METABOHUB

Genotoul Metatoul

@metatoul

BARSA

Un algorithme pour l'annotation automatique de spectres RMN 2D de matrices complexes

Cécile Canlet, Marie Tremblay Franco

CONTEXTE

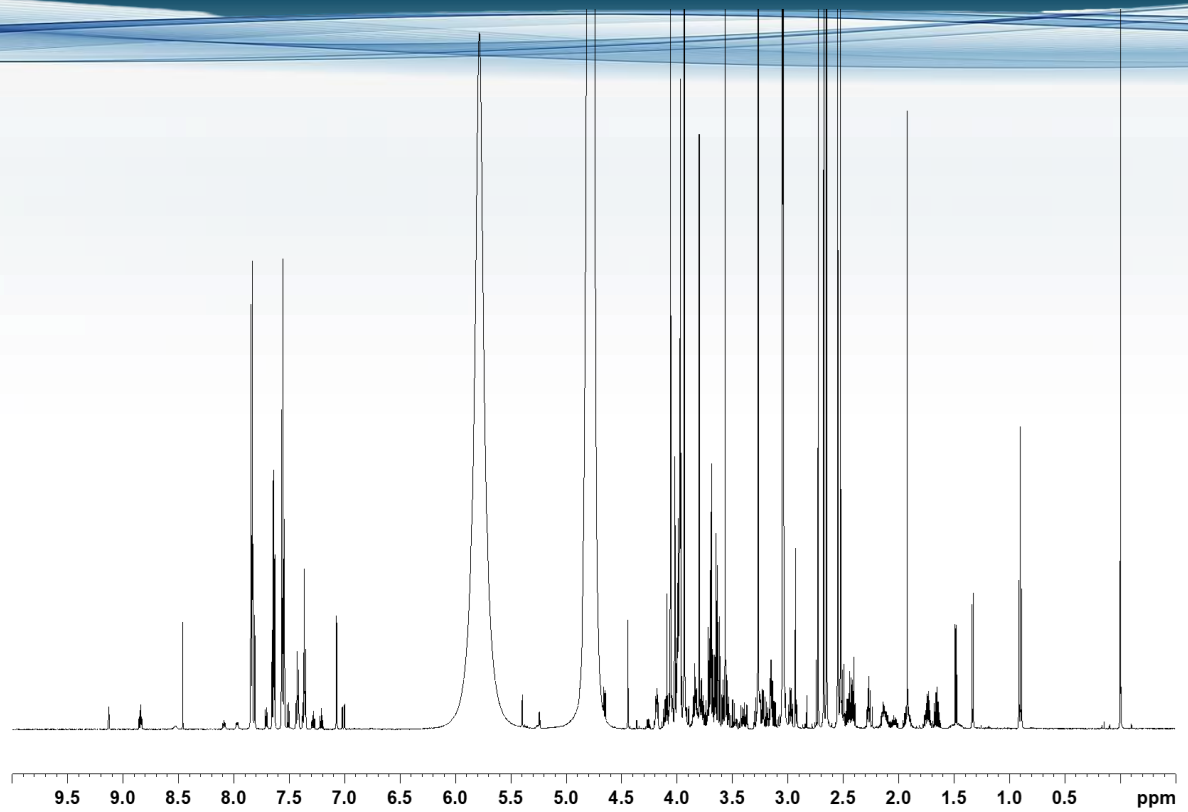
- But des études de métabolomique → mettre en évidence des perturbations métaboliques suite à l'exposition à certains stimuli (par exemple maladies, traitement toxiques, etc) et identifier des marqueurs d'exposition

⇒ Identification des voies métaboliques perturbées par le stimulus

- RMN du proton est l'une des techniques analytiques utilisée pour générer les profils métaboliques

- **Identification des métabolites discriminants = étape importante pour pouvoir identifier les voies métaboliques perturbées par le stimulus.**

Spectre RMN 1H d'une matrice biologique



- Pas de séparation des composés avant l'analyse RMN du proton
- Un métabolite peut avoir plusieurs signaux à différents déplacements chimiques
- Beaucoup de signaux se superposent

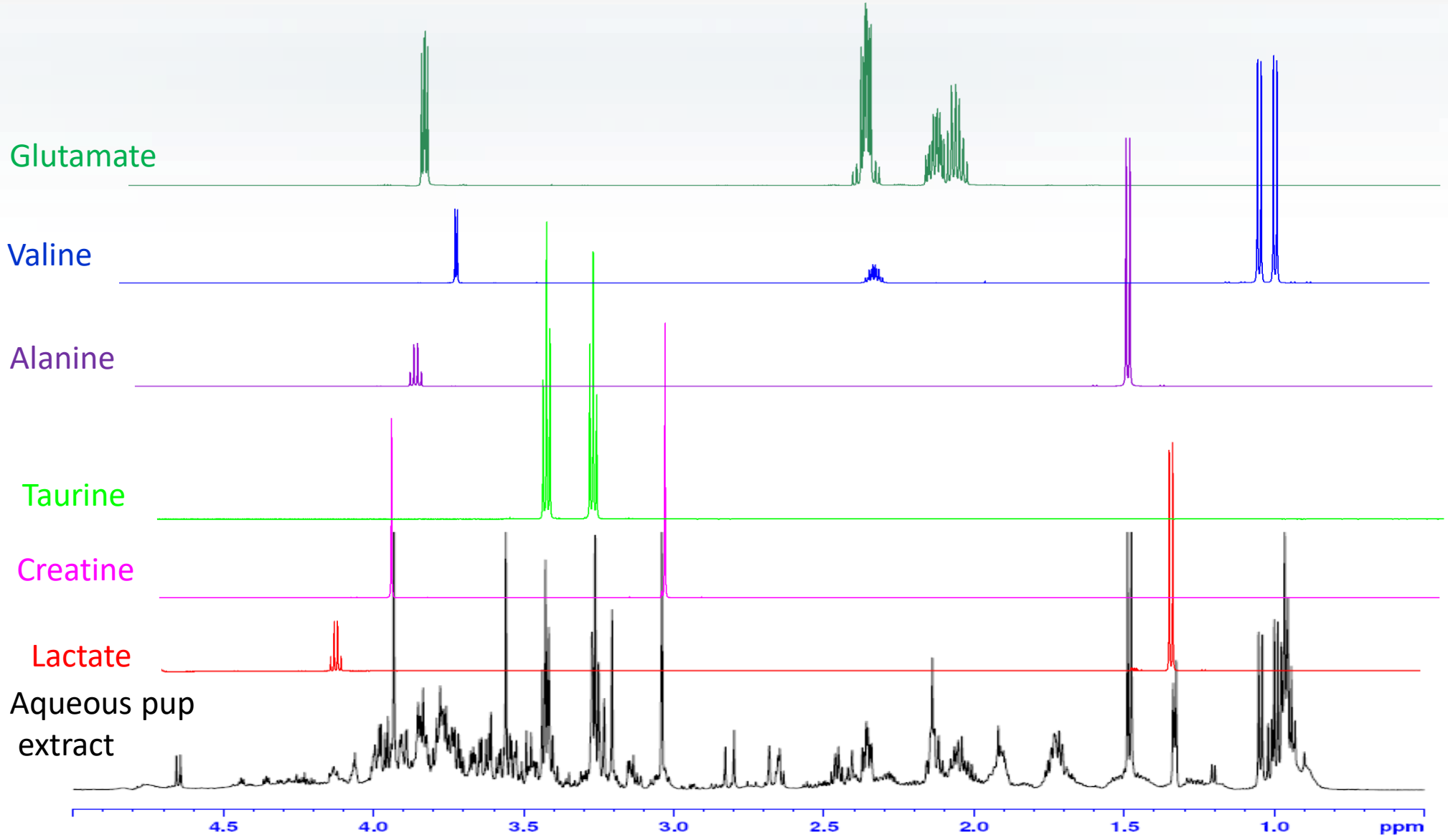
⇒ **Identification des composés dans une matrice biologique est complexe**

Identification des métabolites

- Basée sur les déplacements chimiques 1H, figures de couplage et les constantes de couplage
- Comparaison avec des spectres annotés de matrices similaires dans la littérature
- Spiking : ajout d'un composé de référence dans l'échantillon
- Comparaison avec des spectres RMN de composés de référence :
  - Base de données maison
  - Base de données commerciale (Chenomx)
  - Bases de données libres (HMDB, BRMB)

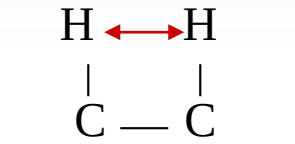
- RMN bi-dimensionnelle (2D)

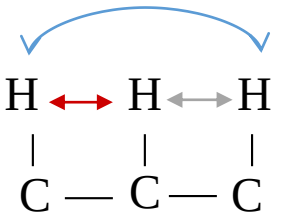
Comparaison avec des spectres RMN 1H de composés de référence



- **Information est éclatée dans les deux dimensions → moins de recouvrement de signaux**

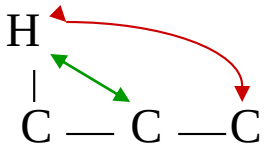
- JRES (J-RESolved) : séparation du déplacement chimique (dimension f2) et couplage scalaire (dimension f1)
- 1H-1H COSY (Correlation SpectroscopY) : corrélations 1H-1H via 3 liaisons



