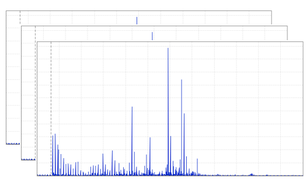
Molecular Networking and *in-silico* MS/MS Database : a workflow to dereplicate and visualize results

I LC-MS/MS analysis of crude extract

- use ramp energy gradient or combined energies (ex: 15,30 and 45 eV) for optimal MS/MS spectrum coverage

II Conversion of proprietary to .mzXML format

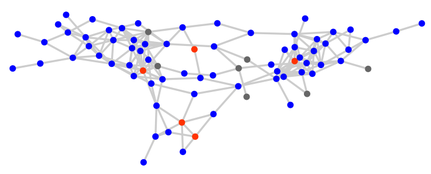
- use ProteoWizard (http://proteowizard.sourceforge.net/)



proteowizard

III Generate Molecular Networks on GNPS

- GNPS server : http://gnps.ucsd.edu
- follow instructions at https://bix-lab.ucsd.edu/display/Public/Molecular+Networking+Documentation for optimal parameters and MN visualization in Cytoscape
- Cytoscape is available at http://www.cytoscape.org/



IV Fetch clustered data from the MN on GNPS

- In the GNPS results page, hit «Download Clustered Data». You will get a folder containing files as described on **Fig. 1**

- the MN attributes file appears as an .out file in the folder. Let's call it **cytoscape_attributes.out**
- the clustered spectra appears as a .mgf file in the folder. Let's call it **your_spectra.mgf**

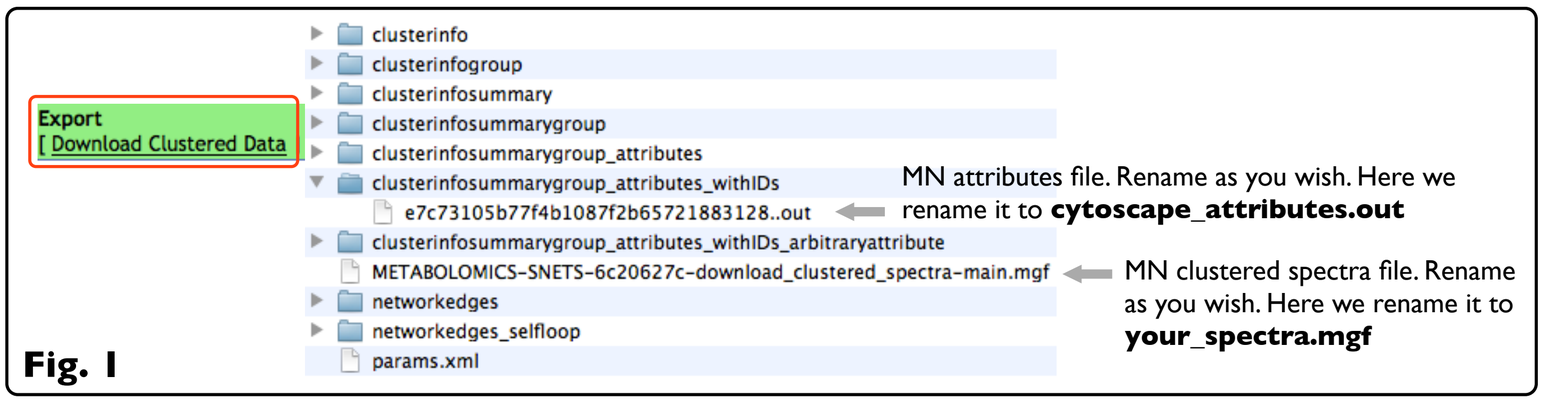


Fig. 1

V Library search in the ISDB

- The first step is to make sure you have a Linux based system, since Tremolo, wich is used for the spectral matching stage, runs under Linux. If you don't have a Linux-based system you can easily install Ubuntu on your Windows or Mac OS via a virtual machine. See instructions here : http://www.wikihow.com/Install-Ubuntu-on-VirtualBox

- Download the UNPD-ISDB and scripts at http://oolonek.github.io/ISDB/

- in order to easily perform library search using Tremolo and merge results with the MN attributes file, we wrote a python script (treat.py) and a bash script (run.sh). These scripts and the UNPD_ISDB files should be placed in your Linux system (respect folders names) as described in **Fig. 2**

- to adjust the library search parameters you nead to edit the **run.sh** file (open with text editor)

The important parameters to edit are seen on **Fig. 3**

- TOLERANCE: ± tolerance for parent mass search in Da. Set a small tolerance for dereplication using parent ion mass as prefilter, keeping in mind the resolution of your data. Increase to the wanted range for variable dereplication search (ex: 100 or 200 Da) Caution as this will also increase calculation times !

