Rapid Identification of Known Natural Compounds in Mixtures

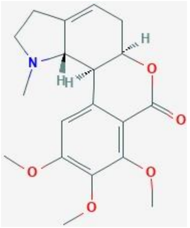
Mariacaterina LIANZA and Jean-Marc NUZILLARD

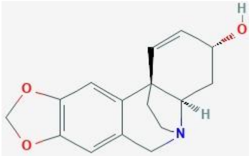
Institute for Molecular Chemistry in Reims (ICMR), UMR CNRS 7312
University of Reims Champagne Ardenne, France
University of Bologna, Italie

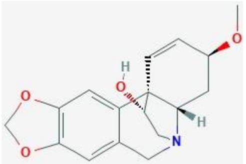
jm.nuzillard@univ-reims.fr

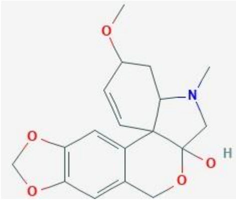
Initial aim of the study

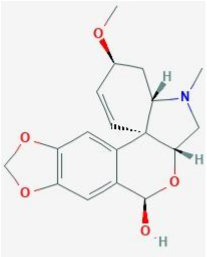
Urceolina peruviana (Amaryllidaceae) bulbs

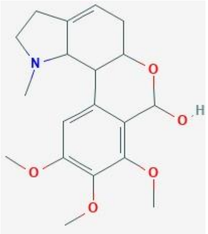






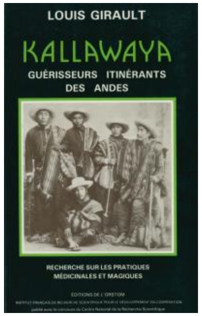






And more…

Louis Girault (ORSTOM/IRD), in his book "Kallawaya, guérisseurs itinérants des Andes: recherches sur les pratiques médicinales et magiques" indicates that bulbs of ***U. peruviana*** are mixed with pork or llama fat to prepare an ointment for healing **tumours and abscesses**.



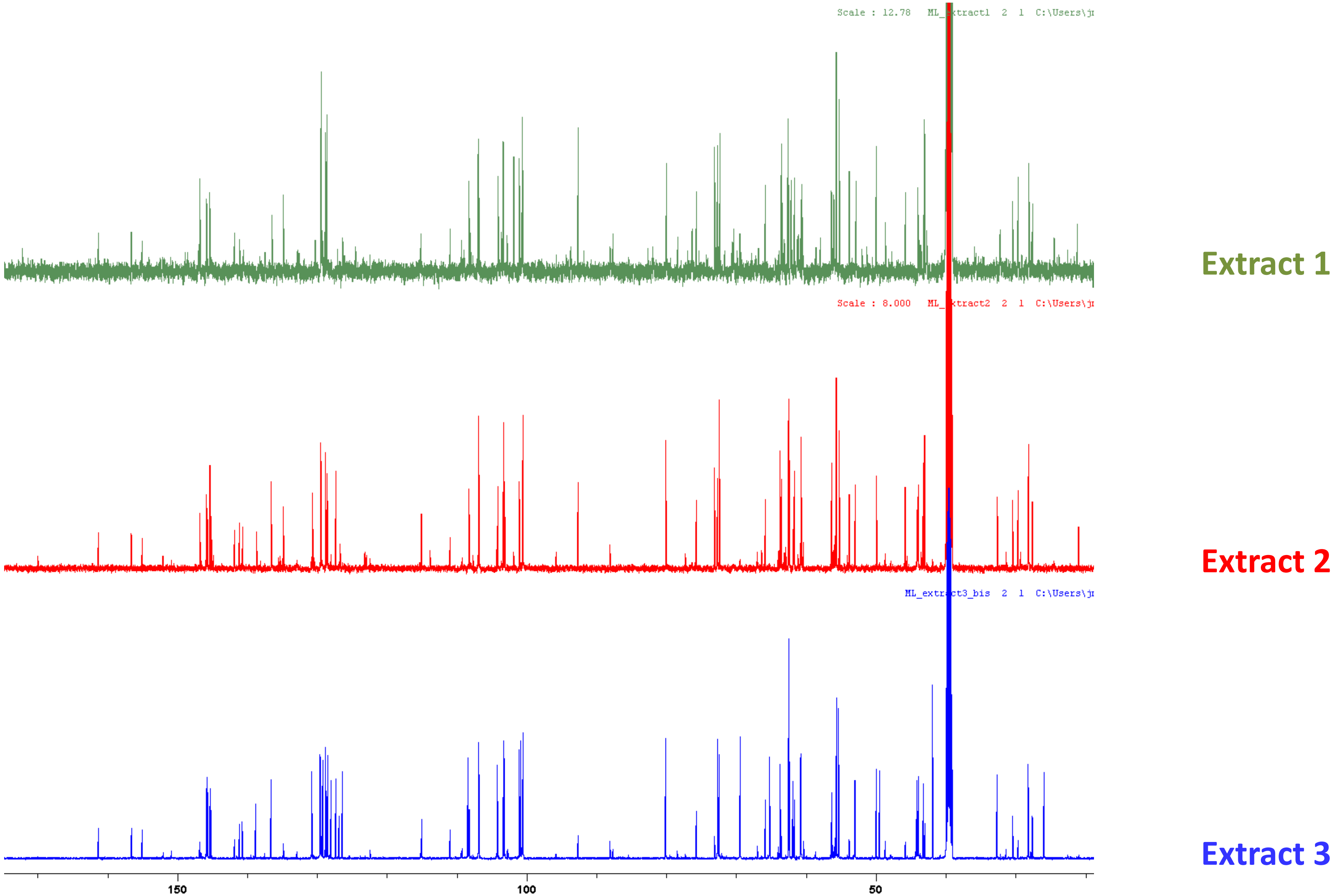
Extraction

- Starting material: bulbs, freeze-dried and crushed
- Two extraction protocols

  - Method I, weakly selective
    - *Natural product research* **2014**, *28*(10), 704-710
    - A single bulb (1,3 g) -> Extract « 1 » (61 mg)

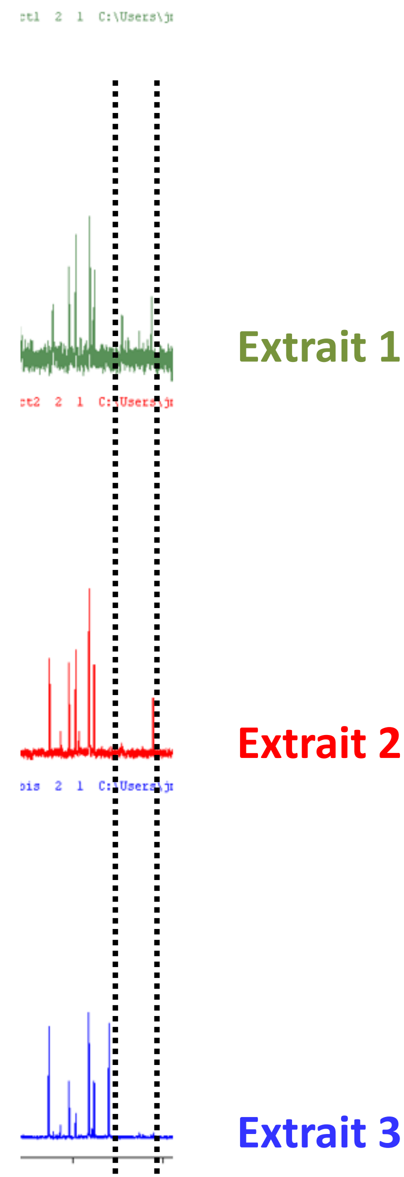
  - Method II, specific to **alkaloids**
    - Patent WO **2006**/064105 A1, preparation of galanthamine
    - A single bulb (1,3 g) -> Extract « 2 » (20 mg)
    - Carried out on 270 g -> Extract « 3 » (2,74 g)

Extracts, ^{13}C NMR in DMSO-d_6



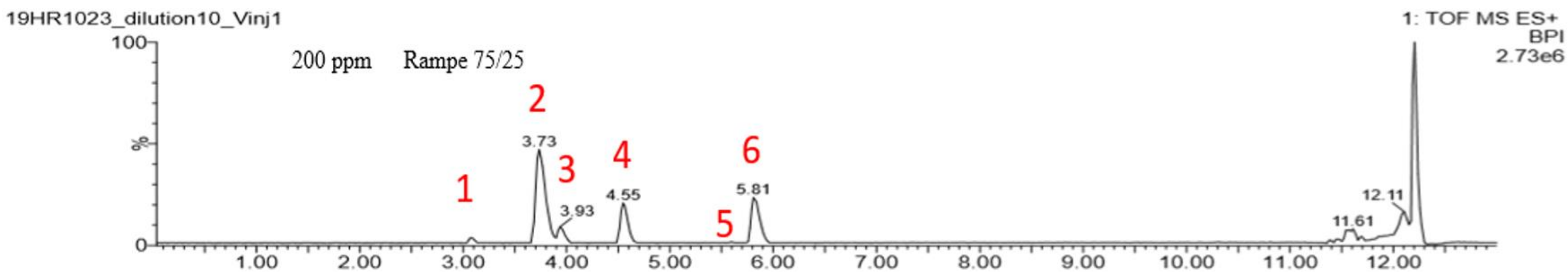
Extracts, ^{13}C NMR in DMSO-d_6

- Extracts 1 and 2 :
  - Produced from the same mass of bulbs but by different methods
- Extracts 2 et 3 :
  - Produced by the same method but starting from different masses
- => Importance of extraction on the nature of the isolated compounds.



