UMSA

8. UMSA

Description

UMSA is an untargeted mass-spectrometry analysis service from RECETOX (Research Centre for Toxic Compounds in the Environment at Masaryk University) in the Czech Republic. The service is evolving to a key component of the emerging EIRENE ESFRI. By means of the integration in EOSC, uniform access to data and computing resources are provided, scaling the service to the target European-wide user community. Typically, mass spectrometry is done in a targeted way to confirm or disprove the presence of a specific compound in a sample. On the contrary, we aim at processing data to correlating the whole spectra (ie. all the present compounds) with other data (social, medical, other sample analyses, etc.) to work with more complex hypotheses of environmental impacts on human health.

The data are unrecoverable, original samples cannot be re-acquired, therefore long-term data storage (even decades) is required, together with appropriate data curation. Tracking provenance of the secondary (derived) datasets(what was the exact process of generating them from the original source data), is fairly critical, as the results may differ dramatically with different settings. Similarly, the exact links between datasets and physical samples they originate from must be maintained.

The current release provides a Galaxy workflow based on re-factored tools originating from Emory university (apLCMS and xMSAnnotator), which detect peaks in the input spectra and matches those to metabolite and pathway databases. Extending the workflow with auto-tuning peak picking parameters (based on the original xMSAnalyzer tool) is in progress. The workflow continues with filtering false positives by predicting their chromatographic retention time (adapted Retip tool) based on computing chemical descriptors and machine learning with respect to a large database of known compounds (doi: 10.1038/s41467-019-13680-7). With respect to the previous release, the service was extended with a set of tools supporting gas chromatography MS (XCMS, RamclustR, MatchMS) and several tools to support conversions among various chemical identifier standards.

Architecture

The service is deployed as a virtual cluster using an Infrastructure Manager RADL recipe. The cluster consists of a single head node, running Galaxy frontend and the Slurm server, and an arbitrary number of Slurm worker nodes. Data is shared among the nodes over NFS.

The deployment of the head node registers a configurable dynamic DNS name (umsa.dyn.cerit-sc.cz currently) to point to its assigned IP address. The well known service endpoint (umsa.cerit-sc.cz) is expected to be an alias pointing to the dynamic one. In the next step a Let's encrypt certificate is acquired to allow smooth https connection. The RADL recipes also require simple provider-specific customization (cloud network names and base image identifiers in particular).

Authentication of the end users is managed by ELIXIR AAI via its corresponding Galaxy module. Configuration of Elixir AAI is the only manual step in the service deployment; the policy of introducing a new service to Elixir AAI requires human approval and exchange of secrets which cannot be automated so far. In the current version, the service can be accessed bona-fide -- all users who pass ELIXIR AAI authentication are allowed. However, we plan a trivial registration procedure (using the ELIXIR tools) to restrict the access.

The payload of the service are several tools in Galaxy. In order to keep strict control on the complex software dependencies, we wrap all the software to Docker containers; Galaxy and Slurm are configured to execute them in this way only. The tools themselves are installed from standard Galaxy toolshed to follow common procedures.

The diagram in figure 26 shows the essential mass-spec data processing workflow. Besides the user input (.mzML spectra files) the workflow needs to query a metabolite database. In order to achieve reproducibility, timestamped snapshots of the online databases are used. They are downloaded and processed only from time to time, and the snapshots are available to all users. Most of the intermediate files passed between the tools are some kind of tabular data, using the HDF5 or Apache feather format.

