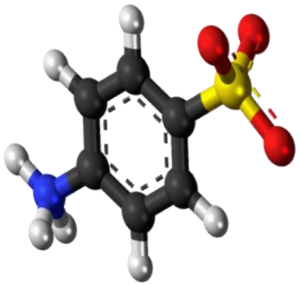
SFSM sfeap RFMF

Atelier 3



Outils d'analyse des spectres de fragmentation pour l'identification des petites molécules

Alexis DELABRIERE, (alexis.delabriere@cea.fr)
Yann GUITTON (yann.guitton@oniris-nantes.fr)
David TOUBOUL (david.touboul@cnrs.fr)
Etienne THEVENOT (etienne.thevenot@cea.fr)

cnrs ICSN cea LABERCA

Au menu aujourd'hui

Introduction : la MS/MS pour l'annotation structurale

I. Prédiction d'une structure chimique à partir d'un spectre MS2

IA) Prétraitement des spectres MS2 (MzMine)

IB) Prédiction d'une structure (CFM-ID, CSI:FingerID, MS-Finder, MetFrag)

II. Exploration structurale d'une collection de spectres MS2

IIA) Réseaux moléculaires (GNPS)

IIB) Recherche de motifs (MS2LDA, MineMS2)

Introduction

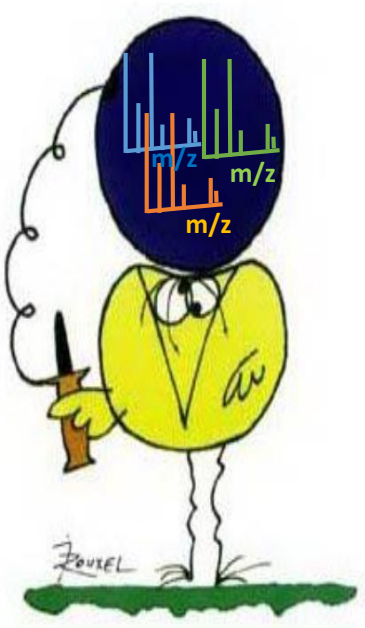
Les données OMICS et la problématique du « big data »

L'étude des systèmes biologiques conduit à la nécessité de gérer des données de plus en plus conséquentes de données afin d'évaluer un plus grand nombre de variables (pertinentes ou non).

Deux approches sont possibles et complémentaires:

- Approches ciblées (connaissance *a priori* de l'objet à étudier ou hypothèses initiales plus ou moins crédibles)
- Approches non ciblées (génération de « big data »)

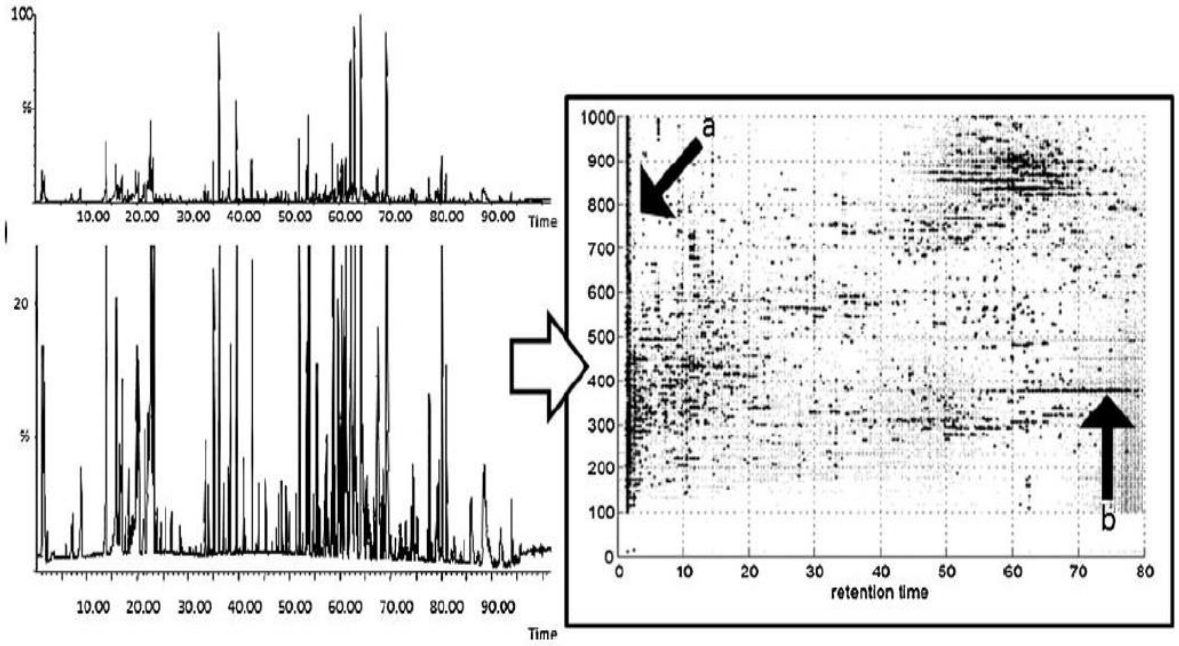
- Comment les générer ?
- Comment les stocker ?
- Comment les analyser ?



D. Touboul

Les données OMICS et la problématique du « big data »

- Un cas compliqué : la métabolomique non-ciblée

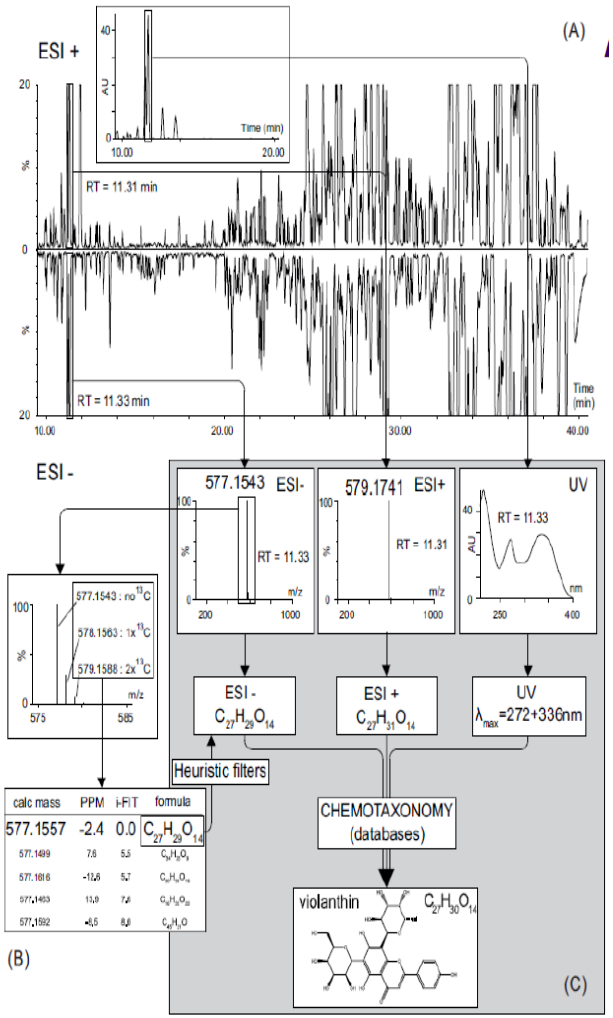


Pas plus de complexité qu'en protéomique en terme de nombre de molécules détectées
Mais plus compliqué d'interprétation car:

- un grand nombre de fonctions chimiques et de squelettes carbonés
- base de données incomplètes

D. Touboul

Du spectre MS(/MS) à la structure



D. Touboul

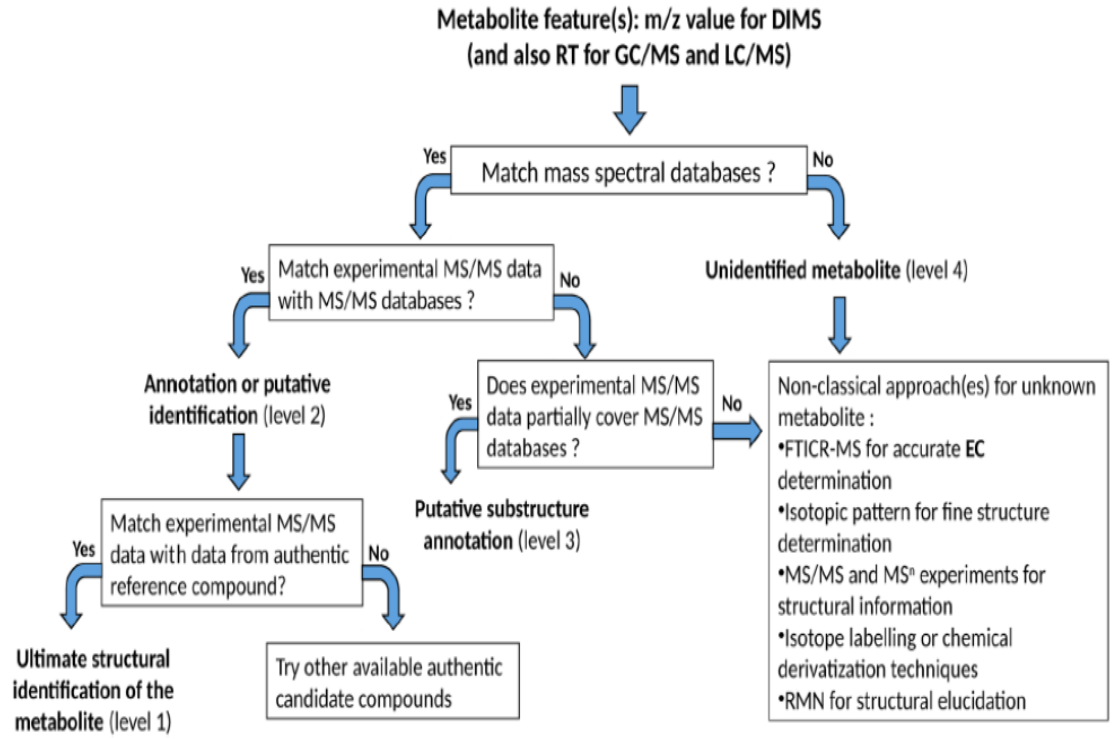
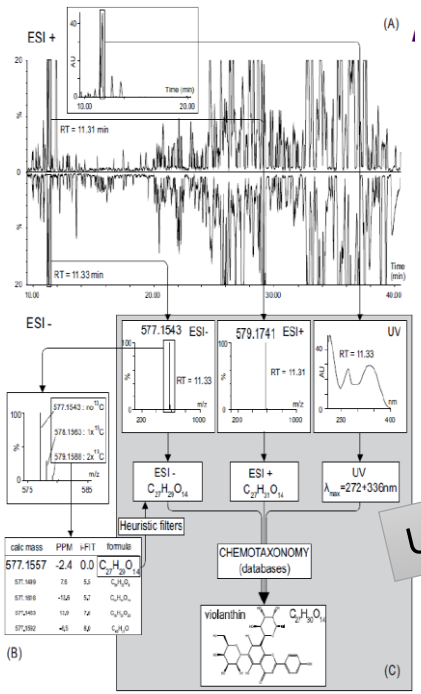


Fig. 3 Strategy for putative metabolite annotation or identification (built from Sumner et al. 2007)

DOI 10.1007/s11306-015-0882-8

Du spectre MS(/MS) à la structure

- Les étapes



- 1 pic → 1 jour ? … mais parfois bien plus
  - m/z, massif isotopique, MS/MS, MSn, UV….
    Purification, RMN

- Une forêt de pics → ? Besoin d'automatisation
  - Quelles stratégies ?
    - Base de données spectrales (NIST, Massbank, Metlin…)
    - Approches prédictives (in-silico)

Une vielle histoire

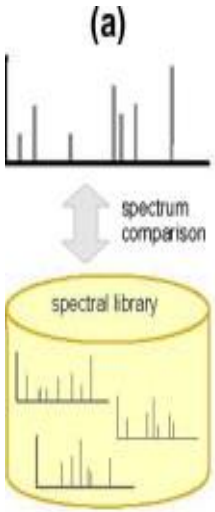
First rule-based approaches for predicting fragmentation patterns, as well as explaining experimental mass spectra with the help of a molecular structure, were developed as part of the **DENDRAL project that started back in 1965** […] "However, it is sad to say that, in the end, the DENDRAL **project failed** in its major objective of **automatic structure elucidation by mass spectral data,** and research was discontinued.

Scheubert *et al. Journal of Cheminformatics* 2013, **5**:12
http://www.jcheminf.com/content/5/1/12

Du spectre MS(/MS) à la structure - Stratégies

- Les stratégies

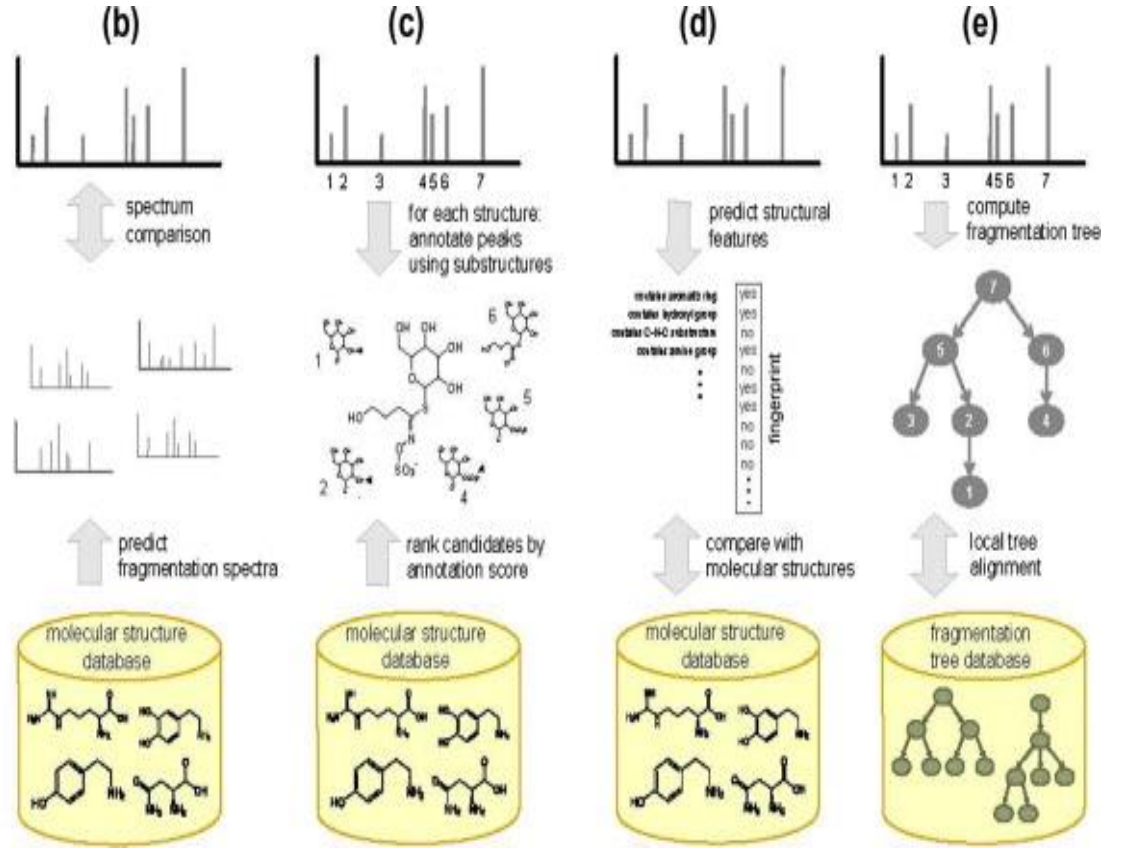
Les bases de données spectrales



- NIST
- MassBank (JP, MoNa, ..)
- METLIN
- In-House

- GNPS

Pproches prédictives (in-silico)



- MetFrag
- CFM-ID

- MassFrontier
- Fragmenter

- Magma

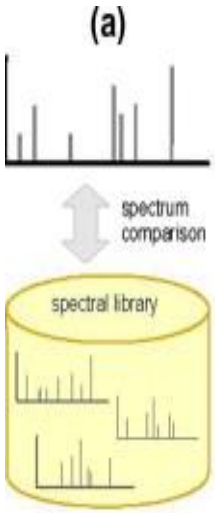
- CSI FingerID

- Sirius

F. Hufsky et al./Trends in Analytical Chemistry 53 (2014) 41–48

Les bases de données spectrales

Les bases de données spectrales



- NIST
- MassBank (JP, MoNa, ..)
- METLIN
- In-House

- GNPS

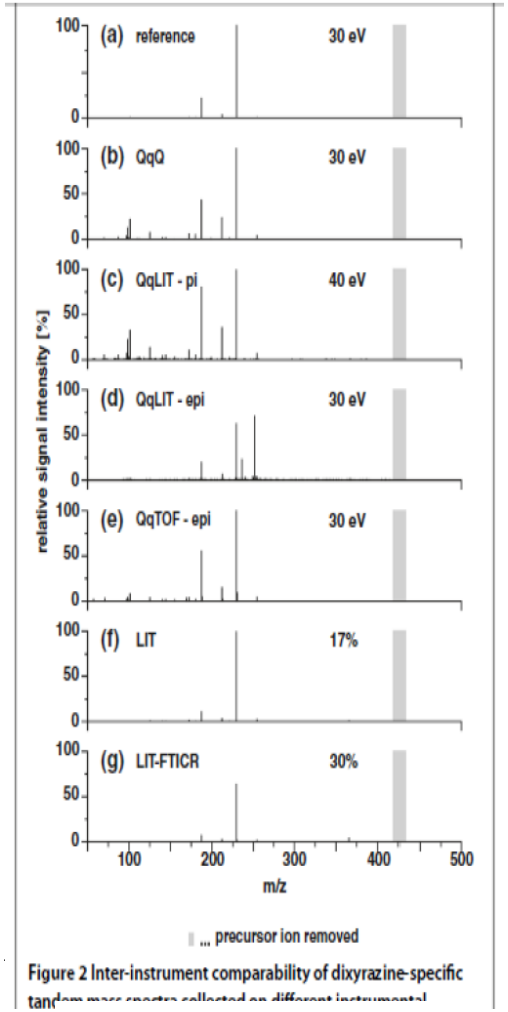


Figure 2 Inter-instrument comparability of dixyrazine-specific tandem mass spectra collected on different instrumental

Scheubert *et al. Journal of Cheminformatics* 2013, **5**:12
http://www.jcheminf.com/content/5/1/12

Un composés, des spectres de référence…
→ Quels critères d'identification retenir ?

Table 1
MS/MS data from electrospray ionization in NIST 14

Mass Analyzer	# spectra	# compounds	# ions
Ion trap	~39,000	~6000	~39,000
Q-TOF (CID)	~42,000	~3000	~4000
QqQ (CID)	~27,000	~1000	~3000
Orbitrap (HCD)	~69,000	~3000	~6000
Ion trap with FTMS	~5000	~2500	NA

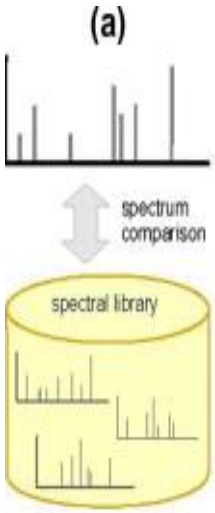
M. Vinaixa et al./Trends in Analytical Chemistry 78 (2016) 23–35

Toujours de grands chiffres mais au final combien de composés ?

F. Hufsky et al./Trends in Analytical Chemistry 53 (2014) 41–48

Les bases de données spectrales

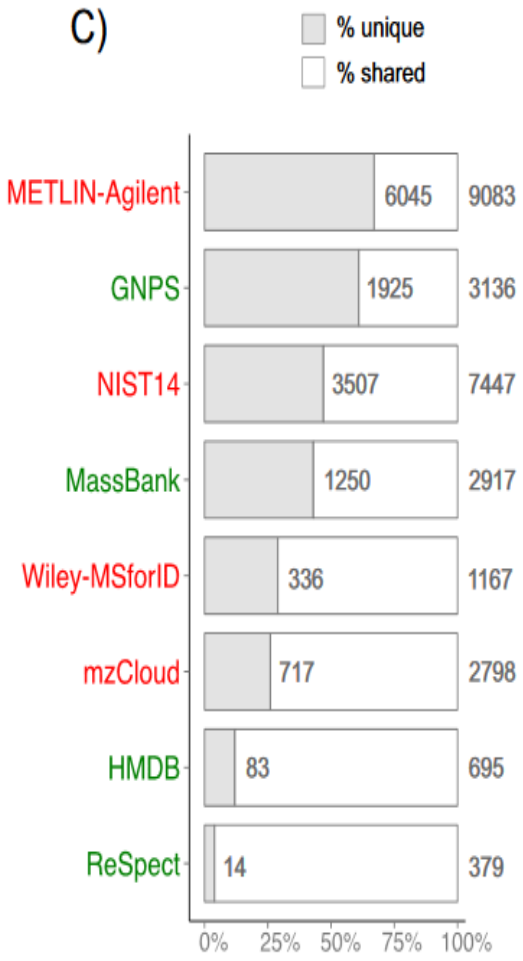
Les bases de données spectrales



- NIST
- MassBank (JP, MoNa, ..)
- METLIN
- In-House

- GNPS

Composé unique avec spectre MS/MS



M. Vinaixa et al./Trends in Analytical Chemistry 78 (2016) 23–35

Des bases de données de spectres MS/MS difficilement exhaustives. **Seul 5-10 %** des composés connus

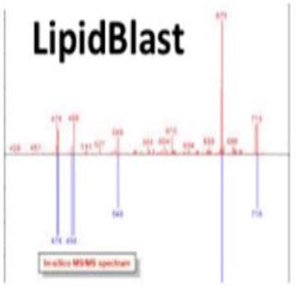
Les approches prédictives (in-silico)

→ Base de données théorique de spectres basée sur des règles issues de standards (Lipidomique)

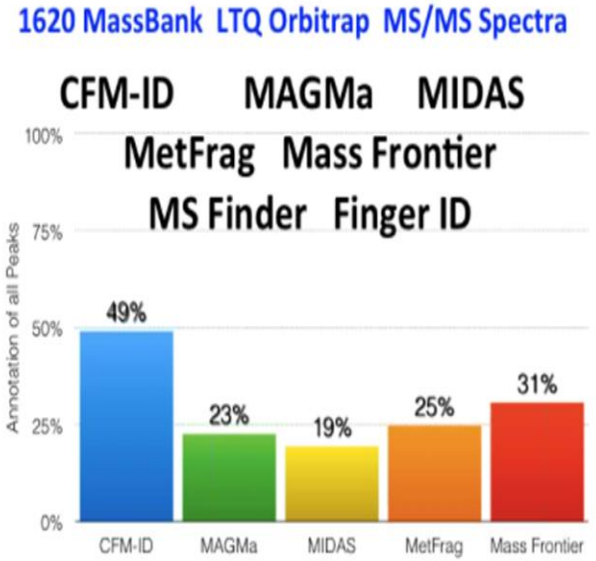
→ Spectres issus de calculs (Règles de chimie et/ou purement informatiques)

Comparaison des outils permettant de générer des spectres MS/MS Théoriques

Mass spectral libraries



- 200,000 MS/MS spectra
- Rules from authentic spectra
- Fully validated



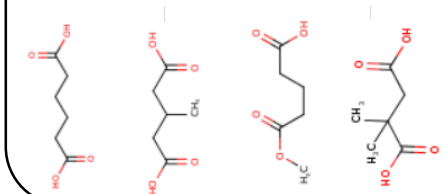
Program	CFM ID	MAGMa	MIDAS	MetFrag	Mass Frontier
CFM ID	100 %	8.2 %	7.3 %	0.3 %	3.3 %
MAGMa	4.1 %	100 %	61.7 %	47.0 %	10.3 %
MIDAS	3.1 %	53.0 %	100 %	48.4 %	8.9 %
MetFrag	0.1 %	51.8 %	62.1 %	100 %	11.7 %
Mass Frontier	2.1 %	13.8 %	13.8 %	14.2 %	100 %

Tobias Wermuth

I. Prédiction d'une structure chimique à partir d'un spectre MS2

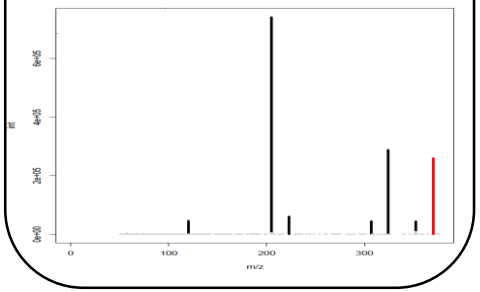
Fonctionnement d'un logiciel d'annotation

Base de données moléculaires



Input :

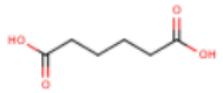
- Un spectre de masse
- La masse du précurseur

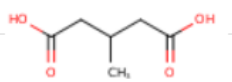


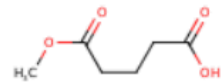
Outil d'annotation*

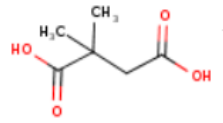
Modèle appris sur une base de données
CFM-ID : Probabilités de transitions
CSI-FingerID : Noyau et SVM.

Output :
Molécules scorées


0,98


0,85


0,63


0,82

- CSI-FingerID (Shen, 2014)
- CFM-ID (Allen, 2014)
- MetFrag (Wolf, 2010)
- Etc…

Bases de données

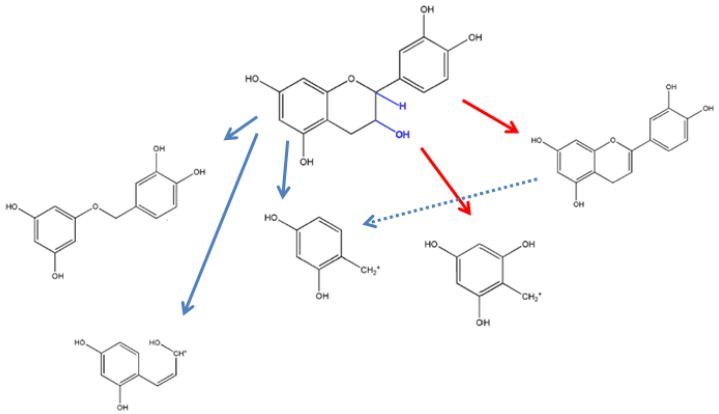
Les bases de données que l'on interroge sont donc très importantes, et à part les principales varient beaucoup d'un software à l'autre.

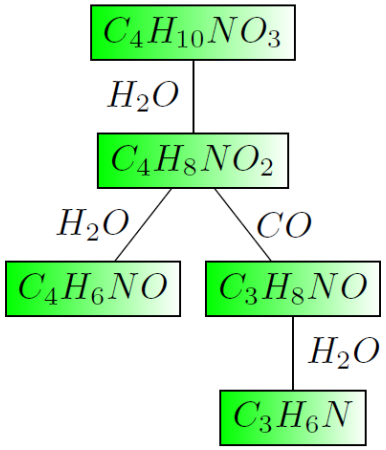
Software	Database
MS-Finder	BMDB, ChEbi, DrugBank, ECMDB, FooDB, HMDB, KNApSAcK, PlantCyc, PubChem, SMPDB, T3DB, UNPD, YMDB, MINE, STOFF
MetFrag	PubChem, KEGG, ChemSpider, MetaCyc, FOR-IDENT, Lipids-Map, CheBi, HMDB
CFM-ID	HMDB, KEGG
CSI:FingerID	PubChem, BioCyc, CheBi, GNPS, HMDB, HSDB, KEGG, KNApSAcK, MeSH

Stratégie d'annotation

Tous les outils d'annotation essayent de reconstituer le processus de fragmentation :

En cassant systématiquement TOUTES
les liaisons dans le graphe moléculaire





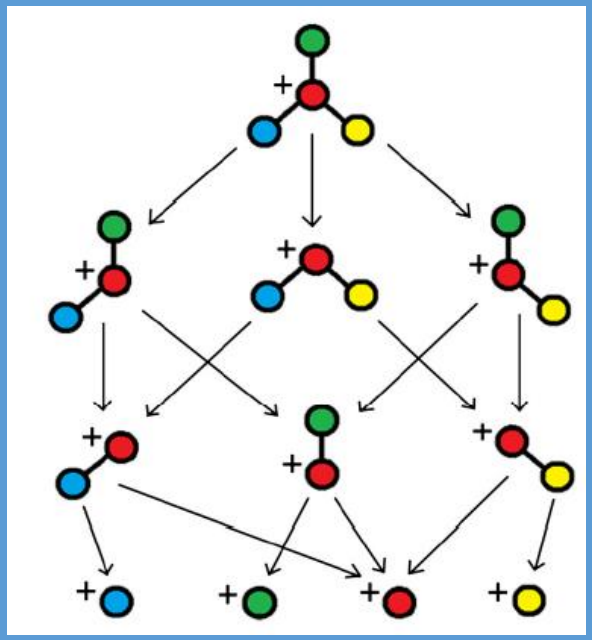
En générant toutes les sous-formules de
formules candidates

Certains peuvent inclure des réarrangements, ou des règles chimiques.

Stratégie d'annotation

Ces fragmentations peuvent ensuite soit être stockées dans un modèle statistique, ou directement scorées en utilisant un modèle physique.

Modèles probabilistes



Ex : CFM-ID

Modèle physique

TABLE 7.1 Average Bond Dissociation Energies, *D* (kJ/mol)[a]

H—H	436[a]	C—H	410	N—H	390	O—H	460	F—F	159[a]
H—C	410	C—C	350	N—C	300	O—C	350	Cl—Cl	243[a]
H—F	570[a]	C—F	450	N—F	270	O—F	180	Br—Br	193[a]
H—Cl	432[a]	C—Cl	330	N—Cl	200	O—Cl	200	I—I	151[a]
H—Br	366[a]	C—Br	270	N—Br	240	O—Br	210	S—F	310
H—I	298[a]	C—I	240	N—I	—	O—I	220	S—Cl	250
H—N	390	C—N	300	N—N	240	O—N	200	S—Br	210
H—O	460	C—O	350	N—O	200	O—O	180	S—S	225
H—S	340	C—S	260	N—S	—	O—S	—		
Multiple covalent bonds[b]									
C=C	611	C≡C	835	C=O	732	O=O	498[a]	N≡N	945[a]

Dépendant des données d'apprentissage

Limité par nos connaissances théoriques

Résultats du CASMI

http://casmi-contest.org/2017/index.shtml

Le CASMI est un concours d'identification de spectres inconnus à partir de molécules candidates **qui existent** dans les bases de données.

Bon indicateur de l'efficacité des logiciels.