Preliminary study of Extract 2

- Waiting for Extract 3
- UPLC, UPLC-HRMS



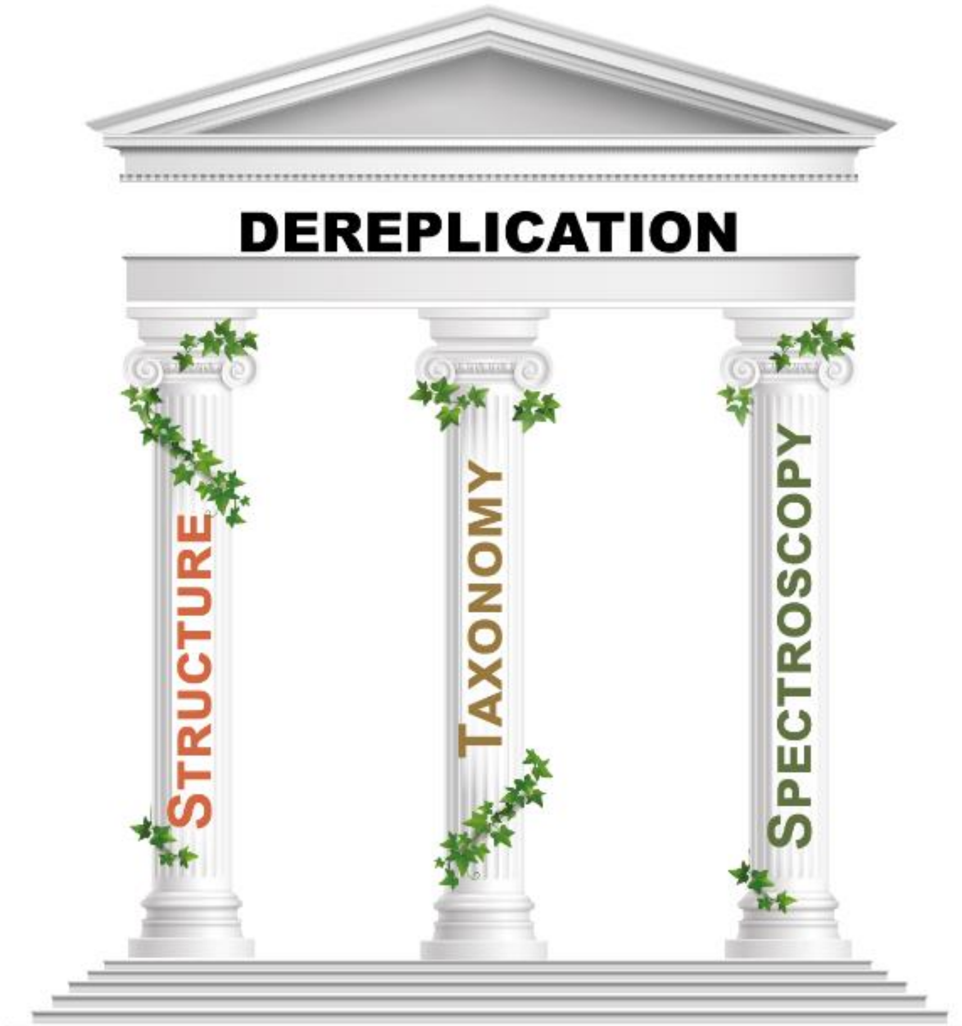
- Proposals for molecular formula from $[M+H]^+$

1: $C_{19}H_{25}NO_5$ 2: $C_{18}H_{21}NO_5$ 3: $C_{16}H_{17}NO_3$

4: $C_{17}H_{19}NO_4$ 5: $C_{18}H_{21}NO_4$ 6: $C_{19}H_{23}NO_5$

How should the known compounds be identified?

The three pillars of dereplication



Creation of a database containing the structures of the molecules that were already identified in the Amaryllidaceae

- PubChem (>103 million compounds)
- Natural Products Only
  - Dictionary of Natural Products
  - Specialized databases, Dictionary of Alkaloids
  - CH-NMR-NP (JEOL)
  - ZINC « Natural Products »
  - UNPD (as included in ISDB, In-silico DataBase, for MS) that was reworked to produce PNMRNP (>200,000 compounds)
  - COCONUT : Sorokina *et al. J. Cheminform.* **2021**, *13*, 2. doi:10.1186/s13321-020-00478-9…

Structure + Taxonomy

- KNApSAcK
  - Plant Cell Physiol. 2012 Feb;53(2):e1.
  - http://www.knapsackfamily.com/knapsack_core/top.php

Select by ...

(•) ALL Types ( ) Organism ( ) Metabolite ( ) Molecular formula
( ) C_ID ( ) CAS_ID ( ) INCHI-KEY ( ) INCHI-CODE ( ) SMILES

[ ] List Clear

last update	2020/01/06
metabolite	51179 entries
metabolite-species pair	116314 entries
species	22943 entries

KNApSAcK, searching for crinine

input type = metabolite , input word = crinine

Number of matched data :7

C_ID	CAS ID	Metabolite	Molecular formula	Mw
C00024357	80665-67-4	6alpha-Hydroxycrinine	C16H17NO4	287.11575804
C00024358	80665-68-5	6beta-Hydroxycrinine	C16H17NO4	287.11575804
C00024372	23367-61-5	(-)-Cherylline (S)-(-)-Cherylline Cherylline Cheryllin Crinine(C17 alkaloid) Crinine	C17H19NO3	285.13649348
C00024384	510-67-8	Crinine Crinidine	C16H17NO3	271.12084342
C00024416	93452-26-7	3-O-Acetylcrinine Krepowine O-Acetylcrinine Krepowine	C18H19NO4	313.1314081
C00025196	4684-32-6	Picrinine Deacetyldeformopicraline	C20H22N2O3	338.16304258
C00027665	82260-04-6	12-Demethoxytabernulosine 10-Methoxypicrinine	C21H24N2O4	368.17360727

All rights reserved. © 2007 NARA INSTITUTE of SCIENCE and TECHNOLOGY

Automated search in KNApSAcK

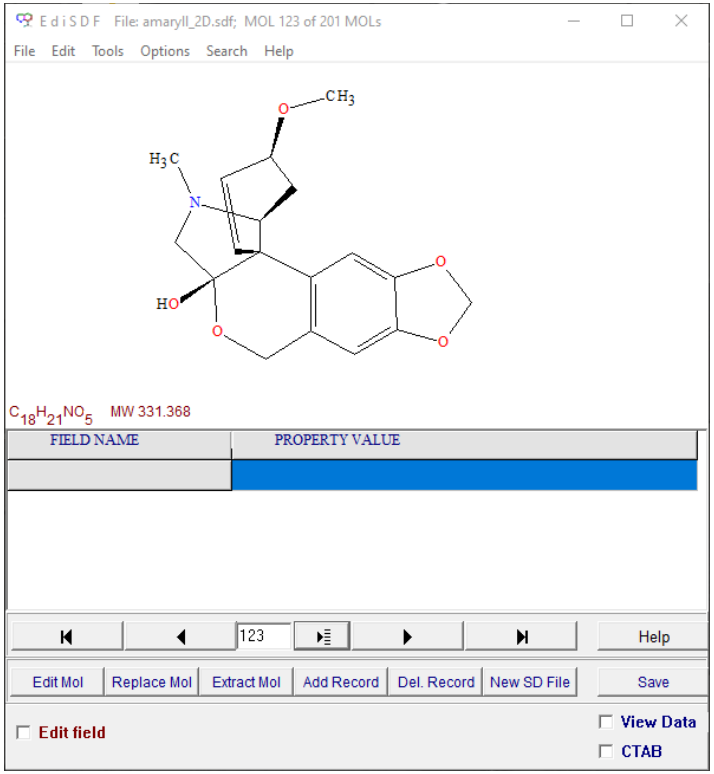
- Create a list of genera relative to a family
  - Example : All Amaryllidaceae genera (*Amaryllis, Narcissus*, …)
  - From Wikipedia or from the «NCBI taxonomy browser»
- Search for (compound, *Genus species*) pairs
  - Example : (C00001576, *Clivia miniata*)
- Create (compound, list of *Genus species*)
  - Example : (C00001567, *Zephyranthes carinata*||*Zephyranthes grandiflora*)
- Search ID for the ID cards of all compounds
  - Molecular Formula, SMILES, InChI, InChIKey, Molar mass, …
- KNApSAcK contains about 50.000 compounds
  - and is therefore not exhaustive (300,000 NPs?)

Drawing 2D structure graphs

- 2D coordinate generation from SMILES
- RDKit (rdkit.org/)
- Python (python.org/)

Open-Source Cheminformatics
and Machine Learning

- Viewer : EdiSDF (free)



Tazettine

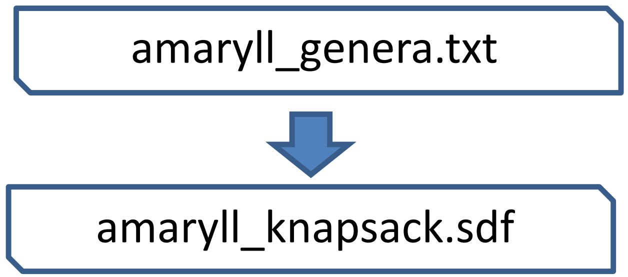
c12c(cc3c(c1)[C@]14[C@](OC3)(CN([C@H]1C[C@@H](C=C4)OC)C)O)OCO2

Spectroscopy: ^{13}C NMR

- Why ^{13}C NMR?
  - Carbon atoms are everywhere in organic molecules (by definition)
  - 1 carbon atom -> 1 peak and 1 peak -> 1 carbon atom (unless symmetry or accident occurs)
  - Narrow peaks (~1 Hz) compared to SW (~50 kHz), unlikely peak collisions
  - Sensitivity : NMR @600 (150 MHz for ^{13}C) and cryoprobe
- ^{13}C NMR Data of known compounds?
  - From published works.
    - Not easy to collect, incomplete, not always reliable
  - From prediction software
    - ACD/CNMR Predictor
    - CSEARCH/NMRPREDICT
    - ChemDraw
    - NMRShiftDB