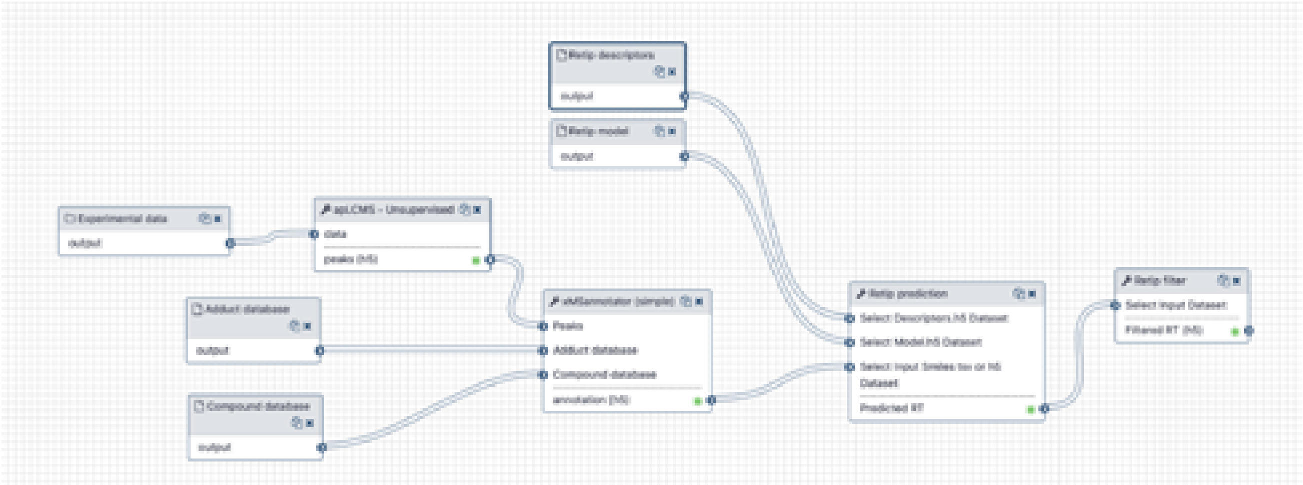


Figure 26 - UMSA LC/MS Data processing Workflow

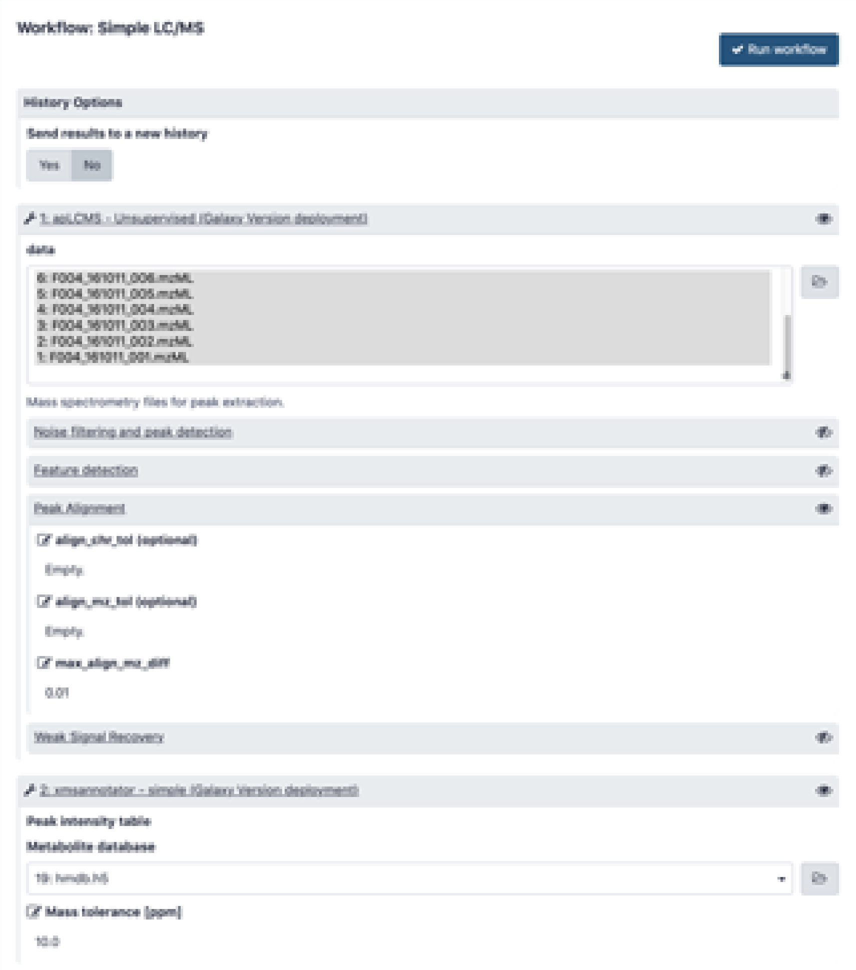
8. UMSA

Service Endpoint

The principal service endpoint is https://umsa.cerit-sc.cz/. The user interface is standard Galaxy with minimalistic visual branding. Login with Elixir AAI credential is required as described above.