- SCORE_THRESHOLD: should be kept low when using *in-silico* DB. Typically 0.2 to 0.3.

- TOP_K_RESULTS: Defines the maximal number of results returned

To launch the search open a terminal window and navigate to the results folder. Type:

```
bash run.sh your_spectra.mgf cytoscape_attributes.out results.out
```

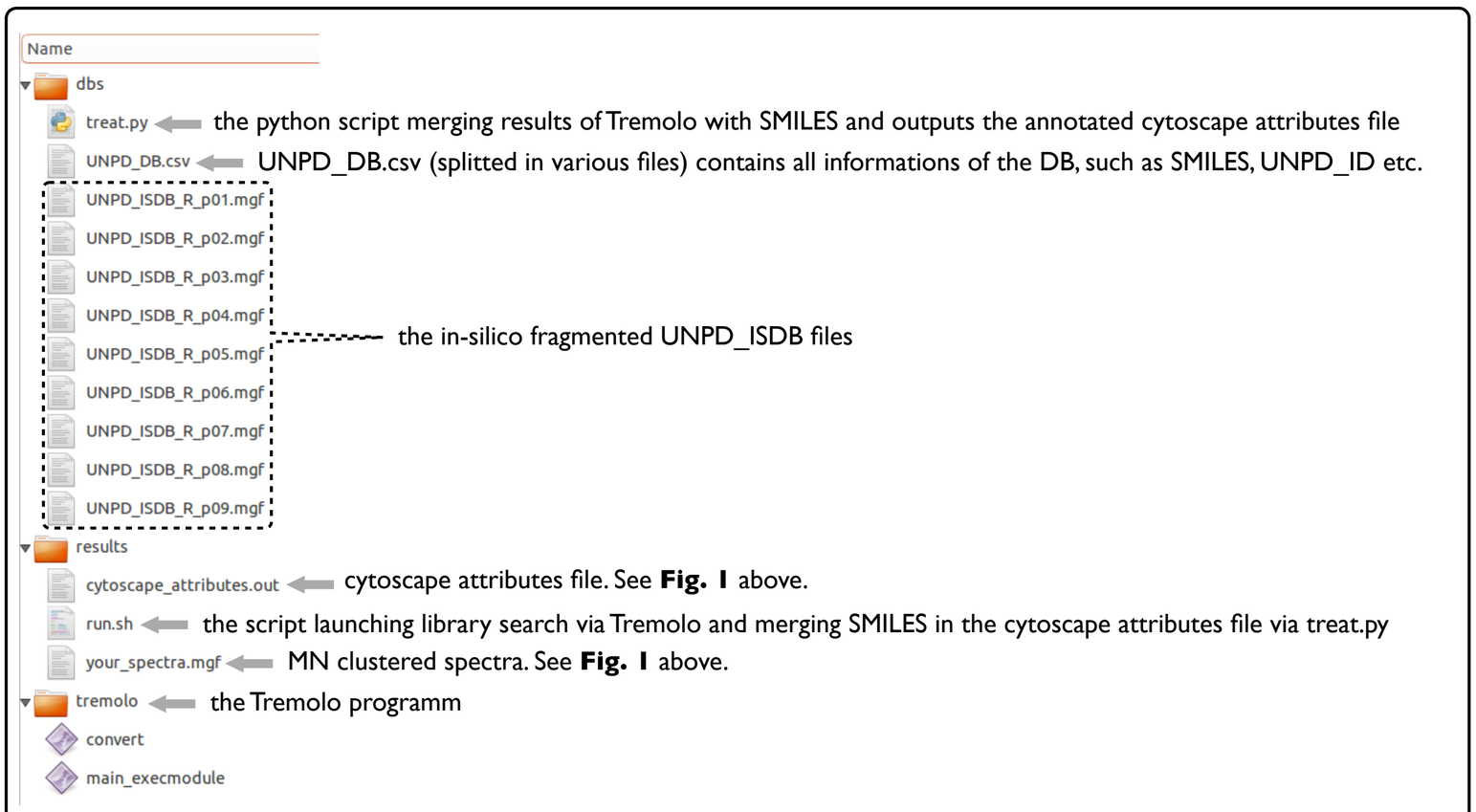


Fig. 2 Organization of the files & folders to run the ISDB library search

```
# Set the tolerance to be used
TOLERANCE=0.005
# Score threshold in Da
SCORE_THRESHOLD=0.2
# Top K results
TOP_K_RESULTS=5
```

Fig. 3 Library search parameters to edit in **run.sh**

VIII Visualize dereplication results in Cytoscape

- Install chemViz plugin for Cytoscape 2.8 (http://apps.cytoscape.org/apps/chemviz) or chemViz2 for Cytoscape 3.x (http://apps.cytoscape.org/apps/chemviz2)
- In Cytoscape load your network then load the **results.out** file as attribute file with corresponding SMILES and IDs
- select nodes of interest, right click and under *Chemoinformatic tools* select *Show structures window*.