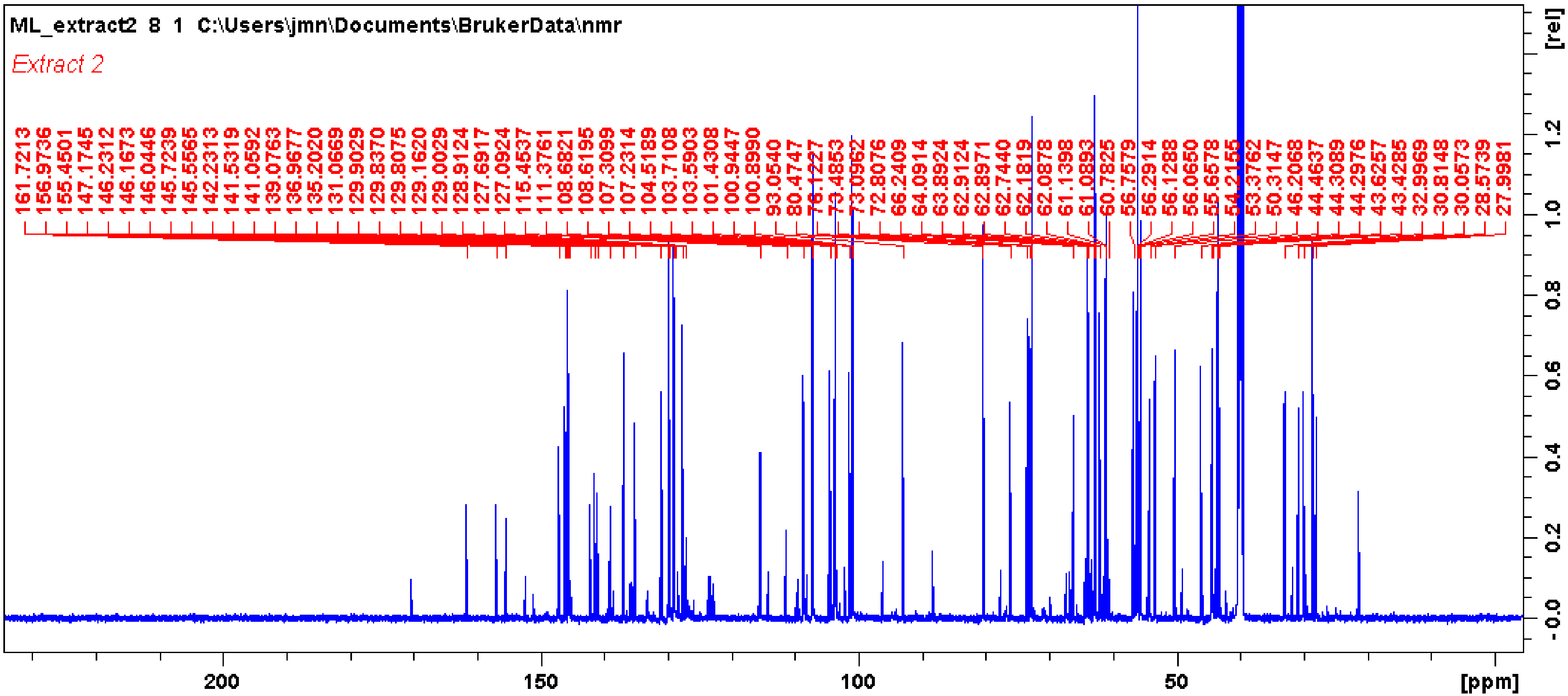
KnapsackSearch : Structure + Taxonomy + Spectroscopy

- https://github.com/nuzillard/KnapsackSearch/



- Can be used if the KNApSAcK website does not change the format of the HTML code it sends back to the web browser.
- File *familyname*_knapsack.sdf contains
  - A 2D structure (with configurations of asymmetric centers) of the compounds related to the initial list of genera.
  - The binomial names of the living beings that produced these compounds
  - The ^{13}C NMR data for each compound as predicted by **nmrshiftdb**

Extract 2: « Naive » dereplication



Extract 2: « Naive » dereplication

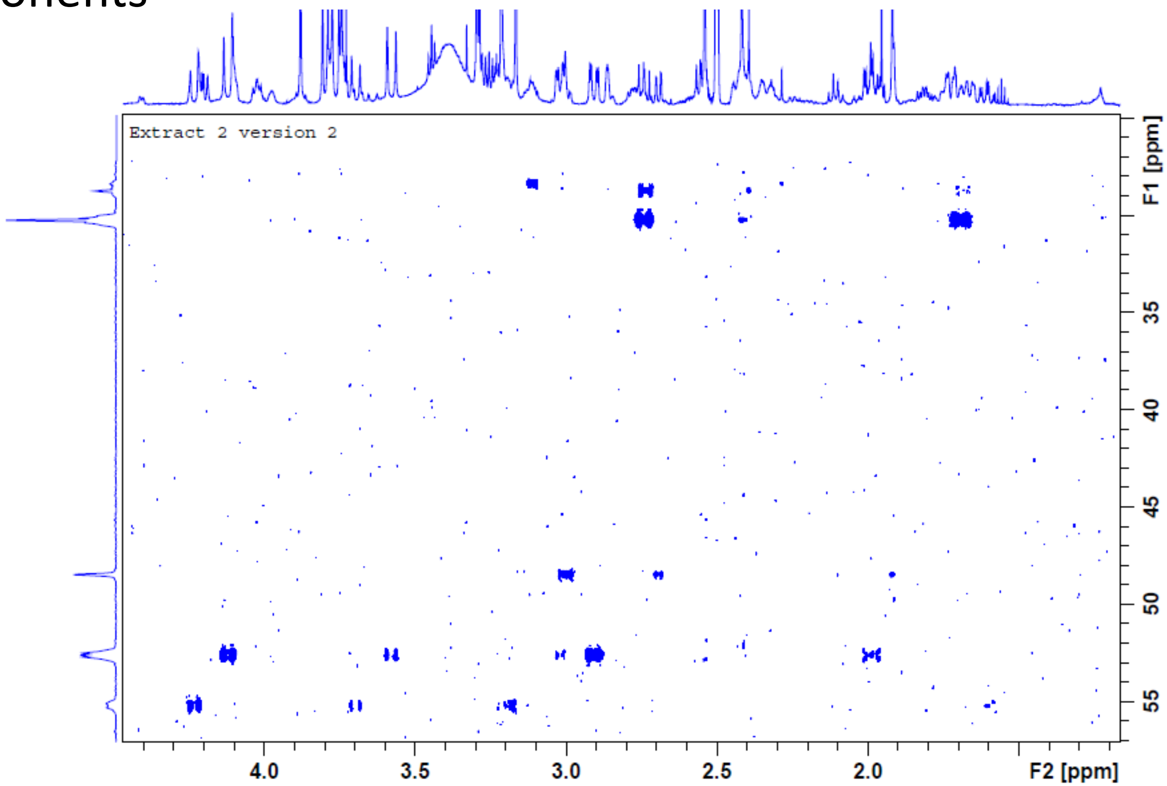
- Python script « matchPP_MFs.py »
  - Peak peaking: ^{13}C NMR chemical shift values (δ_C), no intensity
  - Create six files named *Formula*.sdf (such as C18H21NO4.sdf) from amaryll_knapsack.sdf, each corresponding to a molecular formula proposed by UPLC-HRMS
  - For each compound in each file *Formula*.sdf:
    - Determine N_{ok}, the number of predicted δ_C values that fit with the list of experimental δ_C values from the spectrum of the extact.
    - Calculate score N_{ok} / N_C, with N_C the number of carbon atoms in the current compound
  - Fitting of δ values is defined by the comparison of the absolute value of a difference (predicted *vs* experimental) with a threshold (1.5 ppm)
  - Sort the content of each *Formula*.sdf file by decreasing score.
  - Look at the first structures in each file…

Extract 2: « Naive » dereplication

Formula	Number of compounds	Compound	Score	Formula	Number of compounds	Compound	Score
$C_{16}H_{17}NO_3$	3	Crinine	1.000	$C_{18}H_{21}NO_5$	12	Pseudolycorine 1-acetate	1.000
		Vittatine	1.000			Pseudolycorine 2-acetate	1.000
$C_{17}H_{19}NO_4$	11	Crinamine	1.000			Steinbergine	1.000
		Haemanthamine	1.000			Tazettine	1.000
		Hippamine	1.000			Criwelline	1.000
		Montanine	1.000	$C_{19}N_{23}NO_5$	2	Albomaculine	0.947
$C_{18}H_{21}NO_4$	5	Norpluviine 1-acetate	1.000	$C_{19}H_{25}NO_5$	2	Ungvedine	0.895
		Oduline O-Me	1.000				

Extract 2, Urceolina peruviana, ^{1}H-^{15}N HMBC

- The structure of Amaryllidaceae alkaloids contains only a single nitrogen atom
  - The ^{1}H-^{15}N HMBC spectrum of an extract reveals the major mixture components



Dereplication by « CARAMEL »

- CARAMEL : CARActérisation de MELanges
- *Anal. Chem.* **2014**, *86*, 2955-2962. doi: 10.1021/ac403223f
- Method :
  - Fractionation by CPC (Centrifugal Partition Chromatography)
  - ^{13}C NMR spectra
  - « Peak picking » and « bucketing »
  - Search for δ_C clusters that share the same chromatographic profile
  - Associate δ_C clusters to chemical structures
- The success of the CARAMEL procedure led to the creation of the Nat'Explore company (https://nat-explore.com/)
- The CARAMEL database contains about 4000 compounds and was created with ACD/Labs « C+H NMR Predictor and DB »

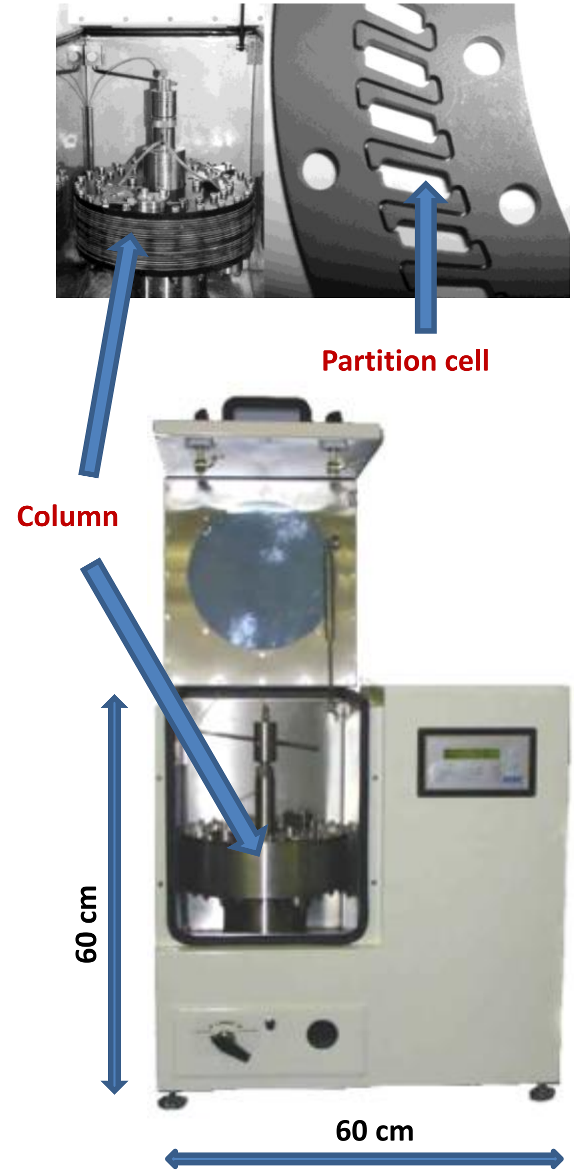
ACD/Labs « C+H NMR Predictor and DB »