The individual tools are available in PeakPicking and Annotation sections, both exist in simple and advanced forms to address the needs of different user experience levels. The in-line documentation provides extensive description of the meaning of numerous input parameters, as well as appropriate references to web pages with further documentation of the tools, and the essential journal papers.

Figure 27 shows a typical input form of a simple workflow connecting the tools together.



Workflow: Simple LC/MS

History Options

Send results to a new history

Yes [✓ No]

1: apLCMS - Unsupervised (Galaxy Version deployment)

data

6: F004_161011_006.mzML
5: F004_161011_005.mzML
4: F004_161011_004.mzML
3: F004_161011_003.mzML
2: F004_161011_002.mzML
1: F004_161011_001.mzML

Mass spectrometry files for peak extraction.

Noise filtering and peak detection

Feature detection

Peak Alignment

align_chr_tol (optional)

Empty.

align_mz_tol (optional)

Empty.

max_align_mz_diff

0.01

Weak Signal Recovery

2: xmsannotator - simple (Galaxy Version deployment)

Peak intensity table

Metabolite database

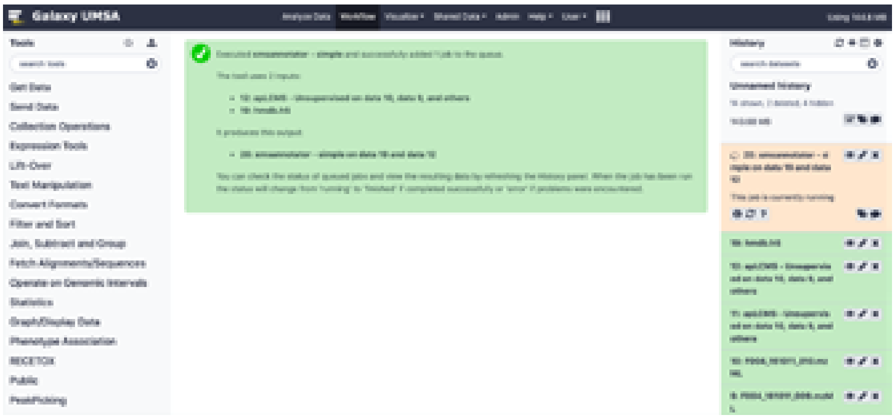
19: hmdb.h5

Mass tolerance [ppm]

10.0

Figure 27 - Screenshot of an input form of a simple workflow in UMSA

Figure 28 shows execution of this workflow in progress, generating its output files in Galaxy history.



Executed umsaannotator - simple and successfully added 1 job to the queue.

The tool uses 2 inputs:

- 12: apLCMS - Unsupervised on data 10, data 8, and others
- 18: hmdb.txt

It produces this output:

- 20: umsaannotator - simple on data 18 and data 12

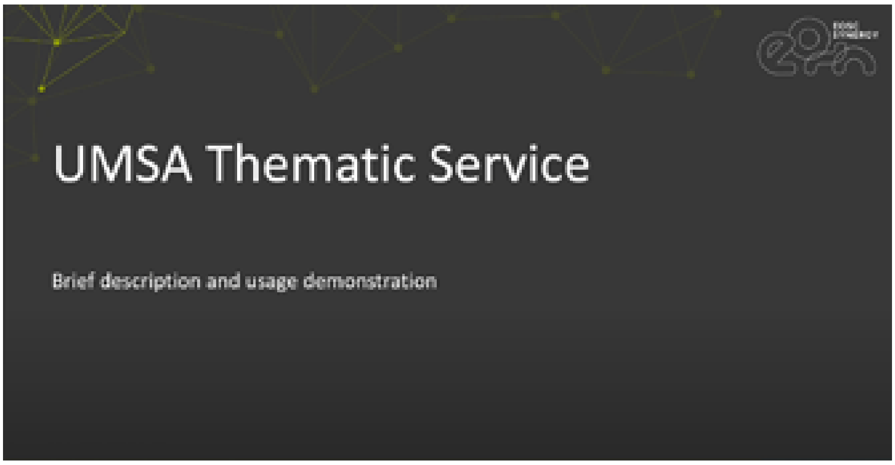
You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

This job is currently running

Figure 28 - Screenshot of the execution of the previous workflow in progress. The peak-picking step (apLCMS) has already finished, annotation is running

Demonstration Video

Click the image to view the video



EOSC SYNERGY

UMSA Thematic Service

Brief description and usage demonstration