Table 2 Results summary for Categories 2 and 3: medal tally and other statistics

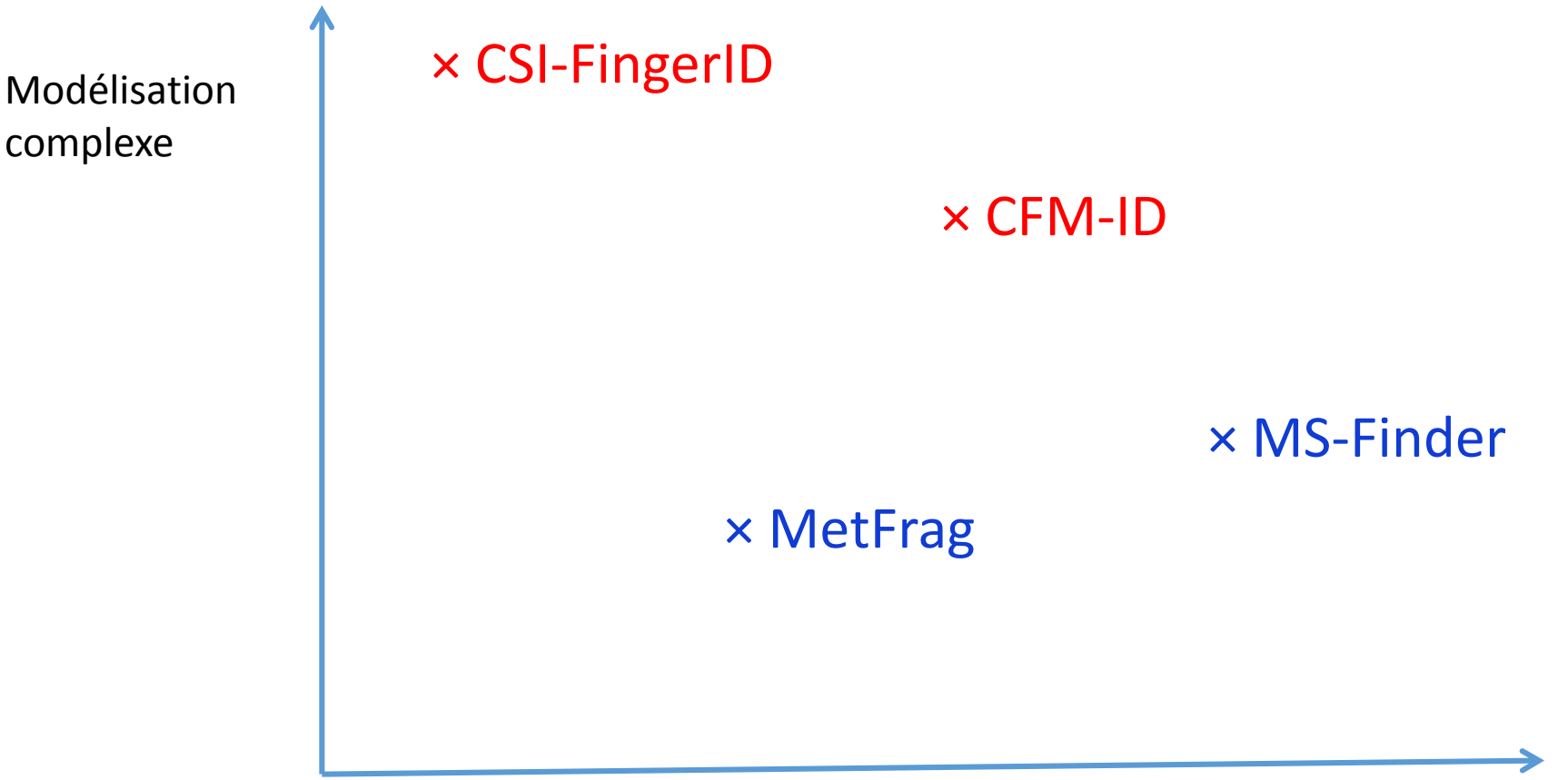
	Category 2					Category 3	
	Allen CFM orig	Brouard CSI: IOKR_A	Dührkop CSI:FID	Vaniya MS–FINDER	Verdegem MAGMa+	Allen CFM retrain +DB	Kind MS–FINDER +MD
Gold	63	**86**	82	70	44	156	**159**
Silver	**71**	50	21	26	53	**52**	38
Bronze	40	31	11	35	**65**	0	0
Gold (neg)	26	20	0	**33**	24	61	**64**
Gold (pos)	37	66	**82**	37	20	**95**	**95**
Top 1 (neg)	12	9	0	**14**	8	47	**59**
Top 1 (pos)	27	53	**70**	32	16	73	**47**
Top 1	39	62	**70**	46	24	120	**146**
Top 3	77	**93**	90	79	59	160	**162**
Top 10	**123**	118	100	101	105	**182**	174
Mean rank	47.98	127.34	25.17	**19.75**	70.79	13.72	**6.4**
Median rank	6	5.2	**1**	3	9.8	**1**	**1**
Mean RRP	0.906	0.874	**0.945**	0.804	0.88	**0.971**	0.904
Median RRP	0.987	0.988	**1**	0.922	0.972	**1**	**1**
Formula 1	1957	**2276**	2156	1867	1524	3861	**4011**
Medal Score	275	375	**396**	305	195	700	**766**

CSI-FingerID est assez loin devant.

En négatif les softs statistiques semblent avoir des difficultés probablement à cause du manque de données.

Resutats CASMI : Schymanski, 2017

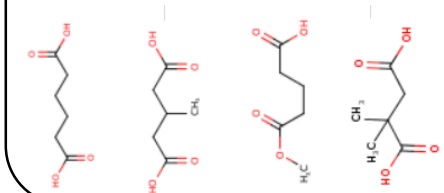
Fragmentation *in silico*



bleu : modèle physique, rouge : modèle statistique

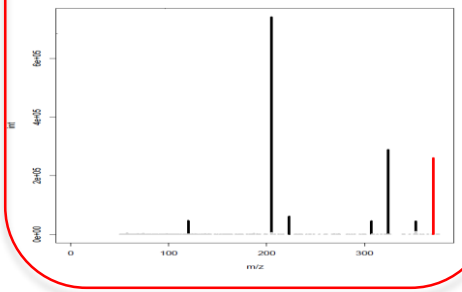
Fonctionnement d'un logiciel d'annotation

Bases de données moléculaires



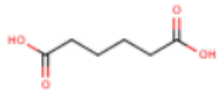
Input :

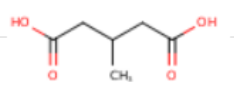
- Un spectre de masse
- La masse du précurseur

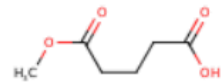


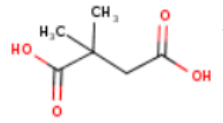
Outil d'annotation*

Output :
Molécules scorées


0,98


0,85


0,63


0,82

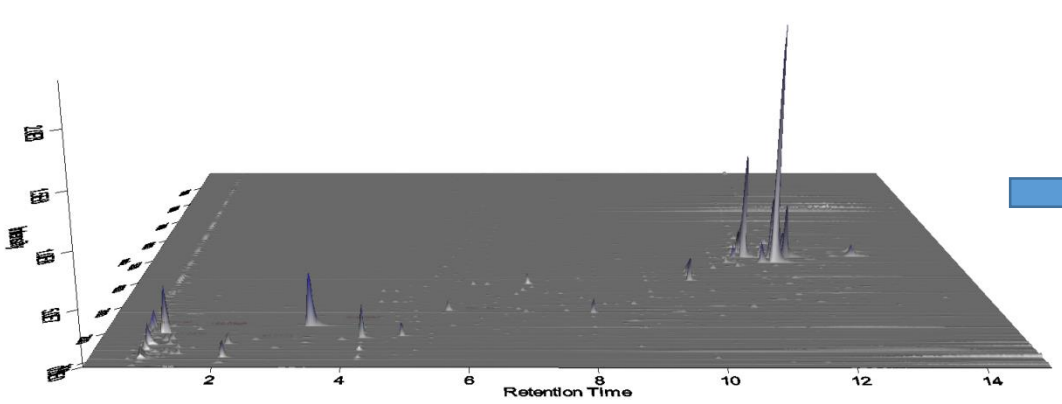
- CSI-FingerID (Shen, 2014)
- CFM-ID (Allen, 2014)
- MetFrag (Wolf, 2010)
- Etc…

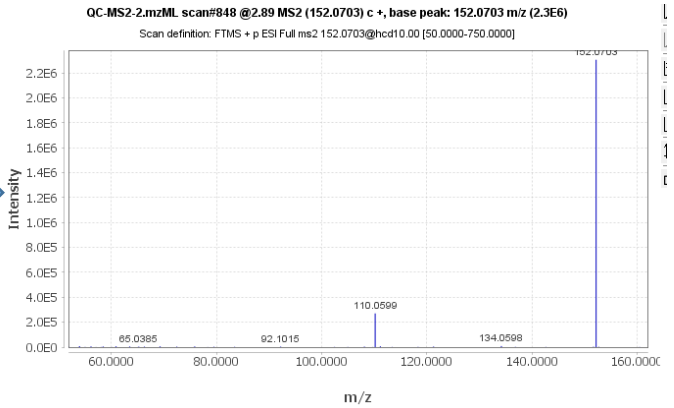
Modèle appris sur une base de données
CFM-ID : Probabilités de transitions
CSI-FingerID : Noyau et SVM.

IA) Prétraitement des spectres MS2

Passer des données brutes au mgf

La première question est de passer des données brutes à un spectre MS-MS utilisable



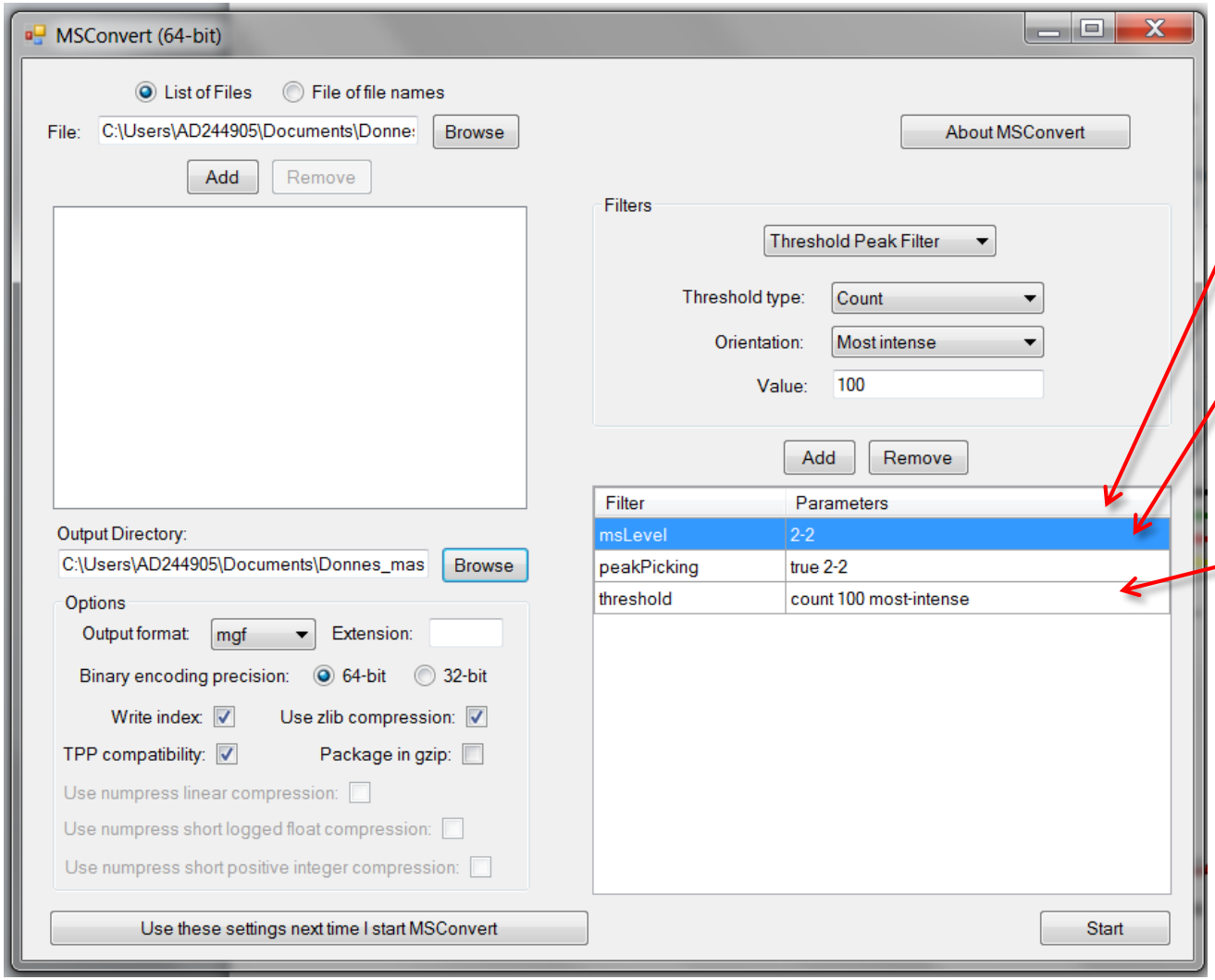


- Solution constructeur (Non évoquée ici)
- Solution open source :

Très facile : MSconvert (proteoWizard)
Facile : MzMine
Plus avancé (programmation) : RMassBank, XCMS

```
BEGIN IONS
FEATURE_ID=925
PEPMASS=162.1122
SCANS=925
RTINSECONDS=50.35
CHARGE=1+
MSLEVEL=2
60.080780029296875 17483.728515625
102.09111785888672 5824.17578125
103.03885650634766 21899.4140625
162.04971313476562 10322.103515625
```

Exemple : MSConvert



Uniquement les spectres MS2

Détection de pic sur les scans MS2

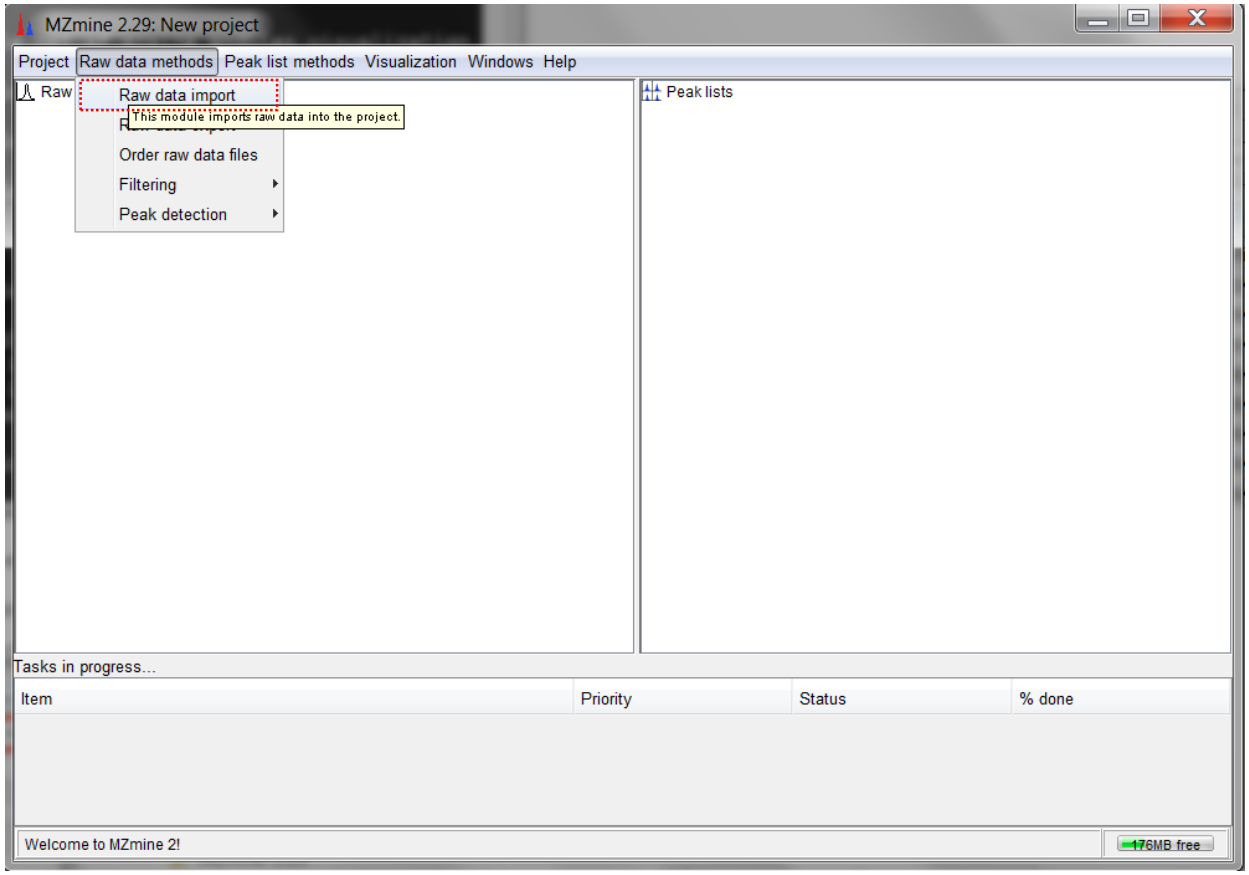
(Option) On garde uniquement 100 pics par spectre MS2

Autre option : Filtre en intensités brutes.

Exemple : MzMine (1)

On doit ouvrir le fichier
Raw Data Methods/ Raw Data import

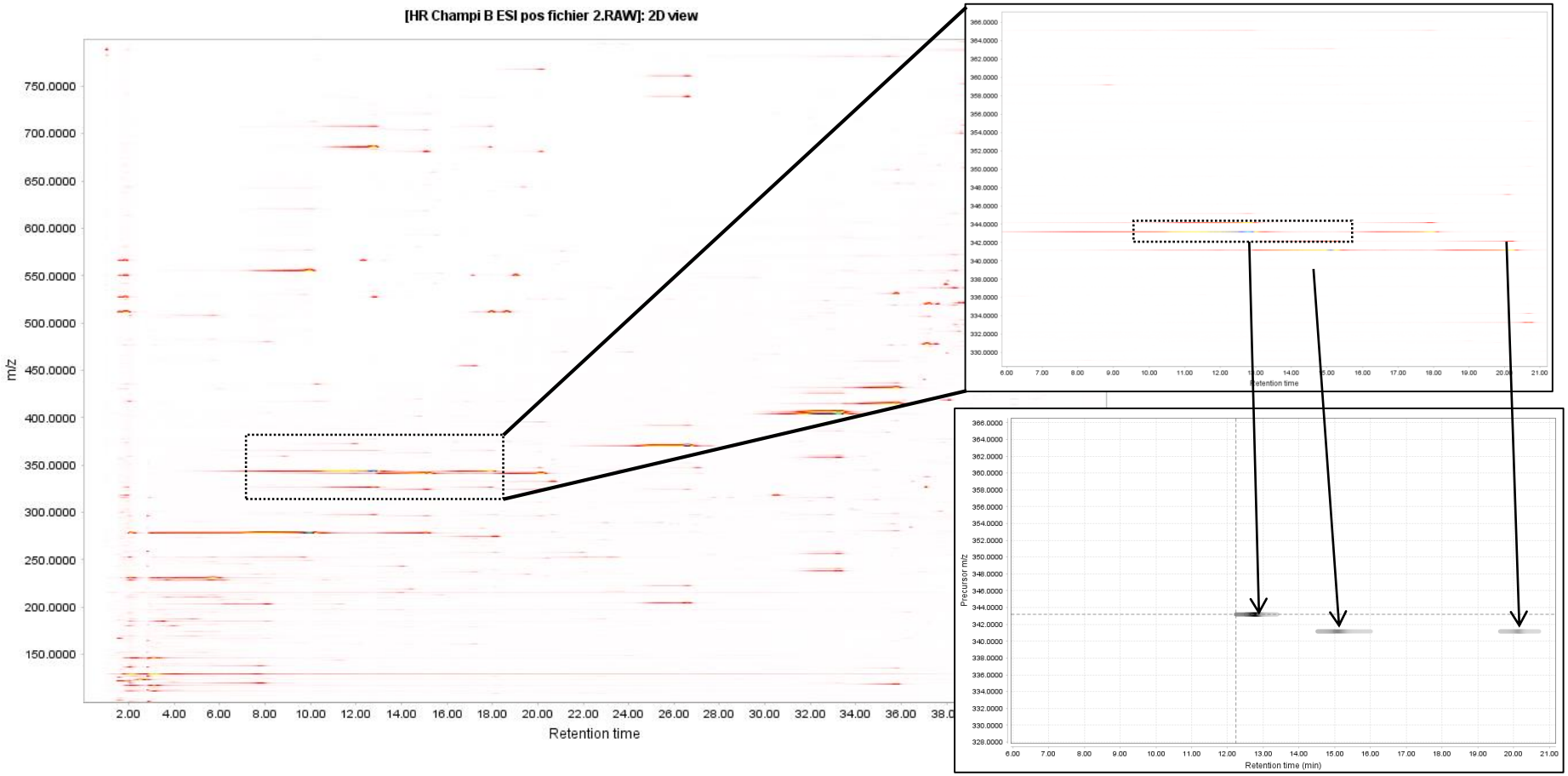
MzMine lit beaucoup de formats différents : mzML, mzXML, CDF, Thermo .RAW, Waters. RAW



Exemple : MzMine (Visu)

Une acquisition,

- Plusieurs spectres MS2
- Plusieurs spectres MS2 pour une molécule

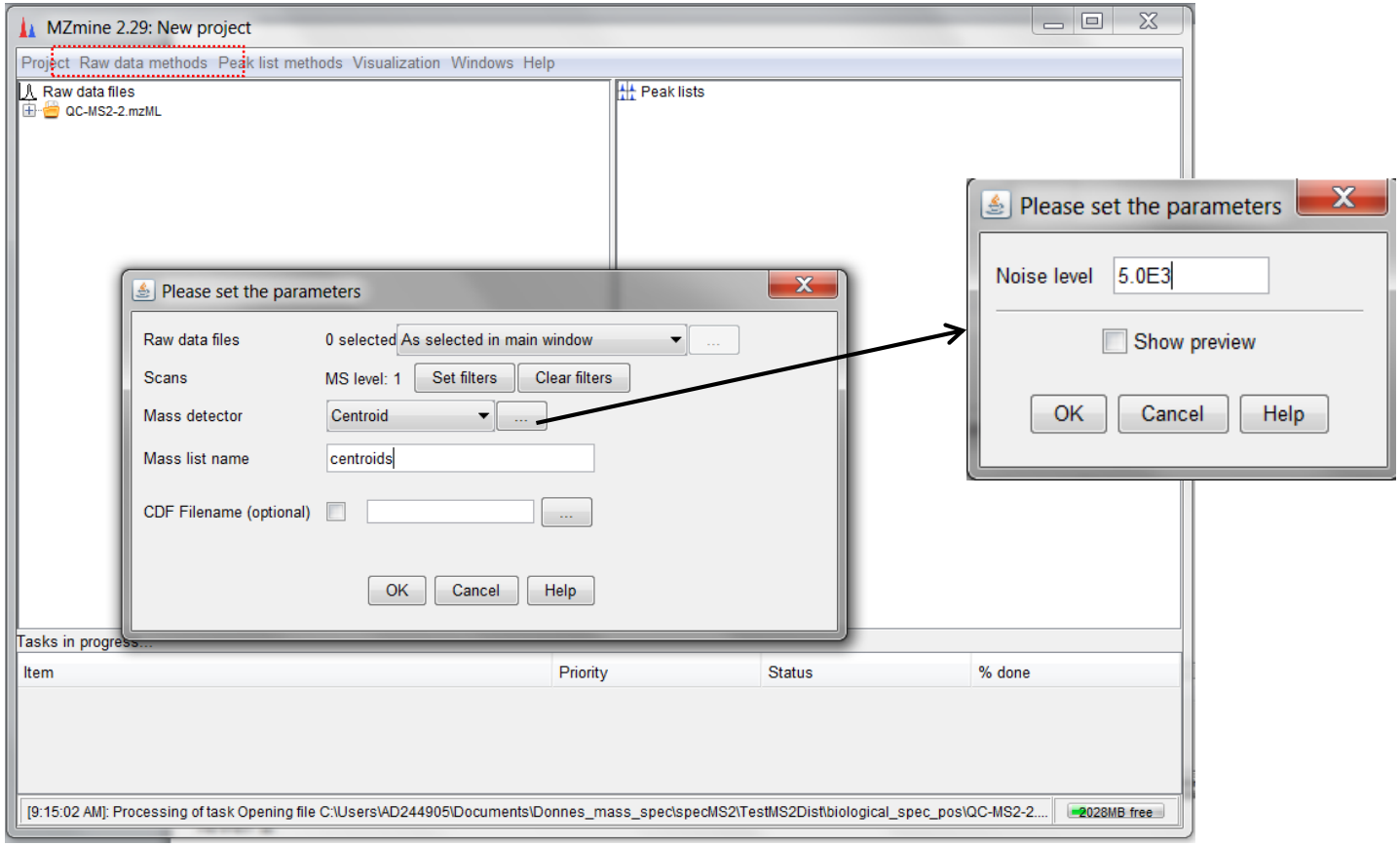


Exemple : MzMine (2)

Performing mass detection, peak detection in mass dimension :

On MS1:
Raw data methods / Mass detection / Set filter : MS level 1

On MS2:
Raw data methods / Mass detection / Set filter : MS level 2



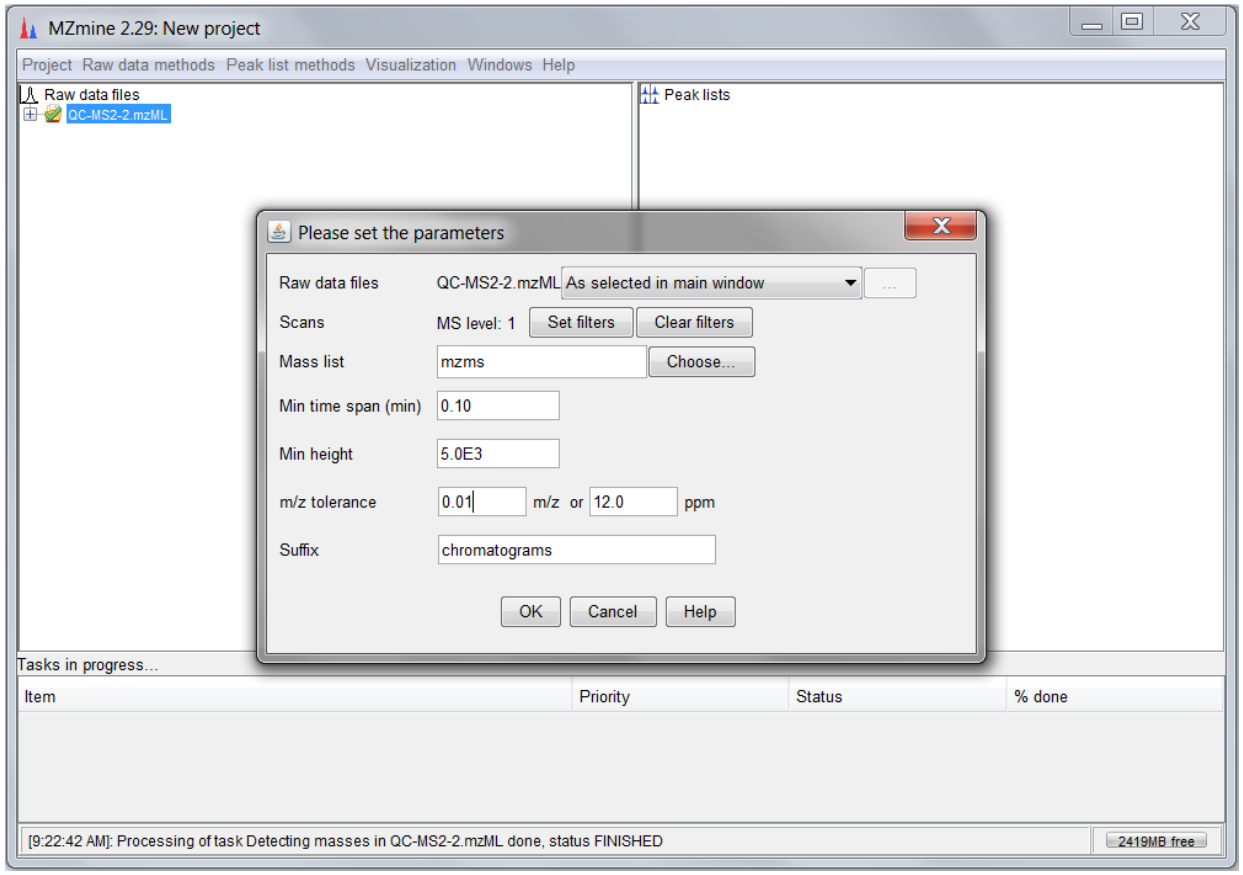
Two mass detections because noise on MS-MS spectra is often < to noise on MS scans

Exemple : MzMine (3)

Construction des chromatogrammes

Raw data methods / Chromatogram builder

Lien entre les points de masses similaires danns les différents scans

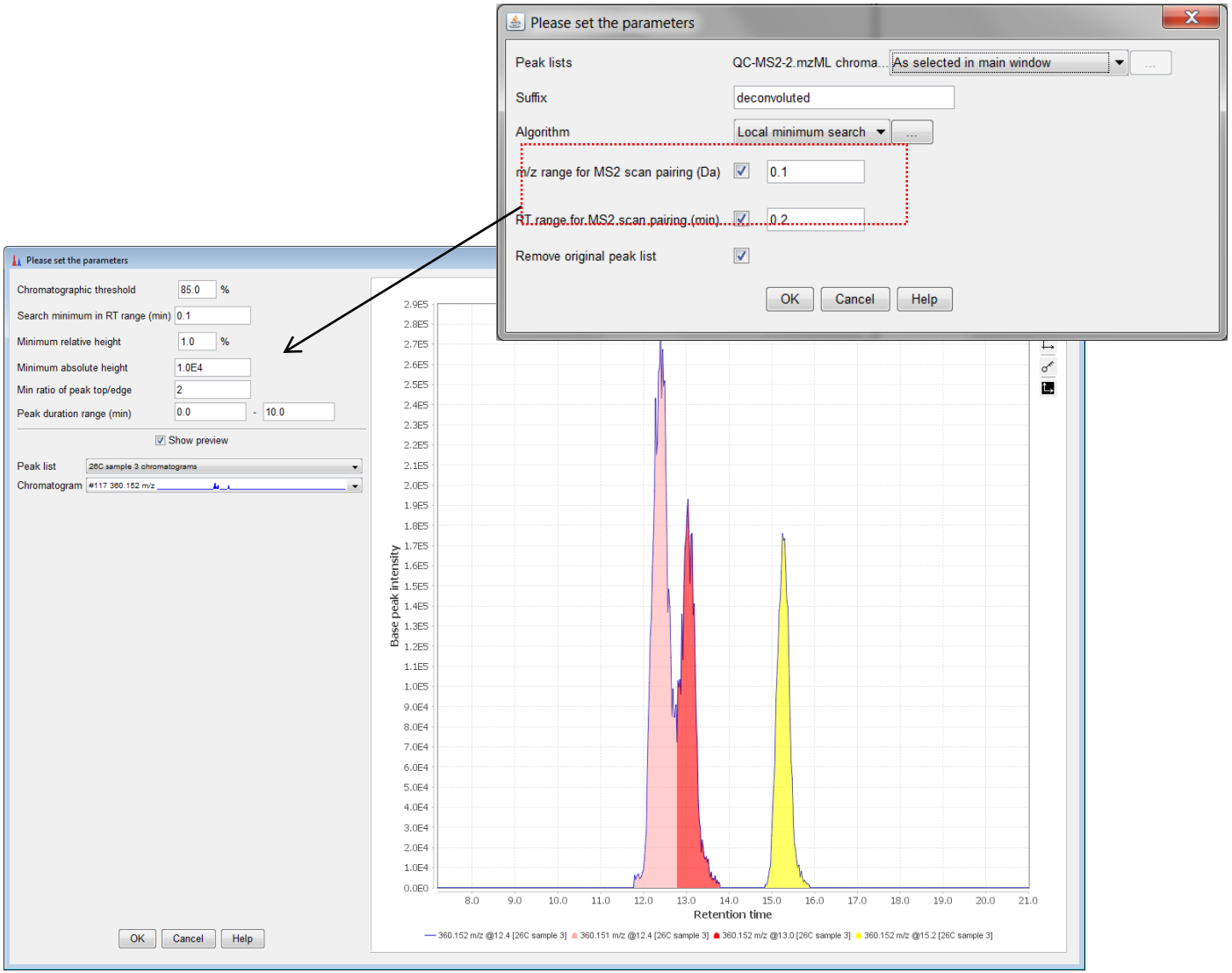


Exemple : MzMine (4)

Detection des pics sur les chromatogrammes

Peak list methods / Peak detection / Chromatogram deconvolution

C'est à cette étape que les spectres MS et MS2 sont reliés en utilisant les masses des précurseurs.



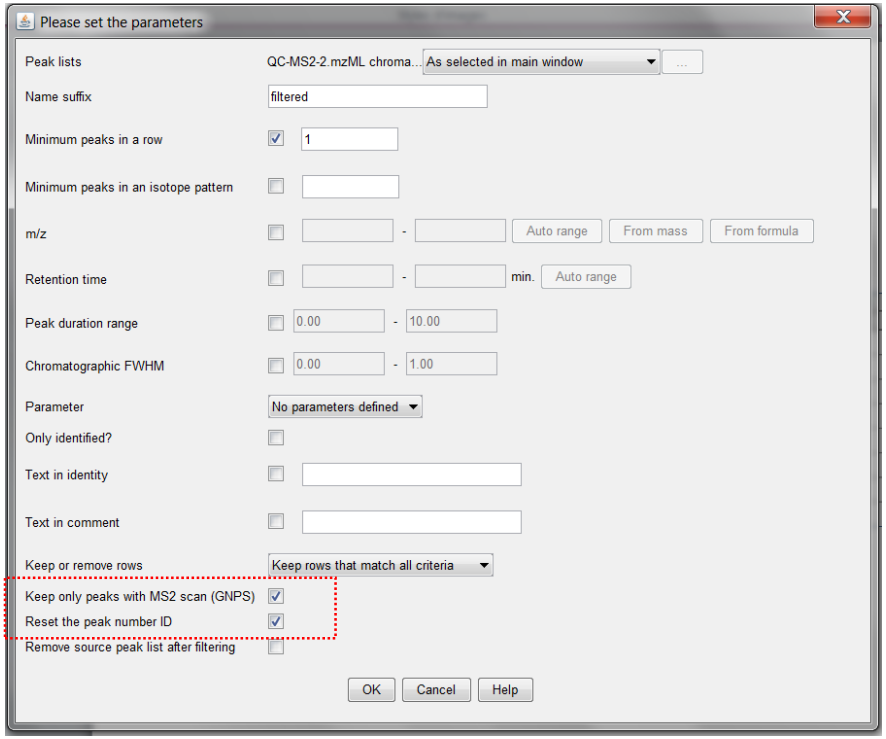
Exemple : MzMine (5)

A cette étape on a une peak table.

