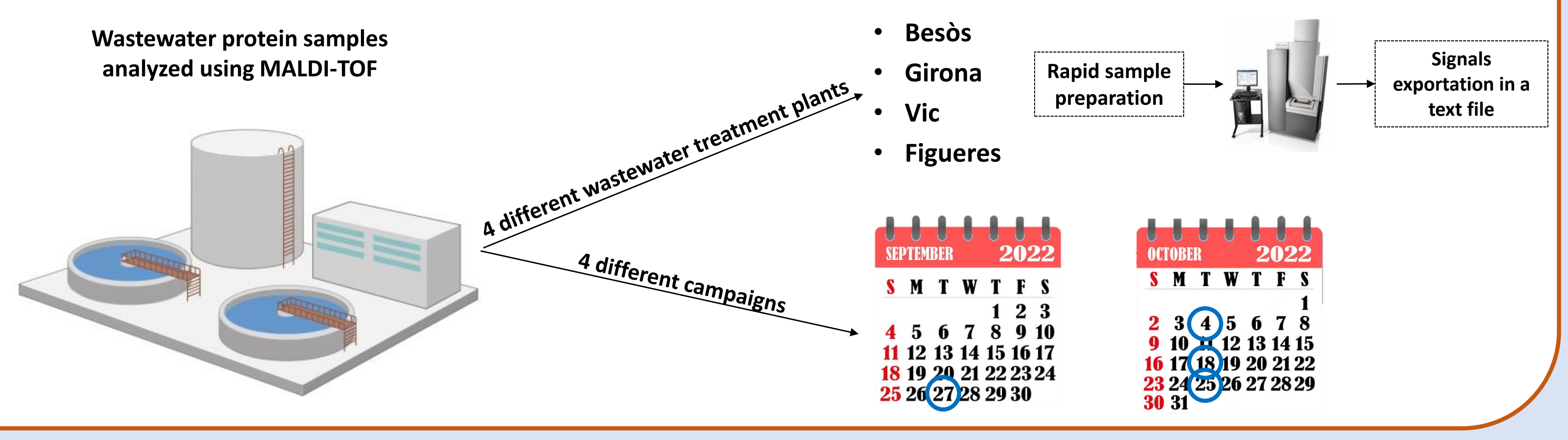
Application functions

- Characterize new samples.
- Classify samples depending on their proteomic profile (in the case of a multisampling study).
- Add or delete ~~spectral~~ examples to the database.
- Add or delete biomarkers from database.

Application Output

- Score of the presence of the biomarker in the sample.
- Total number and sequences of the peptides identified for each biomarker (including the number of unique peptides).
- PCA resulting from the proteomics profile of the samples (in the case of a multisampling study).

Program testing


Wastewater protein samples analyzed using MALDI-TOF

4 different wastewater treatment plants:

- Besòs
- Girona
- Vic
- Figueres

Rapid sample preparation → Signals exportation in a text file

4 different campaigns

RESULTS

Database construction

Database summarize:

- Samples used to build database: **30 mix samples + 18 standard samples** of some proteins (P08835, P19121, P02768, P49065, P07724, P02770).
- Total number of m/z signals identified in the samples: **1825.**
- Total number of different peptides: **229.**
- Number of unique peptides among the total identified peptides: **85.**

Table 3: number of peptides (and unique peptides) in reference spectrum for each biomarker