Start exploring the network !

Examples of results visualization

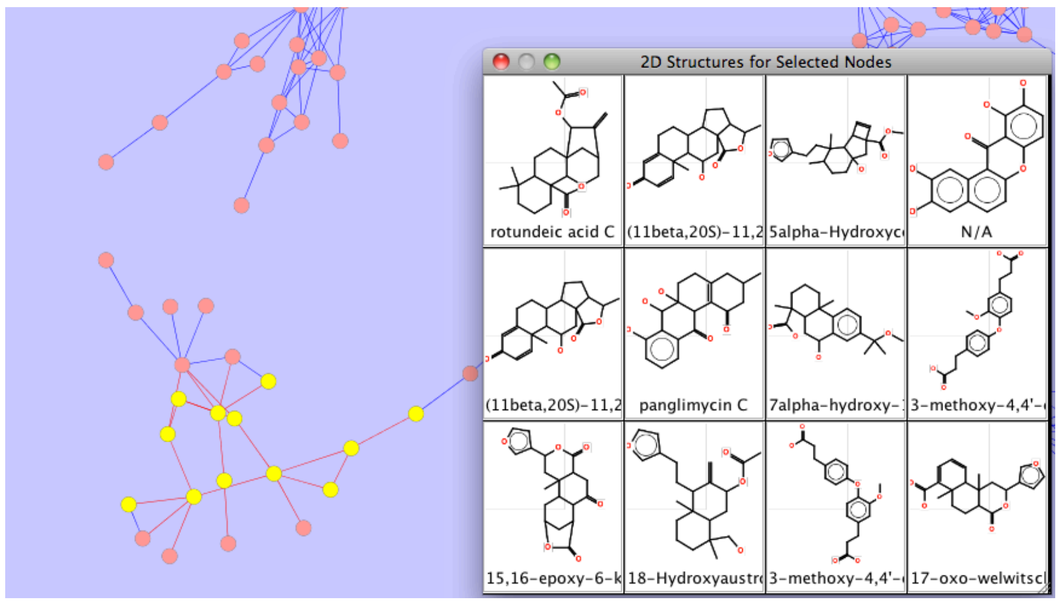


Fig. A - variable dereplication against the UNPD-ISDB indicates the possible presence of diterpenoids for this extract of *Salvia* sp.

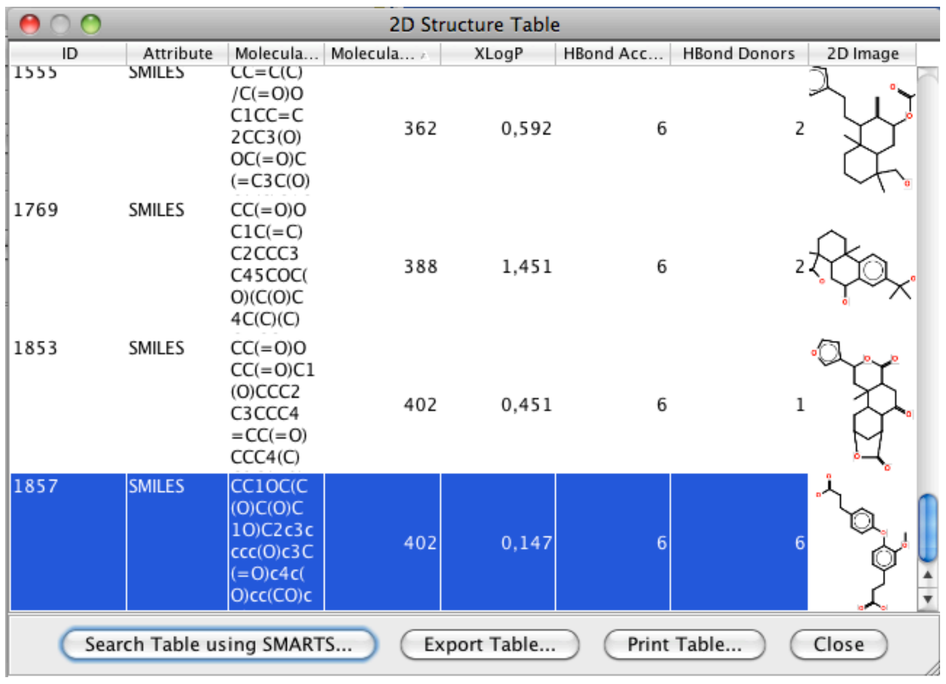


Fig. B - structures can be viewed as a table wich can in turn be searched for substructures or exported.