ID	Average		Identity	Comment	Peak shape	QC-MS2-2.mzML		
	m/z	RT				Status	Height	Area
31	131.0489	2.11					6.6E5	2.1E6
32	132.0805	4.24					6.1E5	1.7E6
33	132.1016	1.39					6.2E6	3.5E7
34	135.0788	9.90					3.3E5	6.6E5
35	136.0866	1.18					4.3E6	1.4E7
36	138.0659	4.88					1.1E5	1.8E5
37	139.0499	1.08					3.1E5	1.4E6
38	142.1224	0.92					2.1E6	1.2E7

On filtre les spectres qui n'ont pas de spectres MS2 associés :

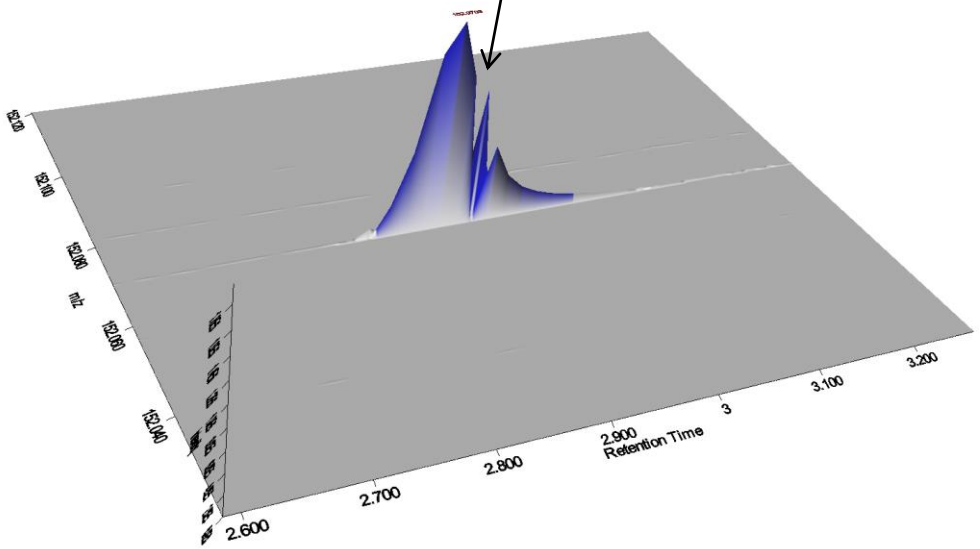
Peak list methods / Filtering / Peak list row filter



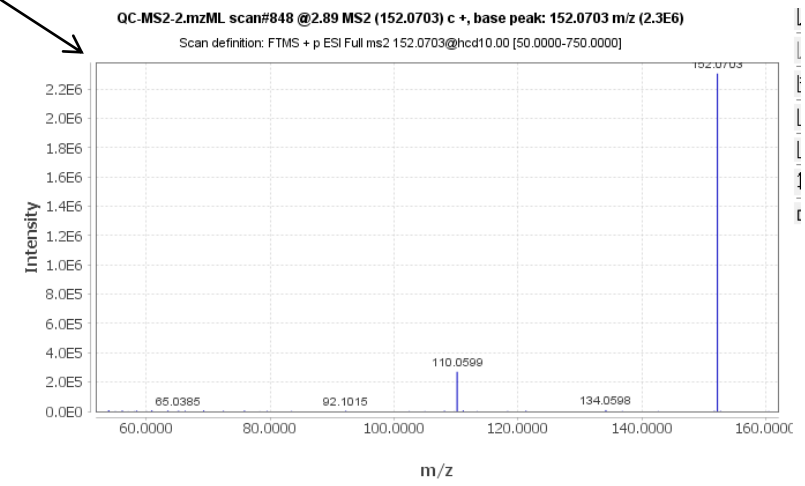
Exemple : MzMine (Visu)

A cette étape on peut voir les pics détectés et les spectres associés

ID	Average		Identity	Comment	Peak shape	QC-MS2-2.mzML		
	m/z	RT				Status	Height	Area
1	139.0499	1.08					3.1E5	1.4E6
2	152.0703	2.88					2.0E6	9.2E6
3	166.0860	2.11					9.8E6	3.9E7
4	167.0560	1.45					9.0E4	1.9E5
5	182.0750	1.18					6.7E6	2.1E7
6	205.0967	4.24					5.7E6	1.6E7
7	260.1851	6.19					7.3E5	1.5E6



Pic en 3D



Spectre MS-MS

Exemple : MzMine (6)

Pour exporter tous les spectres MS2 dans un seul fichier :

Peak list methods / Export / Export for GNPS

Le format MGF et un format texte simple à lire.

6 étapes dans cet exemple. Dans les cas général il y en a un peu plus, je vous invite à vous référer au workflow d'analyses de GNPS :
GNPS data analysis workflow 2.0
Par Louis Félix Nothias

MGF format

```
BEGIN IONS
FEATURE_ID=925
PEPMASS=162.1122
SCANS=925
RTINSECONDS=50.35
CHARGE=1+
MSLEVEL=2
60.080780029296875 17483.728515625
102.09111785888672 5824.17578125
103.03885650634766 21899.4140625
162.04971313476562 10322.103515625
162.07627868652344 10549.431640625
162.11215209960938 796389.125
END IONS

BEGIN IONS
FEATURE_ID=970
PEPMASS=165.0543
```

97000 lignes dont au moins 10 fois le même spectre en utilisant MSconvert
200 pour 10 spectres en utilisant MzMine

IB) Prédiction d'une structure

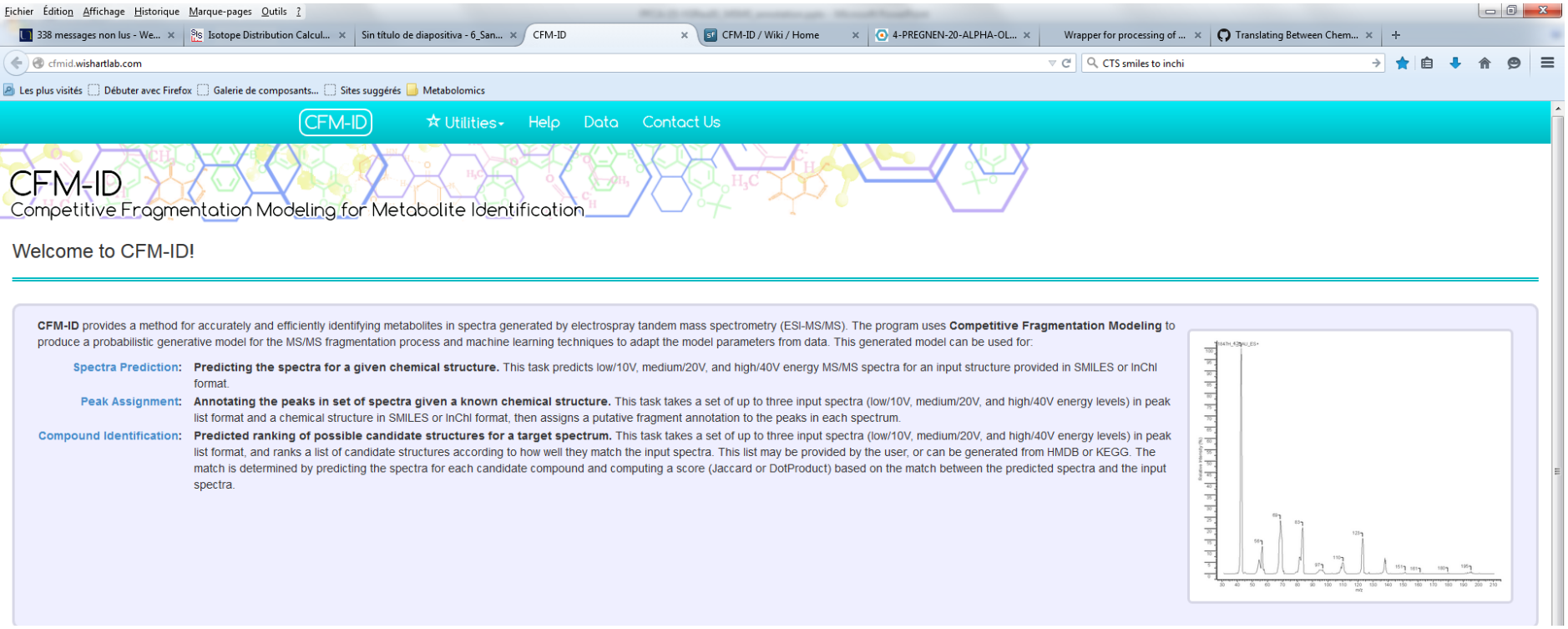
Explication des outils

- CFM-ID
- CSI-FingerID
- MS-Finder
- MetFrag
- Bilan

CFM-ID : Principe

Qui fait quoi?

- CFM-ID: prédiction de spectre MS2 à partir des structures
  assignation de structures aux fragments (pour molécule connues)
  Identification putative sur la base du MS2 et d'une liste de candidats



CFM-ID : Principe

Qui fait quoi?

- CFM-ID: prédiction de spectre MS2 à partir des structures

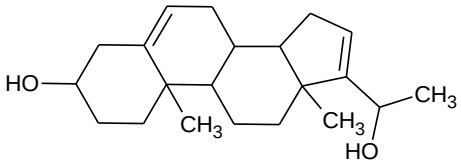
  assignation de structures aux fragments (pour molécule connues)

  Identification putative sur la base du MS2 et d'une liste de candidats

Smiles ou InChi

CC(C1=CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C)O

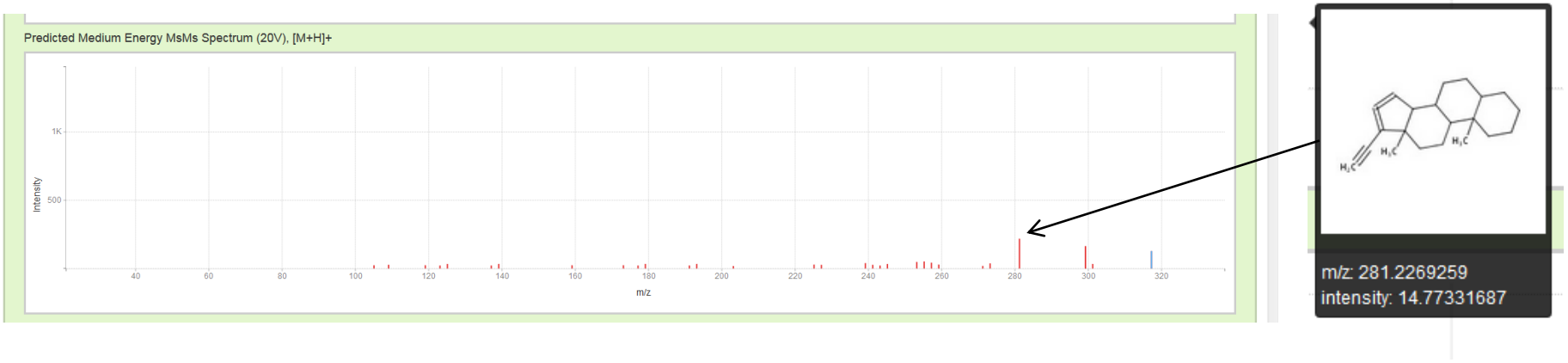
InChI=1/C21H32O2/c1-13(22)17-6-7-18-16-5-4-14-12-15(23)8-10-20(14,2)19(16)9-11-21(17,18)3/h4,6,13,15-16,18-19,22-23H,5,7-12H2,1-3H3



+

Polarité, type d'adduits fragmenté

Spectre MS2 théorique à 3 énergies (pratique quand y a pas de MS2 en base de donnée)



CFM-ID : Principe

Qui fait quoi?

- CFM-ID: prédiction de spectre MS2 à partir des structures

assignation de structures aux fragments (pour molécule connues)

Identification putative sur la base du MS2 et d'une liste de candidats

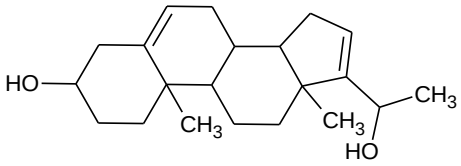
Smiles ou InChi

+

Spectre MS2 observé

CC(C1=CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C)O

InChI=1/C21H32O2/c1-13(22)17-6-7-18-16-5-4-14-12-15(23)8-10-20(14,2)19(16)9-11-21(17,18)3/h4,6,13,15-16,18-19,22-23H,5,7-12H2,1-3H3



CFM-ID : Principe

Qui fait quoi?

- CFM-ID: prédiction de spectre MS2 à partir des structures

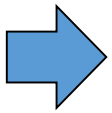
assignation de structures aux fragments (pour molécule connues)

Identification putative sur la base du MS2 et d'une liste de candidats

Spectre MS2 observé

+

Liste de candidats
IDS SMILES



1 0.11087749 84 CC(C1=CCC2C1(CCC3C2CC=C4C3(CCC(C4)O)C)C)O
2 0.11087749 519 C[C@@H](C1=CC[C@@H]2[C@@]1(CC[C@H]3[C@H]2CC=C4[C@@]3(CC[C@H](C4)O)C)C)O
3 0.11072997 166 CC1\2CCC3C(C1CC/C2=C/CO)CC=C4C3(CCC(C4)O)C
4 0.11072997 34 CC12CCC3C(C1CCC2=CCO)CC=C4C3(CCC(C4)O)C
5 0.11008478 400 C/C=C\[C@]12CC[C@@H](CC1=CC[C@@H]3[C@@H]2CC[C@]4([C@H]3CC[C@@H]4O)C)O
6 0.11008478 388 C/C=C/[C@]12CC[C@@H](CC1=CCC3C2CC[C@]4(C3CC[C@@H]4O)C)O
7 0.11008478 386 C/C=C\[C@]12CC[C@@H](CC1=CCC3C2CC[C@]4(C3CC[C@@H]4O)C)O
8 0.11006221 180 C[C@@H](C1CCC2C1(CCC3C2=CC=C4C3(CC[C@@H](C4)O)C)C)O

Liste triée avec scores

Comment avoir une liste de candidats ?

- recherche dans des bases de données ayant les smiles ou les InChi
- formatage du fichiers .txt avec Ids[space]SMILES

Pas toujours simple selon les bases de données

Attention si plusieurs DB utilisées retirer les redondances (là encore pas si simple)

CFM-ID : Principe

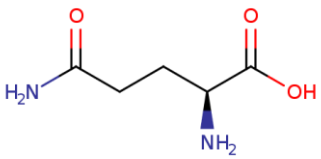
- Principe : CFM-ID apprend la probabilité de passer d'un fragment à un autre

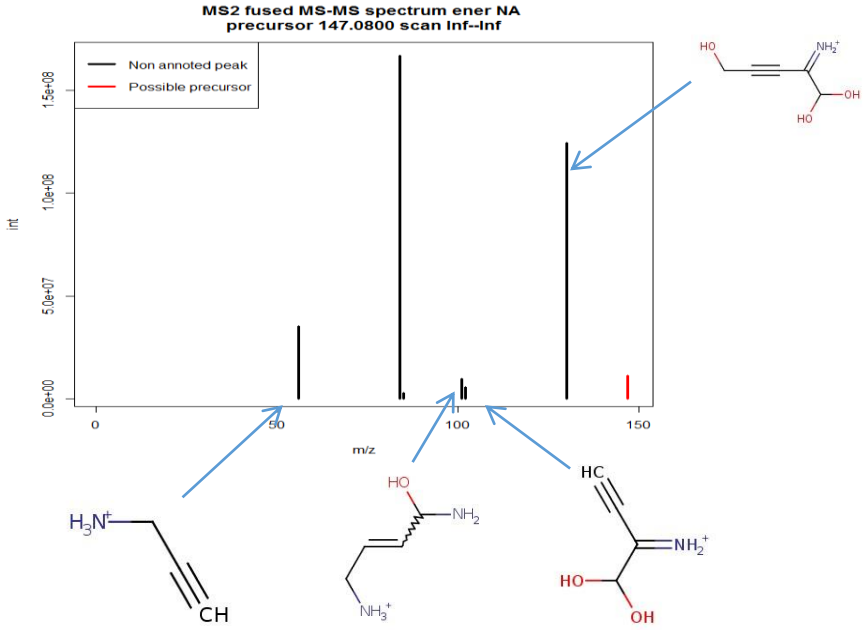
- Le modèle est appris à partir de couples spectre/ structure

CFM-ID : Apprentissage

Un graphe de fragmentation est construit en cassant toutes les liaisons d'une molécule

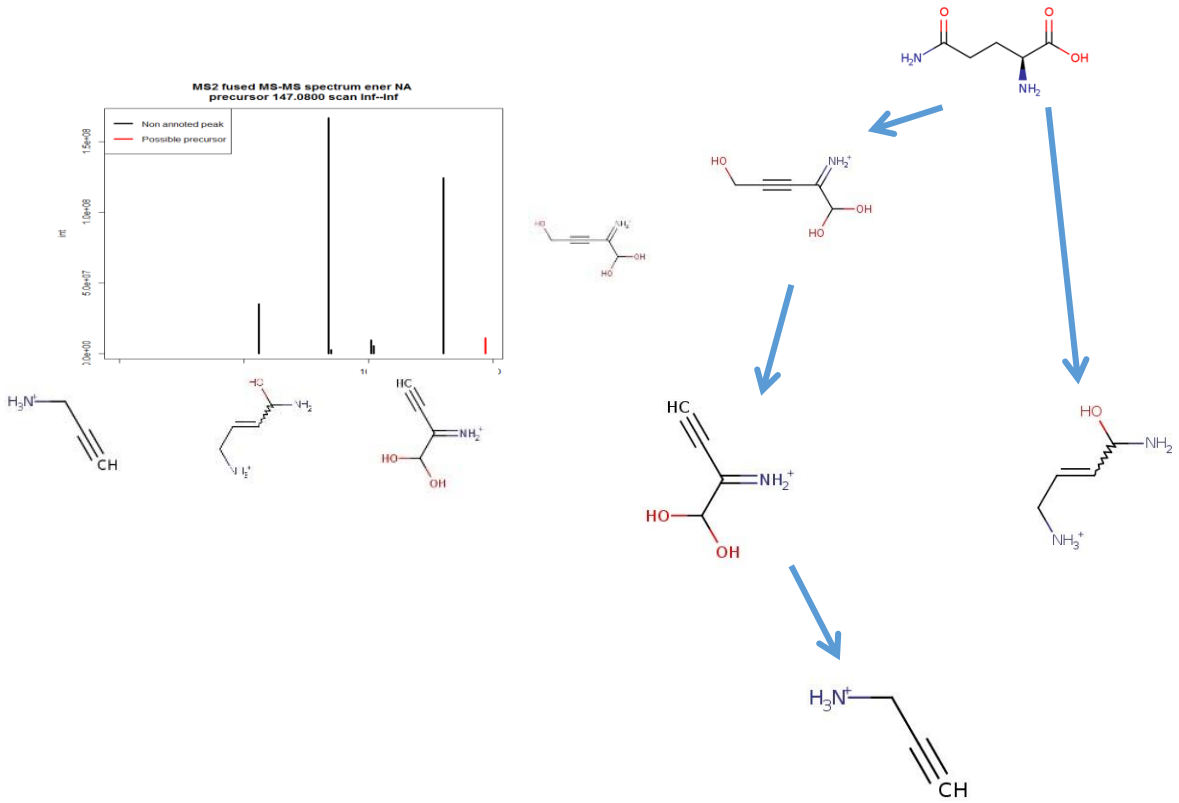
Exemple L-Glutamine





CFM-ID : Apprentissage

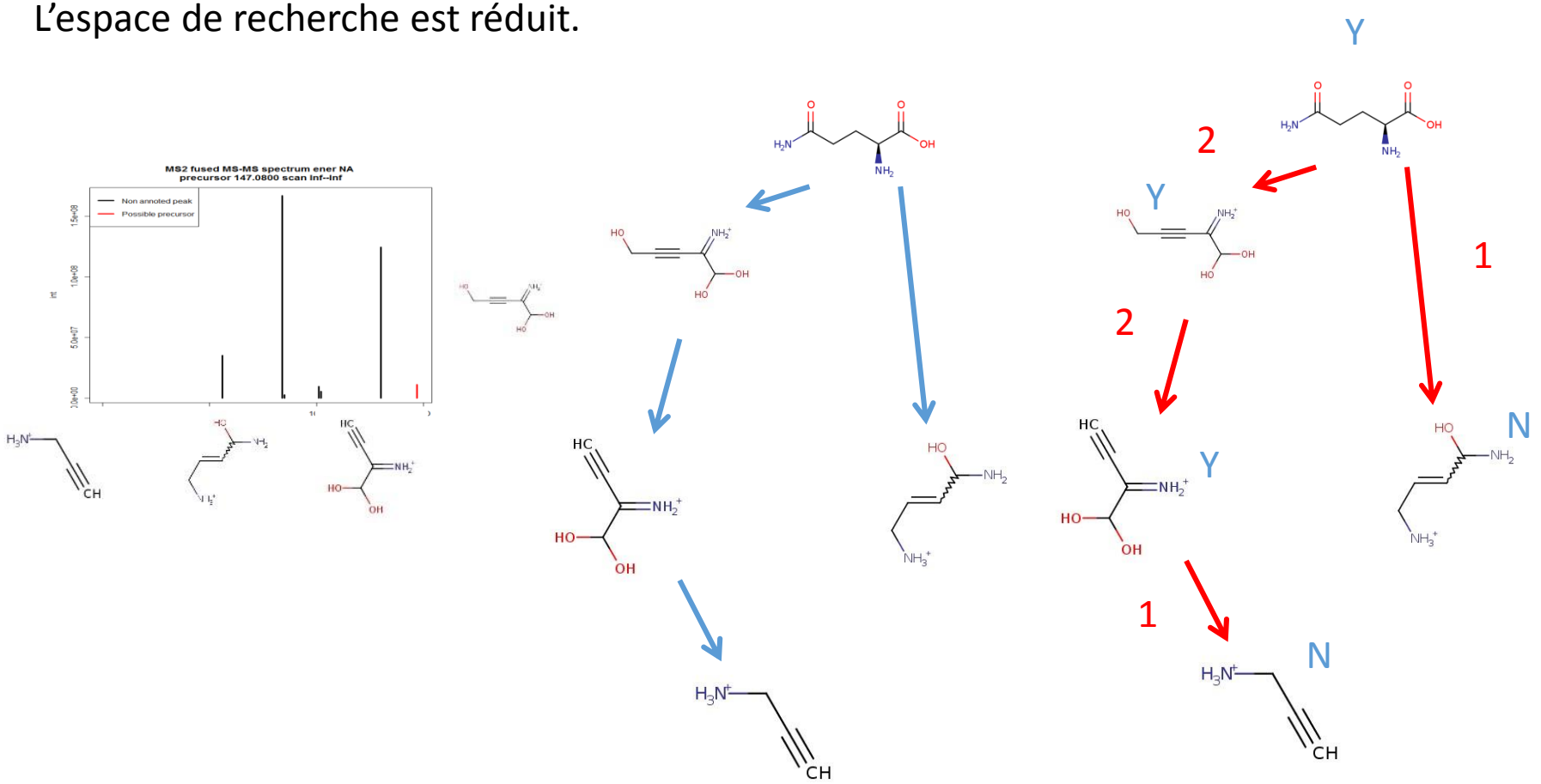
Un graphe de fragmentation est construit en cassant toutes les liaisons d'une molécule



Pour connaitre toutes les transitions il faudrait connaitre tous les fragments, ce qui est trop pour un ordinateur. On doit réduire la dimension du modèle.

CFM-ID : Apprentissage

L'espace de recherche est réduit.

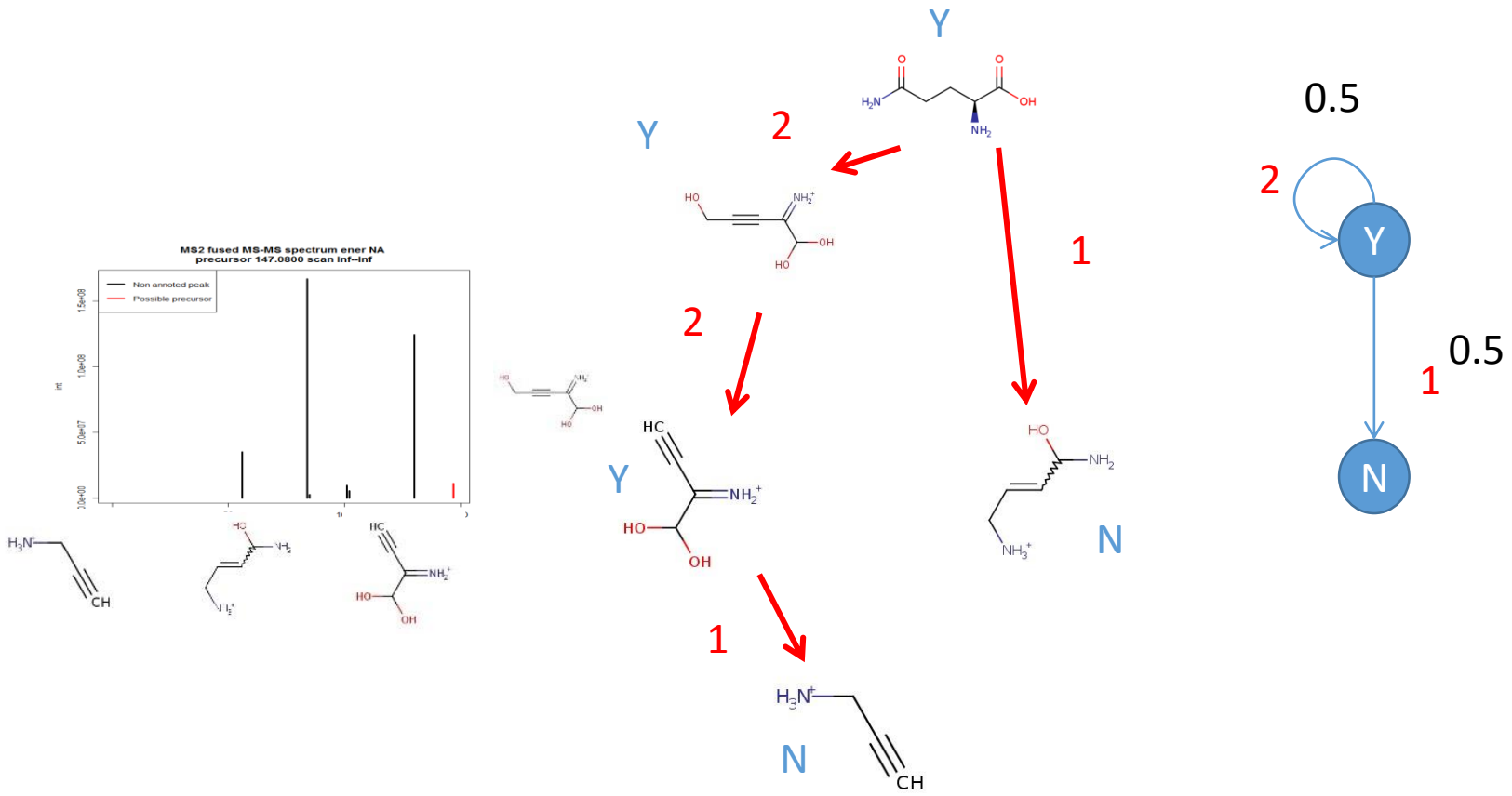


Pour connaitre toutes les transitions il faudrait connaitre tous les fragments, ce qui est trop pour un ordinateur. On doit réduire la dimension du modèle.

En considérant les fragmentations avec hydroxyl ou sans, on passe de 4 types de fragmentation à 2.

CFM-ID : Apprentissage

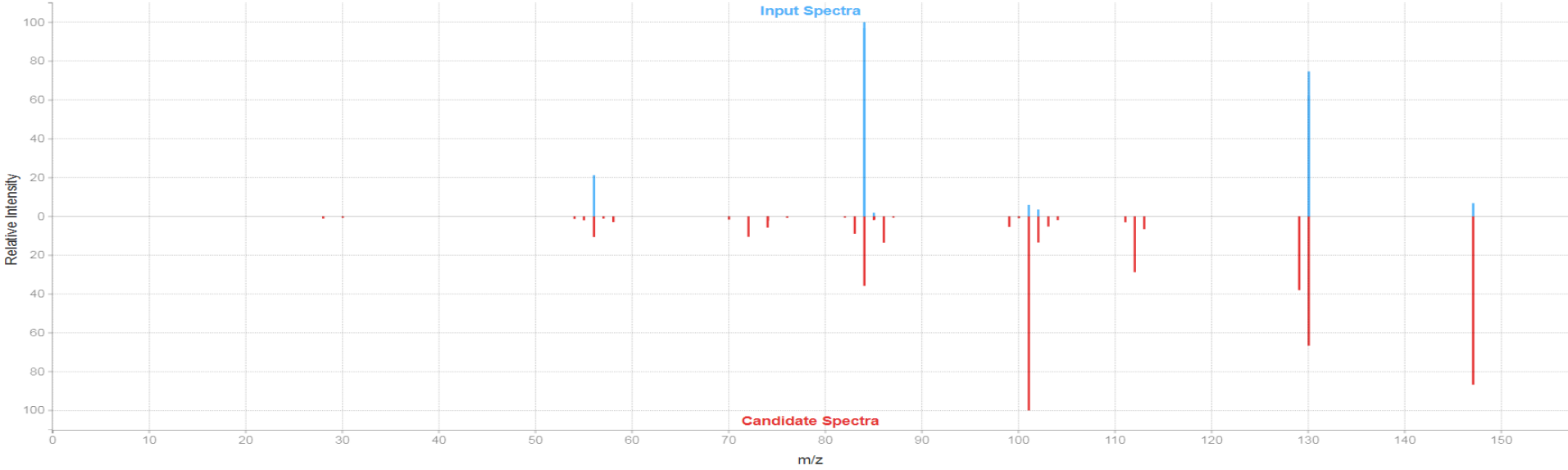
Un modèle probabiliste (chaine de Markov) est associé à ce graphe.



En pratique il y a plus de 2000 descripteurs d'une liaison

CFM-ID : Prédiction

Pour prédire un spectre, toutes les molécules sont cassées et passées dans le modèle.
On obtient un mélange de gaussiennes. Leurs hauteurs sont leur intensités, et leurs positions la masse.



Les spectres sont ensuite scorés en utilisant la similarité gaussienne ou le Jaccard index.

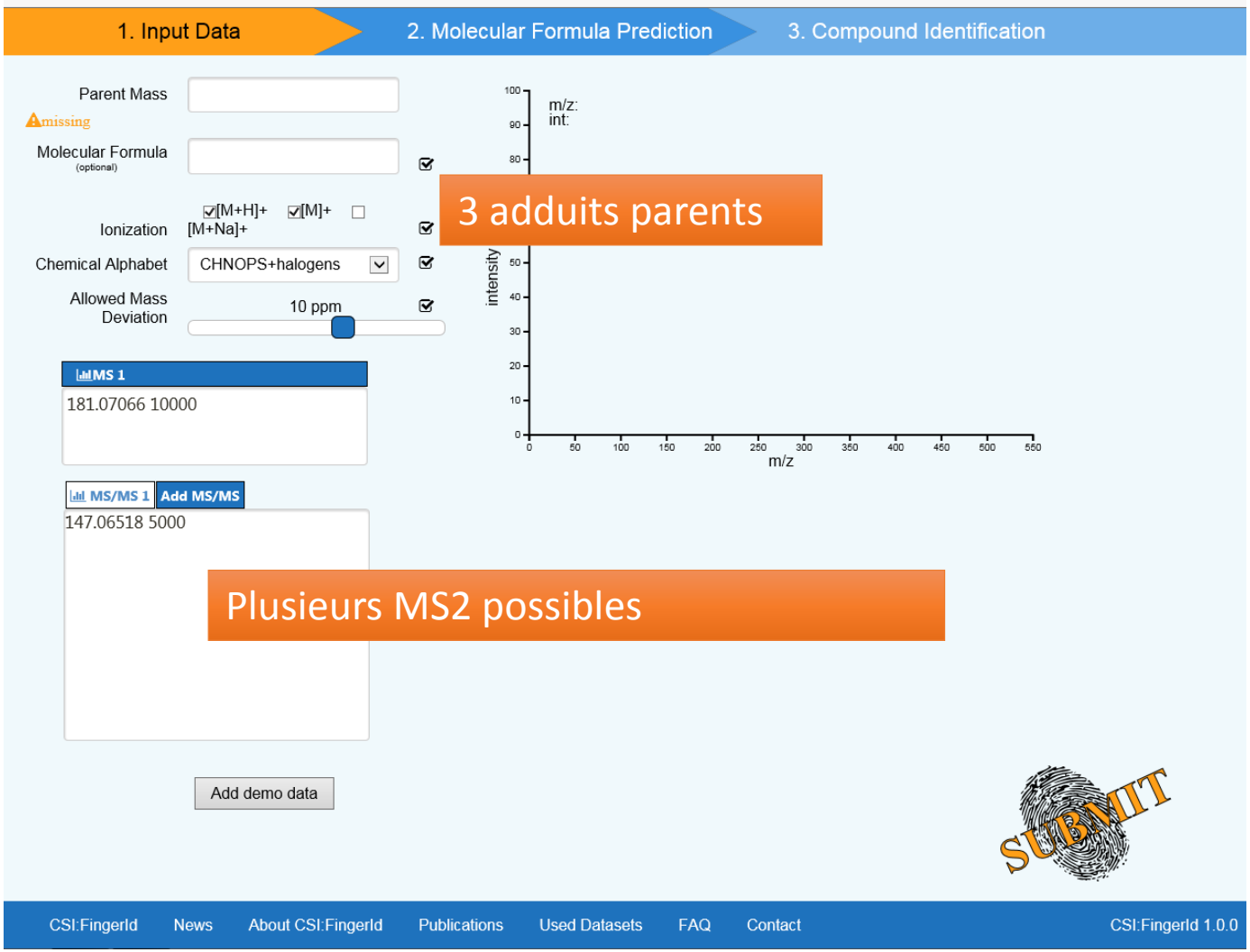
CFM-ID : Résumé

PRO	CONS
• Bonne performance au CASMI, le plus de candidats dans le top 10 (mais pas dans le top 3) • Disponible en ligne et simple à utiliser. • Possibilité de prédire le spectre de chaque candidat • Nombreux usages	• Prédiction plutôt lente • Pas la méthode la plus performante en terme de top-candidat • Pics trop nombreux sur le spectre prédit à cause de la réduction de dimension.