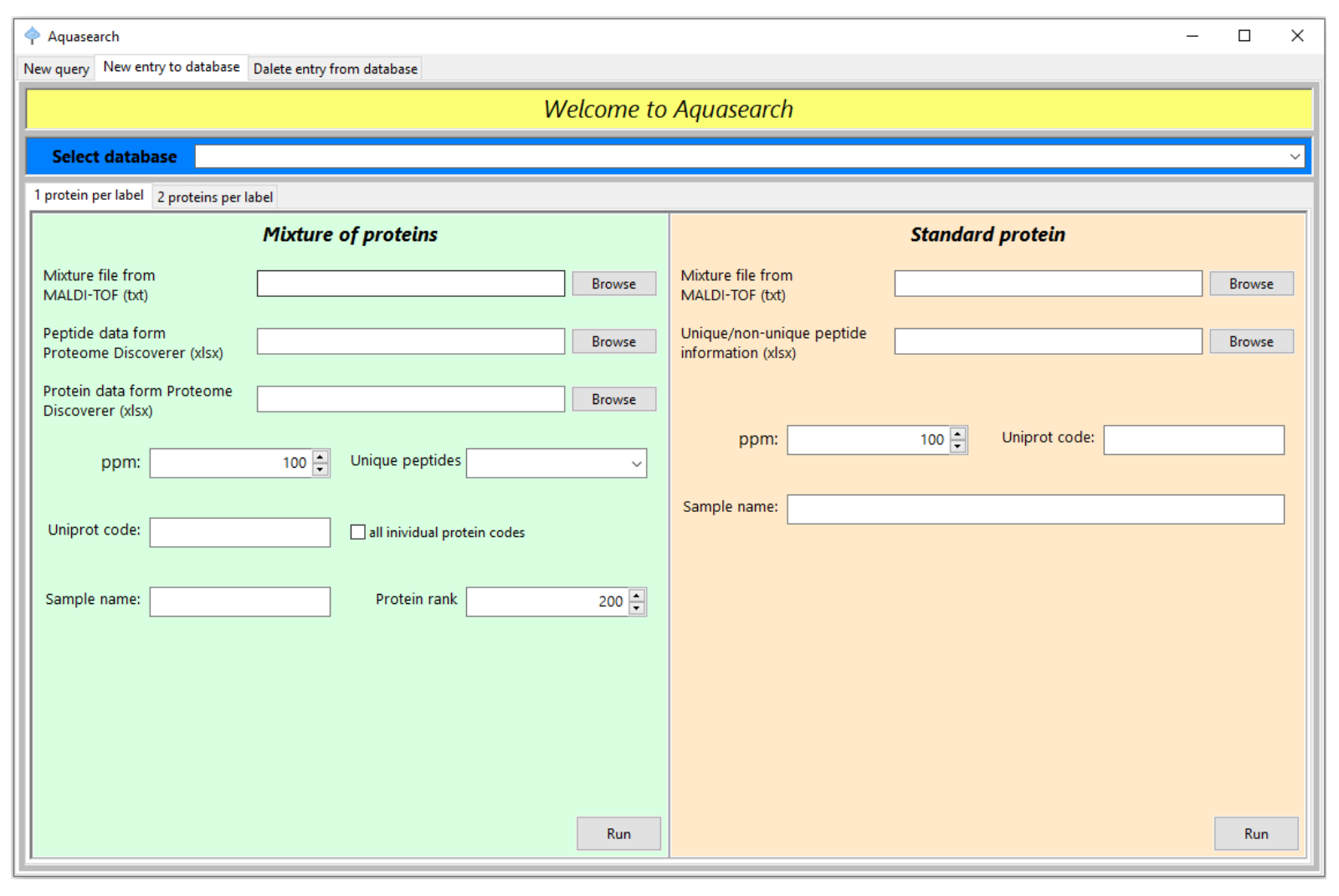
Protein code	Nº Peptides	Nº unique peptides
P04746	39	4
P08835	51	23
P0DUB6	43	3
P19961	41	0
P02769	40	8
P19121	35	22
P02768	35	7
P14639	19	0
P35747	6	1
P49065	16	8
P07724[a]	19	8
P02770[a]	-	-
P00687[b]	11	1
P00689[b]	-	-

[a] peptides enclosed in 'Murid albumin (P07724;P02770)'
[b] peptides enclosed in 'Murid pancreatic (P00687;P00689)'