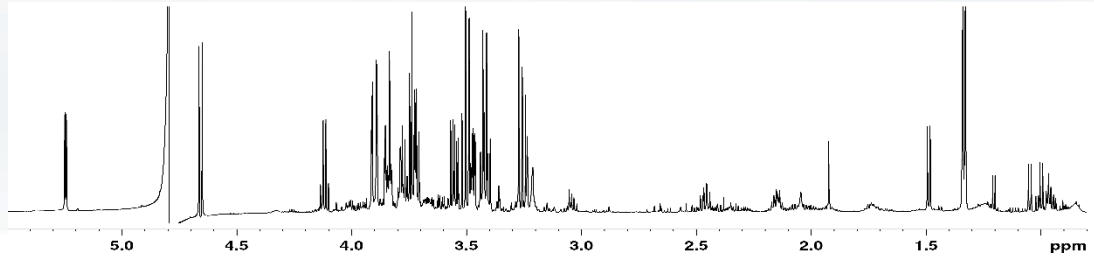


- 1H-1H TOCSY (TOtal Correlation SpectroscopY) : corrélations entre tous les protons d'un même système de spins
- 1H-13C HSQC (Heteronuclear Single Quantum Coherence) : corrélations 1H-13C via 1 liaison

- 1H-13C HMBC (Heteronuclear Multiple Bond Coherence) : corrélations 1H-13C via 2, 3 or 4 liaisons

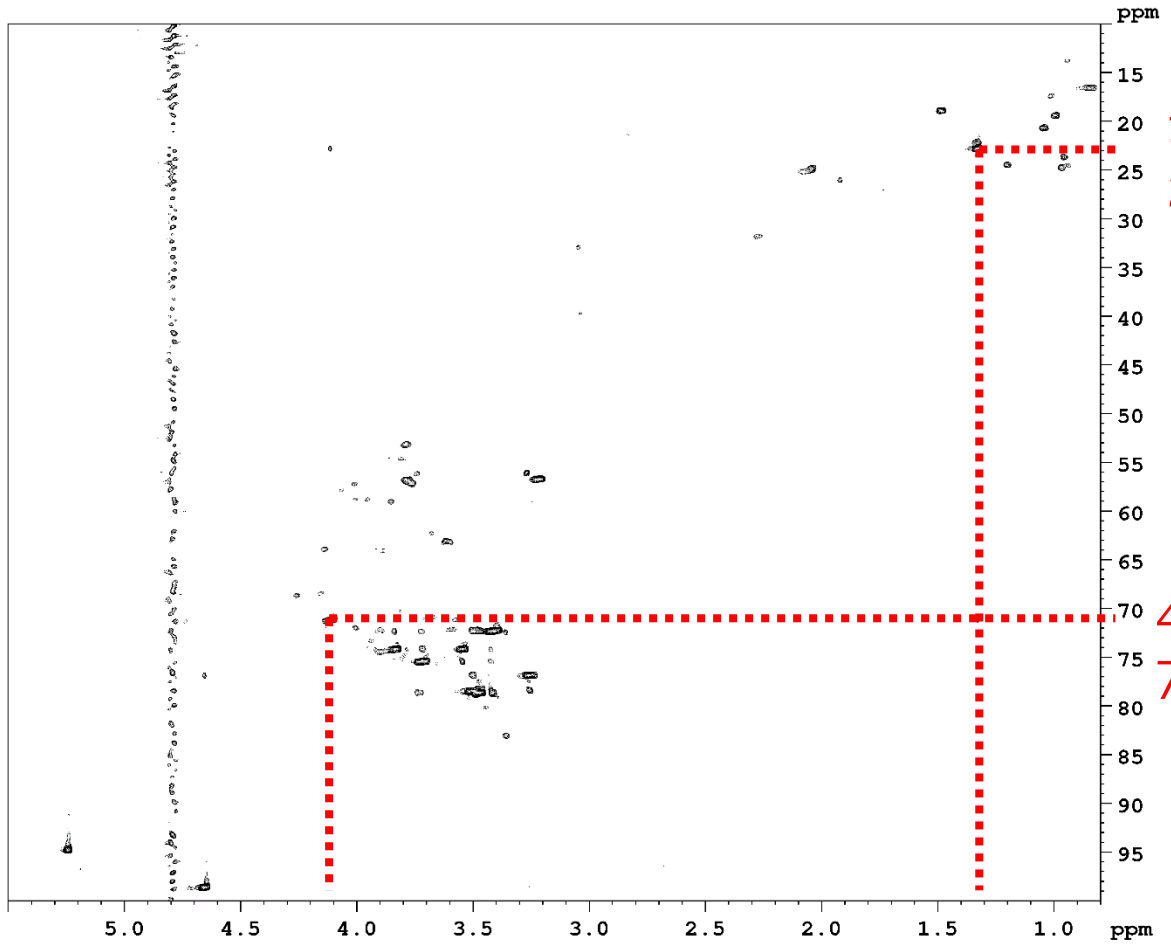


HSQC



Lactate :