- We started to use ACD/Labs « C+H NMR Predictor and DB » for at least four reasons:
  - Easy handling of molecule collections
  - Prediction of ^{1}H and ^{13}C NMR chemical shift values with good reputation
  - Compound selection according to various criteria, including chemical shift value comparison and correlations between chemical shifts (2D NMR).
  - No need for computer code writing or command typing, ready to use for non-coding users

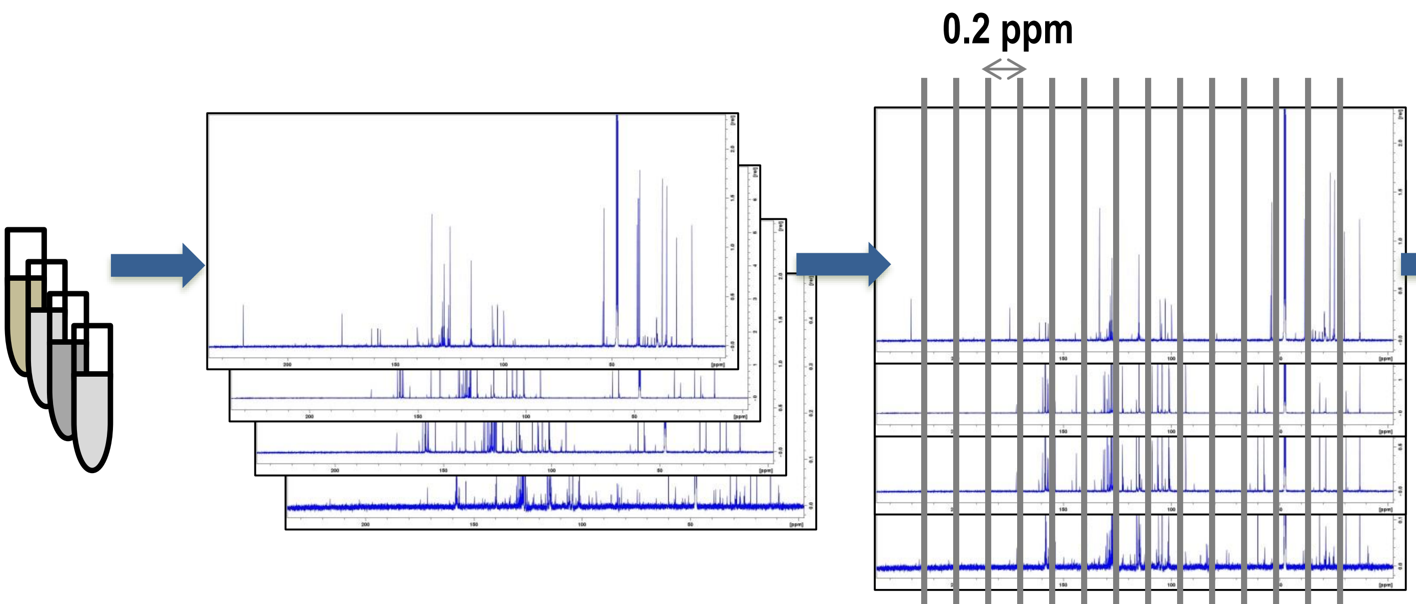
CPC

- **Centrifugal Partition Chromatography**
- Partition of analytes between two liquid phases
- The « column » contains hundred of partition cells
- The stationary phase is maintained by centrifugal force
- The analytes are injected in column head
- The mobile phase percolates through the stationary one
- No irreversible absorption on a solid phase
- All what is injected is recovered
- Modes: elution (isocratic or graduated) and displacement
- High flow rate, typically 20 mL/min
- Possible injection of 5 g in a 200 mL column

- **Preparative technique**



« CARAMEL » Dereplication

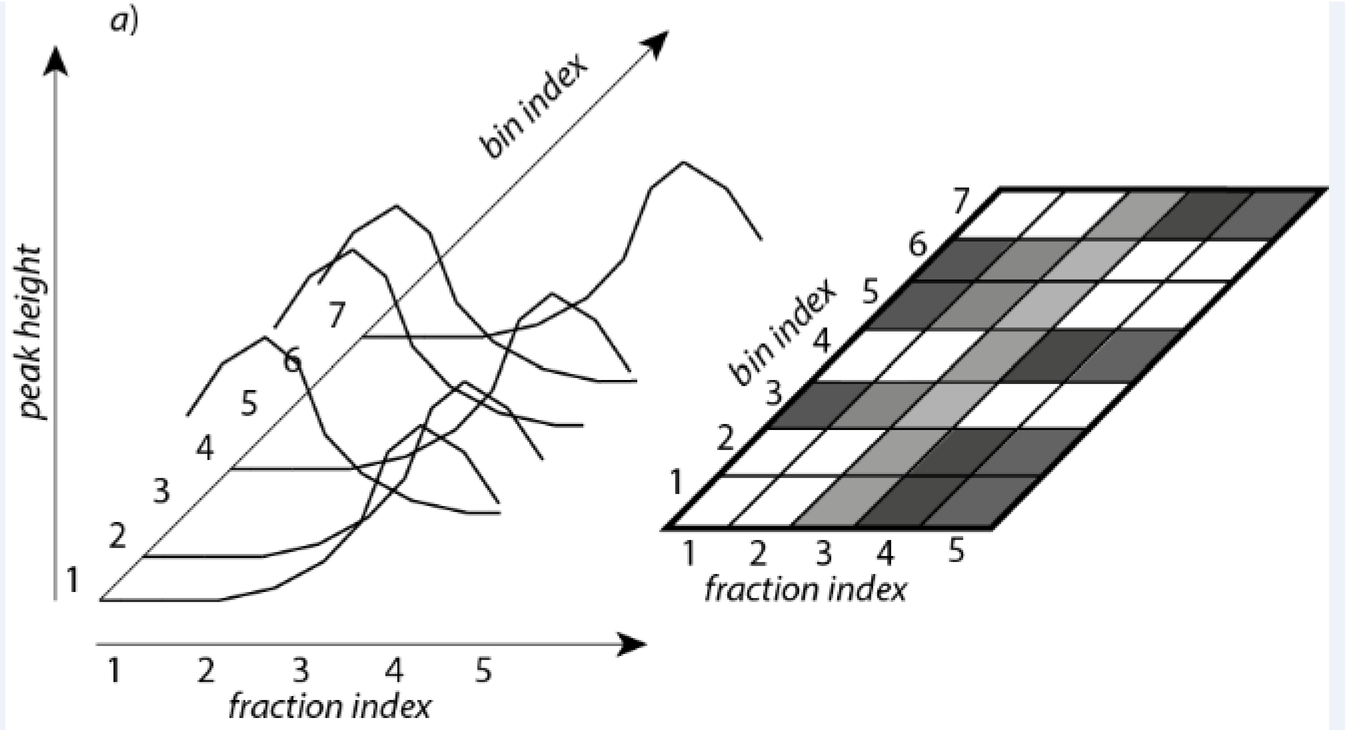


^{13}C NMR spectra

ppm	f1	f2	...	fx
16.3	2E+08	1E+08	...	0
17.5	0	0	...	0
18.7	1E+08	1E+08	...	0
⋮	⋮	Intensity of NMR peaks		
176.1	3E+07	0	...	0
177.7	0	0	...	4E+07
177.9	6E+07	6E+07	...	0
199.5	7E+07	5E+07	...	0

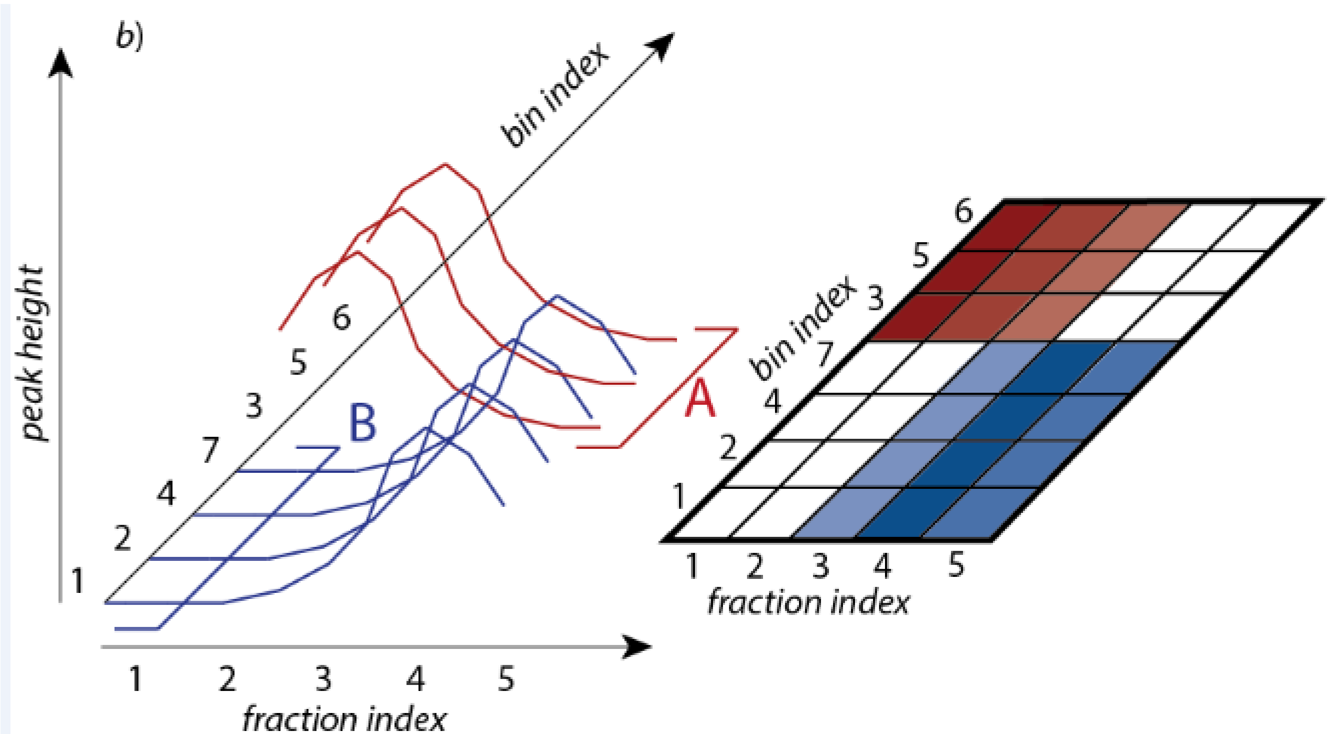
- Peak picking
- File format change
- Bucketing