CSI:FingerID

Qui fait quoi?

-CSIFingerID

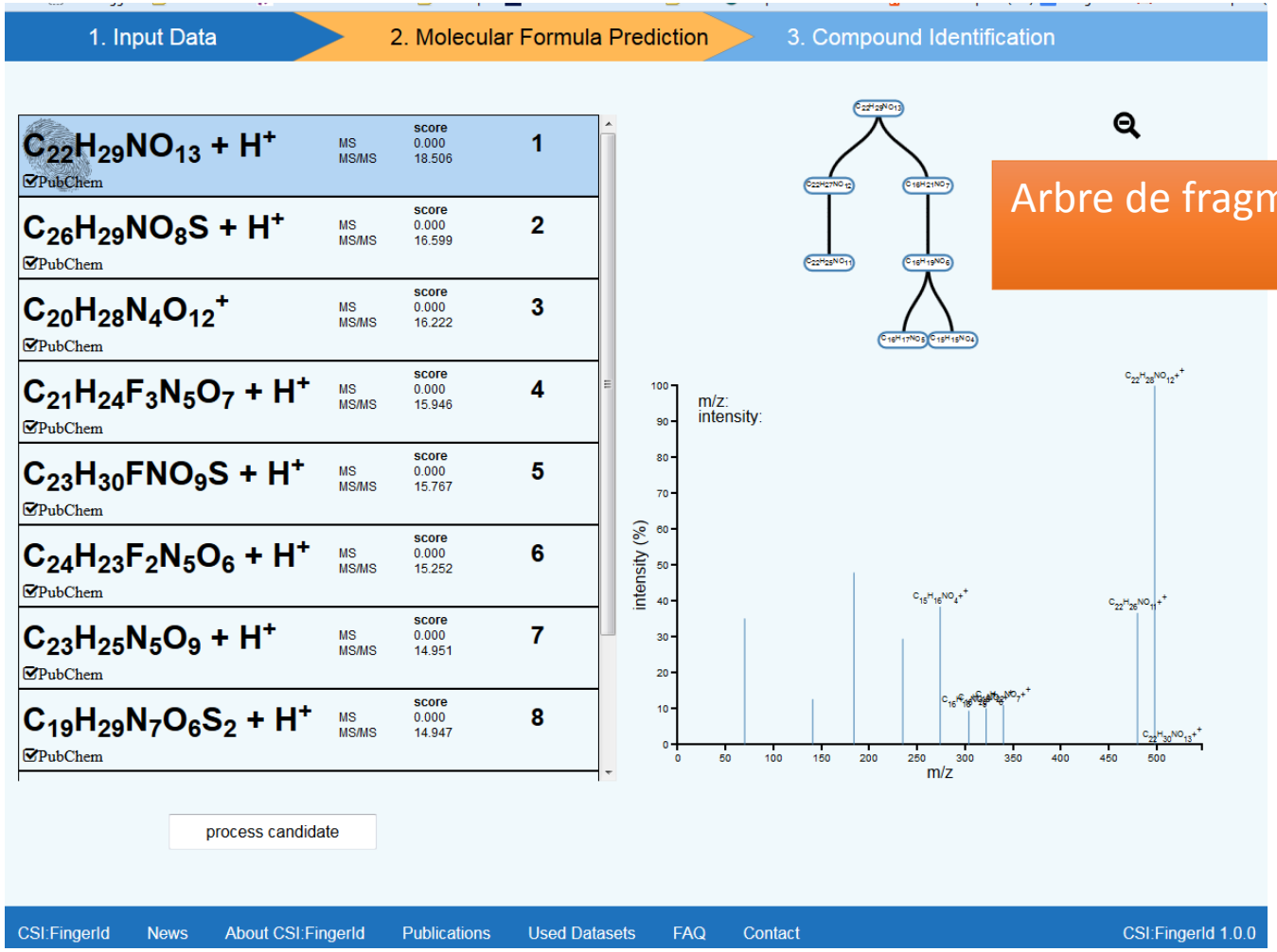


CSI:FingerID

Qui fait quoi?

-CSIFingerID

scores



Arbre de fragmentation

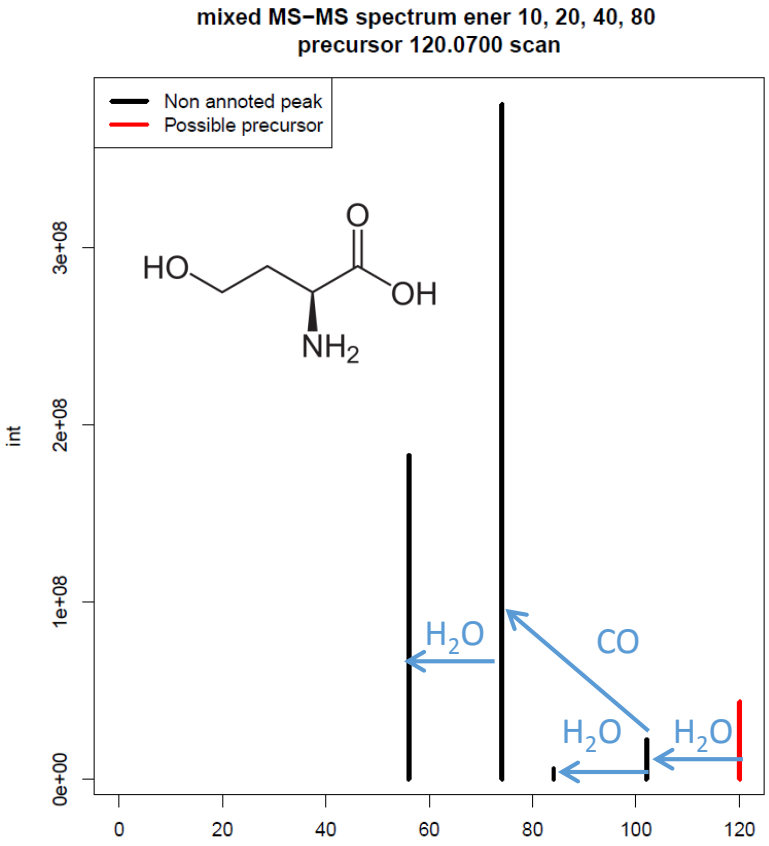
CSI:FingerID : Résumé

- Principe : CSI:FingerID utilise une méthode à noyau (distance) combiné entre des spectres et des arbres de fragmentation.

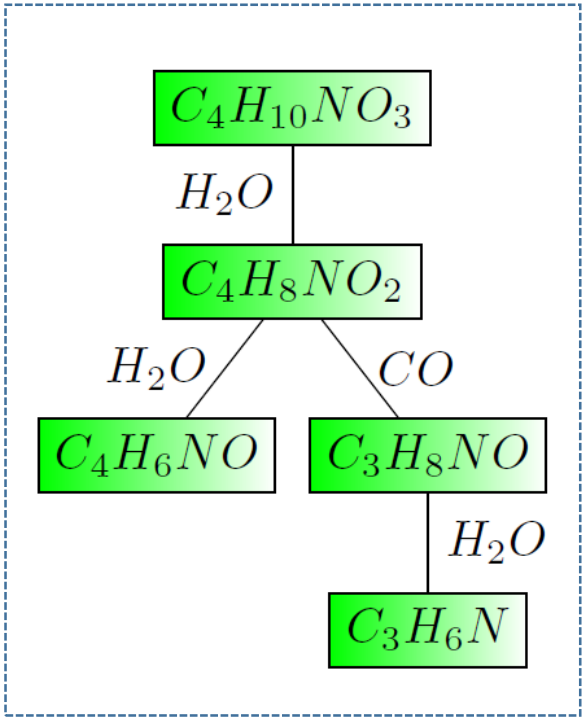
- Le modèle est appris à partir de couples spectre/ structure

CSI-FingerID : Apprentissage

- Pour chaque spectre, un arbre de fragmentation est construit (cf. SIRIUS)



Arbre de fragmentation basé sur des formules.



CSI-FingerID : SIRIUS

Sirius 3.0 : génération de formule brutes en tenant compte du MS et des MS2

MS spectra.txt
156.25 100
157.26 2
158.26 0.2

MS2 spectra.txt
86.25 100
57.26 20
108.26 4

1.) C21H32O2	score: 62.59	tree: +62.59	iso: 11.00	peaks: 17	91.80 %
2.) C19H30N3O	score: 49.39	tree: +49.39	iso: 8.00	peaks: 15	76.73 %
3.) C17H35NO2P score: 48.89	tree: +48.89	iso: 0.00	peaks: 17	91.80 %	
4.) C15H33N4OP score: 39.18	tree: +39.18	iso: 0.00	peaks: 15	76.73 %	
5.) C14H32N6S	score: 9.33	tree: +9.33	iso: 0.00	peaks: 6	4.69 %

Sans le MS

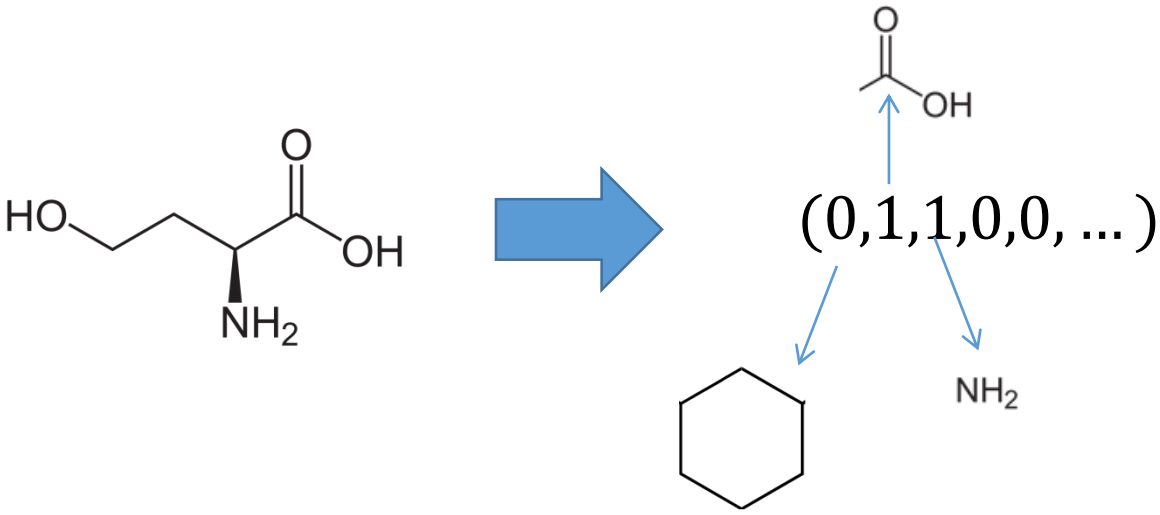
MS2 spectra.txt
86.25 100
57.26 20
108.26 4

1.) C21H32O2	score: 62.59	tree: +62.59	iso: 0.00	peaks: 17	91.80 %
2.) C19H30N3O	score: 49.39	tree: +49.39	iso: 0.00	peaks: 15	76.73 %
3.) C17H35NO2P score: 48.89	tree: +48.89	iso: 0.00	peaks: 17	91.80 %	
4.) C15H33N4OP score: 39.18	tree: +39.18	iso: 0.00	peaks: 15	76.73 %	
5.) C14H32N6S	score: 9.33	tree: +9.33	iso: 0.00	peaks: 6	4.69 %

Pas de MS => pas de score sur le profil isotopique

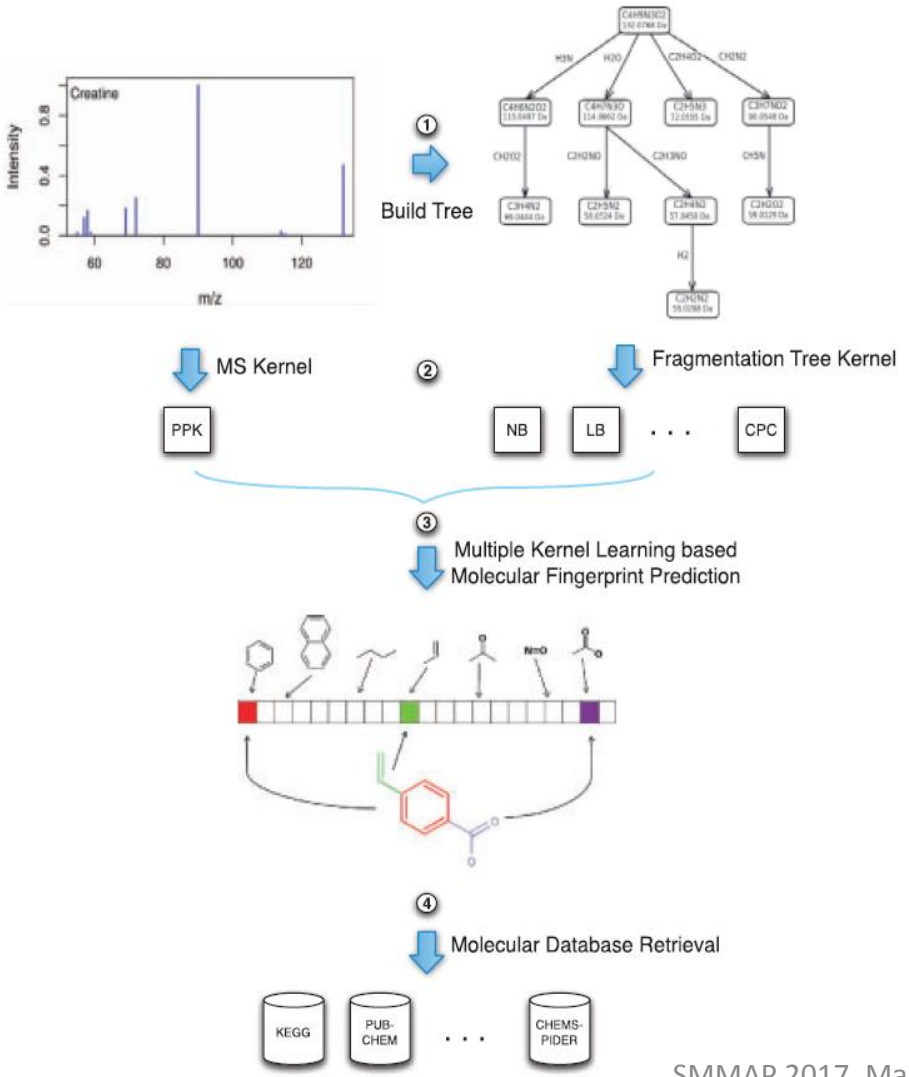
CSI-FingerID : Apprentissage

- Pour chaque molécule un vecteur binaire (fingerprint) indique l'absence ou la présence de certaines sous-structures

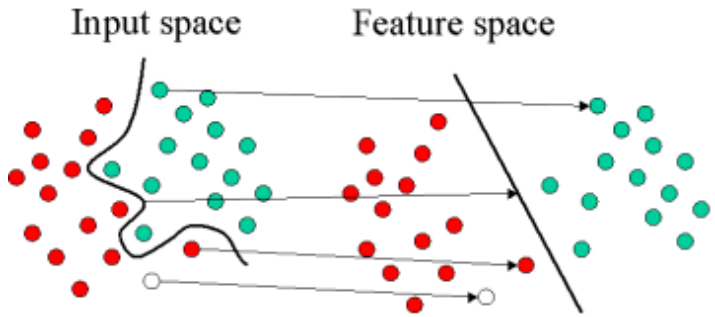


C'est cette représentation qui va être prédite.

CSI-FingerID : Apprentissage



- Un noyau (distance) combiné est calculé entre les spectres/arbres de fragmentation, en combinant des distances plus simple.
- Ce noyau est ensuite utilisé pour prédire la fingerprint en utilisant des SVM.



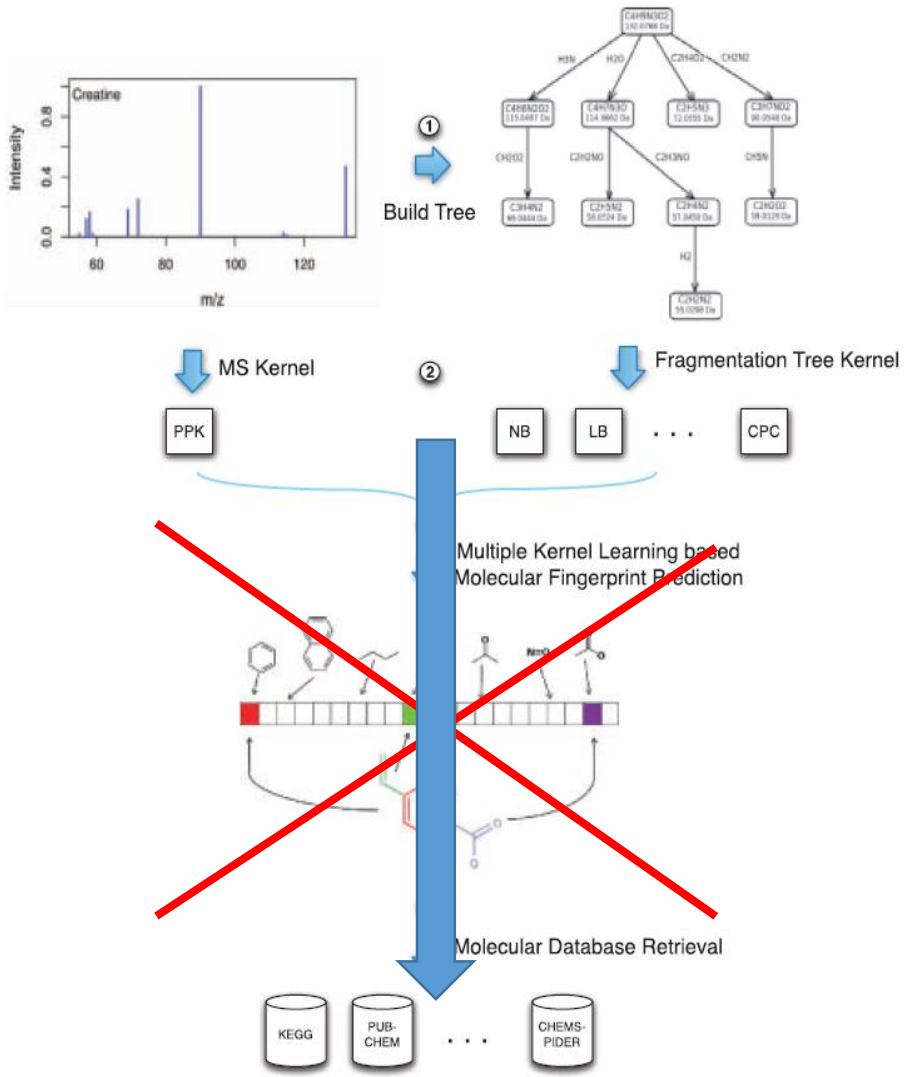
Support Vector Machine

CSI-FingerID : Identification

- Les molécules candidates sont retrouvées et leur fingerprint est calculée.
- La fingerprint correspondant au spectre est calculée en utilisant le spectre fourni en entrée et l'abre calculé.
- On cherche ensuite les molécule qui répondent le mieux à la fingerprint.

CSI-FingerID : Variante

- Variante actuelle (2016)
- Meilleure performance au CASMI



CSI-FingerID : Résumé

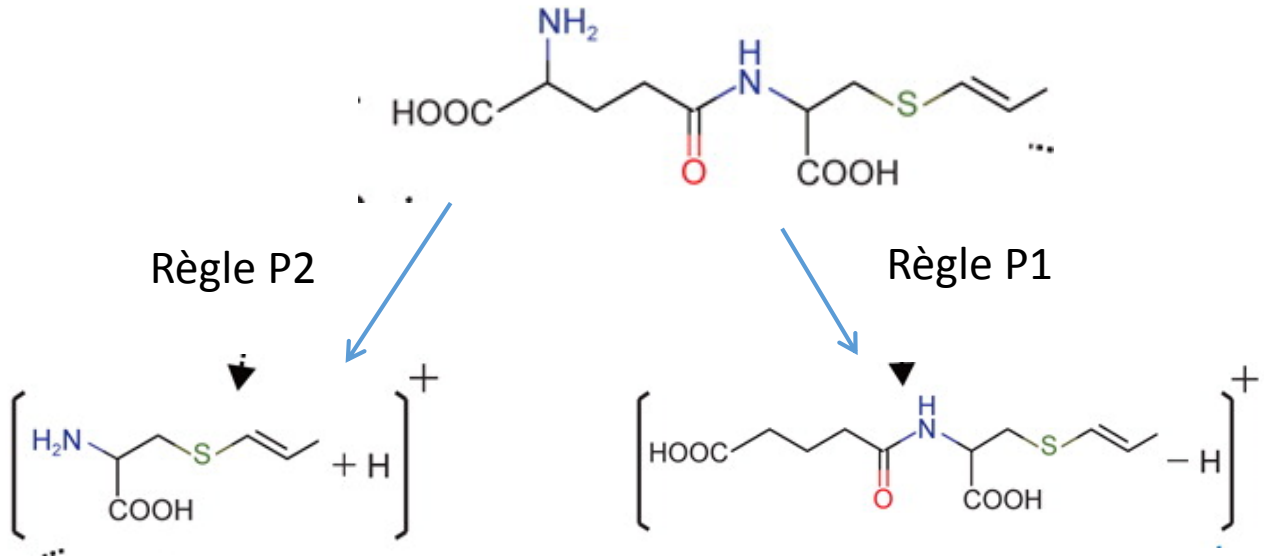
PRO	CONS
• Meilleure performance au CASMI. • Disponible en ligne et simple d'utilisation. • Possibilité de voir les arbres de fragmentation.	• Modèle boite noire non interprétable.

Basé sur la prédiction de l'arbre de fragmentation

MS-Finder : Principe

MS-Finder est basé sur des règles determinées sur des spectres bien curés de MassBank

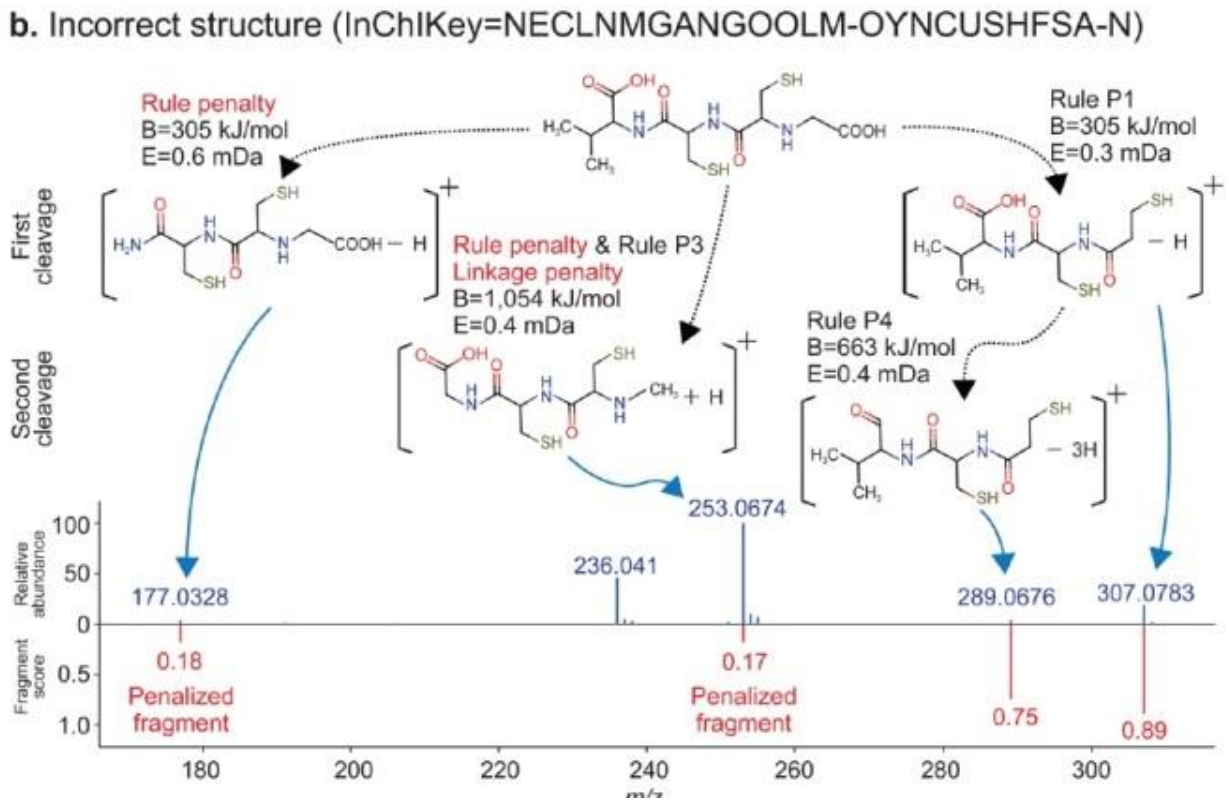
4 règles principales en mode positif
5 en mode négatif



Exemple de règle : « In low-energy CID, a rearrangement of C-bond cleavage adds no hydrogen (rule P1), whereas a rearrangement of N- or O-bond cleavages adds two hydrogens (rule P2) »

MS-Finder : Scoring

- Les candidats sont fragmentés en utilisant les règles.

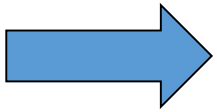


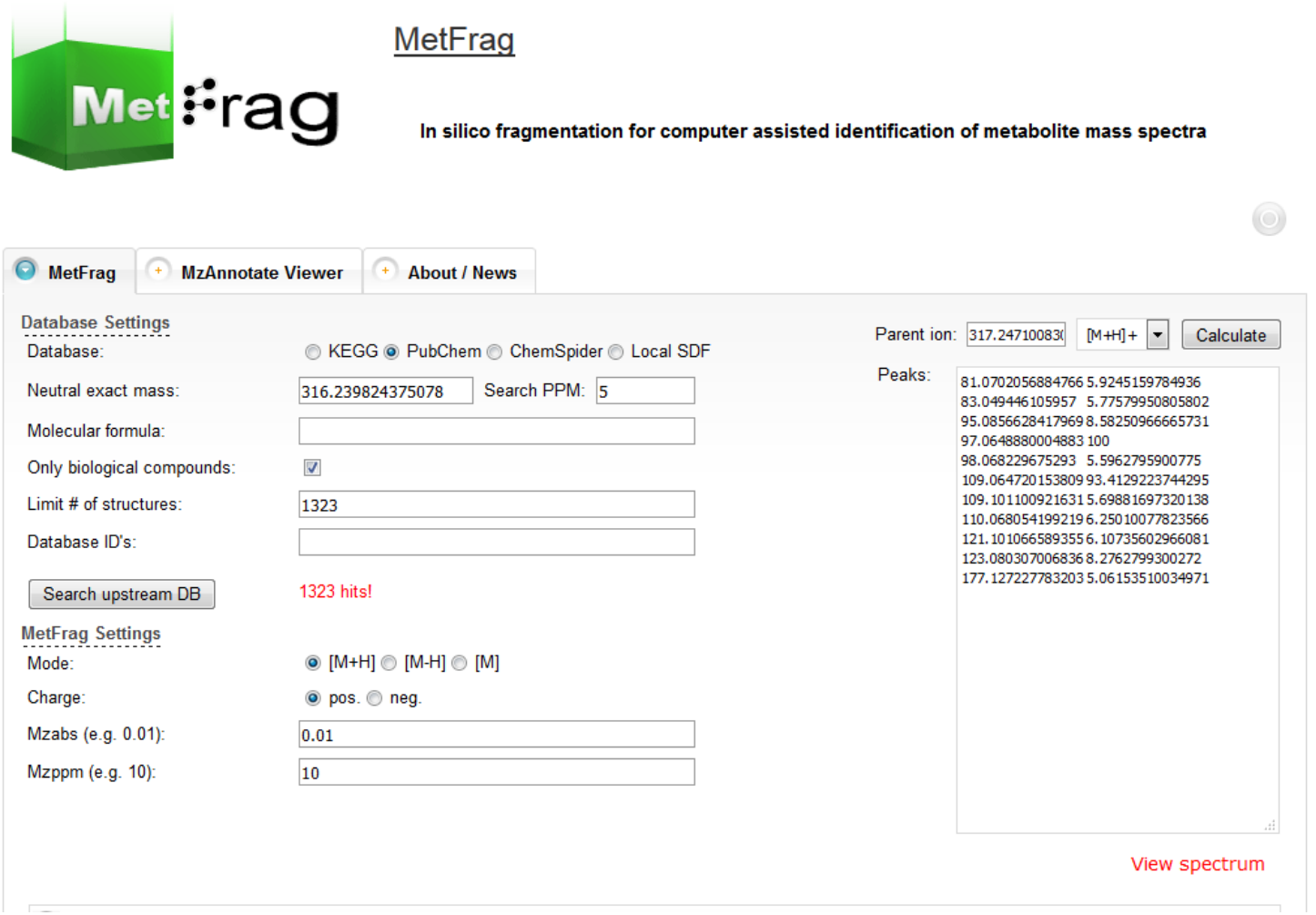
MS-Finder : Résumé

PRO	CONS
• 2nd en terme de médaille d'or au CASMI. • Basé sur des règles simples • Plus facile à interpréter • Grande variété de bases de données interrogées (24)	• Pas de modèle statistique • Non disponible en ligne