CH_3 – CH (OH) - COOH



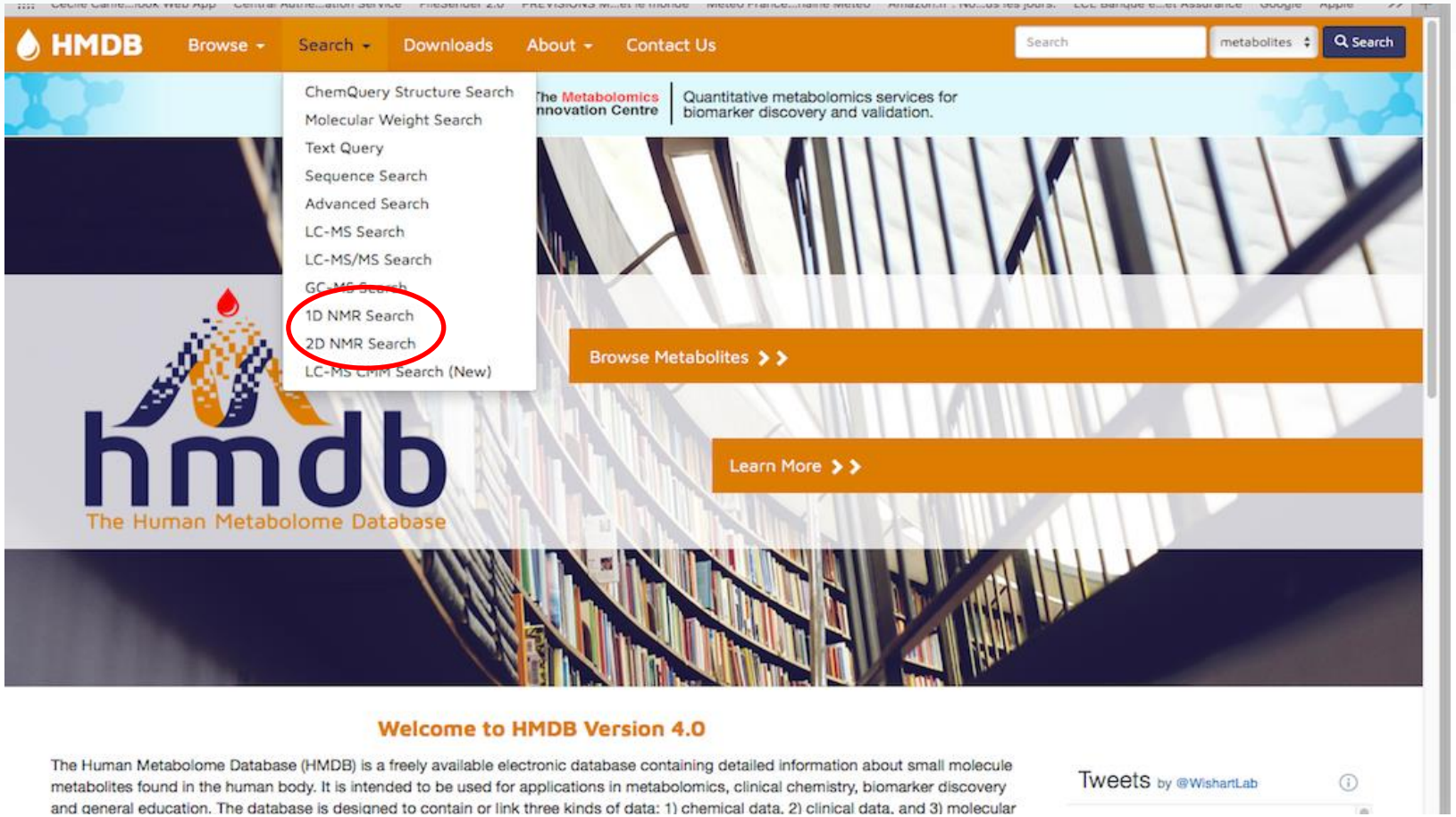
1.33 ppm
22.95 ppm

4.11 ppm
71.25 ppm

BASE DE DONNEES

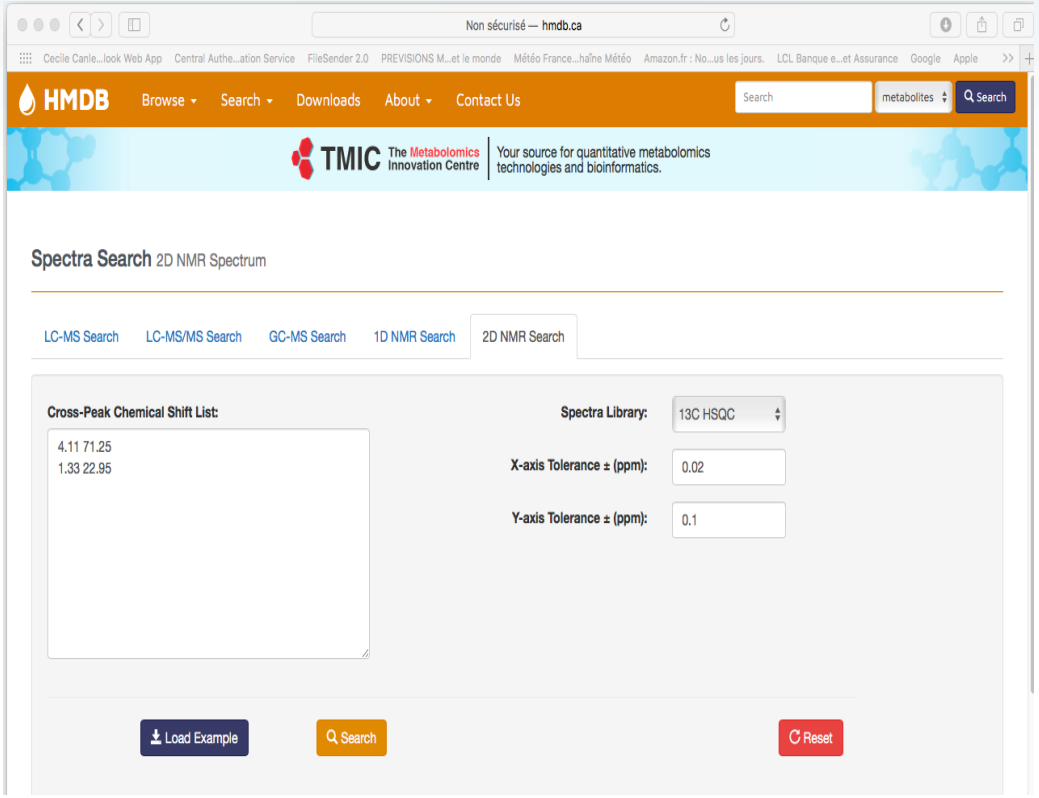
HMDB (The Human Metabolome DataBase) (1)

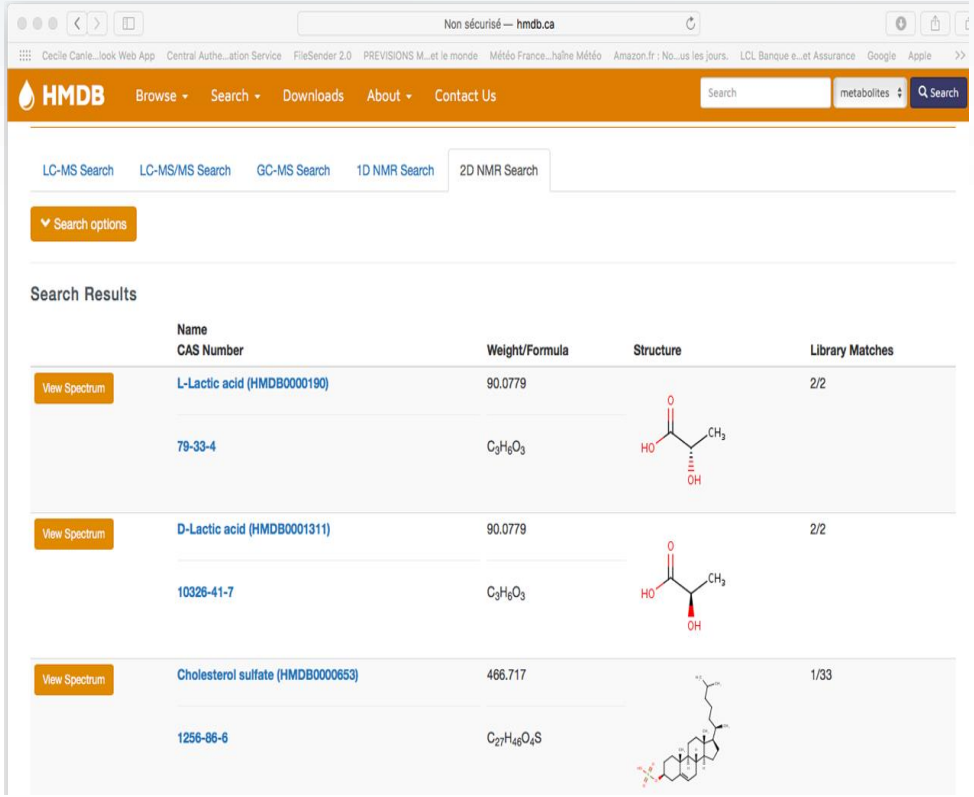
- www.hmdb.ca/→ consulter spectres RMN de standards (1494 composés) ou interroger la base de données



HMDB (2)

Peak-matching 2D → interrogation par composé

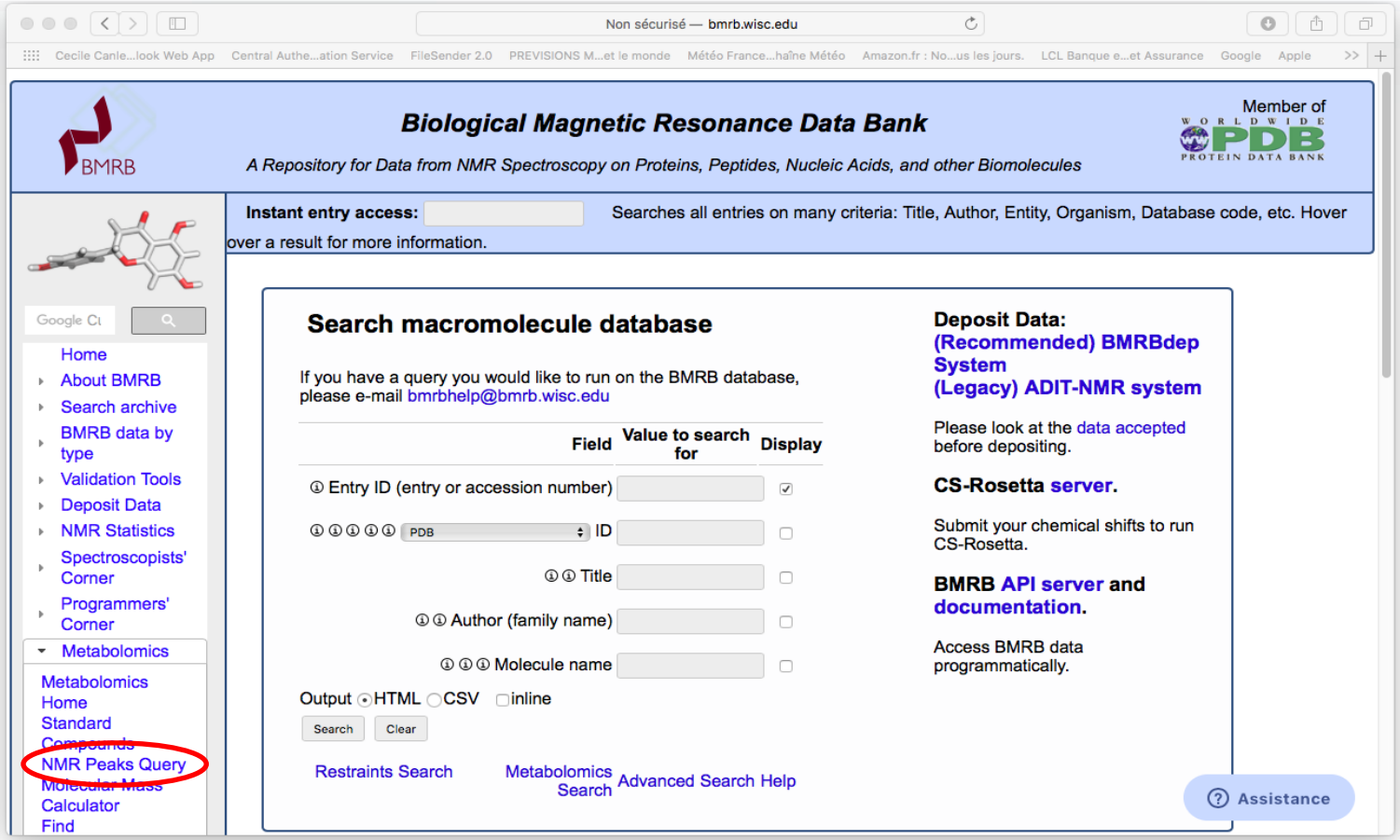




⇒ Lactic acid

BMRB (Biological Magnetic Resonance Data Bank)

- http://www.bmrb.wisc.edu/ → consulter spectres RMN de standards ou interroger base de donnée



PEAKFOREST 2.0.3

- Base de données développée dans le cadre de Metabohub
- Spectres de composés de référence LC-HRMS, GC-HRMS, RMN 1D et 2D
- Spectres RMN enregistrés à différents champs magnétiques (500; 600; 800 MHz) et à différents pH (6 et 7)
- Environ 100 composés avec 8 séquences enregistrés à 600 MHz et pH 7 (noesypr1d; cpmg; C13; JRES; COSY; TOCSY; HSQC; HMBC)
- Outils de peak-matching : LC-MS; MSMS; RMN 1D
- Web Service REST Requests pour extraire des sous-bases d'intérêt

Annotation manuelle de matrices biologiques

- Chronophage
- Compliquée
- Certains signaux ne sont pas identifiés
- Un outil d'annotation automatique de spectres RMN 1D et 2D serait utile
- Outils d'annotation automatique de spectres RMN 1D :
  - Package R Batman : pas facile à utiliser et temps de calcul très longs
  - BAYESIL (http://bayesil.ca/) : interface web (limité au sérum, plasma, et CSF)
  - Chenomx : logiciel commercial
  - ASICS

⇒ Beaucoup de faux positifs → utilisation de la RMN 2D pourrait diminuer le nombre de faux positifs