Clustering of chromatographic profiles



BEFORE

Clustering of chromatographic profiles

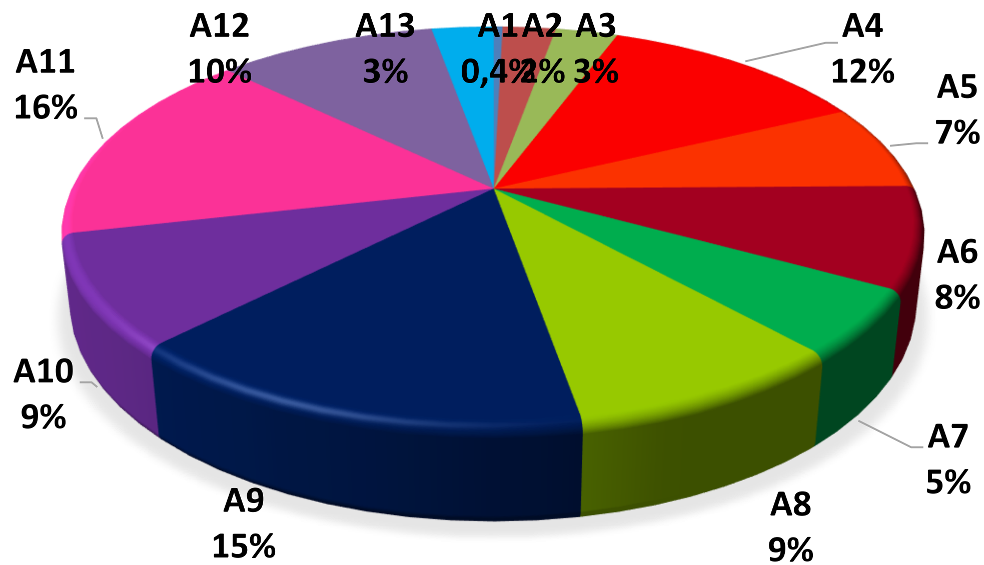


AFTER

« CARAMEL », Extract 3 of *U. peruviana*

- 13 fractions by CPC, « pH-zone refining » (displacement) mode
  - Biphasic solvent system MtBE, CH_3CN et H_2O, 5:2:3 (v/v)
  - Injection, **1 g**, in the aqueous stationary phase (acidified by 10 mM H_2SO_4)
  - Displacement of the alkaloid by the organic mobile phase (basified by 8 mM NEt_3) according to the analyte pK_a value and on partition coefficients
  - The collected mobile phase fractions are analysed by TLC and grouped by similarity
  - Fractions A1 à A13.

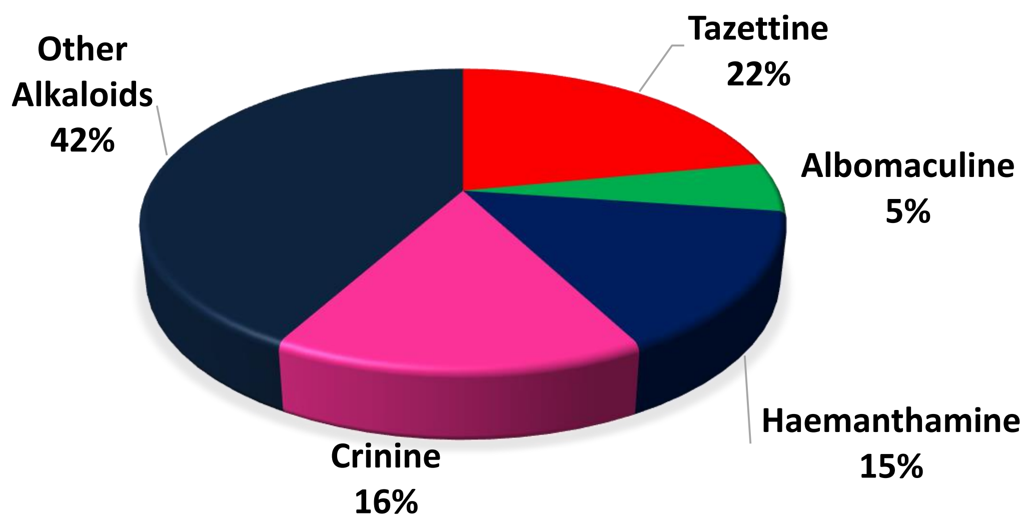
Mass repartition



« CARAMEL », Extract 3 of U. peruviana

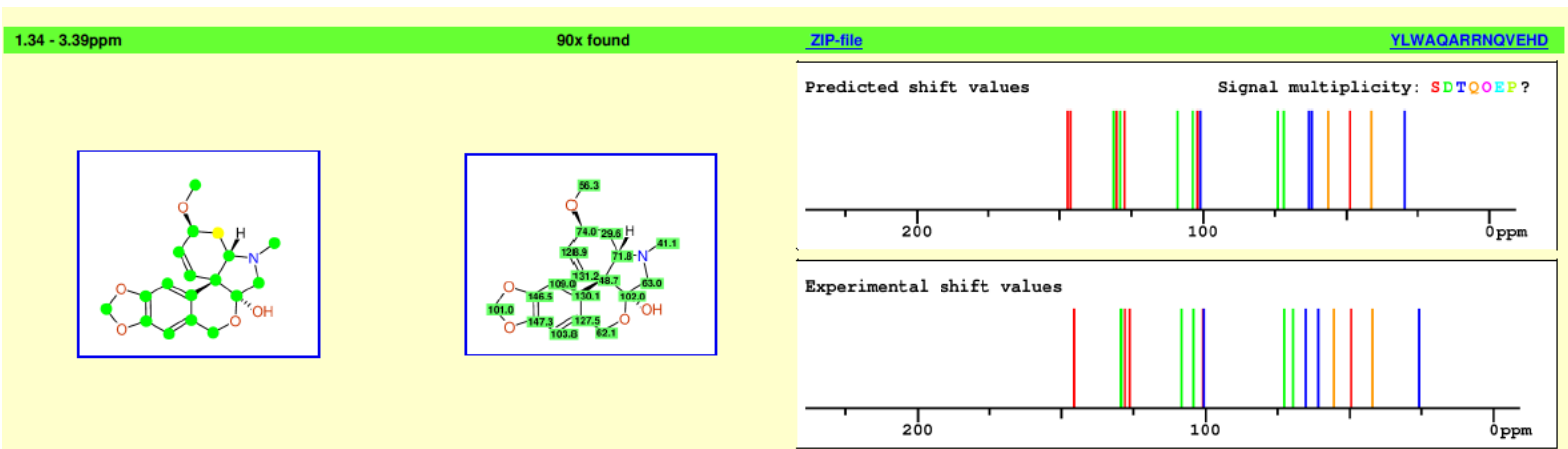
- NMR Analysis
  - ^{1}H, ^{13}C, ^{1}H-^{1}H COSY, ^{1}H-^{13}C HSQC, ^{1}H-^{13} HMBC, ^{1}H-^{1}H ROESY
  - Fractions A3 to A5 are « almost » identical and pure
  - Fractions A7 and A9 « almost » pure
  - Fraction A11 contains a highly major compound
  - **Compounds in fractions A4, A7, A9, A11 can be « readily » identified**
  - Fractions A2, A6, A8, A10, A11, and A12 (transitions) are highly complex

Mass repartition



Fraction A4

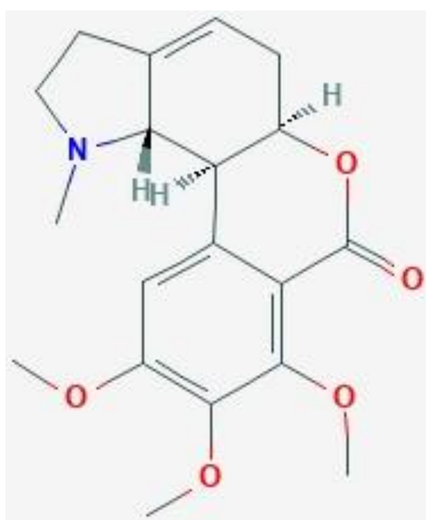
- 18 δ_C values picked in the ^{13}C NMR spectrum of fraction A4
- The HSQC spectrum helps to associate a multiplicity (Q, CH, CH_2, CH_3) to each δ_C value
- CSEARCH web interface