Application development

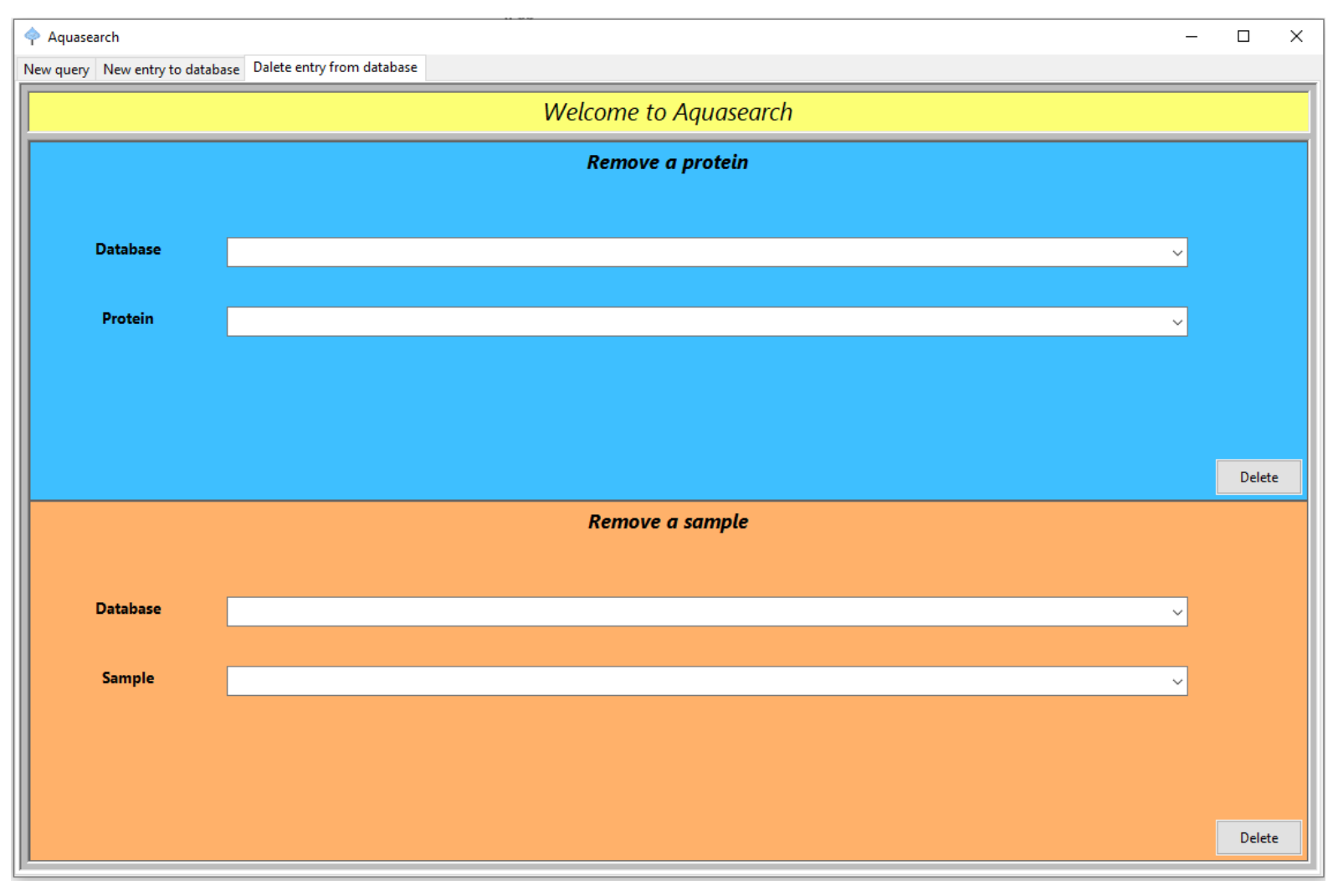
New entry to database

- Depending on the number of proteins:
  - 1 Protein per label
  - 2 Protein per label
- Depending of the sample nature:
  - Mix of proteins
  - Standard


Aquasearch
New query | New entry to database | Dalete entry from database
Welcome to Aquasearch
Select database
1 protein per label | 2 proteins per label

Mixture of proteins
Mixture file from MALDI-TOF (txt): Browse
Peptide data form Proteome Discoverer (xlsx): Browse
Protein data form Proteome Discoverer (xlsx): Browse
ppm: 100; Unique peptides
Uniprot code: ☐ all inividual protein codes
Sample name:; Protein rank 200
Run

Standard protein
Mixture file from MALDI-TOF (txt): Browse
Unique/non-unique peptide information (xlsx): Browse
ppm: 100; Uniprot code:
Sample name:
Run


Aquasearch
New query | New entry to database | Delete entry from database
Welcome to Aquasearch
Remove a protein: Database; Protein; Delete
Remove a sample: Database; Sample; Delete