OUTILS ANNOTATION 2D

COLMAR (Complex Mixture Analysis by NMR; Bingol et al. 2016)

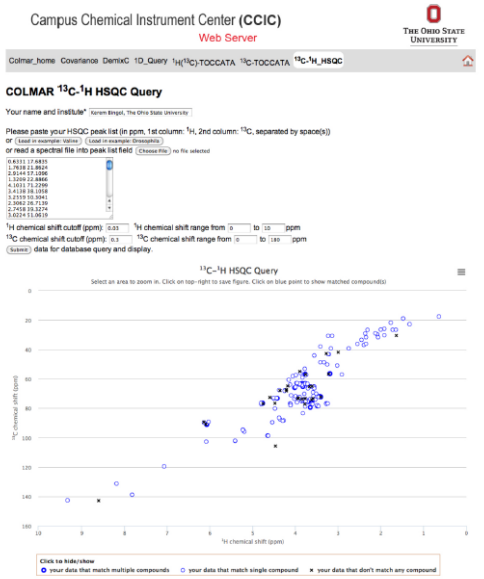
- Interface web : http://spin.ccic.ohio-state.edu/index.php/colmar
- Matrice(s) / fluide(s) : Sérum humain
- Séquence(s) RMN 2D : HSQC / TOCSY
- Bases de données utilisée(s) pour les composés de référence : BRMB et HMDB
- Nombres de composés : 701 pour HSQC / TOCSY
- Pas de combinaison entre les 2 séquences HSQC et TOCSY

Returned list of compounds

List of matched compounds:

Clear all	Metabolite	1H	13C	Matching_ratio	Uniqueness
Show Me	1_Acetic_acid	0.007	0.0527	1	1/1
Show Me	1_Alanine	0.0071	0.0845	1	2/2
Show Me	1_Phosphoethanolamine	0.0061	0.0424	1	2/2
Show Me	1_beta_Alanine	0.0071	0.0686	1	2/2
Show Me	1_L_Arginine	0.0028	0.0655	1	4/5
Show Me	1_Choline	0.0041	0.0276	1	2/3
Show Me	1_D_Fructose	0.0023	0.0161	1	6/7
Show Me	2_D_Fructose	0.0079	0.0833	1	5/6
Show Me	1_L_Glutamine	0.0056	0.0893	1	2/3
Show Me	1_D_Glucose_6_phosphate	0.013	0.107	1	5/6
Show Me	1_Glycerol	0.0042	0.0224	1	2/3
Show Me	1_Glycine	0.0061	0.055	1	1/1
Show Me	1_Gluconic_acid	0.0059	0.0583	1	4/6
Show Me	1_D_Glucose	0.0048	0.0177	1	5/7
Show Me	2_D_Glucose	0.0055	0.0149	1	3/7
Show Me	1_L_Glutamic_acid	0.0057	0.039	1	3/3
Show Me	1_L_Histidine	0.0026	0.0641	1	5/5
Show Me	1_Lactic_acid	0.0028	0.0362	1	2/2
Show Me	1_Methionine_sulfoxide	0.0106	0.0706	1	4/4
Show Me	1_Pyruvic_acid	0.0023	0.0438	1	1/1
Show Me	1_L_Proline	0.0043	0.0455	1	6/6
Show Me	1_L_Serine	0.0064	0.042	1	2/2
Show Me	1_Succinic_acid	0.0077	0.0792	1	1/1
Show Me	1_Taurine	0.0035	0.0209	1	2/2
Show Me	1_D_Trehalose	0.0106	0.0319	1	4/7
Show Me	1_Maltose	0.0051	0.0447	1	7/14
Show Me	2_Maltose	0.0058	0.0253	1	6/14
Show Me	1_Phosphorylcholine	0.0091	0.1641	1	2/3
Show Me	1_Adenosine	0.0089	0.0545	0.75	4/8
Show Me	1_DSS	0.0073	0.0342	0.75	3/4
Show Me	1_IMP	0.0182	0.1272	0.714286	3/7
Show Me	1_Inosine	0.0064	0.0568	0.75	3/8
Show Me	1_NAD	0.0184	0.0382	0.764706	8/17

Graphical User Interface



METABOMINER (Xia et al., 2008)

- Matrice(s) / fluide(s) : Liquide céphalo-rachidien / Urine / Plasma
- Séquence(s) RMN 2D : HSQC / TOCSY
- Bases de données utilisées pour les composés de référence : HMDB / BRMB
- Nombres de composés / spectres : 502 HSQC / 223 TOCSY
- Algorithme :
  - Coefficient unicité = nombre de pic(s) voisin(s) à distances définies (0.01 à 0.05ppm pour H et 0.05 à 0.25ppm pour C)
  - Seuil adaptatif = variation du décalage autorisé en fonction coefficient unicité
  - Signature minimale = définition d'un ensemble pics minimum nécessaires pour identifier de façon unique un composé parmi tous les autres composés inclus dans la BdD de référence. Basée sur le nombre de pics des composés et de leur voisinage (coefficient d'unicité)

SpinCouple (Kikuchi et al. 2015)

- Matrice(s) / fluide(s) : contenu intestinal et muscle de poisson, intestin de sériole adulte japonnaise, contenu de tube digestif de termite, eau de rizière, algue, ciboule et fécès humain
- Séquence(s) RMN 2D : Jres
- Bases de données utilisée(s) pour les composés de référence : maison / Birmingham Metabolite Library
- Nombres de composés / spectres : 598 (dont 155 issus de BML)
- Algorithme :
  - Batch annotations, basées sur le déplacement chimique 1H et les constantes de couplage 1H – 1H
  - Unicité d'un pic de la BdD : 1 / nombre de matches autour du pic de référence quand testé par rapport à la BdD en utilisant les valeurs de tolérance spécifiées
  - Précision : valeurs de tolérance statistiquement significatives pour l'annotation

OBJECTIF

- Outils existants utilisent leur propre base de données ou des bases de données publiques (HMDB, BRMB) → difficile de vérifier la qualité des données ou de rajouter des spectres de composés de référence (pH ou champs magnétiques différents)

- Développement d'un algorithme d'annotation semi-automatique de spectres RMN 2D
  - Permettant l'annotation des métabolites présents dans une matrice biologique complexe avec une probabilité de présence
  - Basé sur l'interrogation de la base de données PeakForest
  - En combinant plusieurs séquences RMN 2D
  - En appliquant des seuils et des conditions d'unicité
  - Pour réduire le nombre de faux positifs

MATERIEL & METHODES

Templates

Composé de référence

1H-1H TOCSY NMR correlations				
peak index	v (F2) [ppm]	v (F1) [ppm]	intensity [rel]	annotation
1	4,0151	2,9134	[illegible]	2-3
2	4,0149	4,0206	[illegible]	2-2
3	2,962	2,9481	[illegible]	3b-3b
4	2,962	4,0182	[illegible]	3b-2
5	2,8813	2,9263	[illegible]	3a-3
6	2,881	4,0182	[illegible]	3a-2

analytical_sample | NMR_analyzer | JRES | COSY | TOCSY | HSQC | HMBC | carbon-1: ...

Matrice complexe

1H-13C HSQC NMR correlations				
peak index	v (F2) (1H) [ppm]	v (F1) (13C) [ppm]	intensity [rel]	annotation
2	0.8999	16.0523	[illegible]	
1	0.9437	13.9434	[illegible]	
4	0.9980	19.5009	[illegible]	
3	1.0160	17.4913	[illegible]	
5	1.0485	20.7414	[illegible]	
8	1.2692	27.2224	[illegible]	
7	1.3359	22.9187	[illegible]	
6	1.3393	22.2392	[illegible]	
13	1.4557	19.4879	[illegible]	
14	1.4564	21.9300	[illegible]	
10	1.4740	25.4259	[illegible]	
11	1.4777	27.1741	[illegible]	
9	1.4885	19.0464	[illegible]	
12	1.5628	22.1441	[illegible]	
15	1.7248	29.0156	[illegible]	
16	1.9227	30.3966	[illegible]	
18	1.9875	38.6313	[illegible]	
17	2.0085	26.4727	[illegible]	
19	2.0698	31.7230	[illegible]	
20	2.1459	29.0702	[illegible]	
21	2.1643	42.3894	[illegible]	
22	2.2798	31.7511	[illegible]	
23	2.3538	31.7511	[illegible]	
25	2.4433	45.7749	[illegible]	
24	2.4585	33.6275	[illegible]	
26	2.6987	39.3423	[illegible]	
27	2.7207	37.3607	[illegible]	
28	2.8183	39.3836	[illegible]	
29	2.8814	37.3194	[illegible]	
30	2.9607	37.4433	[illegible]	
31	3.0193	41.8605	[illegible]	
85	3.1984	86.6795	[illegible]	
36	3.2001	56.6219	[illegible]	
32	3.2075	30.1753	[illegible]	
37	3.2252	56.9673	[illegible]	

analytical_sample | NMR_analyzer | JRES | COSY | TOCSY | HSQC | HMBC | carbon-13-1D_acquisition | CPMG-1D_acquisition

Comparaison des peak-lists de la matrice complexe aux peak-lists des composés de référence

PEAKFOREST (2)

- Interrogation via web services REST
- Recherche sur la (les) séquence(s) d'intérêt
- Le pH
- …
- Exemple séquence HSQC, spectres enregistrés à pH=7

  https://metabohub.peakforest.org/rest/v1/spectra/nmr2d/search?query=tocsy&token=9131jq9l8gsjn1j14t351h716u&max=500