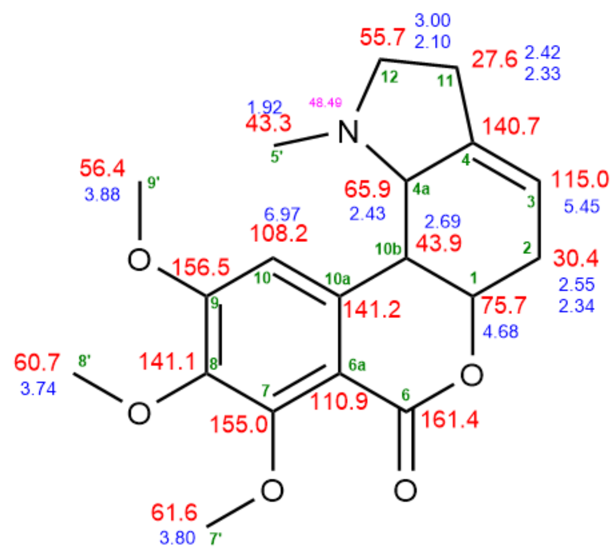


- The structure of tazettine is ranked first

Fraction A7

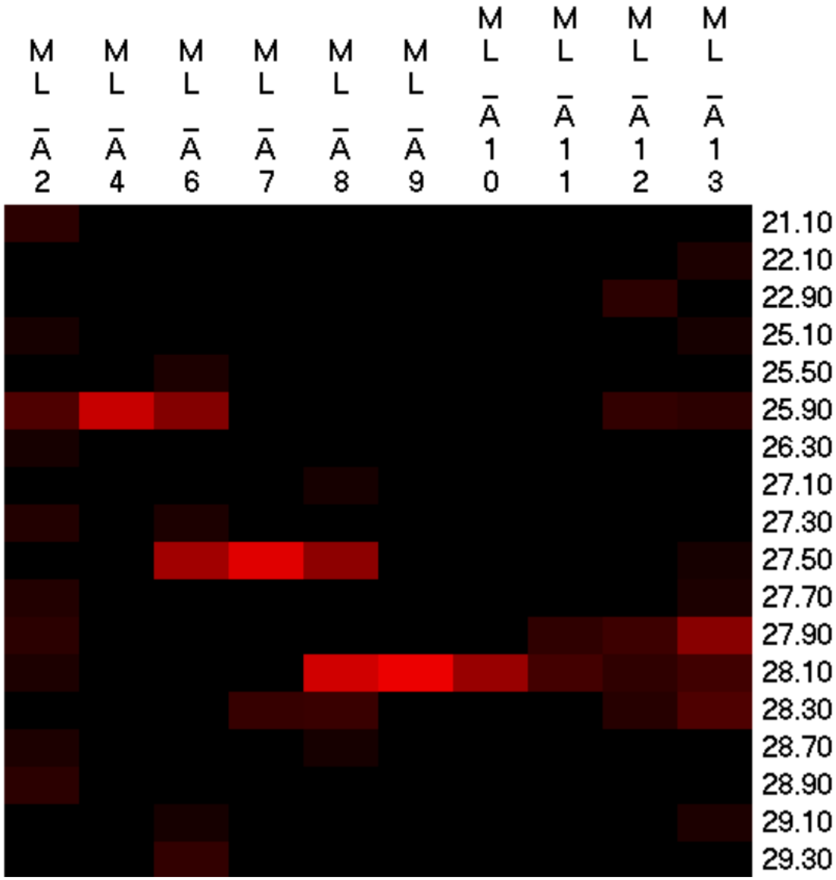
- 19 (δ_C, multiplicity) pairs in ^{13}C and 2D HSQC NMR spectra
- CSEARCH not helpful in this case.
- KNApSAcK contains 2 $C_{19}H_{23}NO_5$ compounds and 2 $C_{19}H_{25}NO_5$ compounds
- The structure contains 3 CH_3-O-Aryl groups (from δ_C, δ_H, and 2D HMBC)
- Only one possibility in KNApSAcK: albomaculine
- Validation by other NMR spectra (1D and 2D)

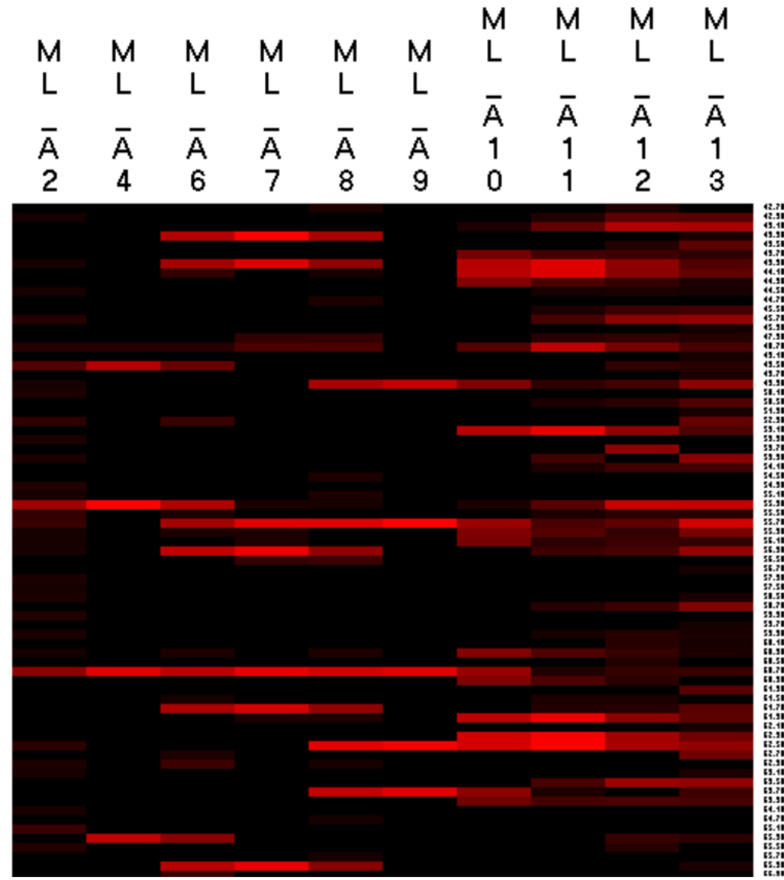




« CARAMEL », Extract 3 of *U. peruviana*

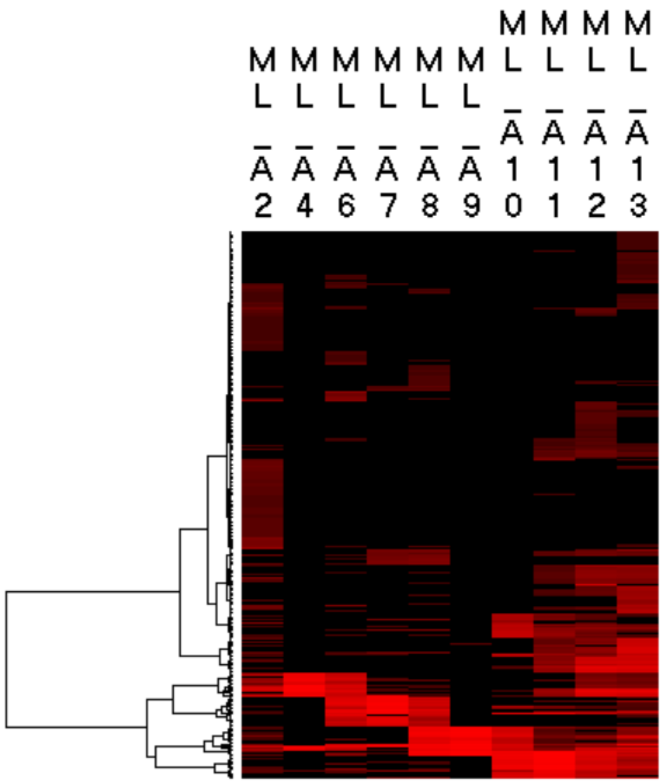
- Data display by « PermutMatrix », *Bioinformatics* **2005**, *21*, 1280-1281.

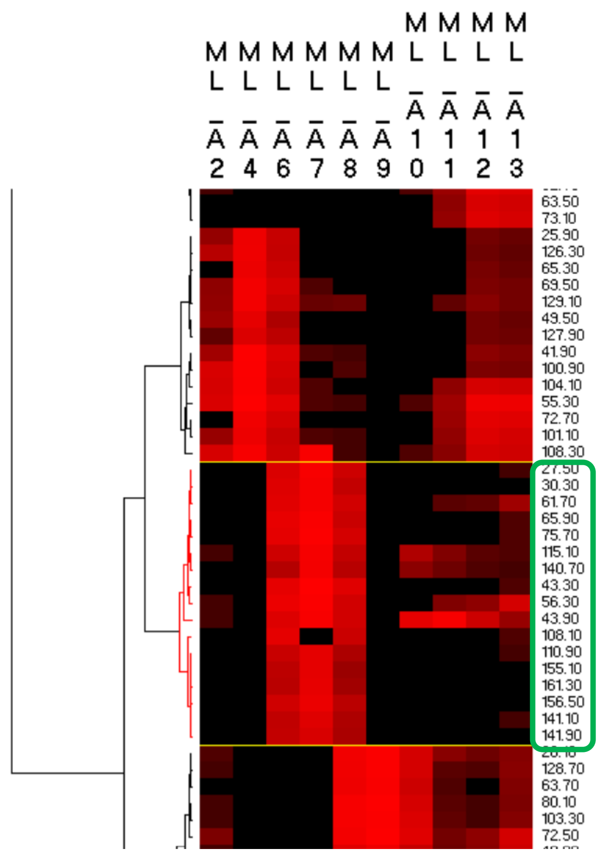




« CARAMEL », Extract 3 of *U. peruviana*

- Classification by « PermutMatrix », *Bioinformatics* **2005**, *21*, 1280-1281.



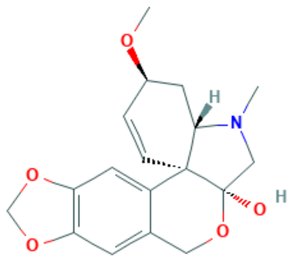


Set of δ values associated to identical chromatographic profiles

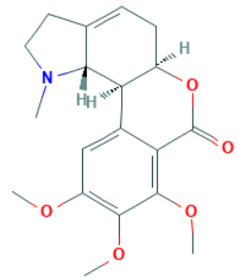
=>

Database search

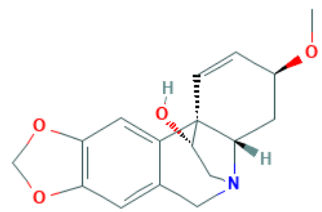
« Work in progress »



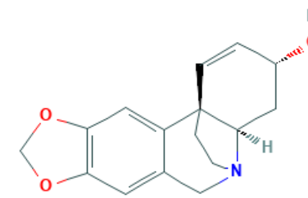
Tazettine



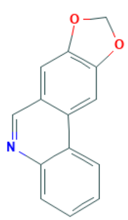
Albomaculine



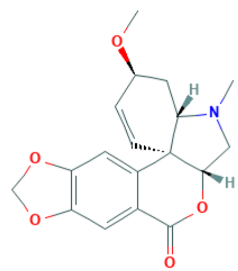
haemanthamine



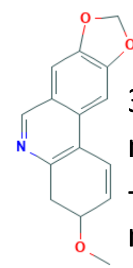
crinine



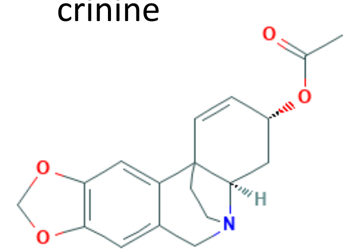
Trisphaeridine



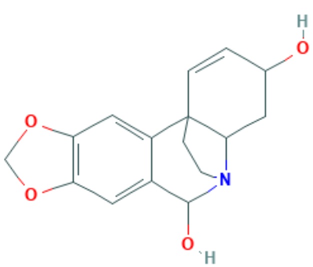
3-Epimacronine



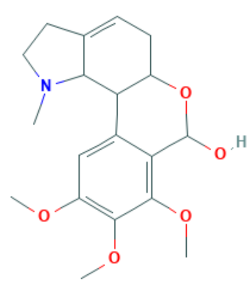
3-Methoxy-8,9-methylenedioxy-3,4-dihydrophe-nanthridine