MetFrag : Principe

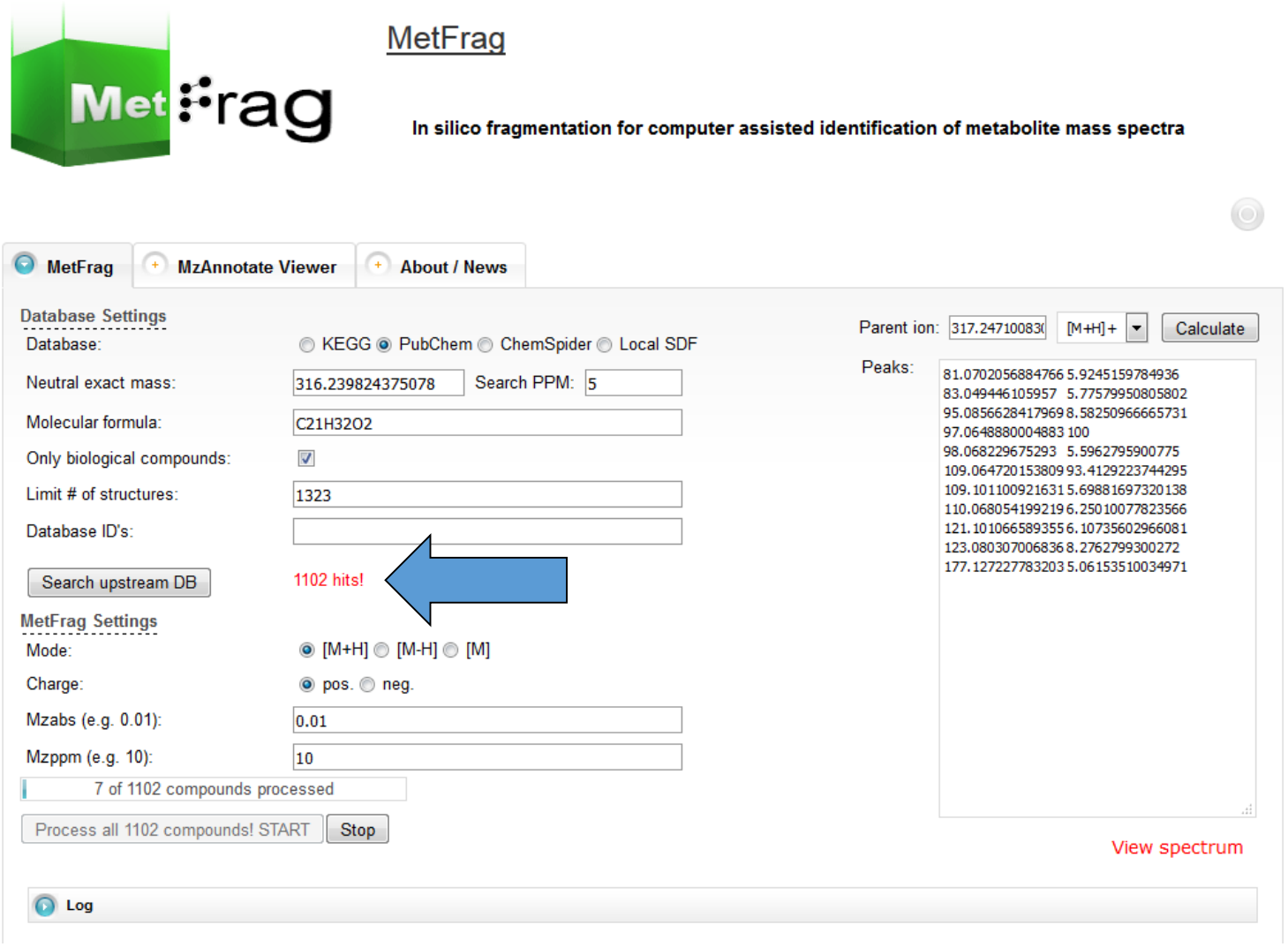
MetFrag - MetFusion





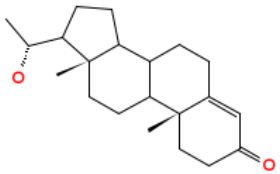
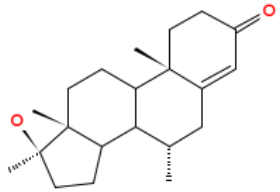
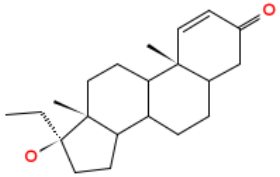
MetFrag : Principe

MetFrag - MetFusion



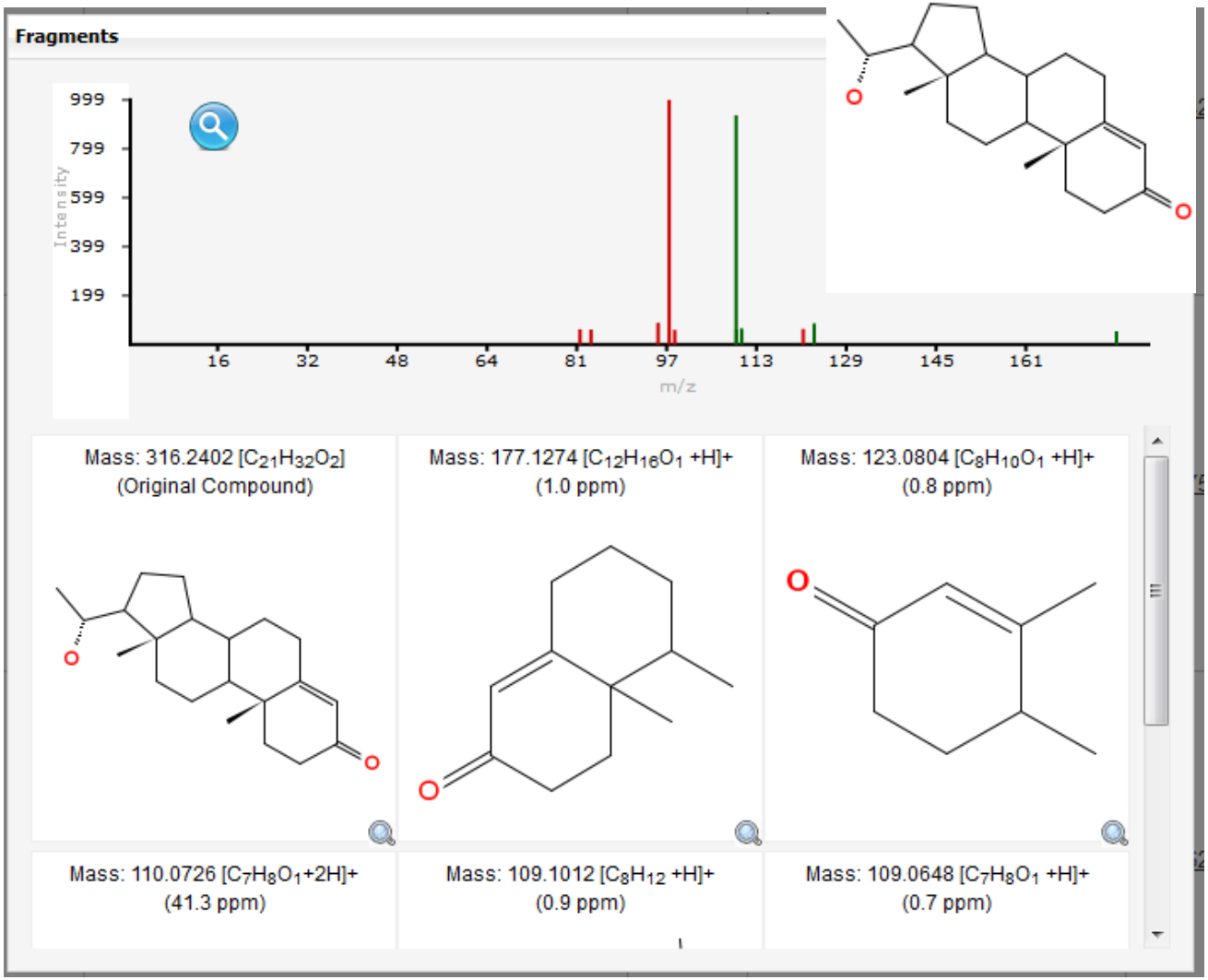
MetFrag : Principe

MetFrag - MetFusion

Score	# Explained Peaks	Trivial Name	Exact Mass	Structure	Database ID	Actions
1.0	5	• 20alpha-Hydroxy-4-pregnen-3-one • 20alpha-Hydroxypregn-4-en-3-one • 20alpha-Hydroxyprogesterone • (S)-20-Hydroxypregn-4-en-3-one	$C_{21}H_{32}O_2$ 316.2402		C04042	Fragments Download
0.965	7	• Bolasterone • 7alpha,17alpha-Dimethyltestosterone • 17beta-Hydroxy-7alpha,17-dimethylandrost-4-en-3-one • U19763	$C_{21}H_{32}O_2$ 316.2402		C14475	Fragments Download
0.965	7	• 17-Hydroxy-5alpha,17alpha-pregn-1-en-3-one	$C_{21}H_{32}O_2$ 316.2402		C14962	Fragments Download

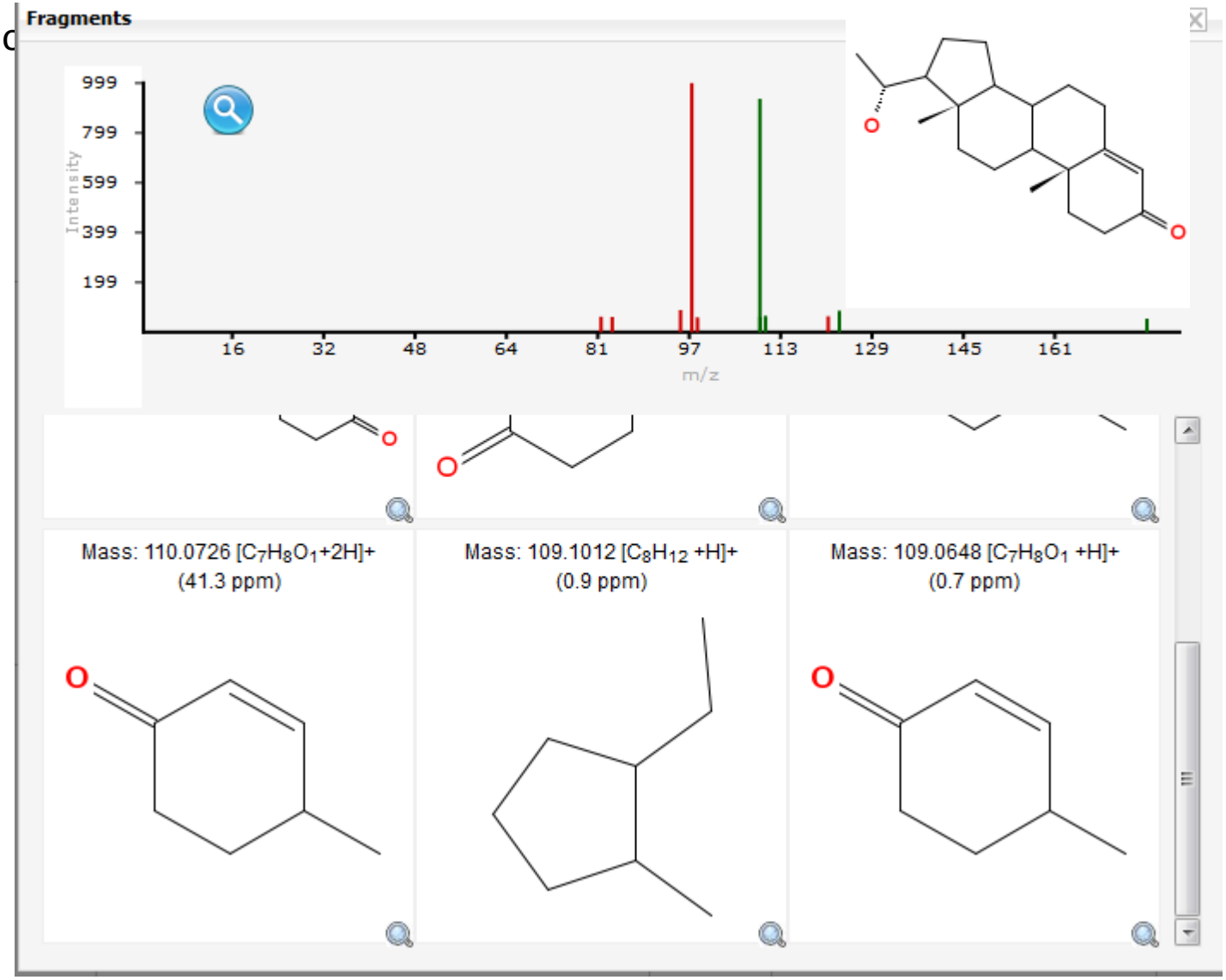
MetFrag : Principe

MetFrag - MetFusion



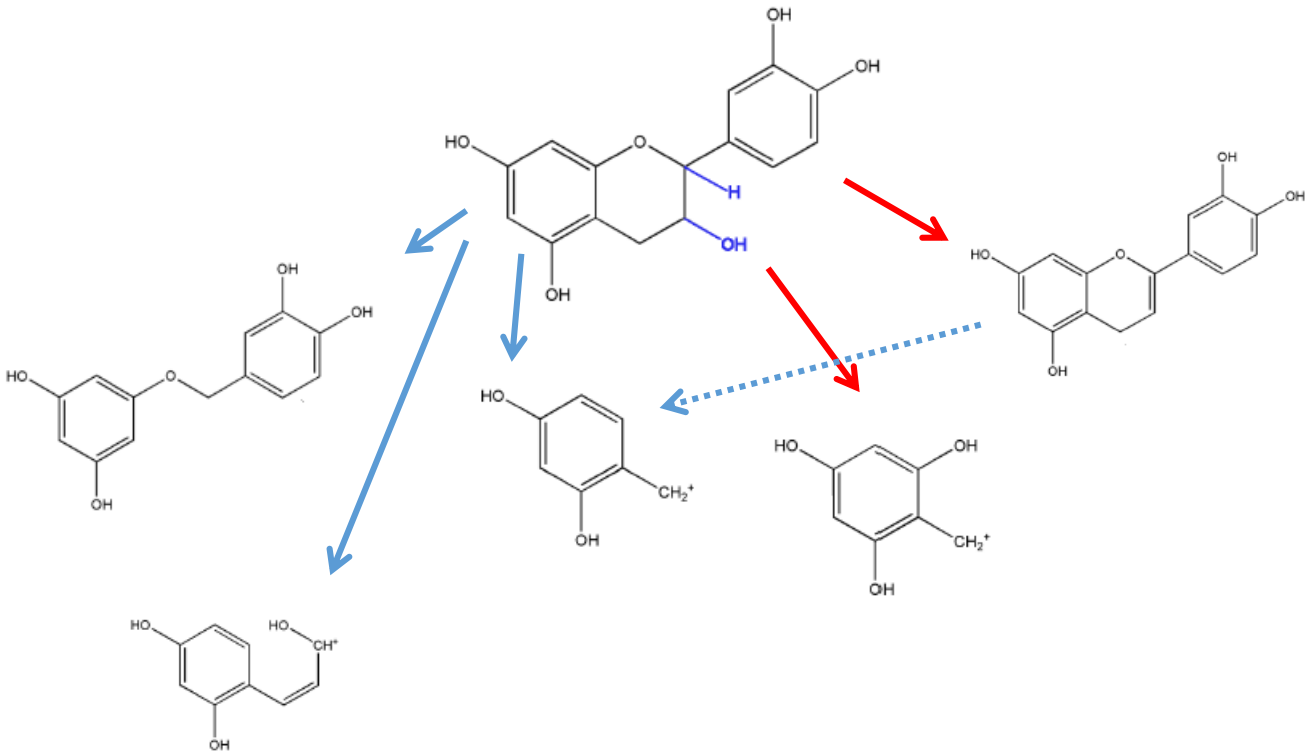
MetFrag : Principe

MetFrag - MetFusic



MetFrag : Principe

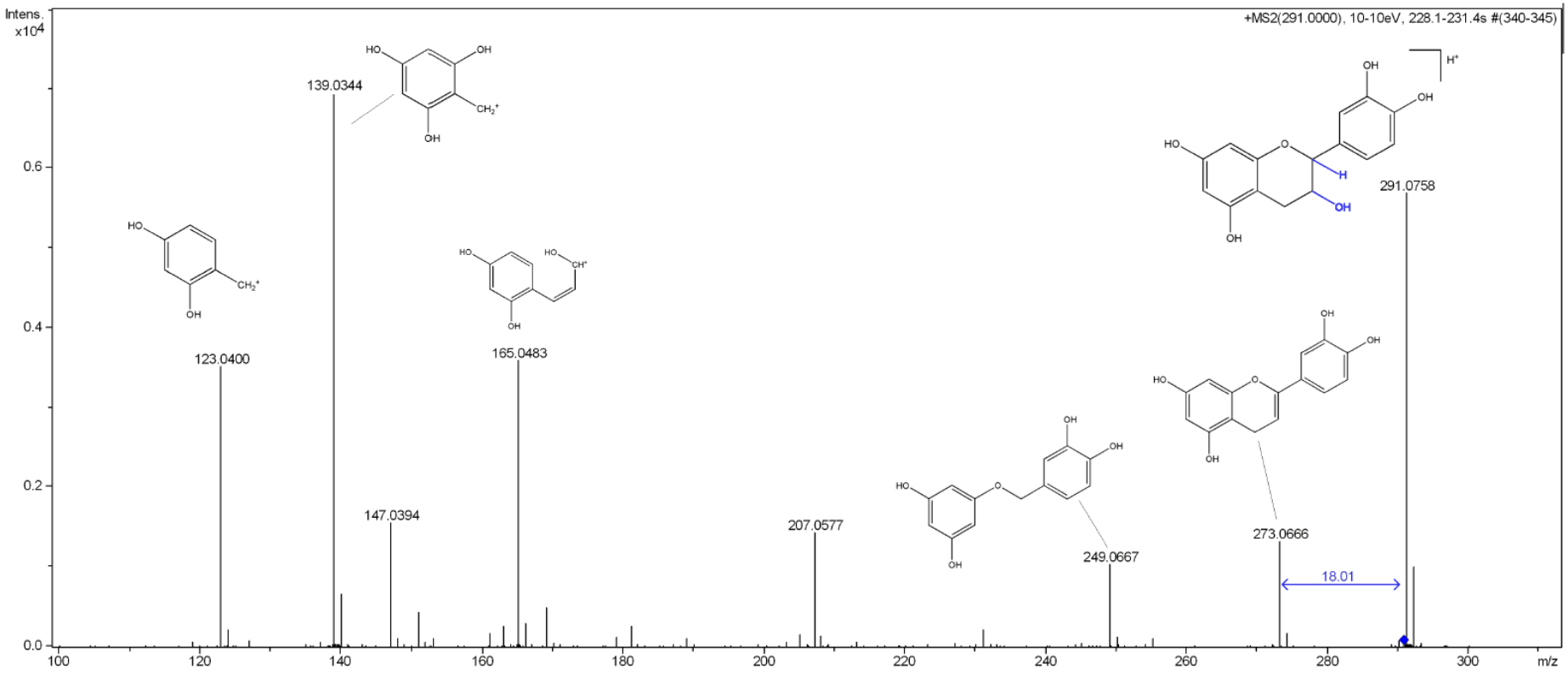
Metfrag est basé sur la fragmentation **exhaustive** *in silico* de molécules. Quelques réarrangements simples sont également pris en compte.



PAS DE REGLES sur la fragmentation, si on peut arriver sur un fragment de plusieurs façons, seul le plus court chemin est pris en compte

MetFrag : Principe

Les pics qui peuvent être expliqués par des fragments sont utilisés pour scorer les molécules en utilisant l'énergie de dissociation des liaisons.



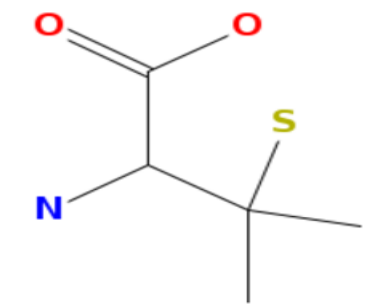
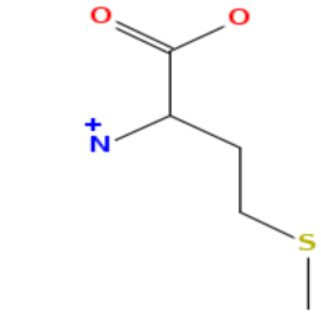
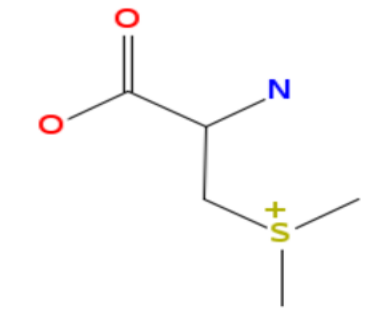
Wolf et al., 2010

MetFrag : Scoring

Les molécules sont scorées en utilisant plusieurs propriétés :

- Pour chaque fragment le BDE (Bond Dissociation Energy) est calculé pour chaque liaison cassée.
- Un terme dérivé de la distance cosine est utilisé pour les intensités.
- Optional : Retention time term, substructure exclusion… in command line version only.

Exemple de résultats de MetFrag pour un spectre de Methionine.

1.0	7	$C_5H_{11}N_1O_2S_1$ 149.051		20593829
0.997	5	$C_5H_{12}N_1O_2S_1$ 150.0583		5249997
0.946	6	$C_5H_{12}N_1O_2S_1$ 150.0583		18347837

MetFrag : Résumé

PRO	CONS
• Disponible en ligne et facile à utiliser • Les fragmentations sont faciles à interpréter • Rapide • Pas de modèle requis	• Pas de bonnes performances au CASMI. • Les fragmentations calculées ne reflètent pas forcément les phénomènes physiques.

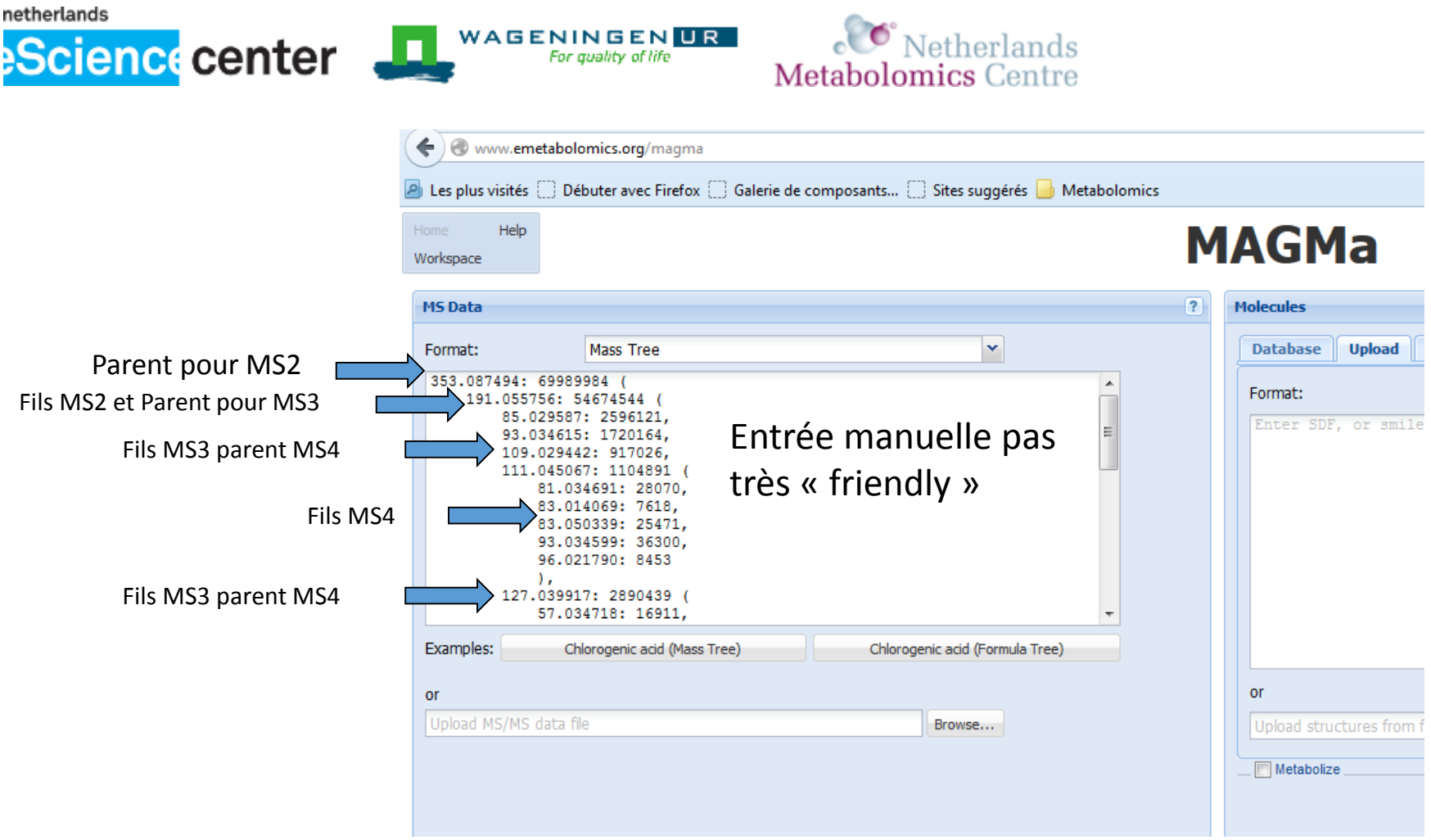
Récapitulatif des outils de prédiction

Nom	CASMI 2017	Interface graphique	Online	Offline	Physique/ Statistique	Interprétabilité	DBs
CFM-ID	3rd	×	✓	✓	Statistique	Moyenne	2
CSI	1st	✓	✓	✓	Statistique	Moyenne	9
MS-Finder	2nd	✓	✓	✓	Physique	Bonne	15
MetFrag	5th	×	✓	✓	Physique	Faible	8

D'autre outils existent : MaGMA, iMet, etc.

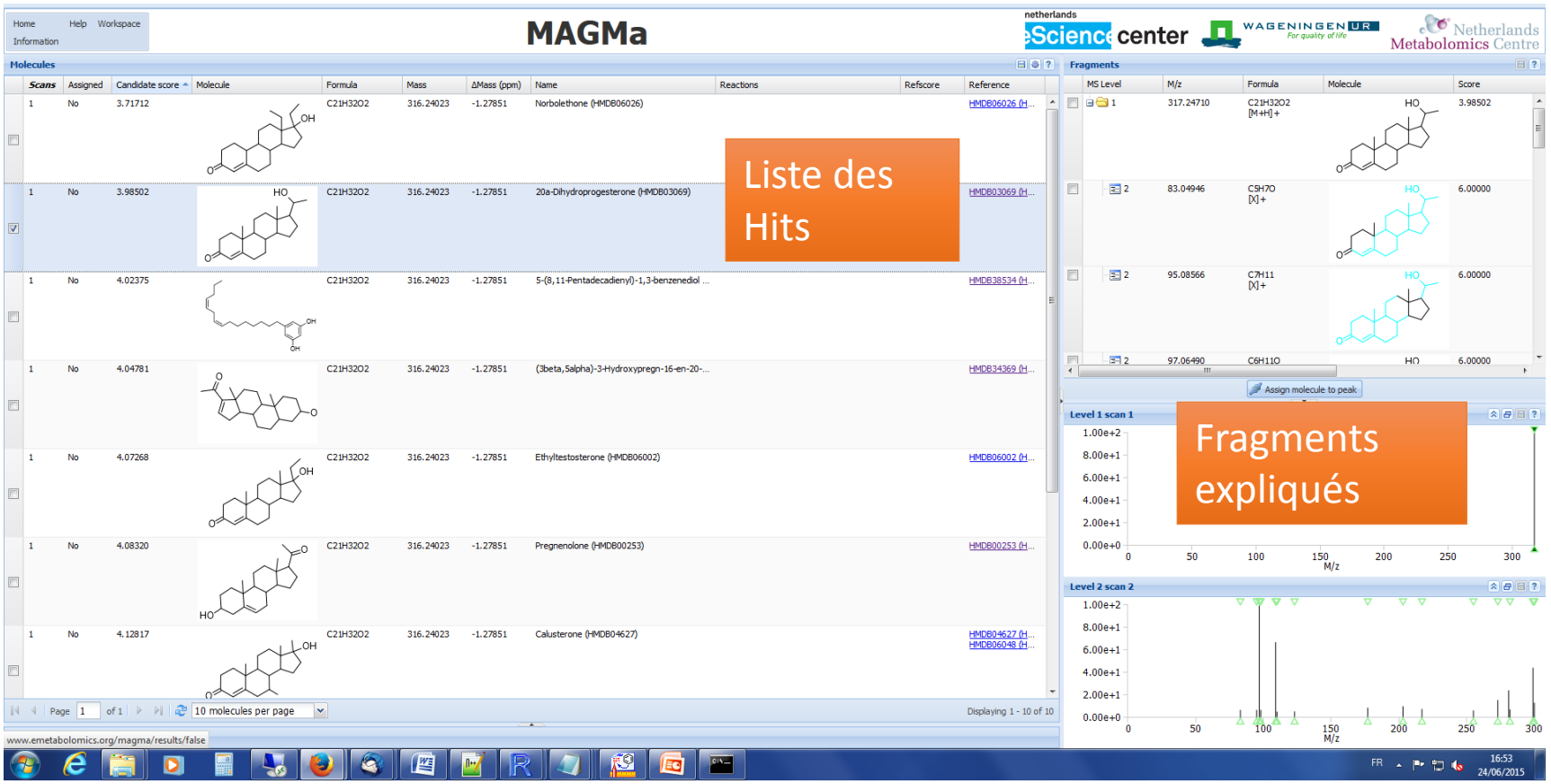
Qui fait quoi?

-Magma: annotation (identification putative) sur la base d'un MSn tree (Kegg, PubChem, HMDB)



Qui fait quoi?

-Magma: annotation (identification putative) sur la base d'un MSn tree (Kegg, PubChem, HMDB)



II. Exploration structurale d'une collection de spectres MS2

IIA) Réseaux moléculaires

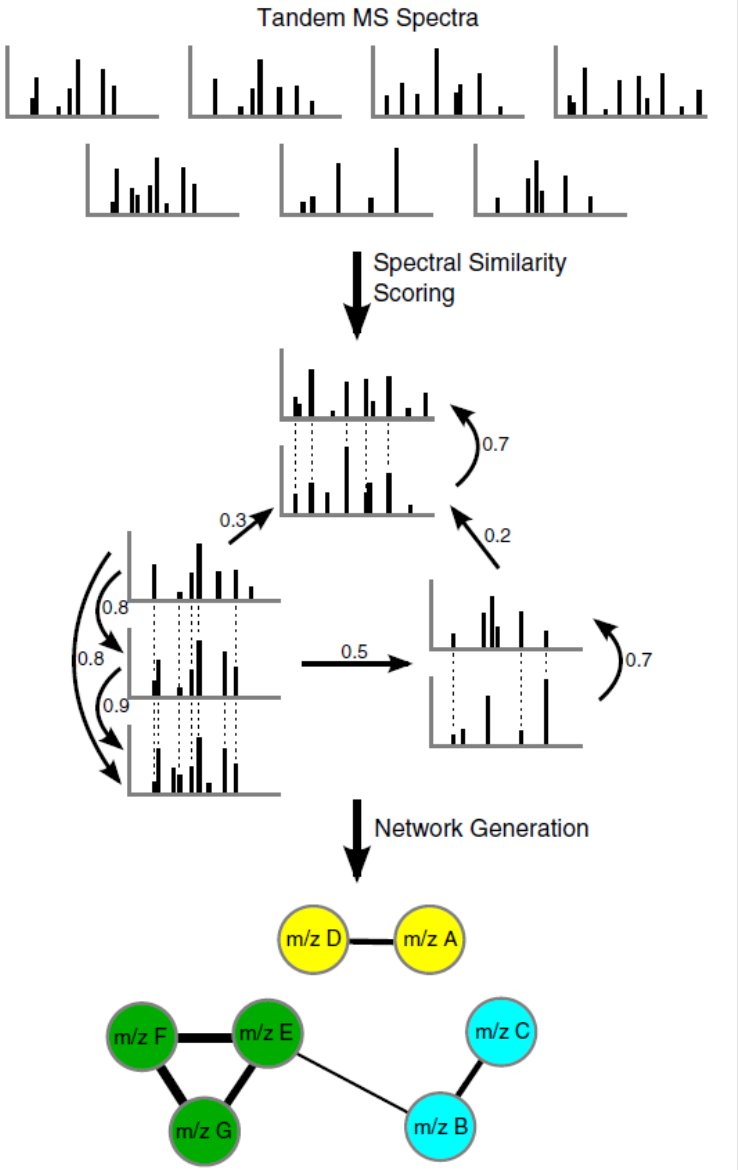


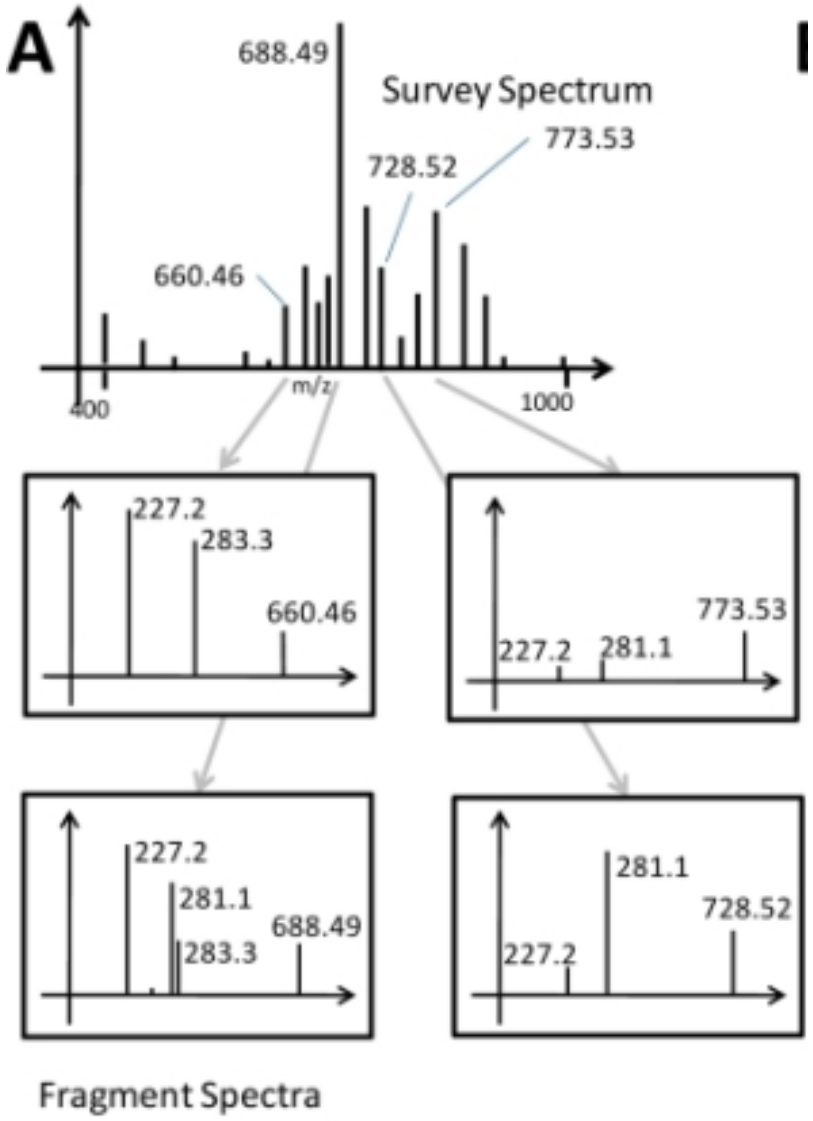
Figure 8 Using spectral alignment of tandem MS data to generate a molecular network. The thickness of the edges indicates the similarity between the spectra. Figure redrawn from Watrous *et al* [230].