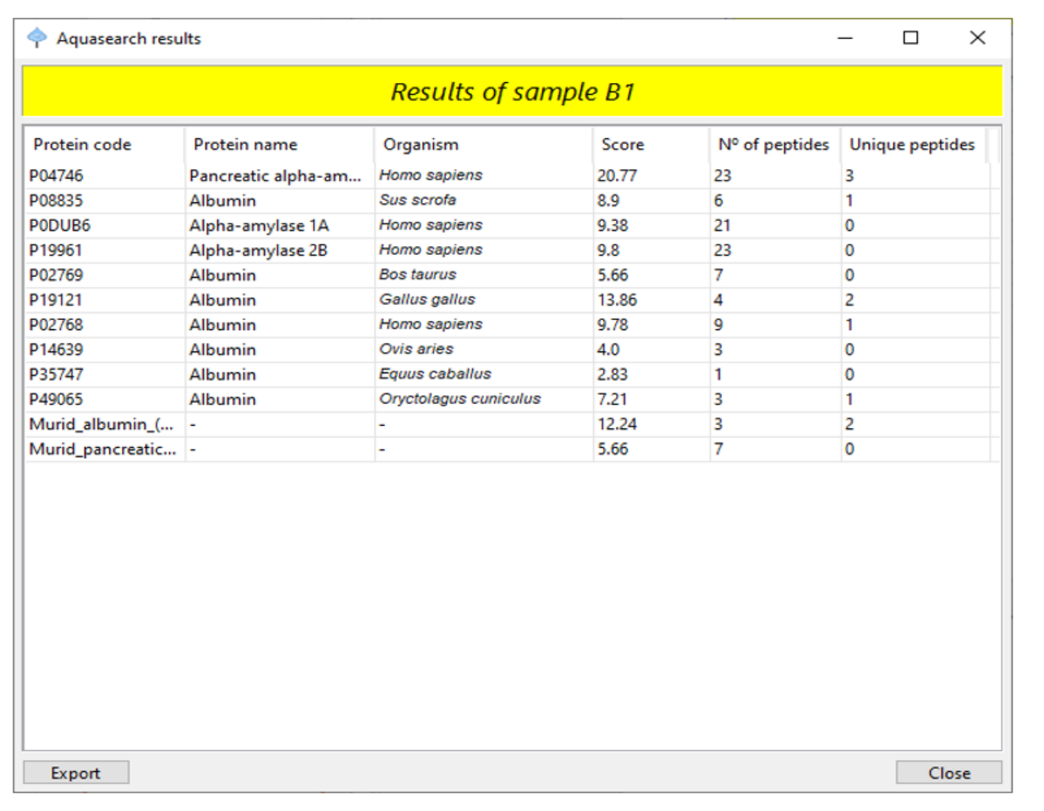
Delete entry from database

- Remove the biomarker signals from a database
- Remove the sample signals from a database

Removing a protein or a sample, all its signals are deleted and the rest of the signals are filtered again and the reference spectra recalculated

Program testing

A)