Metabolite	ppm_f2	ppm_f1
Acetic Acid	1.9225	26.1314
Citric acid	2.5381	48.5330
Creatinine	4.0640	59.2097
D-Glucose	3.2481	76.8668
D-Glucose	3.4250	72.3072
DL-3-aminobutyric acid	3.6156	45.4543
Hippuric acid	7.8476	129.7817
Isoleucine	0.9442	14.0011

BARSA: Bi-dimensionAl nmR Spectra Annotation (1)

Script R

Pour chaque composé de référence

- Comparaison de toutes les paires de pics du composé à la liste de paires de pics de la matrice biologique
- Tolérance sur les déplacements chimiques
- Calcul d'une probabilité de présence
  - # paires de pics annotées /# paires de pics théoriques

BARSA: Bi-dimensionAl nmR Spectra Annotation (2)

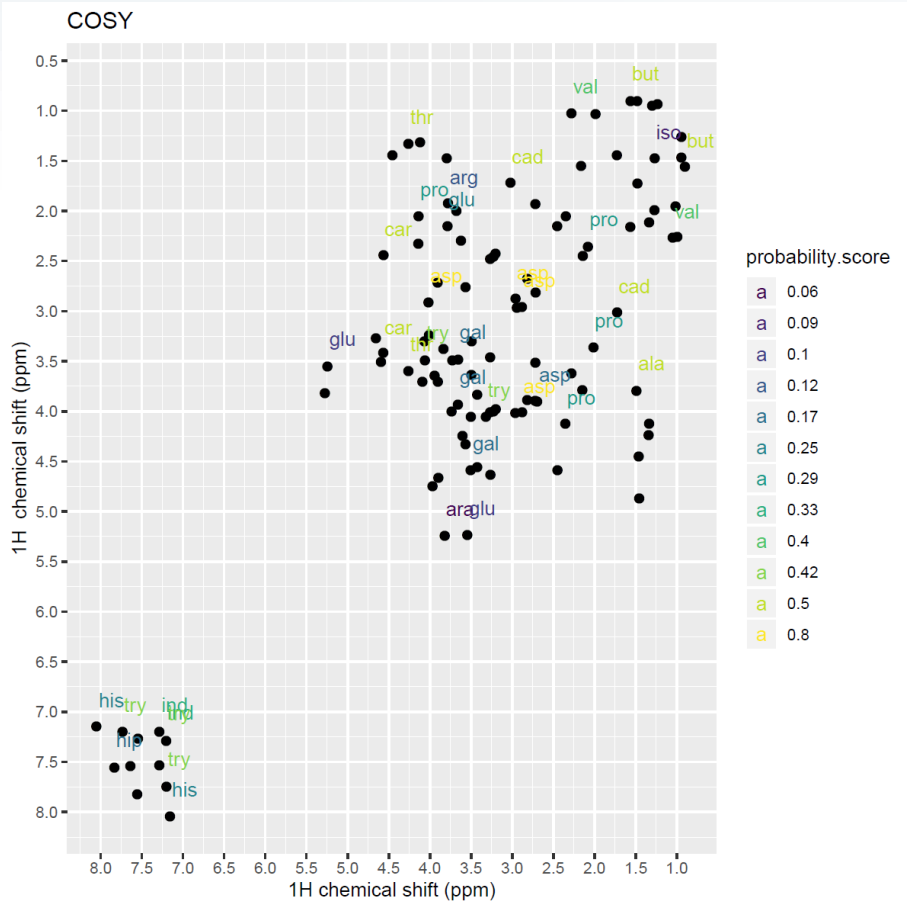
- Application de filtres : réduction du nombre de faux positifs
  - Seuil sur la probabilité de présence
    - Suppression de tous les métabolites dont la probabilité de présence < seuil
  - Condition d'unicité
    - Suppression de toutes les paires de pics assignées à plusieurs métabolites

- Combinaison: possibilité de combiner 2 (ou plus) séquences
  - Score combiné=score moyen

BARSA: Bi-dimensionAl nmR Spectra Annotation (3)

ppm1	ppm2	Metabolite	score
1.487	3.792	alanine	0.5
3.827	5.242	arabinose	0.05887
3.781	1.933	arginine	0.125
2.882	4.018	asparagine	0.8
2.882	2.964	asparagine	0.8
2.962	2.883	asparagine	0.8
4.014	2.913	asparagine	0.8
2.695	3.904	Aspartic_acid	0.1666
0.898	1.567	Butyric acid	0.5
1.566	0.898	Butyric acid	0.5
1.73	3.019	Cadaverine	0.5
3.026	1.729	Cadaverine	0.5
4.572	2.453	carnitine	0.5
4.572	3.431	carnitine	0.5
3.501	4.592	galactose	0.1666
3.659	3.478	galactose	0.1666
3.66	3.931	galactose	0.1666

ppm1	ppm2	commonMetabolitesList
1.73	3.019	Cadaverine Lysine
7.21	7.281	indoxylsulfate tryptophan
7.283	7.205	indoxylsulfate tryptophan
7.205	7.284	tryptophan indoxylsulfate



EVALUATION DE L'ALGORITHME (1)

- Matrice « complexe »: mélange de 23 composés standards connus (à la même concentration 10mMol)
- Critères de validation
  - Sensibilité : capacité à détecter la totalité des métabolites présents dans la matrice, c'est-à-dire les vrais positifs
  - Spécificité : capacité à ne pas identifier des métabolites qui ne sont pas présents dans la matrice biologique, c'est-à-dire les vrais négatifs

		MELANGE		
		Présents	Absents	Total
ALGORITHME	Annotés	Vrais positifs (VP)	Faux positifs (FP)	VP + FP
	Pas annotés	Faux négatifs (FN)	Vrais négatifs (VN)	FN + VN
	Total	23	66	89

EVALUATION DE L'ALGORITHME (2)

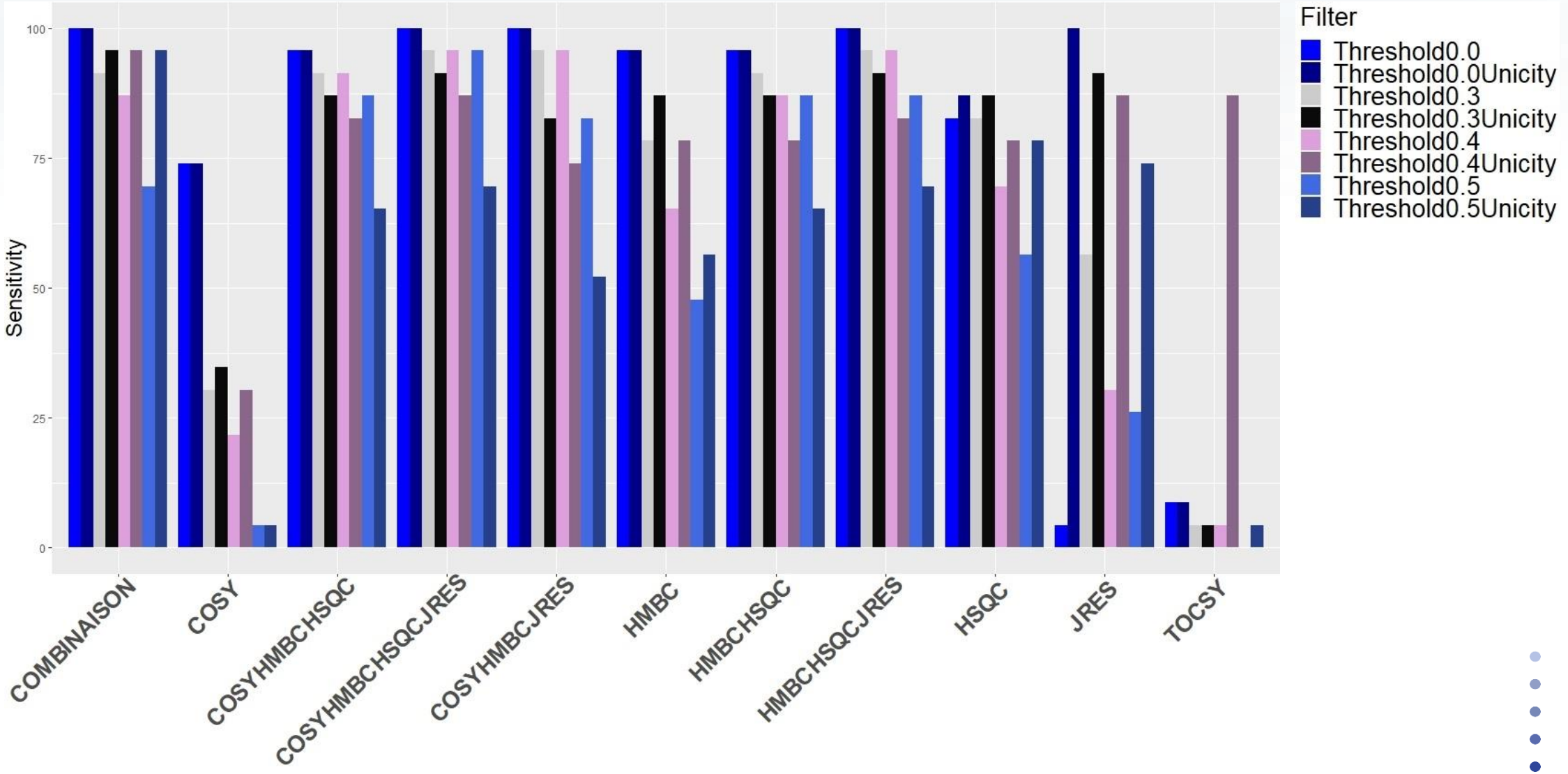
- Matrice « complexe »: urine de synthèse = matrice « blanche » à laquelle ont été ajoutés 33 composés de concentration connue mais différente (100µM – 20mM)