Ce n'est, ni plus ni moins, qu'une manière de classer des spectres de masse tandem par homologie, donc assez similaire des arbres de fragmentation sans décrire réellement les voies de fragmentation!

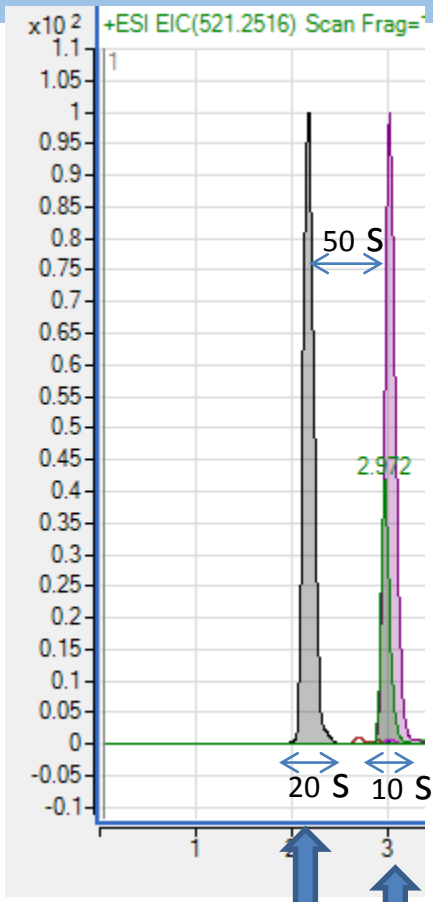
Hypothèse:
Deux molécules possédant des structures proches vont conduire à des spectres de masse similaires.

Mode d’acquisition des données

Data-Dependant Acquisition (DDA)



DDA



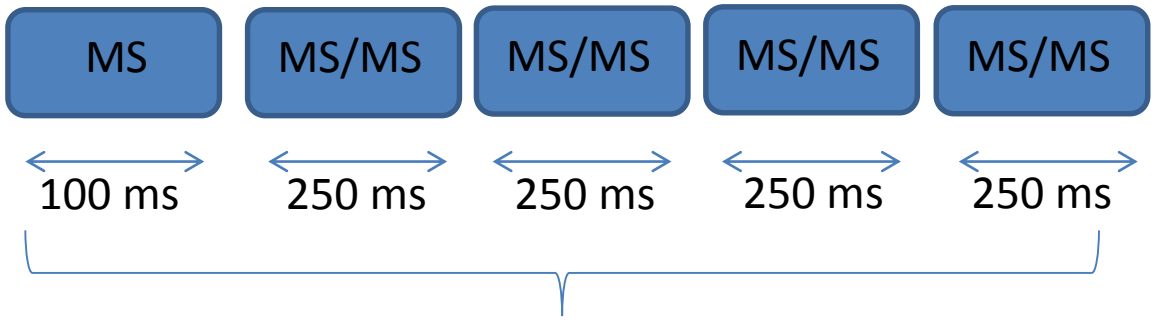
until 16 spectra acquisitions in MSMS

until 8 spectra acquisitions in MSMS

Our condition in DDA mode:

MS 10 spectres/s

MS/MS a maximum of 4 precursor ions per cycle



Total operating time: 1.2 s between 2 scans

Origine des réseaux moléculaires

Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle

Katty X. Wan, Ilan Vidavsky, and Michael L. Gross
Department of Chemistry, Washington University, St. Louis, Missouri, USA

J Am Soc Mass Spectrom 2002, 13, 85–88

$$\cos\theta = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2 \sum_i b_i^2}}$$

Alternatively, a simpler approach using a spreadsheet can be utilized in cases where the number of comparisons to make is small. The approach should be useful for high-throughput applications such as identification of metabolites and characterization of combinatorial libraries.

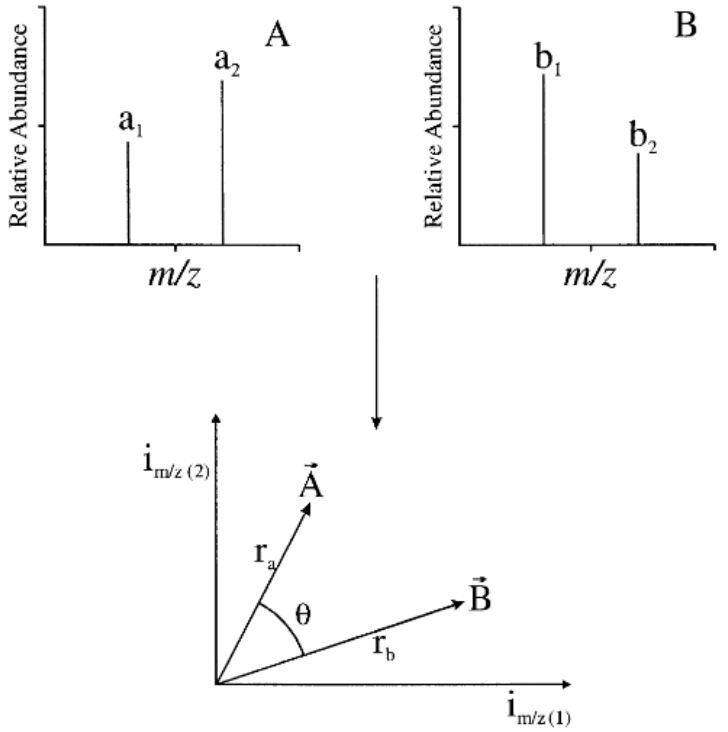


Figure 1. Schematic vector representations of the Spectra A and B.

Initialement …
pour l'analyse de séquences peptidiques

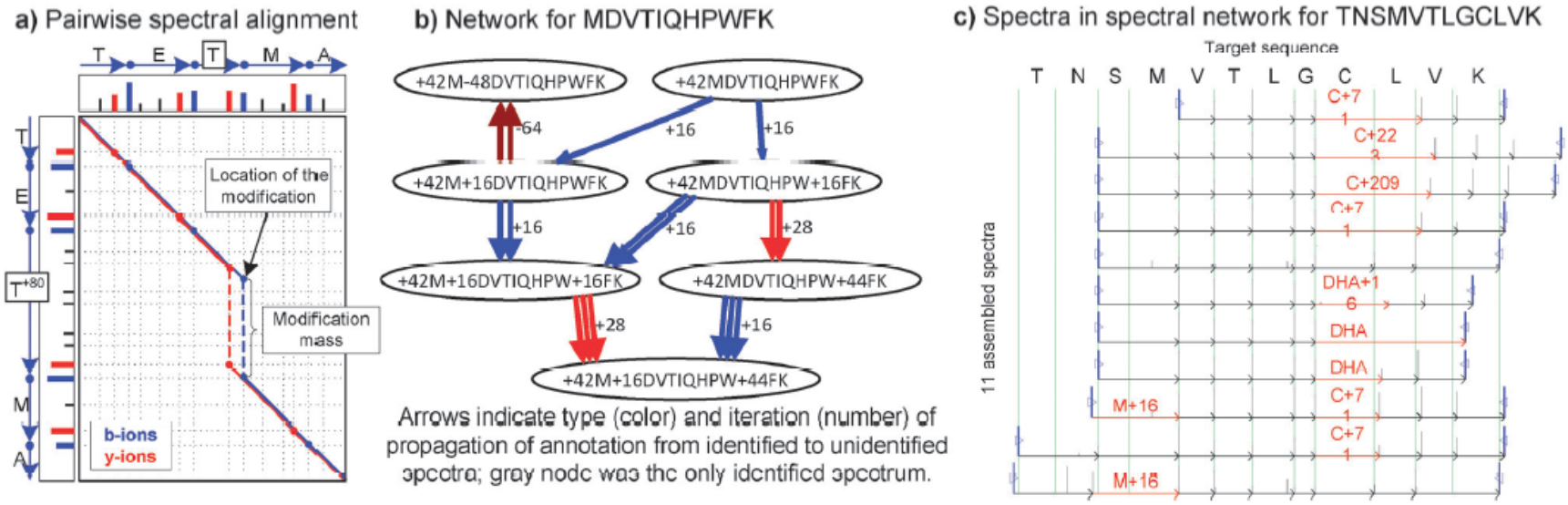
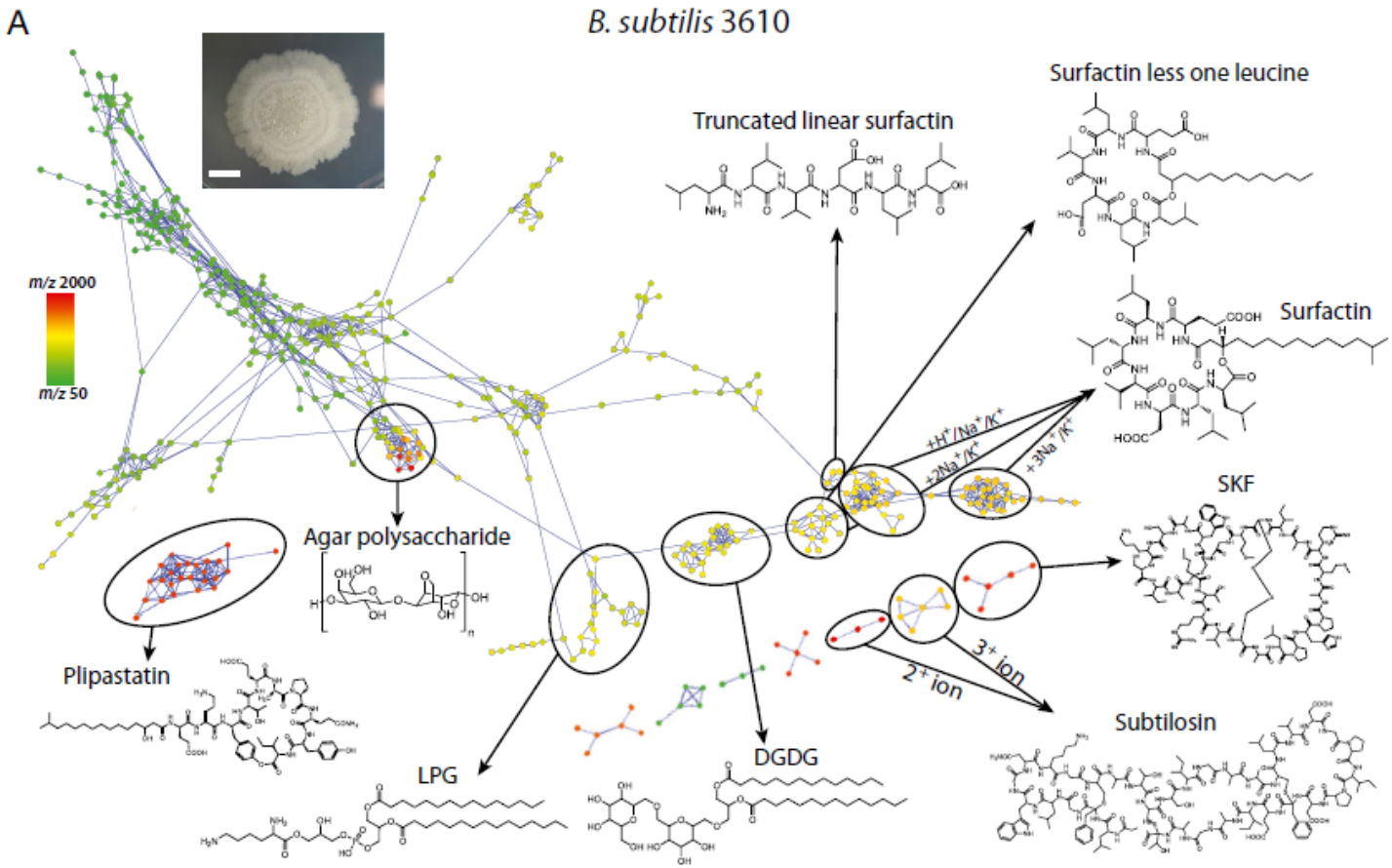


Fig. 1 Discovery and identification of post-translational modifications through spectral networks; (a) Spectral alignment between modified and unmodified variants of the peptide TETMA (*b*-ions shown in blue, *y*-ions in red, blue/red lines track consecutively matched *b*/*y*-ions); (b) Grouped modification states of the peptide MDVTIQHPWFK from a sample of cataractous lenses. Nodes in the spectral network represent individual MS^2 spectra and edges between nodes represent significant spectral alignments such as that shown in part (a); (c) Spectra assembled in the spectral network for TNSMVTLGCLVK with diverse Cysteine modifications on a monoclonal antibody. Each arrow corresponds to the propagation of a sequence and/or PTM from an identified spectrum to an unidentified spectrum (repeated arrows are iterative propagations). Arrow colors correspond to types of modifications transferred.

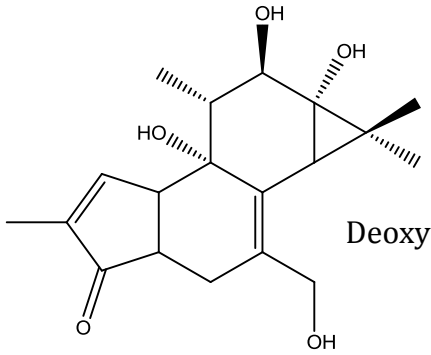
Mol. BioSyst., 2012, **8**, 2535–2544

Initialement …
pour l'analyse de séquences peptidiques



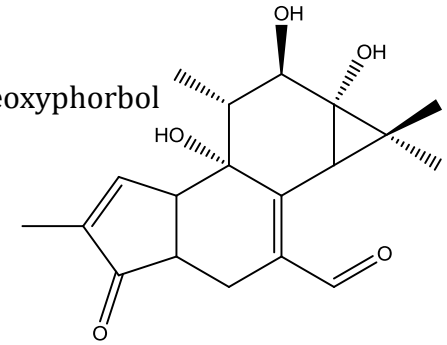
www.pnas.org/cgi/doi/10.1073/pnas.1203689109

Prenols: diterpinoids

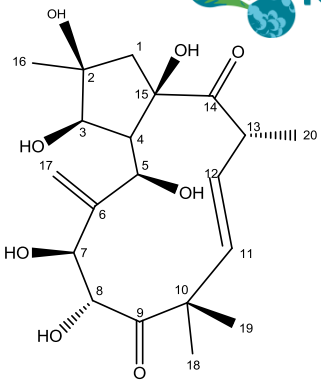


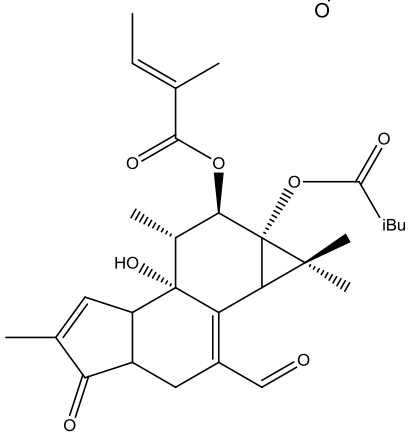
Deoxyphorbol

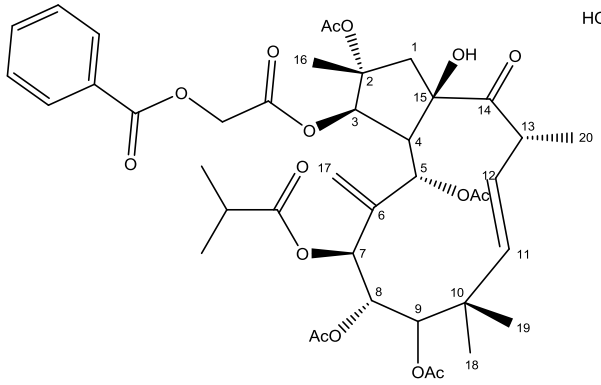
Dideoxyphorbol

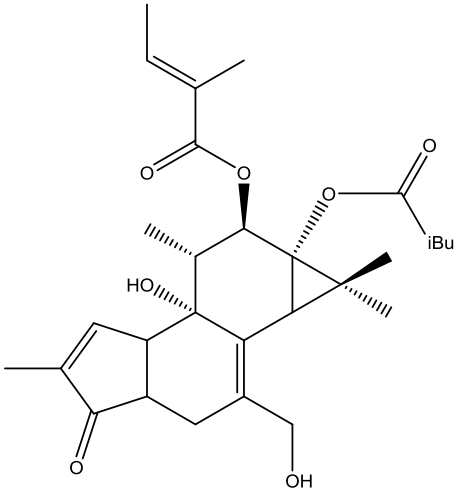


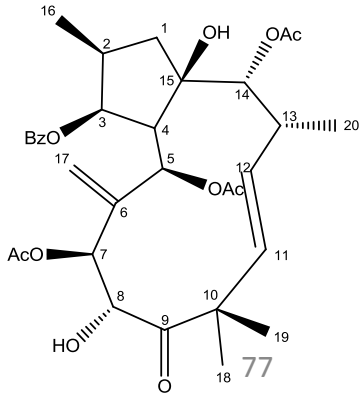
Jatrophane









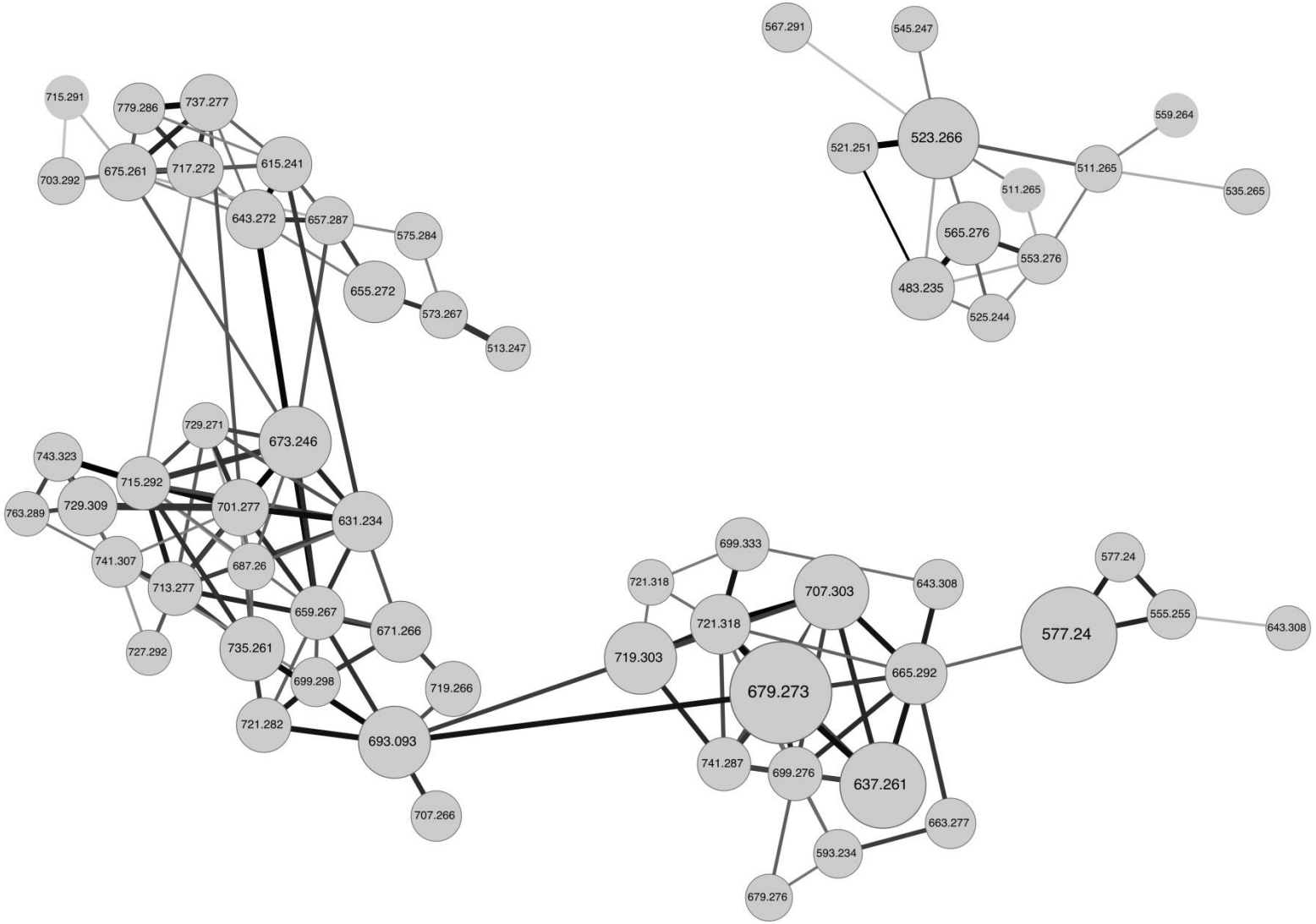


Analyses by LC with strong percentage of organics solvent
30 mL acétonitrile by one LC run

Collaboration with ICSN and University of Corte

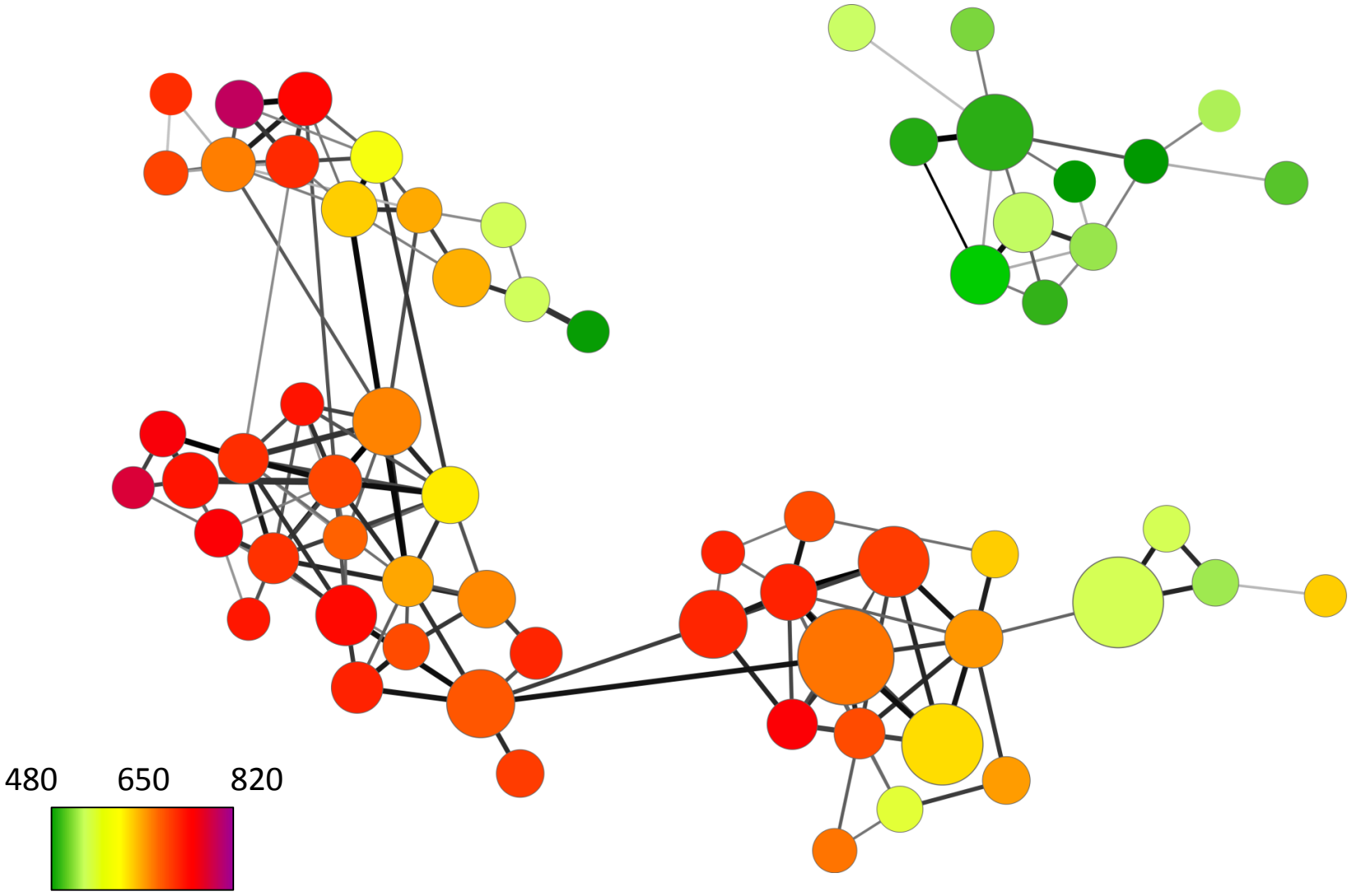
Etude de l'activité anti-CHIKV d'*Euphorbia* de Corse par réseaux moléculaires MS/MS

Représentation des réseaux moléculaires (analyses SFC-qTOF) des fractions anti-CHIKV

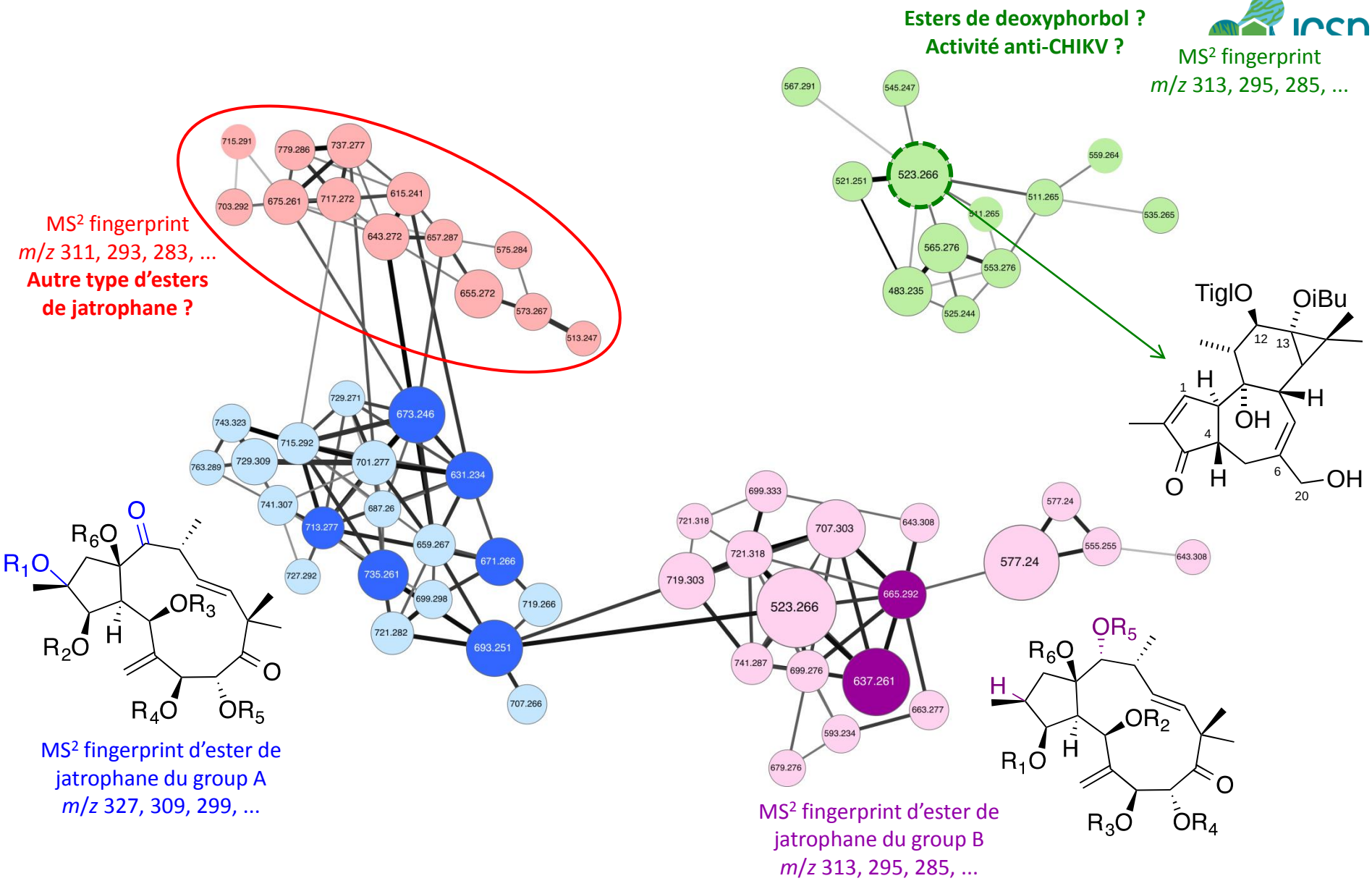


Etude de l'activité anti-CHIKV d'*Euphorbia* de Corse par réseaux moléculaires MS/MS

Représentation avec calque *m/z*

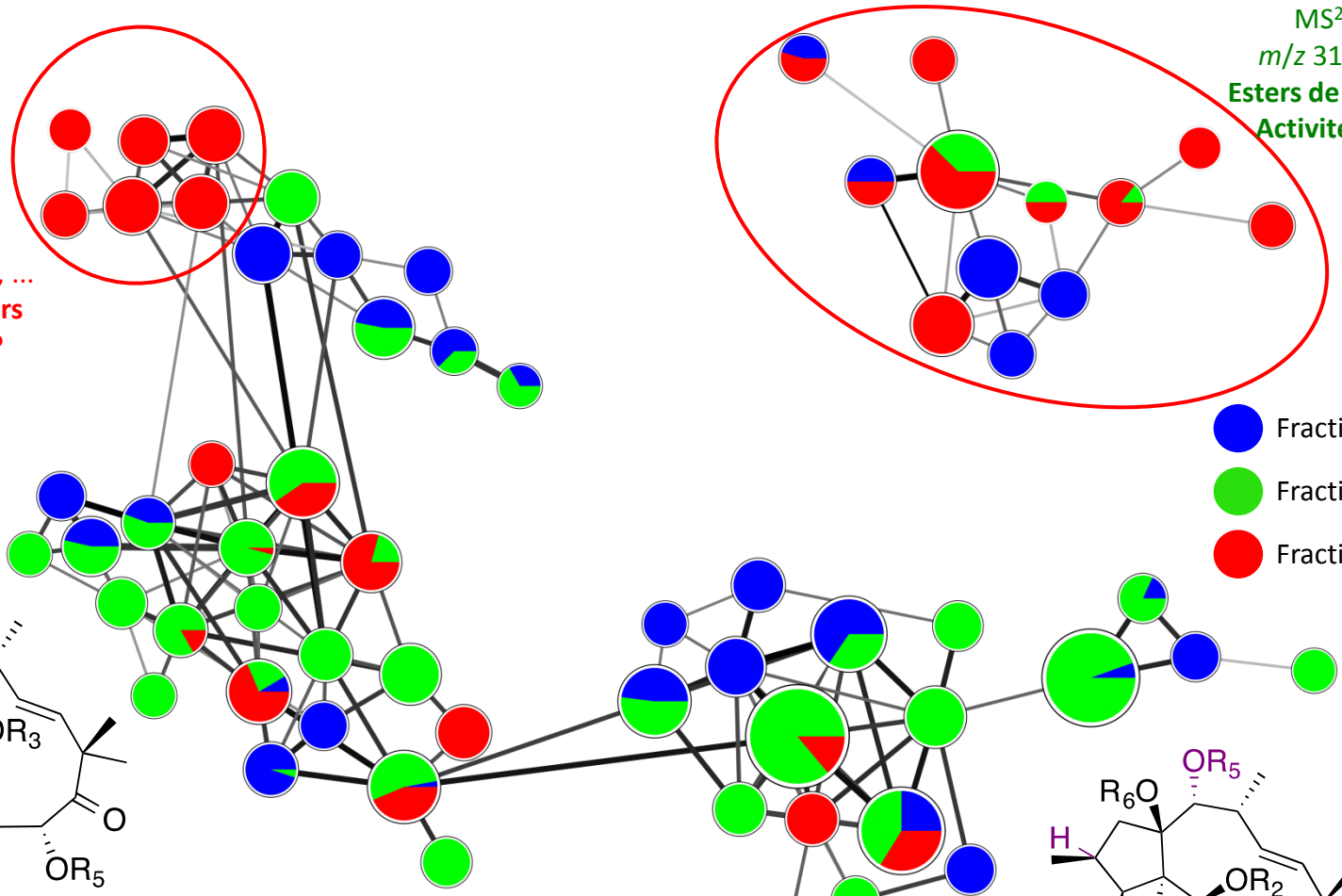


Molecular network



Etude de l'activité anti-CHIKV d'*Euphorbia* de Corse par réseaux moléculaires MS/MS

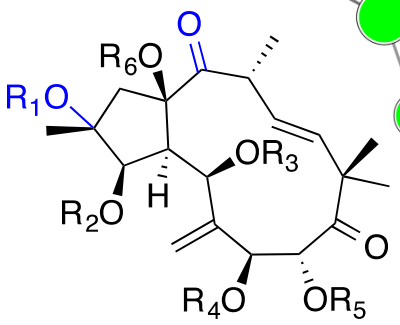
Représentation de la répartition par fraction



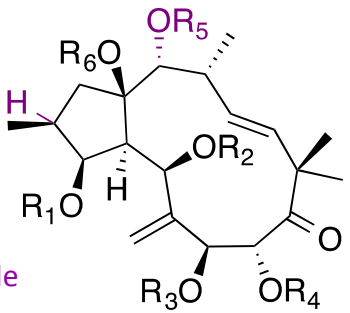
MS² fingerprint
m/z 313, 295, 285, ...
Esters de deoxyphorbol ?
Activité anti-CHIKV ?