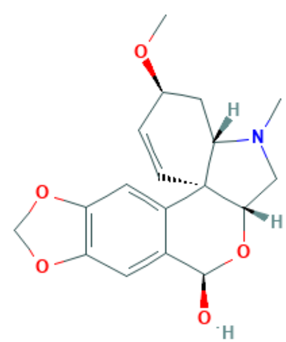
Crinine acetate



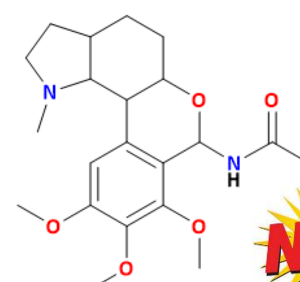
6α-hydroxy-crinine



nerinine



pretazettine



6-dehydroxy-6-acetamido-nerinine

Creation of taxonomy-focused databases

- The CARAMEL procedure involved since its early beginning the identification of compounds by searching a database build with ACD/Labs software.
- The CARAMEL database was incrementally enriched each time a new plant extract was studied and is therefore « naturally » taxonomy-focused
- Each compound is associated with predicted ^{1}H and ^{13}C chemical shifts by means of a tedious, compound-by-compound procedure (about 1 min per compound), but that is quicker and more reliable than DB manual input of experimental values

- **There was a need for an efficient way of creating taxonomy-focused databases searchable by means of ACD/Labs software**

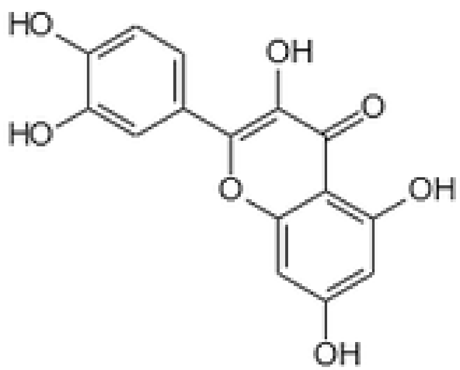
Creation of taxonomy-focused databases

- KNApSAcK used to be the easiest way to associate compound structures and taxonomic data
- KnapsackSearch associates structures, taxonomy and ^{13}C NMR chemical shifts predicted by nmrshiftdb
- The LOTUS database (lotus.naturalproducts.net)
  - Offers an easier access to a greater number of compounds
  - Contains taxonomic and bibliographic data
  - Relies on the framework created for the COCONUT DB coconut.naturalproducts.net

COCONUT
COllection of Open Natural ProdUcTs

LOTUS

A DB with one compound inside: Quercetin



Q409478

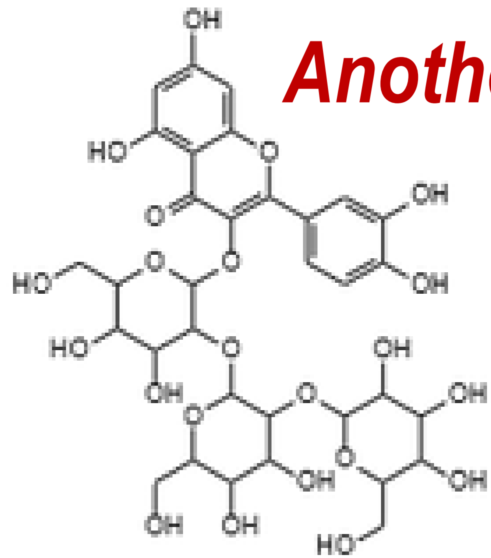
Quercetin

Mol. formula	C15H10O7
Mol. weight	302.24
Tmp. LOTUS id	LTS0004651

- Found in LOTUS, simple search: InChI=1S/C15H10O7/c16-7-4-10(19)12-11(5-7)22-15(14(21)13(12)20)6-1-2-8(17)9(18)3-6/h1-5,16-19,21H

- Looking for « quercetin » alone results in 32 compounds

Result downloaded as SDF file *lotus_simple_search_result.sdf*



Another quercetin in LOTUS…

Q27114778

Quercetin

Mol. formula	$C_{33}H_{40}O_{22}$
Mol. weight	788.66
Tmp. LOTUS id	LTS0205097

lotus_simple_search_result.sdf

```

Actelion Java MolfileCreator 2.0

  0  0  0  0  0  0              0 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 22 24 0 0 0
M  V30 BEGIN ATOM
M  V30 1 O 1.299 0 0 0
M  V30 2 C 0 -0.75 0 0
M  V30 3 C -1.299 0 0 0
M  V30 4 O -1.299 1.5 0 0
M  V30 5 C -2.598 -0.7499 0 0
M  V30 6 C -3.8971 0 0 0
M  V30 7 C -3.8971 1.5 0 0
M  V30 8 C -5.1961 2.25 0 0
M  V30 9 C -6.4951 1.5 0 0
M  V30 10 O -7.7942 2.25 0 0
M  V30 11 C -6.4951 0 0 0
M  V30 12 O -7.7942 -0.7499 0 0
M  V30 13 C -5.1961 -0.7499 0 0
M  V30 14 O -2.598 -2.2499 0 0
M  V30 15 C -1.299 -2.9999 0 0
M  V30 16 C -1.299 -4.4999 0 0
M  V30 17 C 0 -5.25 0 0
M  V30 18 O 0 -6.75 0 0
M  V30 19 C 1.299 -4.5 0 0
M  V30 20 C 1.299 -3 0 0
M  V30 21 O 2.598 -2.25 0 0
M  V30 22 C 0 -2.2499 0 0
M  V30 END ATOM
M  V30 BEGIN BOND
M  V30 1 2 1 2
M  V30 2 1 2 3
```

```
> <DOI>
10.1002/PTR.2650040508

> <DOI>
10.1016/S0031-9422(00)80124-7

> <DOI>
10.1007/S10600-010-9598-1

> <Unordered_taxonomy>
,Persicaria salicifolia,Rhododendron praetervisun
ferruginea,Medicago murex,Paracalyx,Primulaceae,C
chrysantha,Myrsine africana,Dregea,Podophyllum ve
quadrifida,Conyza,Calophyllaceae,Cyperus brevifol
pulchella,Polygonum lapathifolium,Astragalus floc
pseudocicera,Arnica amplexicaulis,Geum,Achillea n
melanantherum,Crataegus,Ericaceae,Scutellaria bai
italica,Juglans,Dolichandra,Rosaceae,Epimedium dc
laevigatus,Populus deltoides,Psittacanthus cuneif
ovatum,Rhododendron nervulosum,Solanum lycopersic
kaki,Tragopogon pratensis,Agaricus,Vismia baccife
aureum,Carthamus,Polypodiaceae,Rothmaleria,Campar
capensis,Eryngium,Arnica nevadensis,Annona cherin
jatamansi,Picradeniopsis pringlei,Medicago monspe
kotoense,Diospyros,Nephrophyllidium,Vincetoxicum
sericea,Senecio subdentatus,Patersonia occidental
sinensis,Filipendula,Fagopyrum cymosum,Purshia,Pł
amphibia,Brassica campestris,Pteridium aquilinum,
corniculatus subsp. corniculatus,Robinsonia macrc
lobophyllum,Warburgia ugandensis,Beta vulgaris,Rc
speciosa,Lathyrus vernus,Cassinia,Polygonum hydrc
```

Cleaning quercetin.sdf

- Cleaning by a series of three « in place » transformations

```
(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m uniqInChI quercetin.sdf
Read: 1 -- Written: 1 -- Discarded: 0

(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m tautomer quercetin.sdf

(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m rdcharge quercetin.sdf

(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>
```

- If two compounds have the same InChI, only the first one is kept in the output file (*uniqInChI.py*, uses RDKit)
- Replaces aliphatic iminols by amides (*tautomer.py*, uses RDKit) to compensate InChI decoding oddities
- Correct data produced by RDKit for electrically charged atoms (*rdcharge.py* uses *sdfrw.py*, in github.com/nuzillard/KnapsackSearch)

^{13}C NMR chemical shifts in quercetin.sdf

```
(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m addnmrsdb quercetin.sdf
predictSdf quercetin.sdf 4 3d 1>C:\Users\jmn\AppData\Local\Temp\tmpt1hadug4.txt 2>errorlog.txt
Running: predictSdf quercetin.sdf 4 3d 1>C:\Users\jmn\AppData\Local\Temp\tmpt1hadug4.txt 2>errorlog.txt
"predictSdf quercetin.sdf 4 3d 1>C:\Users\jmn\AppData\Local\Temp\tmpt1hadug4.txt 2>errorlog.txt" returned with code: 0


(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m fakeACD nmrsdb_quercetin.sdf

(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>_
```

- *addnmrsdb.py* calculates the ^{13}C NMR chemical shifts using nmrshiftdb
- *fakeACD.py* puts them to the ACD/Labs format, as if they were experimental values.

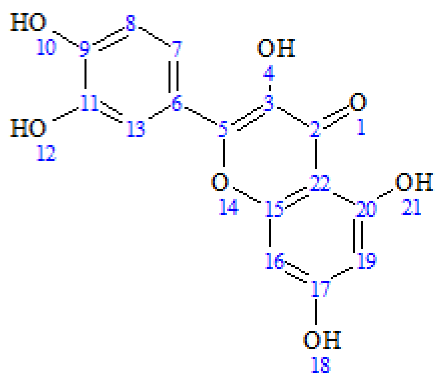
```
>  <CNMR_SHIFTS>
0:2|175.26;1:3|136.20;2:5|144.80;3:6|125.20;4:7|122.72;5:8|115.74;6:9|148.78;7:11|145.46;8:13|116.14;9:15|157.05;10:16|94.02;11:17|1
64.18;12:19|99.18;13:20|161.85;14:22|105.34

>  <NMRSHIFTDB2_ASSIGNMENT>
2, 175.26 \
3, 136.20 \
5, 144.80 \
6, 125.20 \
7, 122.72 \
8, 115.74 \
9, 148.78 \
11, 145.46 \
13, 116.14 \
15, 157.05 \
16, 94.02 \
17, 164.18 \
19, 99.18 \
20, 161.85 \
22, 105.34 \
```

- *addnmrsdb.py* creates *nmrsdb_quercetin.sdf* with a <NMRSHIFDB2_ASSIGNMENT> tag
- *fakeACD.py* creates *fake_acd_nmrsdb_quercetin.sdf* with a <CNMR_SHIFTS> tag

ACD/Labs DB, fake_acd_quercetin.NMRUDB

ACD/C+H NMR Predictors and DB: Database Window - [C:\USERS\JMN\DOCUMENTS\CNRS...ONS\BOLZANO\CNMR_PREDICT\FAKE_ACD_QUERCETIN.NMRUDB]

Database View Record Search Lists Table Training Options ACD/Labs Help



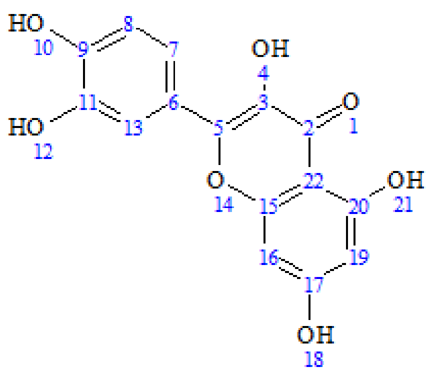
Atom No.	13C Shift	1H Shift
2	175.26	
3	136.2	
5	144.8	
6	125.2	
7	122.72	
8	115.74	
9	148.78	
11	145.46	
13	116.14	
15	157.05	
16	94.02	
17	164.18	
19	99.18	
20	161.85	
22	105.34	

δ_C values calculated by nmrshiftdb

- After importation of *fake_acd_nmrsdb_quercetin.sdf*
- The process of DB production may be stopped at this point because the DB is ready for compound search.