- Critères de validation
  - Sensibilité
  - Spécificité

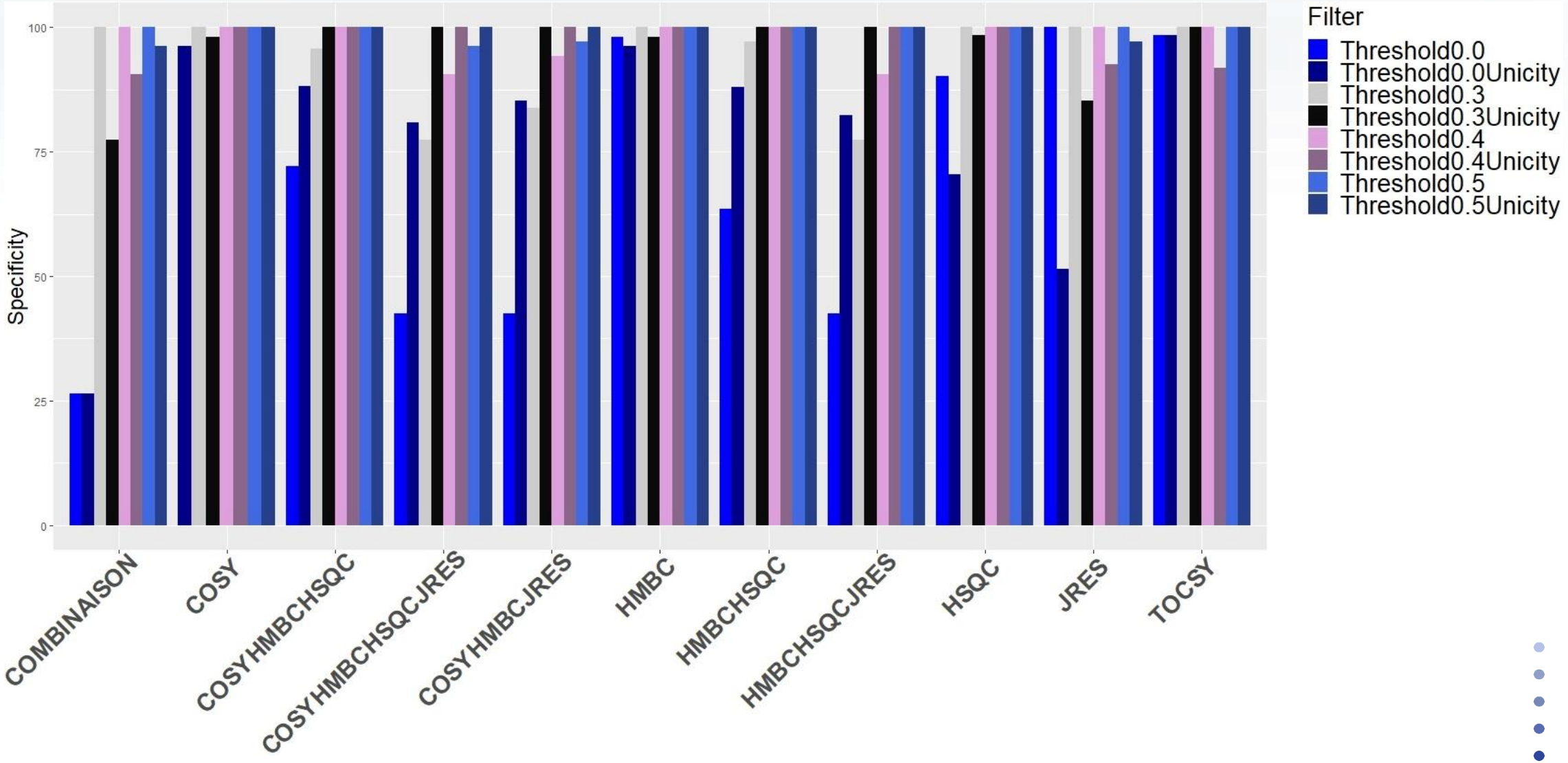
		MELANGE		
		Présents	Absents	Total
ALGORITHME	Annotés	Vrais positifs (VP)	Faux positifs (FP)	VP + FP
	Pas annotés	Faux négatifs (FN)	Vrais négatifs (VN)	FN + VN
	Total	33	66	99