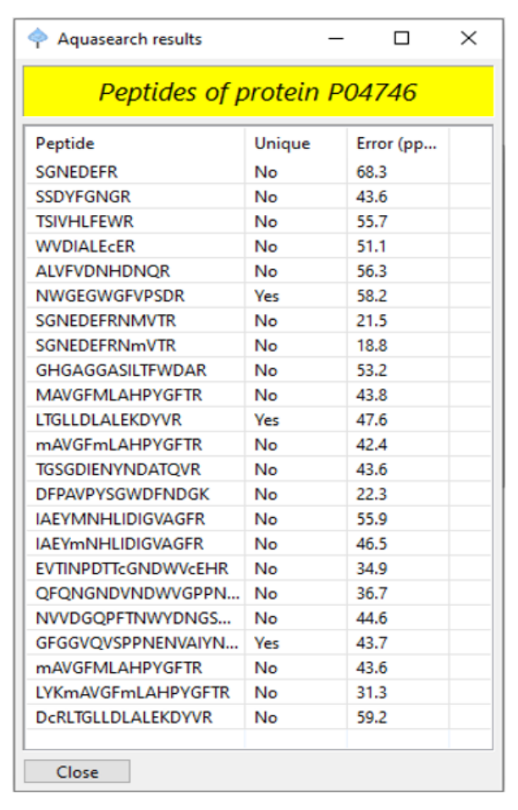

Aquasearch results

Results of sample B1

Protein code	Protein name	Organism	Score	Nº of peptides	Unique peptides
P04746	Pancreatic alpha-am...	*Homo sapiens*	20.77	23	3
P08835	Albumin	*Sus scrofa*	8.9	6	1
P0DUB6	Alpha-amylase 1A	*Homo sapiens*	9.38	21	0
P19961	Alpha-amylase 2B	*Homo sapiens*	9.8	23	0
P02769	Albumin	*Bos taurus*	5.66	7	0
P19121	Albumin	*Gallus gallus*	13.86	4	2
P02768	Albumin	*Homo sapiens*	9.78	9	1
P14639	Albumin	*Ovis aries*	4.0	3	0
P35747	Albumin	*Equus caballus*	2.83	1	0
P49065	Albumin	*Oryctolagus cuniculus*	7.21	3	1
Murid_albumin_(...	-	-	12.24	3	2
Murid_pancreatic...	-	-	5.66	7	0

Export | Close

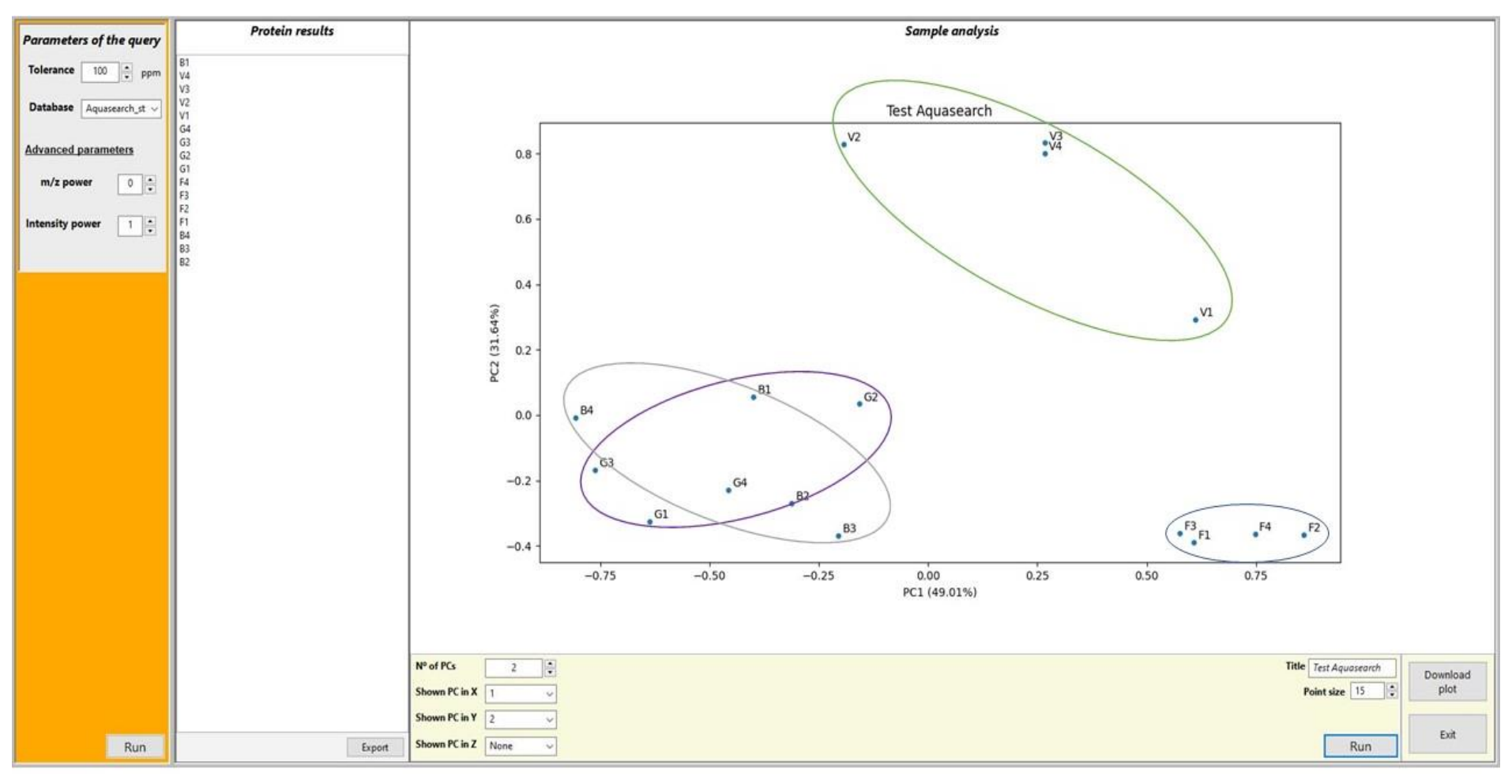
B)