Check NMR in fake_acd_quercetin.NMRUDB

ACD/C+H NMR Predictors and DB: Database Window - [C:\USERS\JMN\DOCUMENTS\CNRS...ONS\BOLZANO\CNMR_PREDICT\FAKE_ACD_QUERCETIN.NMRUDB]



Atom No.	13C Shift	13C Calc	1H Shift
2	175.26	176.74	
3	136.2	135.98	
5	144.8	147.42	
6	125.2	122.54	
7	122.72	120.50	
8	115.74	116.05	
9	148.78	147.63	
11	145.46	145.16	
13	116.14	115.55	
15	157.05	156.52	
16	94.02	93.93	
17	164.18	164.34	
19	99.18	98.65	
20	161.85	160.88	
22	105.34	103.49	

- In green, the δ_C values calculated by the ACD/Labs software for assignment checking.
- **Only a single click is required to check a database, in less than one second per structure on the average**

Export fake_acd_quercetin.NMRUDB

- Export DB *fake_acd_quercetin.NMRUDB* as *fake_acd_quercetin_exported.sdf*

```
>  <NMRSHIFTDB2_ASSIGNMENT>
2, 175.26 \; 3, 136.20 \; 5, 144.80 \; 6, 125.20 \; 7, 122.72 \; 8, 115.74 \; 9, 148.78 \; 11, 145.46 \; 13, 116.14 \; 15, 157.05
\; 16, 94.02 \; 17, 164.18 \; 19, 99.18 \; 20, 161.85 \; 22, 105.34 \

>  <CNMR_SHIFTS>
0:2|175.26;1:3|136.20;2:5|144.80;3:6|125.20;4:7|122.72;5:8|115.74;6:9|148.78;7:11|145.46;8:13|116.14;9:15|157.05;10:16|94.02;11:17|1
64.18;12:19|99.18;13:20|161.85;14:22|105.34

>  <HNMR_SHIFTS>

>  <CNMR_CALC_SHIFTS>
0:Exact = 176.74,ExactErr = 2.04,NN = 176.76,Increm = 176.34;1:Exact = 135.98,ExactErr = 0.72,NN = 136.44,Increm = 136.25;2:Exact
= 147.42,ExactErr = 0.88,NN = 147.24,Increm = 146.94;3:Exact = 122.54,
ExactErr = 1.16,NN = 122.03,Increm = 120.95;4:Exact = 120.5,ExactErr = 1.5,NN = 120.41,Increm = 120.6;5:Exact = 116.05,ExactErr =
0.85,NN = 115.84,Increm = 116.29;6:Exact = 147.63,ExactErr = 0.88,NN =
 148.5,Increm = 146.73;7:Exact = 145.16,ExactErr = 0.54,NN = 146.55,Increm = 145.36;8:Exact = 115.55,ExactErr = 0.95,NN =
 114.65,Increm = 115.63;9:Exact = 156.52,ExactErr = 0.98,NN = 157.72,Increm = 1
57.36;10:Exact = 93.93,ExactErr = 1.07,NN = 94.27,Increm = 95.49;11:Exact = 164.34,ExactErr = 1.66,NN = 165.41,Increm =
164.86;12:Exact = 98.65,ExactErr = 0.95,NN = 99.02,Increm = 99.72;13:Exact = 160
.88,ExactErr = 0.52,NN = 162.79,Increm = 162.59;14:Exact = 103.49,ExactErr = 0.91,NN = 103.64,Increm = 102.7
```

NEW!

- Tag <CNMR_CALC_SHIFTS> reports the calculated ^{13}C NMR chemical shift values that were used for checking.

<CNMR_CALC_SHIFTS> → <CNMR_SHIFTS>

- Make as if the δ_C values calculated by the ACD/Labs software for checking were experimental ones

```
(rdkit3) C:\Users\jmn\Documents\CNRS21\Communications\Bolzano\CNMR_Predict>python -m CNMR_predict fake_acd_quercetin_exp
orted.sdf true_acd_quercetin.sdf
```

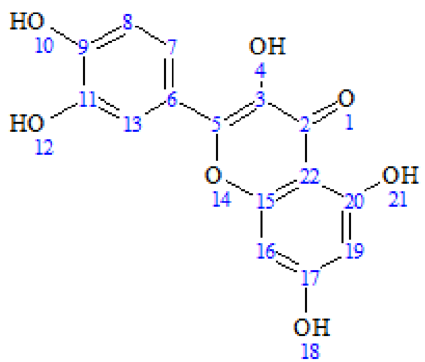
- *CNMR_Predict.py* transforms *fake_acd_quercetin_exported.sdf* into *true_acd_quercetin.sdf* by replacing the δ_C values under tag <CNMR_SHIFTS> (previously predicted by nmrshiftdb) by those under tag <CNMR_CALC_SHIFTS> (predicted by ACD/Labs for checking) for all compounds present in the DB.

Finally, creation of lotus_quercetin.NMRUB

- Create lotus_quercetin.NMRUDB and import true_acd_quercetin.sdf
- Check ^{13}C NMR chemical shifts again (no surprise)

ACD/C+H NMR Predictors and DB: Database Window - [C:\USERS\JMN\DOCUMENTS\CNRS2...TIONS\BOLZANO\CNMR_PREDICT\LOTUS_QUERCETIN.NMRUDB]

Database View Record Search Lists Table Training Options ACD/Labs Help



Atom No.	13C Shift	13C Calc	1H Shift
2	176.74	176.74	
3	135.98	135.98	
5	147.42	147.42	
6	122.54	122.54	
7	120.5	120.50	
8	116.05	116.05	
9	147.63	147.63	
11	145.16	145.16	
13	115.55	115.55	
15	156.52	156.52	
16	93.93	93.93	
17	164.34	164.34	
19	98.65	98.65	
20	160.88	160.88	
22	103.49	103.49	

To sum up

- *fake_acd_quercetin.NMRUDB* contains δ_C values from nmrshiftdb
- *lotus_quercetin.NMRUBD* contains δ_C values calculated by ACD/Labs for DB checking

Atom No.	13C Shift	13C Calc
2	175.26	176.74
3	136.2	135.98
5	144.8	147.42
6	125.2	122.54
7	122.72	120.50
8	115.74	116.05
9	148.78	147.63
11	145.46	145.16
13	116.14	115.55
15	157.05	156.52
16	94.02	93.93
17	164.18	164.34
19	99.18	98.65
20	161.85	160.88
22	105.34	103.49

fake_acd_quercetin.NMRUDB

Atom No.	13C Shift	13C Calc
2	176.74	176.74
3	135.98	135.98
5	147.42	147.42
6	122.54	122.54
7	120.5	120.50
8	116.05	116.05
9	147.63	147.63
11	145.16	145.16
13	115.55	115.55
15	156.52	156.52
16	93.93	93.93
17	164.34	164.34
19	98.65	98.65
20	160.88	160.88
22	103.49	103.49

lotus_quercetin.NMRUBD

Other NMR-based dereplication tools

- We wanted to analyse directly ^{13}C NMR spectra of mixtures, without (CPC) fractionation, by "naive dereplication" assisted by peak intensity analysis.
  - DerepCrude algorithm: *J. Nat. Prod.* **2017**, *80*, 5, 1387–1396.

- The DerepCrude algorithm was reworked (U. of Angers, France) to include the multiplicity information (CH_n with n = 0, 1, 2, 3), without considering ^{13}C NMR peak intensities, leading to the
  - MixONat algorithm: *Anal. Chem.* **2020**, *92*, *13*, 8793–8801.

- An attempt to isolate δ_C and δ_H clusters, compound by compound, on the 2D HSQC and HMBC NMR spectra lead to
  - HMBC networking algorithm: *J. Chem. Inf. Model.* **2018**, *58*, 262–270.

Future works

- *Urceolina peruviana*
  - Report the description and interpretation of currently identified compounds
  - Genrate NMReDATA files (nmredata.org)
  - Possibly identify other compounds
- PNMRNP
  - The Predicted NMR Natural Product (PNMRNP) database, zenodo.org/record/3765243, contains 210,000 compounds. A 3D version has been prepared with Schrödinger LigPrep Software for NP virtual docking with SARS-COV-2 proteins. Publication of PNMRNP-3D is in progress.
- Prediction of ^{1}H NMR chemical shifts (and couplings?)
  - Using the same principle used with ACD/Labs δ_C prediction. Problems of H atom chemical non-equivalence have to be solved.
- Prediction of 2D NMR spectra…

MANY THANKS TO…

- Pr Jean-Hugues Renault
- Dr Jane Hubert
- Dr Alexis Kotland
- Dr Ali Bakiri
- Nicolas Borie

CPC
CARAMEL
ACD

- Dr Simon Rémy
- Carine Machado
- Dominique Harakat

MS

- Agathe Martinez
- Anthony Robert

NMR

Alma Mater Studiorum A.D. 1088

Université de Reims Champagne-Ardenne

ICMR Institut de Chimie Moléculaire de Reims UMR CNRS 7312

cnrs dépasser les frontières

Liberté • Égalité • Fraternité République Française

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation

FEDER

Région Champagne Ardenne

Marne Le Département

Natural Product Team in Reims