RESULTATS & DISCUSSION

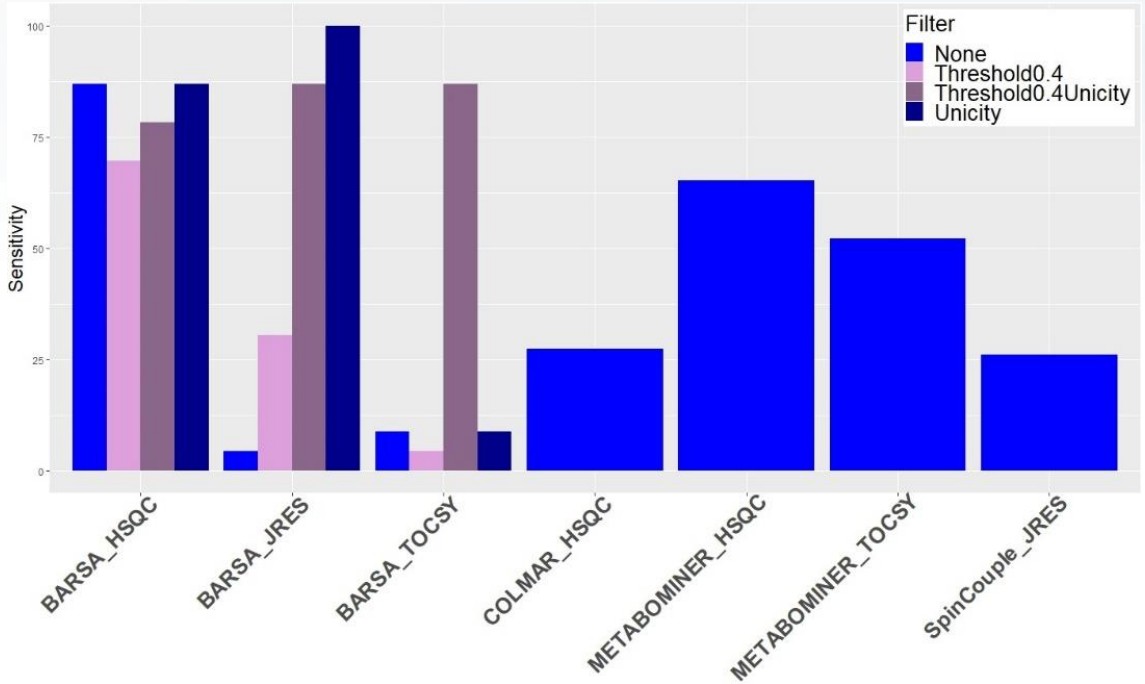
ANNOTATION MATRICE TEST – SENSIBILITE

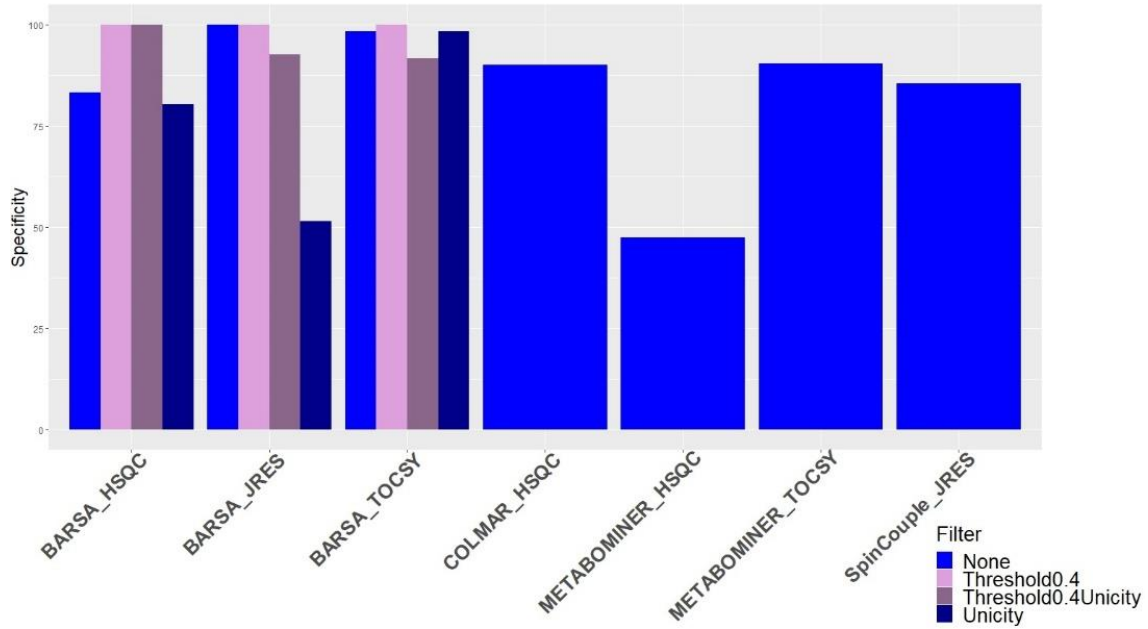


ANNOTATION MATRICE TEST – SPECIFICITE



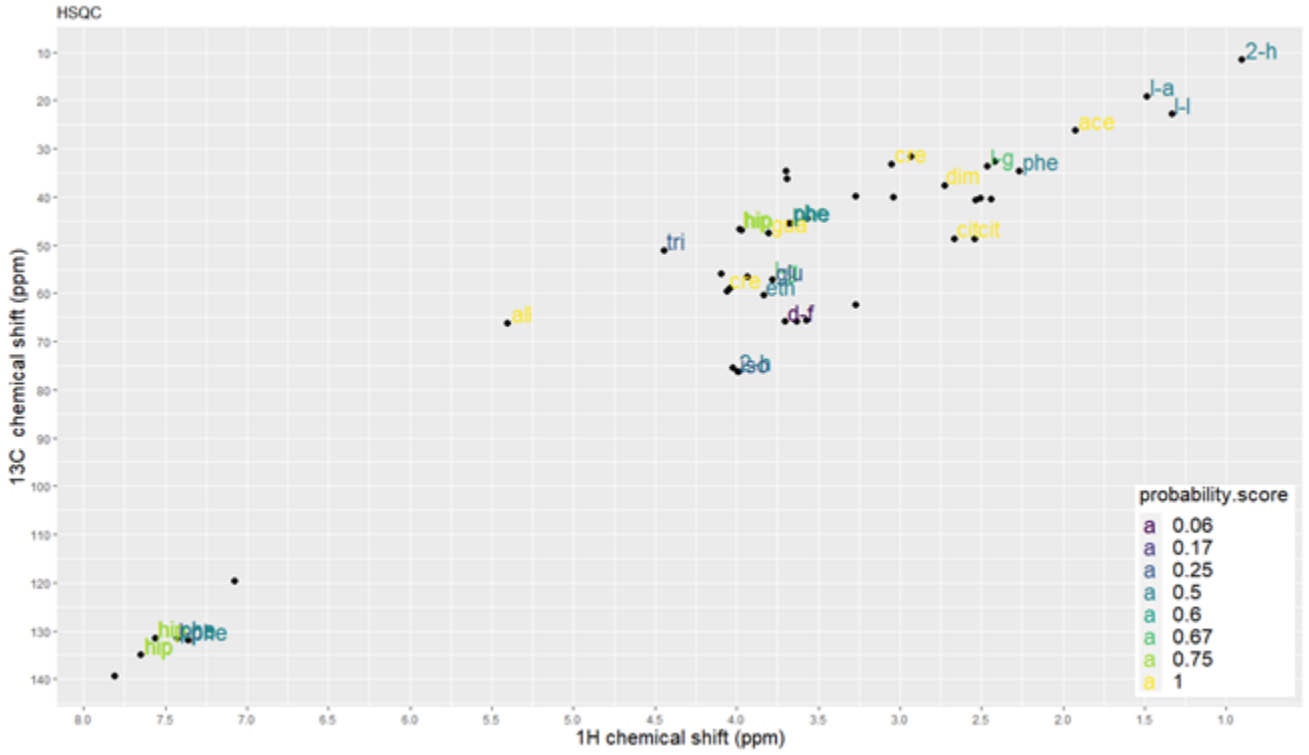
ANNOTATION MATRICE TEST – COMPARAISON AUTRES ALGORITHMES



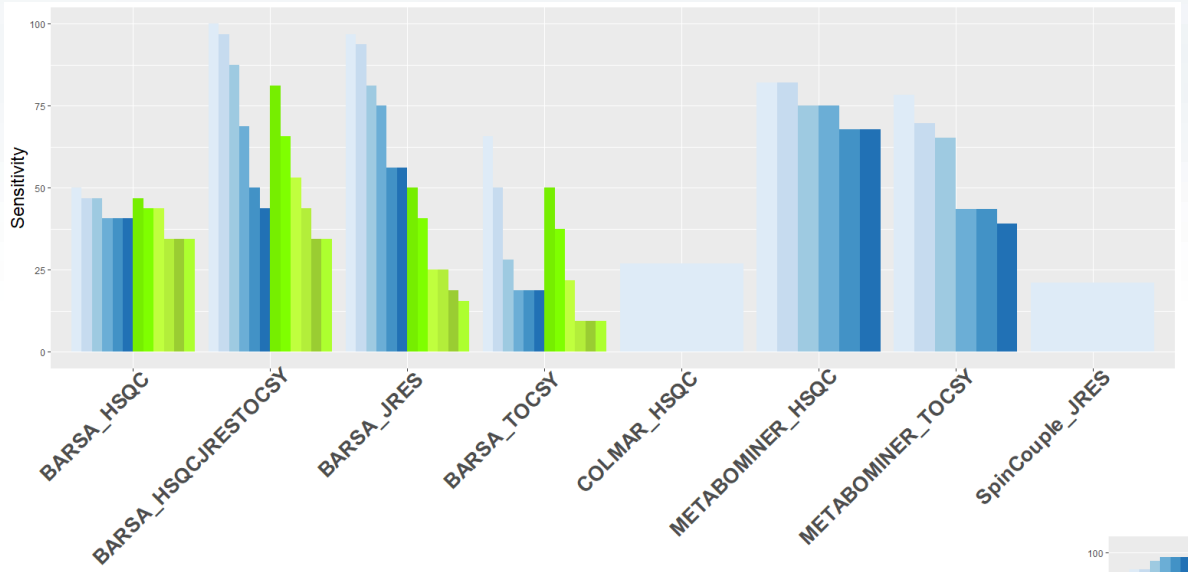


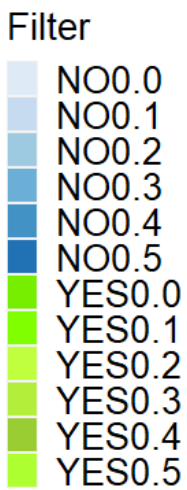
ANNOTATION URINE DE SYNTHESE – ANNOTATION BARSA

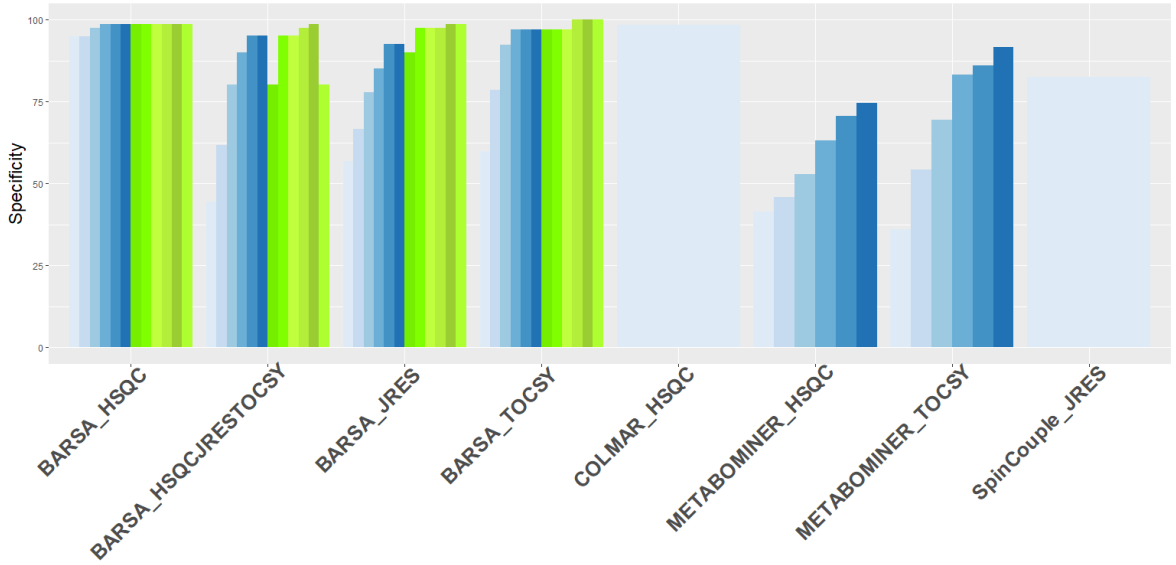
Metabolite	score.HSQC	score.JRES	score.TOCSY	averageScore
2-hydroxybutyric acid	0,5	0,692	0,313	0,502
acetic acid	1	1		1
allantoin	1			1
citric acid	1	1		1
creatinine	1	0,5		0,75
d-fructose	0,059	0,111	0,07	0,08
dimethylamine	1	1		1
ethanolamine	0,5	0,5	0,25	0,417
glutamic acid	0,25	0,167		0,208
guanidineacetic acid	1	1		1
hippuric acid	0,75	1	0,7	0,817
isocitric acid	0,25	0,938		0,594
l-alanine	0,5	0,667	0,75	0,639
l-glutamine	0,667	0,079	0,111	0,286
l-lactic acid (sodium salt)	0,5	0,333	0,75	0,528
l-phenylalanine	0,167	0,65	0,167	0,328
l-phenylalanine	0,167	0,5	0,167	0,278
phenylacetyl-l-glutamine	0,5	0,293	0,278	0,357
phenylacetylglycine	0,6	0,5		0,55
trigonelline hydrochloride	0,25	0,111	0,286	0,216
(e)-3-(1h-imidazol-5-yl)prop-2-enoic acid		0,167	0,1	0,133
3-methyl-2-oxovaleric acid		0,077	0,04	0,058
3-methyl-l-histidine		0,333	0,143	0,238
5-aminopentanoic acid		0,048	0,111	0,079
5-oxo-l-proline		0,111	0,188	0,149
acetone		1	1	1
butyric acid		0,25		0,25
carnitine		0,167	0,2	0,183
cholic acid		0,033	0,054	0,044
choline chloride		0,143	0,2	0,171
créatine		1	0,5	0,75



ANNOTATION URINE DE SYNTHESE– COMPARAISON AUTRES ALGORITHMES







Interprétation des résultats – urine de synthèse

Composition de l'urine de synthèse (33 composés) :