MS² fingerprint
m/z 311, 293, 283, ...
Autre type d'esters de jatrophane ?

Fraction G ++
Fraction H ++
Fraction I +++



MS² fingerprint d'ester de jatrophane du group A
m/z 327, 309, 299, ...



MS² fingerprint d'ester de jatrophane du group B
m/z 313, 295, 285, ...

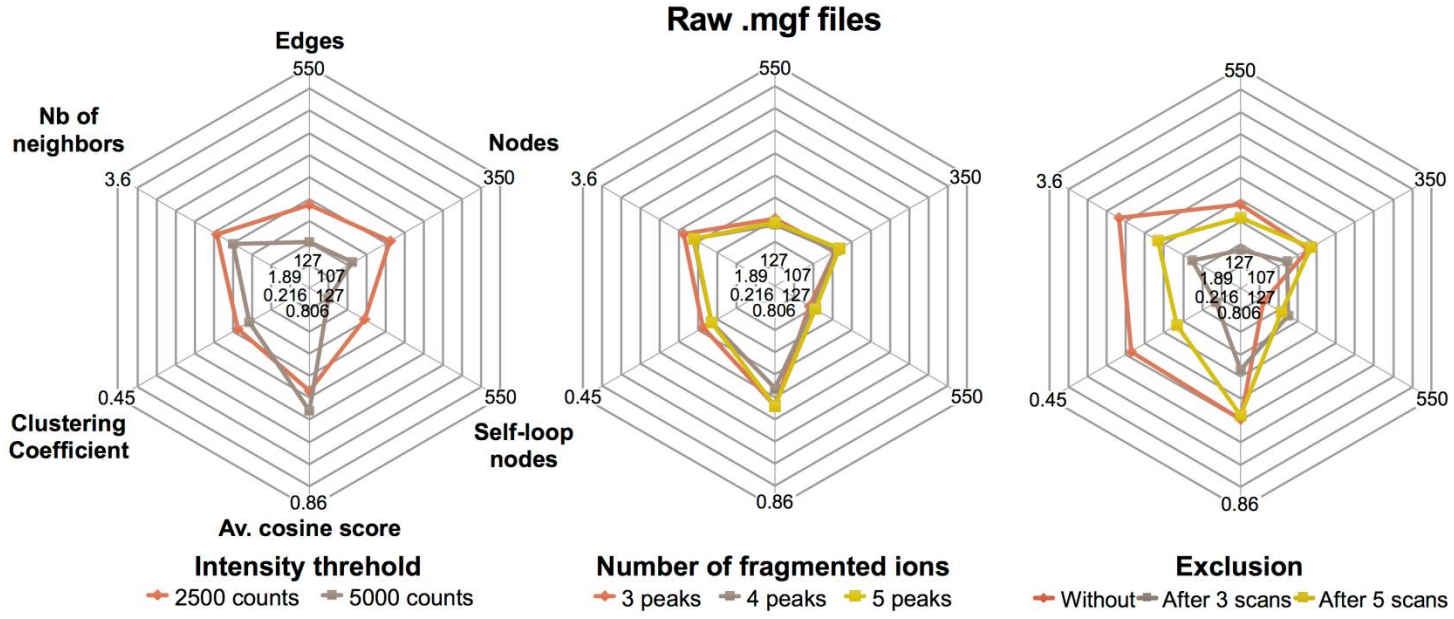
Données redondantes et non robustes …

- Quelles sont les bonnes conditions d'acquisition des données ?
- Comment différencier de réels isomères ?
- Comment avoir des données semi-quantitatives?
- Comment mieux annoter les réseaux ?

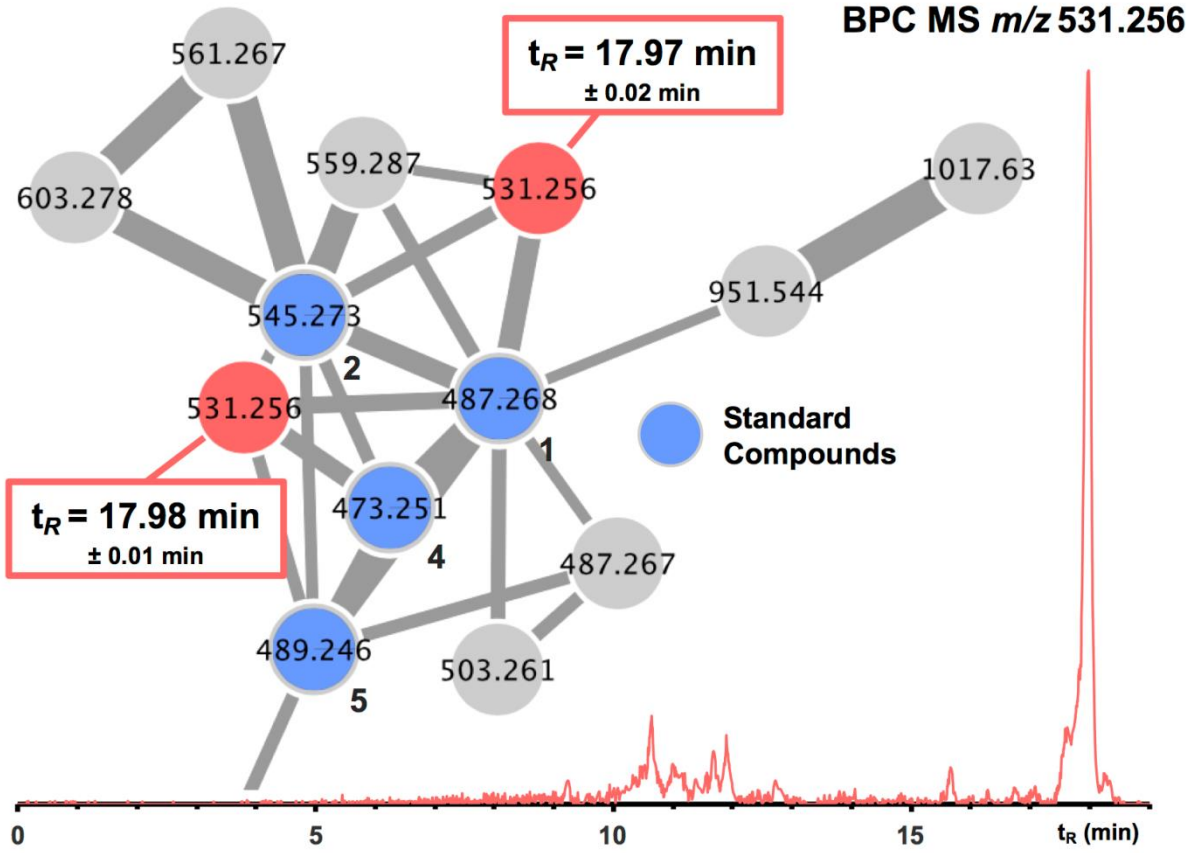
Optimisation des paramètres d'acquisition

Experiment n°	Minimum Intensity	Fragmented peaks	Exclusion	File (Mo)
1	2500	3	No	20.3
2	2500	3	3	4.6
3	2500	3	5	9.9
4	2500	4	No	16.3
5	2500	4	3	7.4
6	2500	4	5	9.7
7	2500	5	No	14.9
8	2500	5	3	7.6
9	2500	5	5	9.6
10	5000	3	No	17.2
11	5000	3	3	4.2
12	5000	3	5	9.9
13	5000	4	No	14.6
14	5000	4	3	5.1
15	5000	4	5	7.9
16	5000	5	No	17.6
17	5000	5	3	4.8
18	5000	5	5	8.4
Mean				10.6
19	10000	3	No	11.4
20	10000	3	3	3.7
21	10000	3	5	7.1
22	10000	4	No	10.6
23	10000	4	3	3.6
24	10000	4	5	4.5
25	10000	5	No	11.1
26	10000	5	3	2.7
27	10000	5	5	4.7

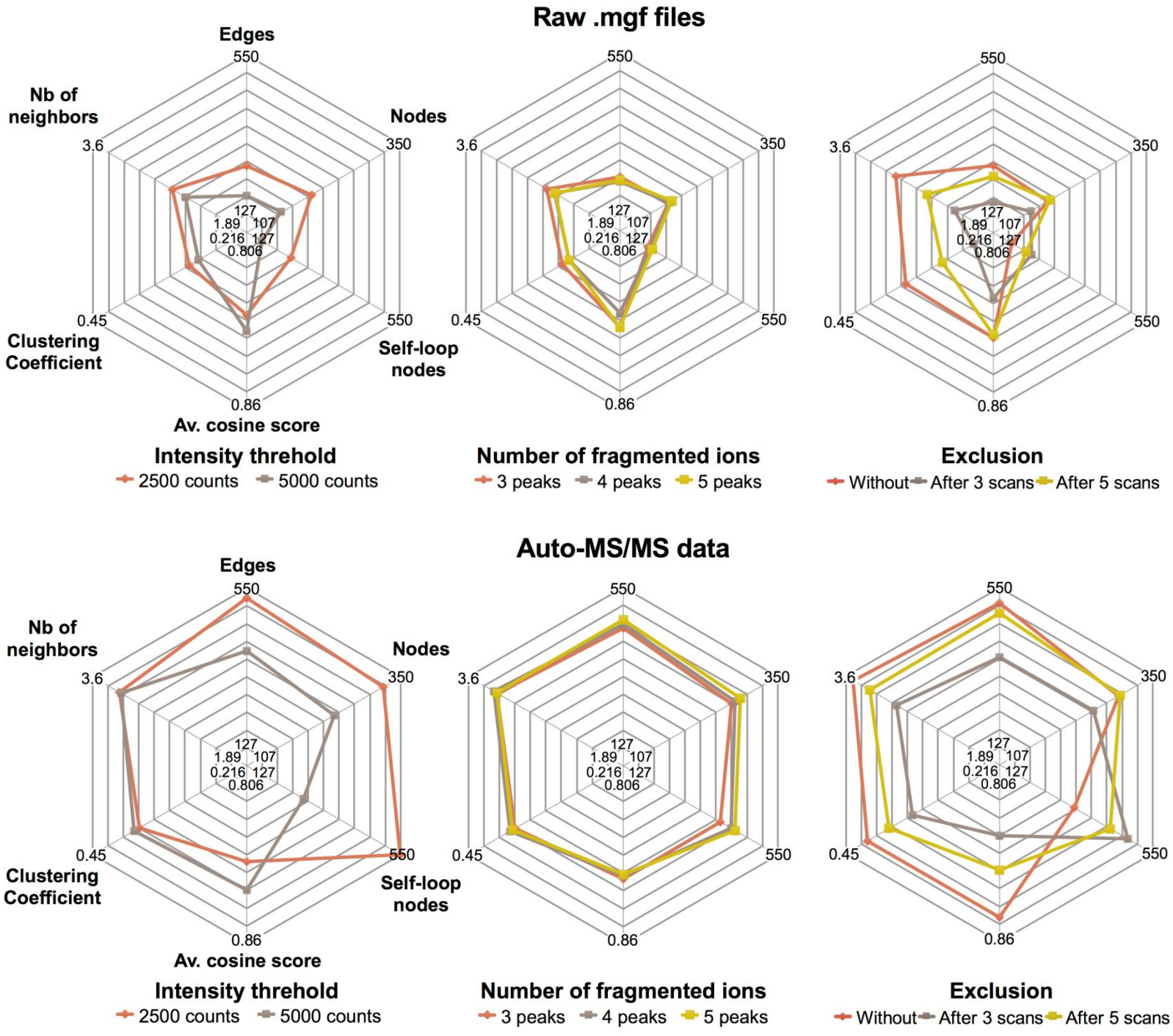
Optimisation des paramètres d'acquisition



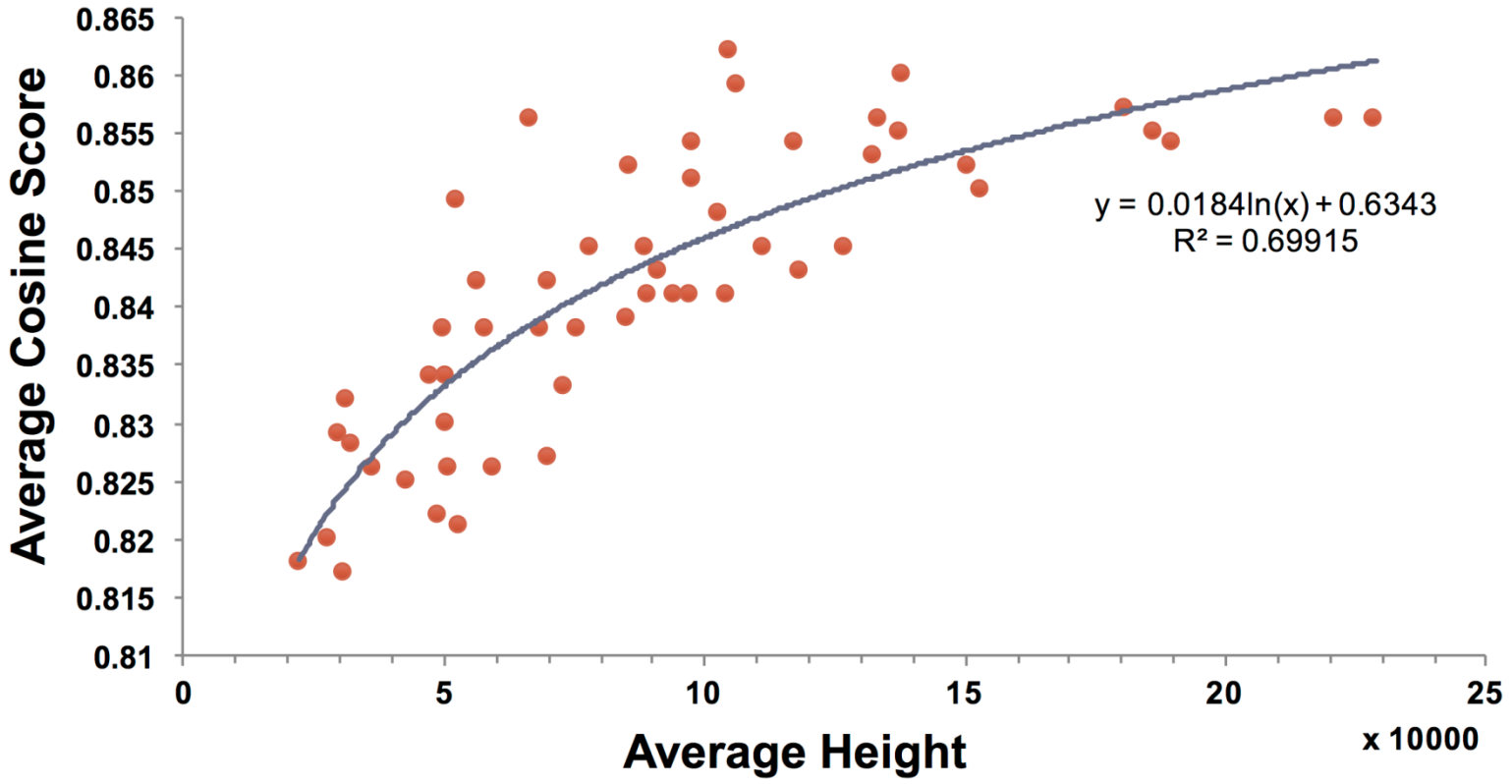
Optimisation des paramètres d'acquisition



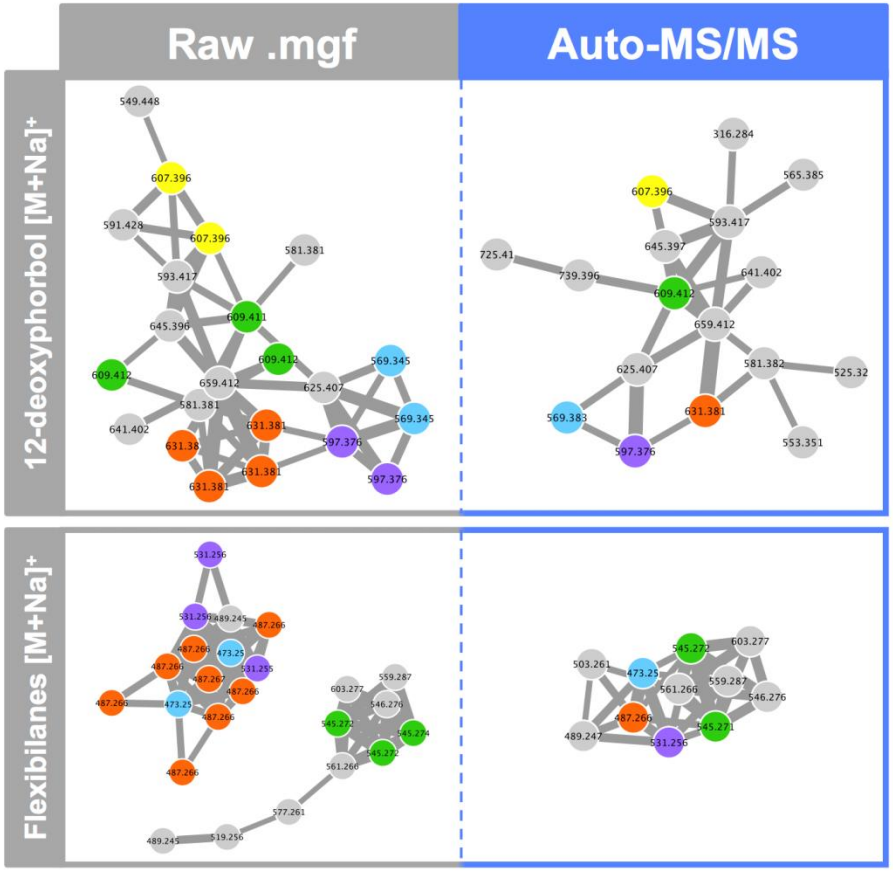
Op


Raw .mgf files
Edges
Nodes
Self-loop nodes
Av. cosine score
Clustering Coefficient
Nb of neighbors
550
350
550
0.86
0.45
3.6
127
1.89 107
0.216 127
0.806
Intensity threhold
2500 counts
5000 counts
Number of fragmented ions
3 peaks
4 peaks
5 peaks
Exclusion
Without
After 3 scans
After 5 scans
Auto-MS/MS data
Edges
Nodes
Self-loop nodes
Av. cosine score
Clustering Coefficient
Nb of neighbors
550
350
550
0.86
0.45
3.6
127
1.89 107
0.216 127
0.806
Intensity threhold
2500 counts
5000 counts
Number of fragmented ions
3 peaks
4 peaks
5 peaks
Exclusion
Without
After 3 scans
After 5 scans

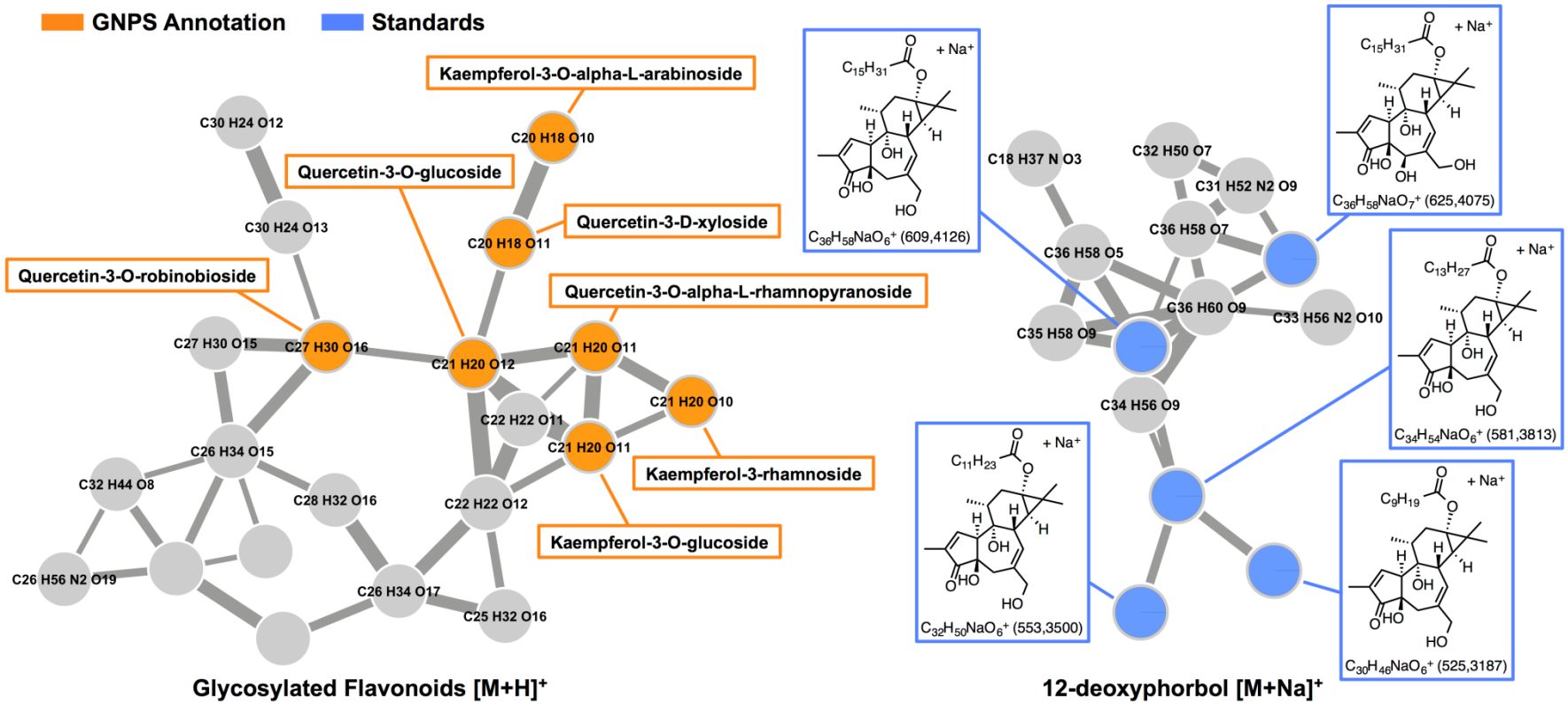
Optimisation des paramètres d'acquisition



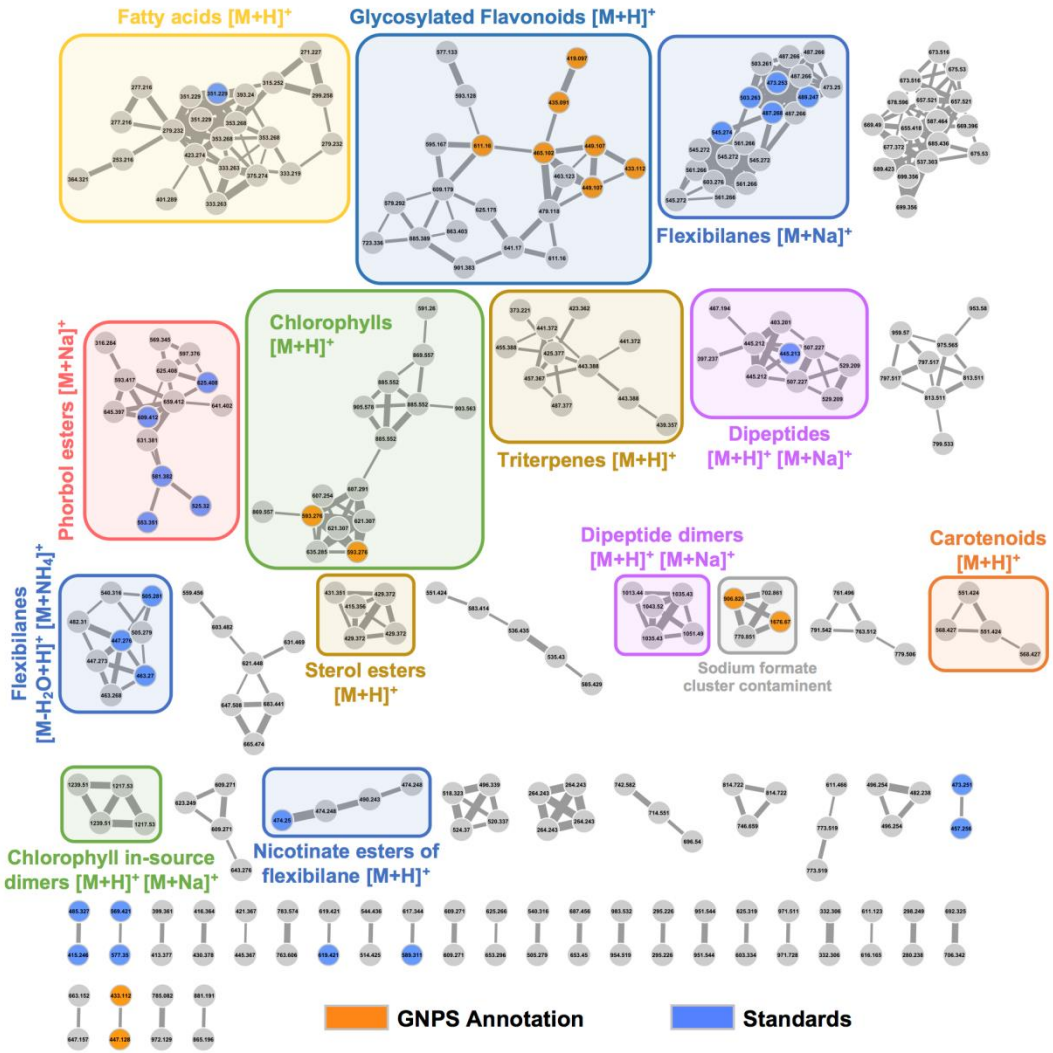
Optimisation des paramètres d'acquisition



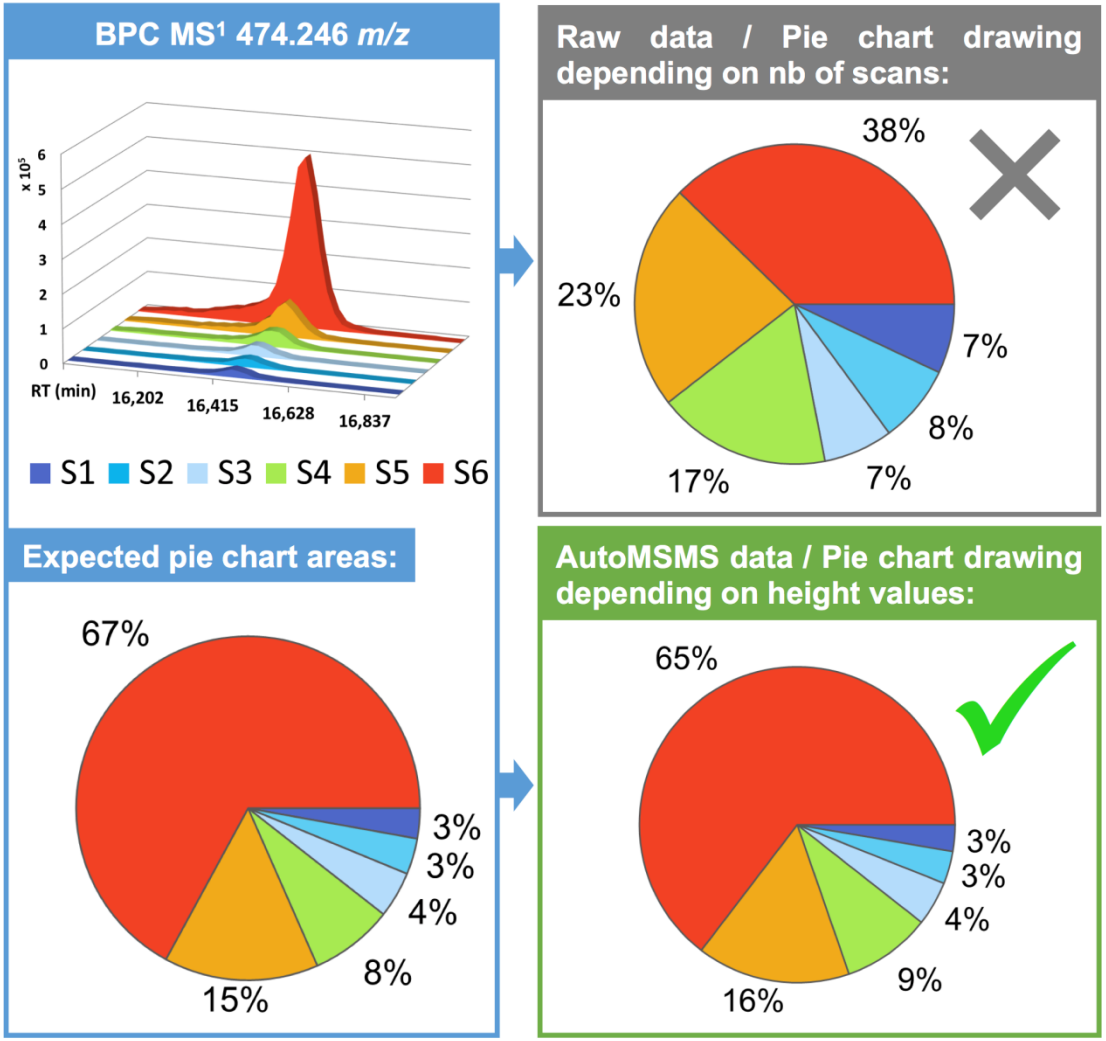
Optimisation des paramètres d'acquisition