Aquasearch results

Peptides of protein P04746

Peptide	Unique	Error (pp...
SGNEDEFR	No	68.3
SSDYFGNGR	No	43.6
TSIVHLFEWR	No	55.7
WVDIALEcER	No	51.1
ALVFVDNHDNQR	No	56.3
NWGEGWGFVPSDR	Yes	58.2
SGNEDEFRNMVTR	No	21.5
SGNEDEFRNmVTR	No	18.8
GHGAGGASILTFWDAR	No	53.2
MAVGFMLAHPYGFTR	No	43.8
LTGLLDLALEKDYVR	Yes	47.6
mAVGFmLAHPYGFTR	No	42.4
TGSGDIENYNDATQVR	No	43.6
DFPAVPYSGWDFNDGK	No	22.3
IAEYMNHLIDIGVAGFR	No	55.9
IAEYmNHLIDIGVAGFR	No	46.5
EVTINPDTTcGNDWVcEHR	No	34.9
QFQNGNDVNDWVGPPN...	No	36.7
NVVDGQPFTNWYDNGS...	No	44.6
GFGGVQVSPPNENVAIYN...	Yes	43.7
mAVGFMLAHPYGFTR	No	43.6
LYKmAVGFmLAHPYGFTR	No	31.3
DcRLTGLLDLALEKDYVR	No	59.2

Close


Parameters of the query: Tolerance 100 ppm; Database Aquasearch_st; Advanced parameters: m/z power 0; Intensity power 1; Run
Protein results: B1, V4, V3, V2, V1, G4, G3, G2, G1, F4, F3, F2, F1, B4, B3, B2; Export
Sample analysis: Nº of PCs 2; Shown PC in X 1; Shown PC in Y 2; Shown PC in Z None; Title Test Aquasearch; Point size 15; Run; Download plot; Exit

*The circles have been manually added to group the samples from the same wastewater treatment plant. The analysis is completely unsupervised

Example of identification

Rsults of the sample from the Besos River (NE Spain) firstly collected (B1) are depicted.

The Score indicates:

- Score > 4: Unlikely presence of the protein
- 4 < Score > 5: The presence of the protein is probable (decide with the other parameters)
- **Score > 5: The protein is in the sample**

Classification of the samples (multisampling studies)

The protein scores shown in above image are used to carry out a Principal Component Analysis (PCA) [4]. As a result of the analysis of the 16 samples (4 samples from 4 wastewater treatment plant), **3 well defined groups are obtained**:

1. Urban areas (Besos and Girona)
2. Poultry activity area (Figueres)
3. Pork activity area (Vic)

CONCLUSIONS

- ✓ The MALDI-TOF analytical technique has a huge potential for a rapid characterization of proteomics in wastewater samples, previous to a more comprehensive analysis with a more expensive and time consuming techniques such as LC-HRMS.
- ✓ The Aquaserch software has built a representative in-house database with some of the biomarkers associated with the presence of animal and human activity to characterize and classify the samples depending on these biomarkers.
- ✓ The score punctuation reported by Aquaserch for each biomarker can identify accurately the presence or absence of the studied biomarkers in the samples and classify them in a multisampling study.
- ✓ Aquasearch is the unique proteomic screening application tested in real wastewater samples. **Aquasearch enables to effectively identify protein contaminations in a rapid and high-throughput way.**

ACKNOWLEDGEMENTS

Carlos Pérez-López acknowledges the predoctoral scholarship FPI 2019–090182 included in the Grant CEX2018-000794-S funded by MCIN/AEI/10.13039/ 501100011033. This work has been financially supported by the Spanish Ministry of Science and Innovation (Project CEX2018-000794-S and PID2020-114065RB-C22 /MCIN/AEI /10.13039/501100011033; WaterPROT).

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. M.Carrascal et al. (2023) SewageProteinInformationMining:Discoveryof Large Biomoleculesas Biomarkersof Populationand IndustrialActivities. Environ. Sci. Technol. 2023, 57, 30, 10929–10939.
4. S. Wold, K. Esbensen, and P. Geladi, 'Principal component analysis', Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52, 1987