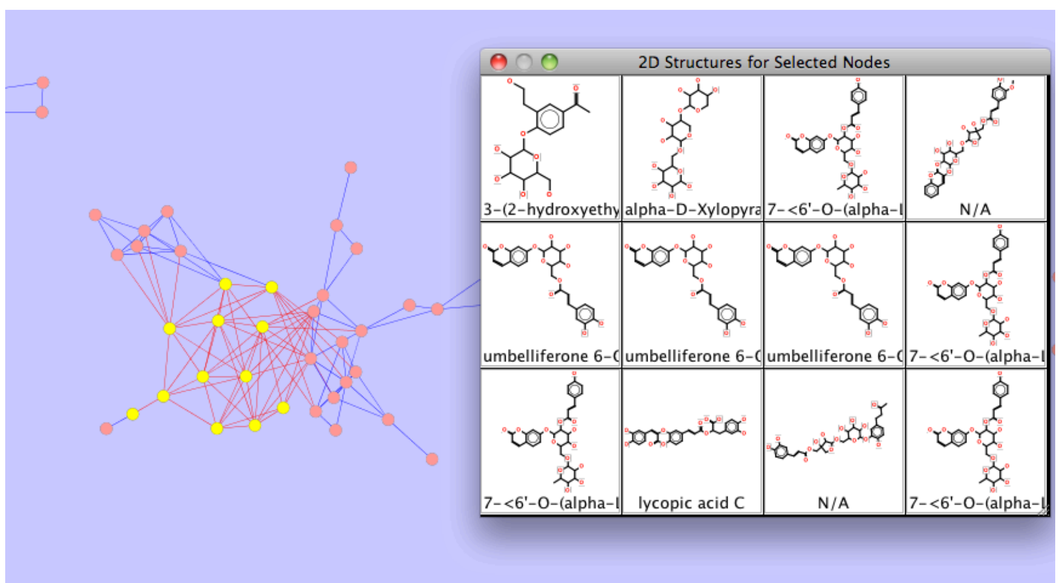


Fig. C - here caffeoyl and coumarin glycosylated compounds seem to hide under this cluster.

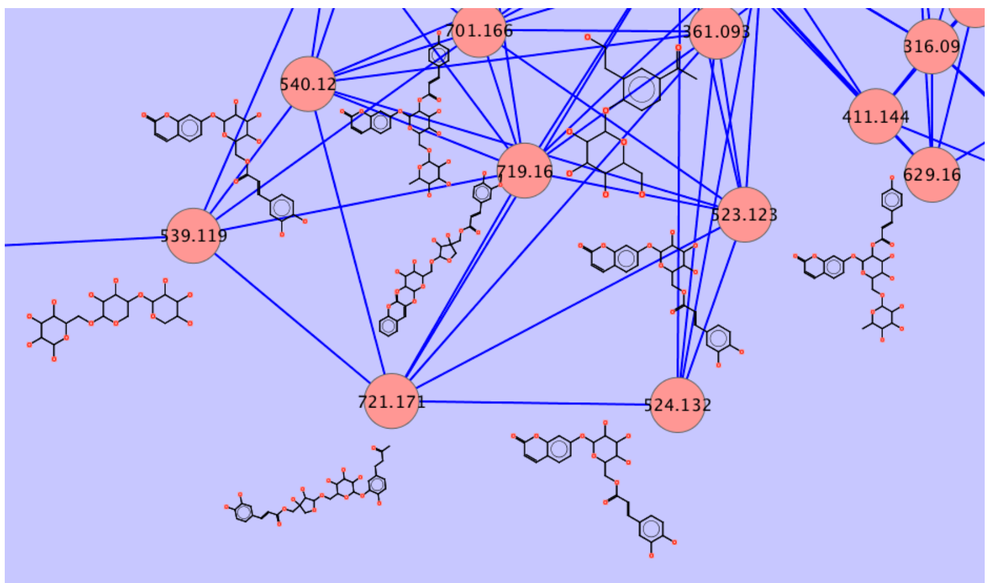


Fig. D - structures can also be directly displayed on top of the nodes

Notes

- Outliers in the panel of proposed structures should be spotted (see **Fig. A** for ex.) and attention should be focussed on compounds of a common class or bearing common structural functionalities that can lead to similar MS/MS fragmentation data (sugars, aliphatic side chains etc ...)
- Merging orthogonal informations : phylogenetic data, comparison of hits logP with experimental retention times, input of exact mass for the molecular formula determination is the key to assess the relevance of the hits.
- More than an strict dereplication tool, using variable dereplication mode against the ISDB should be seen as an exploratory tool allowing to gain a feeling of the chemistry behind a cluster of metabolites.