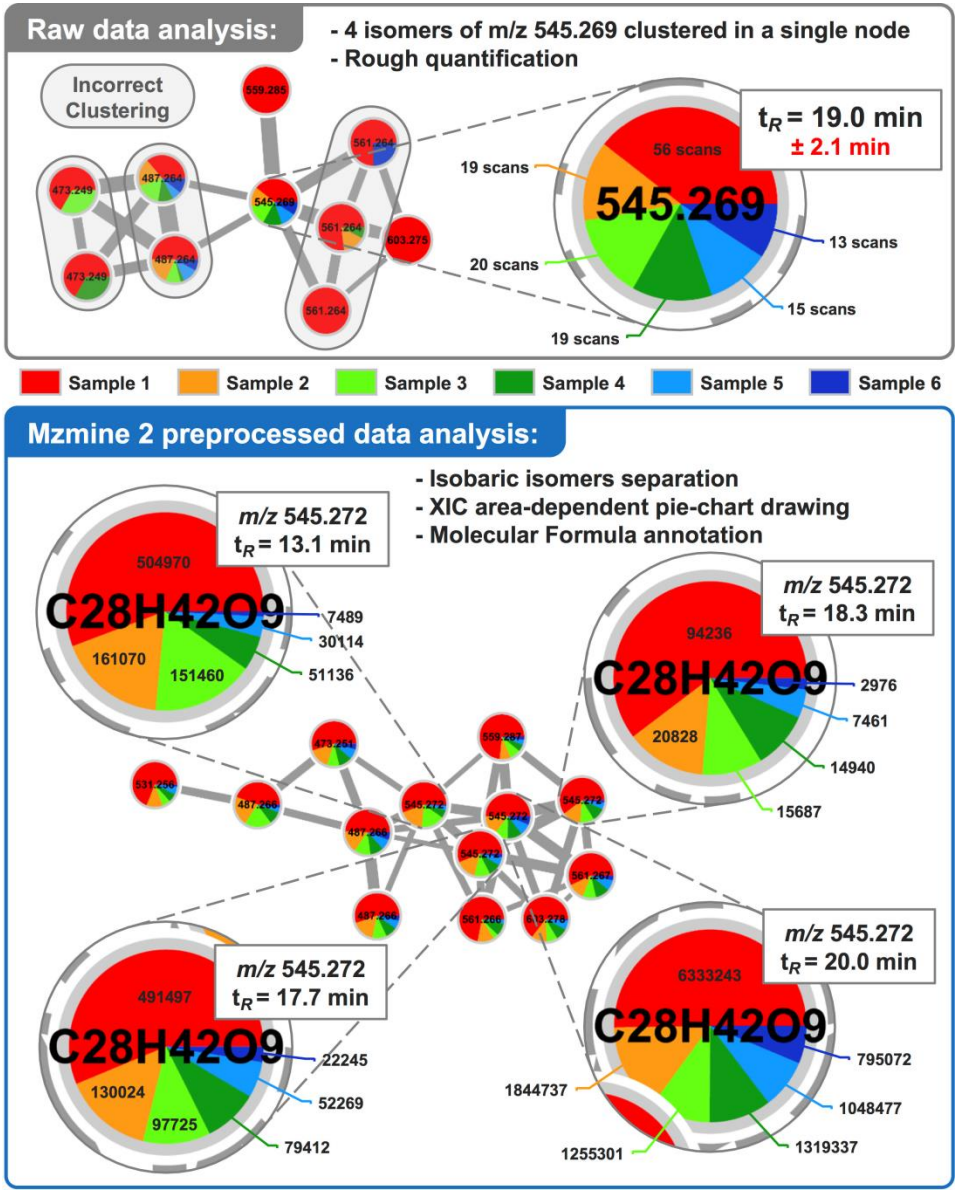
Optimisation des paramètres d'acquisition



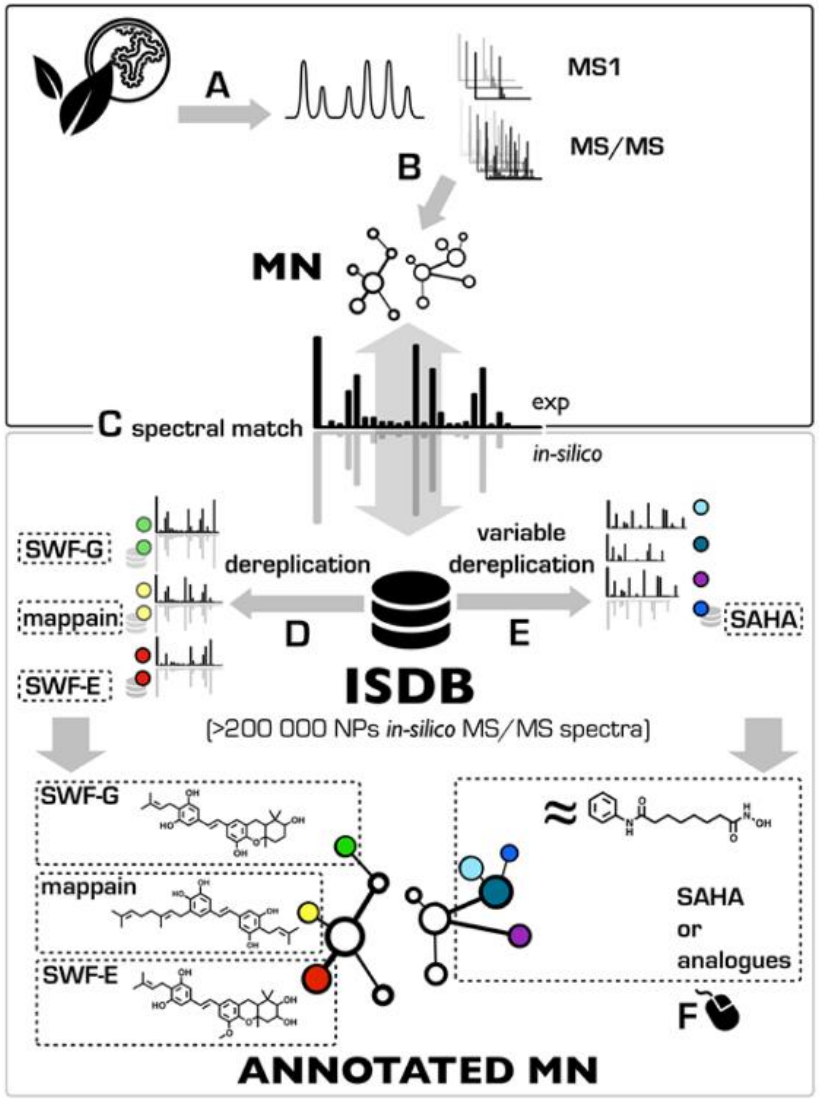
Optimisation des paramètres d'acquisition



Un outil générique



Recherche en base de données *in silico*



Integration of Molecular Networking and *In-Silico* MS/MS Fragmentation for Natural Products Dereplication

Pierre-Marie Allard,[†] Tiphaine Péresse,[‡] Jonathan Bisson,[§] Katia Gindro,[∥] Laurence Marcourt,[†] Van Cuong Pham,[⊥] Fanny Roussi,[‡] Marc Litaudon,[‡] and Jean-Luc Wolfender*[,†]

DOI: 10.1021/acs.analchem.5b04804
Anal. Chem. 2016, 88, 3317−3323

Recherche en base de données *in silico*

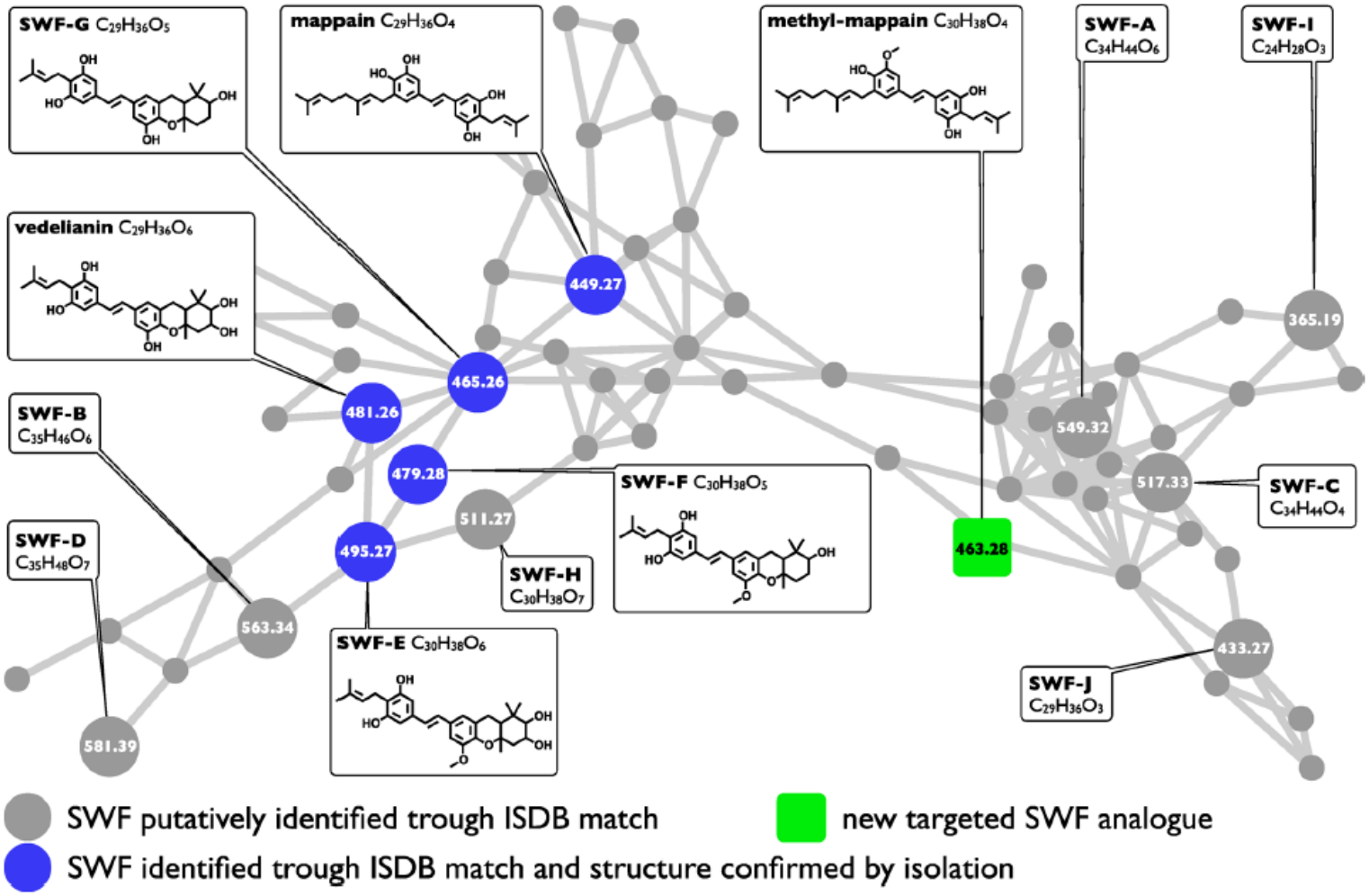


Figure 2. Cluster corresponding to compounds of the SWF (schweinfurthin) family observed in the MN of EtOAc extract of various *Macaranga* species. Match with the ISDB was made using the parent ion mass as filter (D in Figure 1).

Remerciements

ceba

Agence Nationale de la Recherche
ANR

Louis-Felix Nothias-Scaglia

Florent Olivon

Simon Remy

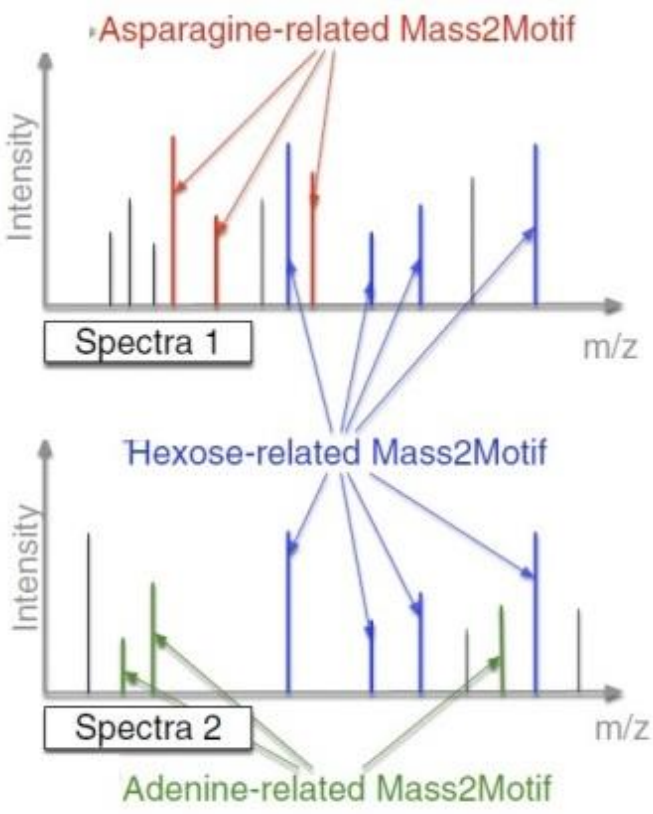
IIB) Recherche de motifs

MS2LDA (Mass2Motifs)

MS2LDA
Unsupervised Substructure Discovery

Mass2Motifs est un approche tirée du texte mining. On cherche à trouver des « blocs » représentatifs de structure spécifiques.

- Les spectres sont vus comme des textes.
- Les pics sont les mots qui composent ces textes.
- Les motifs sont des sujets, des ensembles de mots qui apparaissent souvent ensemble.

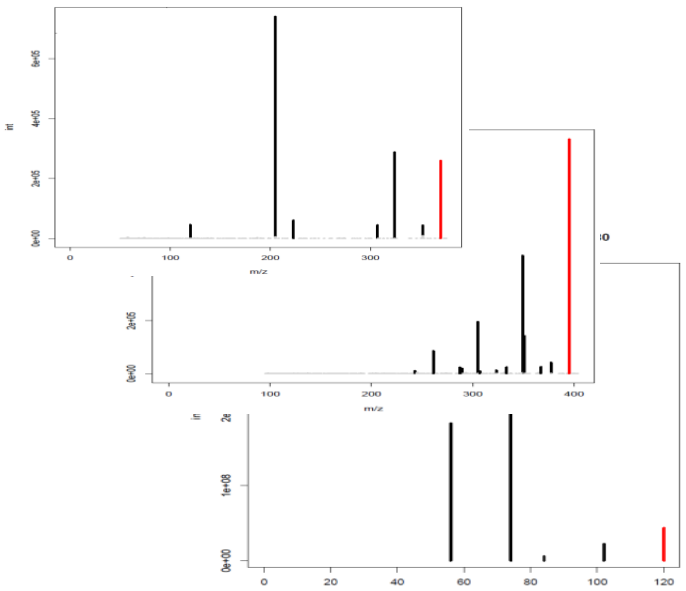


van Der Hooft *et al.* (2016), Topic modeling for untargeted substructure exploration in metabolomics, *PNAS*, **48**, 13738-13743.

MS2LDA : Principe

Input

Un ensemble de spectres MS-MS

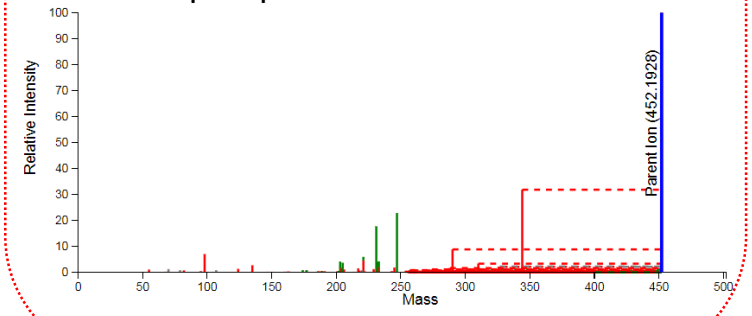


Output

- Un ensemble de motif
- Un ensemble de fragment récurrents

Motif	Feature	Min m/z	Max m/z
motif_31	fragment_109.0775	109.075	109.08
motif_133	fragment_110.0225	110.02	110.025
motif_264	fragment_110.0625	110.06	110.065
motif_146	fragment_110.0725	110.07	110.075
motif_122	fragment_110.0725	110.07	110.075

- Les motifs présent dans chaque spectres.

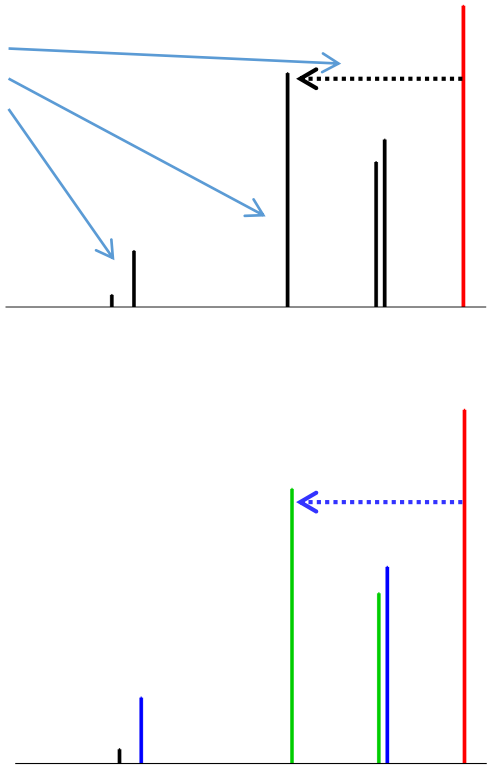


MS2LDA : Principe

- Chaque pic a une certaine probabilité d'appartenir à un motif

Name	Intensity	Mass2Motif (Probability)
fragment_105.0675	1000.0	motif_13 (1.000),
loss_17.0275	1000.0	motif_6 (1.000),
fragment_79.0525	49.0	motif_13 (0.984), motif_433 (0.014),
loss_43.0425	49.0	motif_433 (0.054), motif_127 (0.946),
fragment_103.0525	34.0	motif_19 (0.797), motif_433 (0.202),

Words



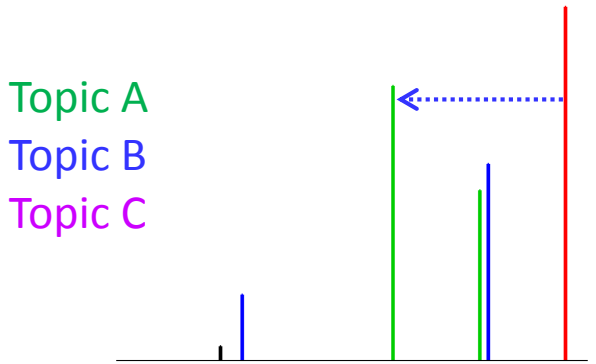
- Chaque motif est compose d'une probabilité de générer chaque mot

Topic A=(0,0,0.4,0.4,0,0.2)

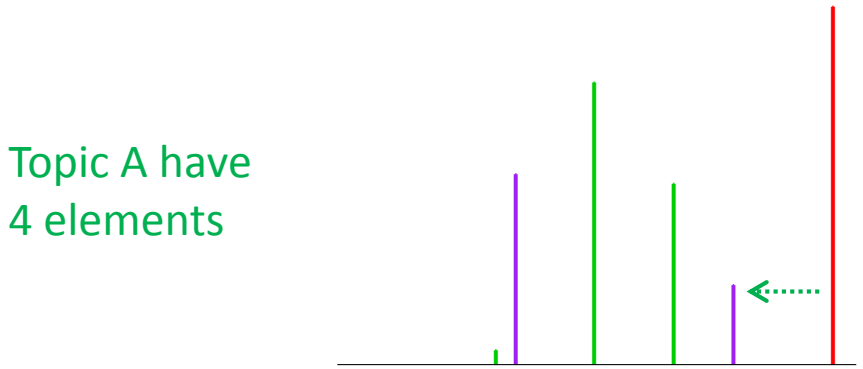
Topic B =(0,0.3,0,0,0.1,0)

MS2LDA : Principe

- Les mots d'un même motif peuvent être différents suivant les spectres.



- Un motif n'a donc pas une définition « stricte », c'est une loi de probabilité, et un spectra est un ensemble de tirage aléatoire parmi ces motifs



MS2LDA : Motif

Pour un motif on peut voir :

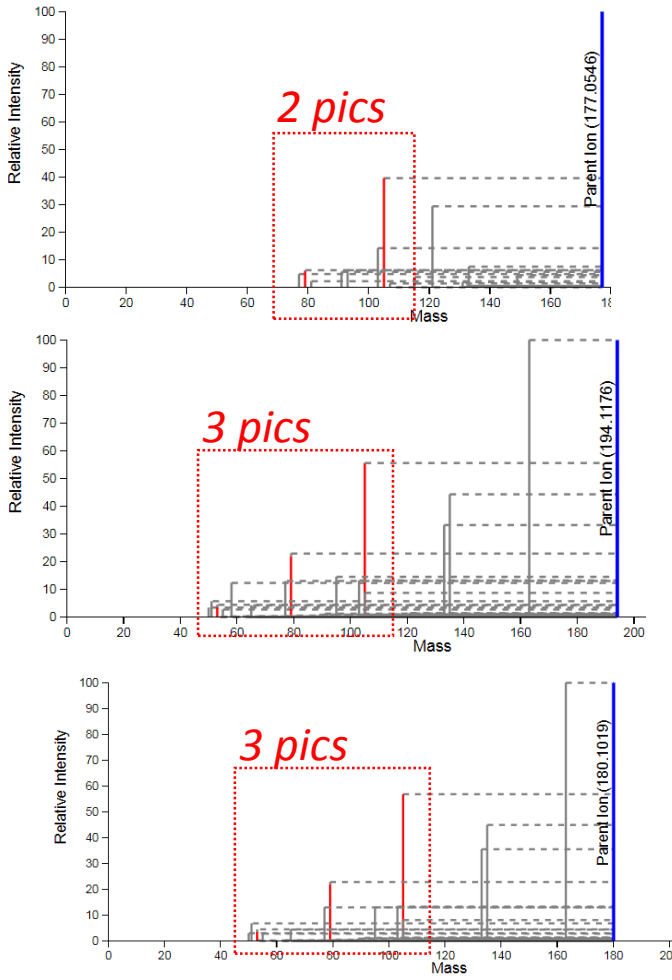
Les fragments ou pertes le composant

Feature	Probability
fragment_105.0675	0.838
fragment_79.0525	0.120
fragment_53.0375	0.042

Les spectres dans lesquels il est présent

Fragmentation spectra	Annotation	Probability	Overlap Score
ufz_0090.ms	ufz_0090	0.452	0.958
eawag_0736.ms	eawag_0736	0.259	0.999
ufz_0253.ms	ufz_0253	0.201	0.957
ufz_0294.ms	ufz_0294	0.127	0.958
eawag_0801.ms	eawag_0801	0.120	0.999

Plus visuellement

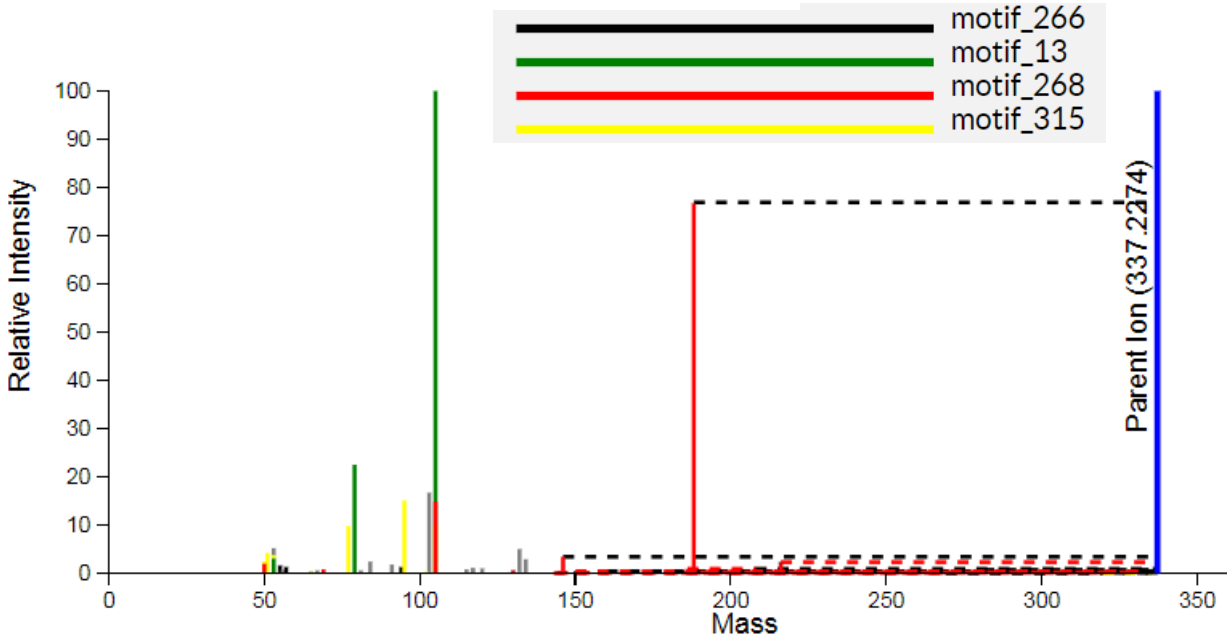


MS2LDA : Spectre

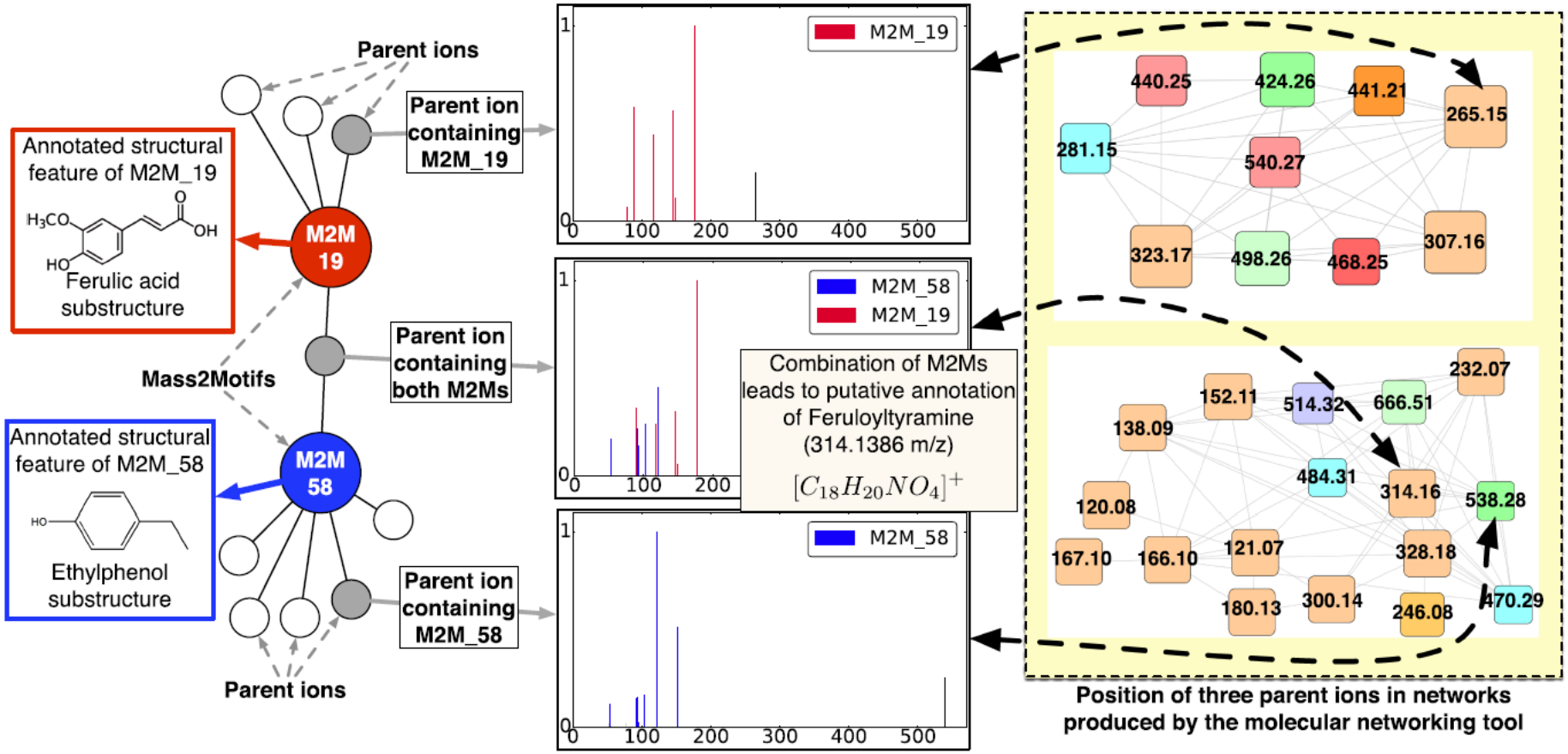
Pour un spectre on peut les voir les motifs qui lui sont associés :

Motif	Probability	Overlap Score
motif_13	0.259	0.999
motif_315	0.084	0.785
motif_266	0.181	0.311
motif_268	0.398	0.112

On peut également voir les pics associés à chaque Motif. (bleu = ion parent, gris = pics non associés à un motif)



MS2LDA : vers les réseaux



Ces motifs peuvent être utilisé pour faire des réseaux et visualiser les points communs

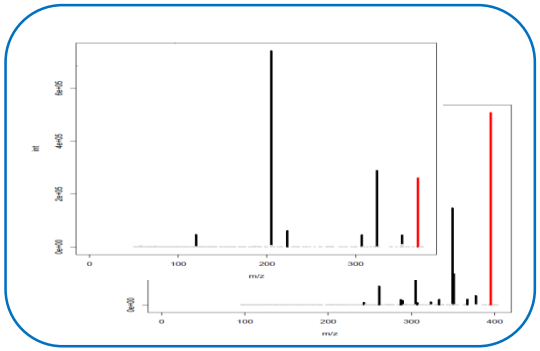
MS2LDA : Résumé

PRO	CONS
• Méthode non supervisée • Disponible en ligne • Prend en compte multiple composants d'une molécule	• Difficile à interpréter • Dépend beaucoup de paramètres difficiles à régler. • Plus efficiace sur les grosse molécules.

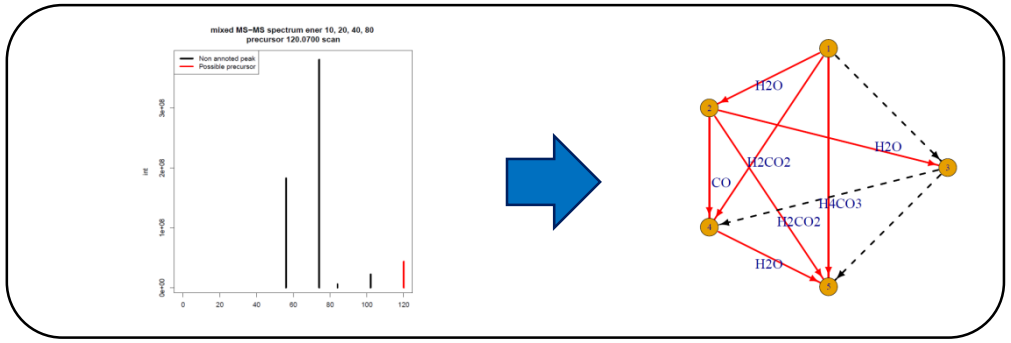
Pas de définition fixé du motif

MineMS2 = Motifs + Fragmentation *in silico*

1.

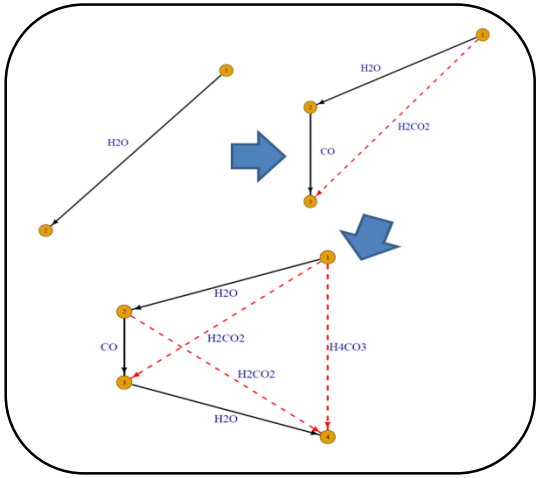


Ensemble de spectres

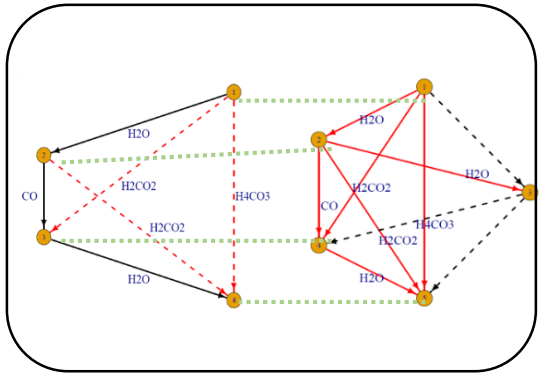


Conversion des spectres en graphes de fragmentation

2.

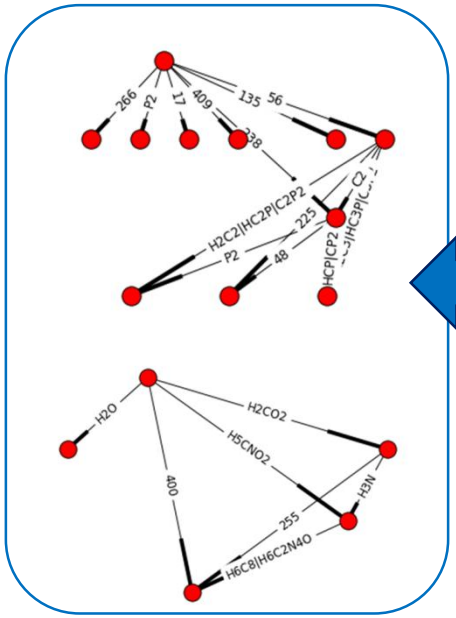


Génération des candidates



Compte des patternes

Frequent subgraph mining (FSM)



Motifs

Poster P79

Thanks for coming!

Questions?