- **7 composés avec une conc > 1 mM** : creatinine ; citrate ; hippurate ; ascorbate ; glycine ; TMAO ; guanidoacetate
- **7 composés avec 0,5 mM < conc < 1 mM** : histidine ; isocitrate; 2-hydroxybutyrate ; phenylacetylglutamine ; creatine; acetate; dimethylamine
- **8 composés avec 0,25 mM < conc < 0,5 mM** : cysteine; glutamine; glucose; pyroglutamate; serine; ethanolamine; alanine; formate
- **11 composés avec conc < 0,25 mM** : glycerol; allantoin; indoxylsulfate; myoinositol; lysine; trigonelline; 1-methylhistidine; 3-methylhistidine; threonine; lactate; fructose
- **TMAO** n'est pas dans la base de données

Résultats sans filtre

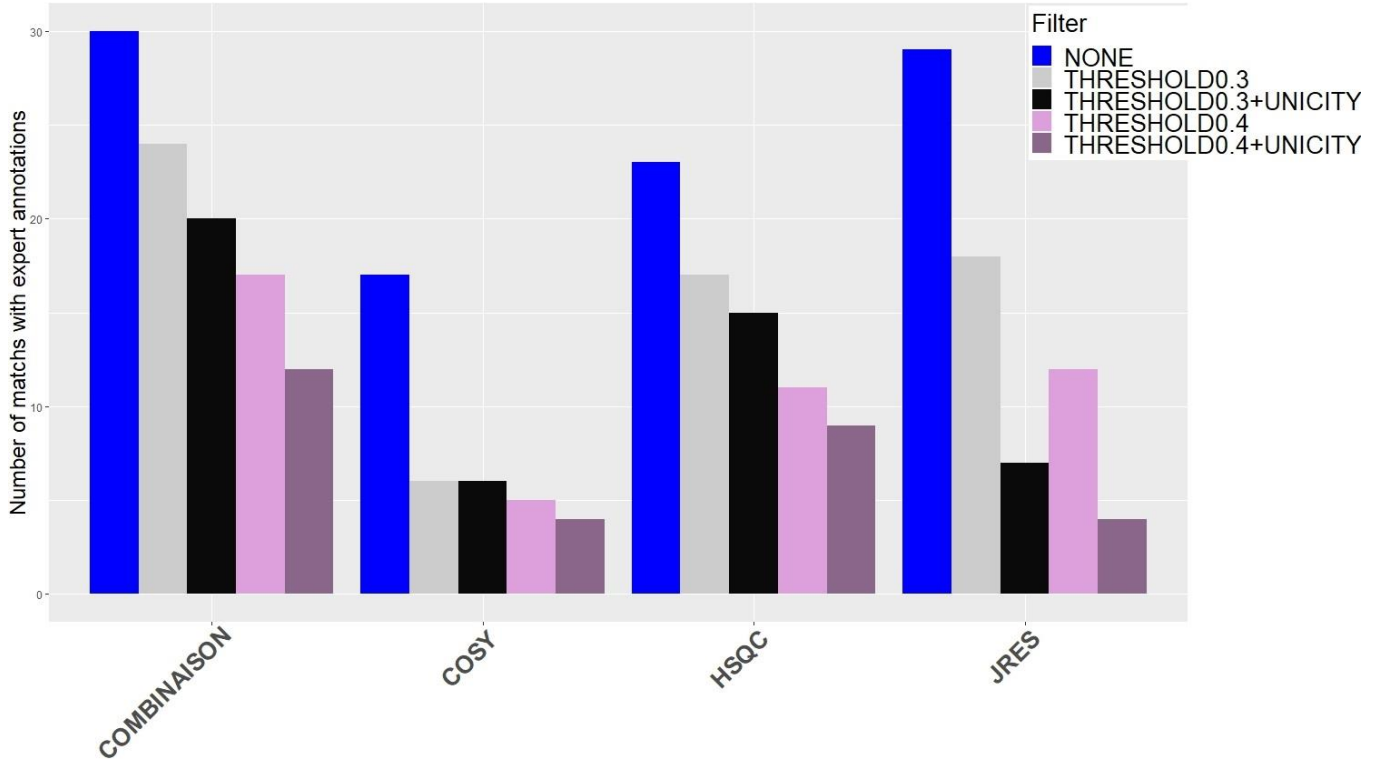
- 77 métabolites détectés (sur 99 composés présents dans la base)
- 32 métabolites présents dans l'urine
- 45 faux positifs → **Non acceptable**

Résultats avec un score moyen supérieur à 0,3 / 0,15

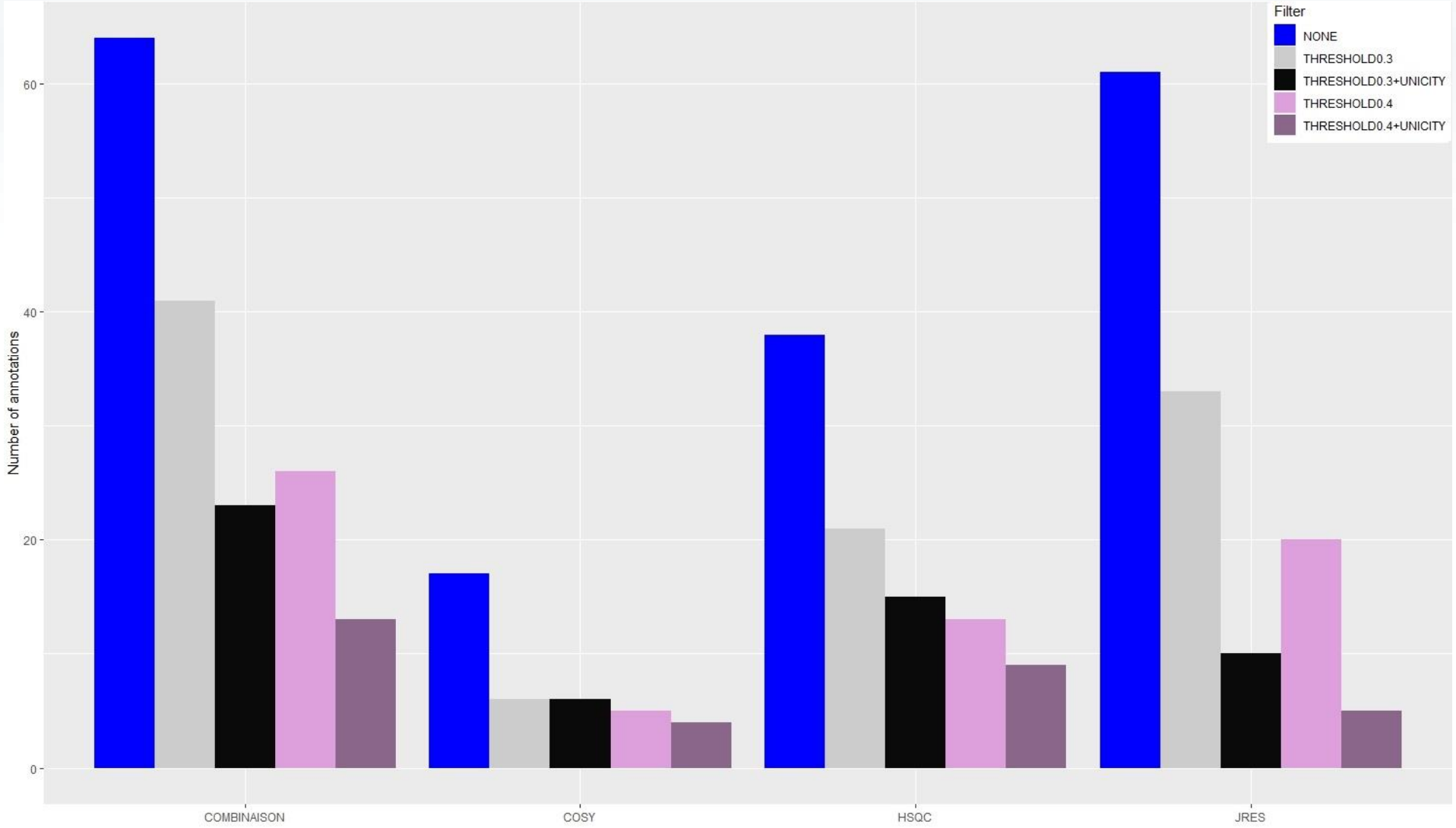
- 31 métabolites détectés / 30 métabolites détectés
- 23 composés dans l'urine / 24 métabolites détectés
- 8 faux positifs / 6 faux positifs → acetone, sarcosine, phenylacetylglycine, phenylalanine (x2), betaine
- 9 composés non trouvés par l'algo : glucose, glutamine, 3-methylhistidine, histidine, trigonelline, cysteine, lysine, pyroglutamate, 1-methylhistidine → conc < 0,5 mM sauf histidine / 8 composés non trouvés dans l'algo : cysteine, ascorbate, lysine, serine, pyroglutamate, fructose, 1-methylhistidine → conc < 0,5 mM
- Ascorbate : non retrouvé en manuel (réducteur)
- Cysteine : transformée en partie en cystine

ANNOTATION PLASMA DE REFERENCE (1)

- Plasma de reference humain (NIST SRM 1950) : utilisation de la COSY, HSQC, la JRes et une combinaison des 3 séquences
- Comparaison métabolites annotés automatiquement à la liste des métabolites annotés par les experts MTH (40)



ANNOTATION PLASMA DE REFERENCE (1)



DISCUSSION/CONCLUSION

- Algorithme développé permet l'annotation automatique de spectres RMN 2D de matrices biologiques complexes

- Réduction du nombre de faux positifs
  - Combinaison de plusieurs séquences
  - Application de seuil sur la probabilité de présence
  - Application condition d'unicité

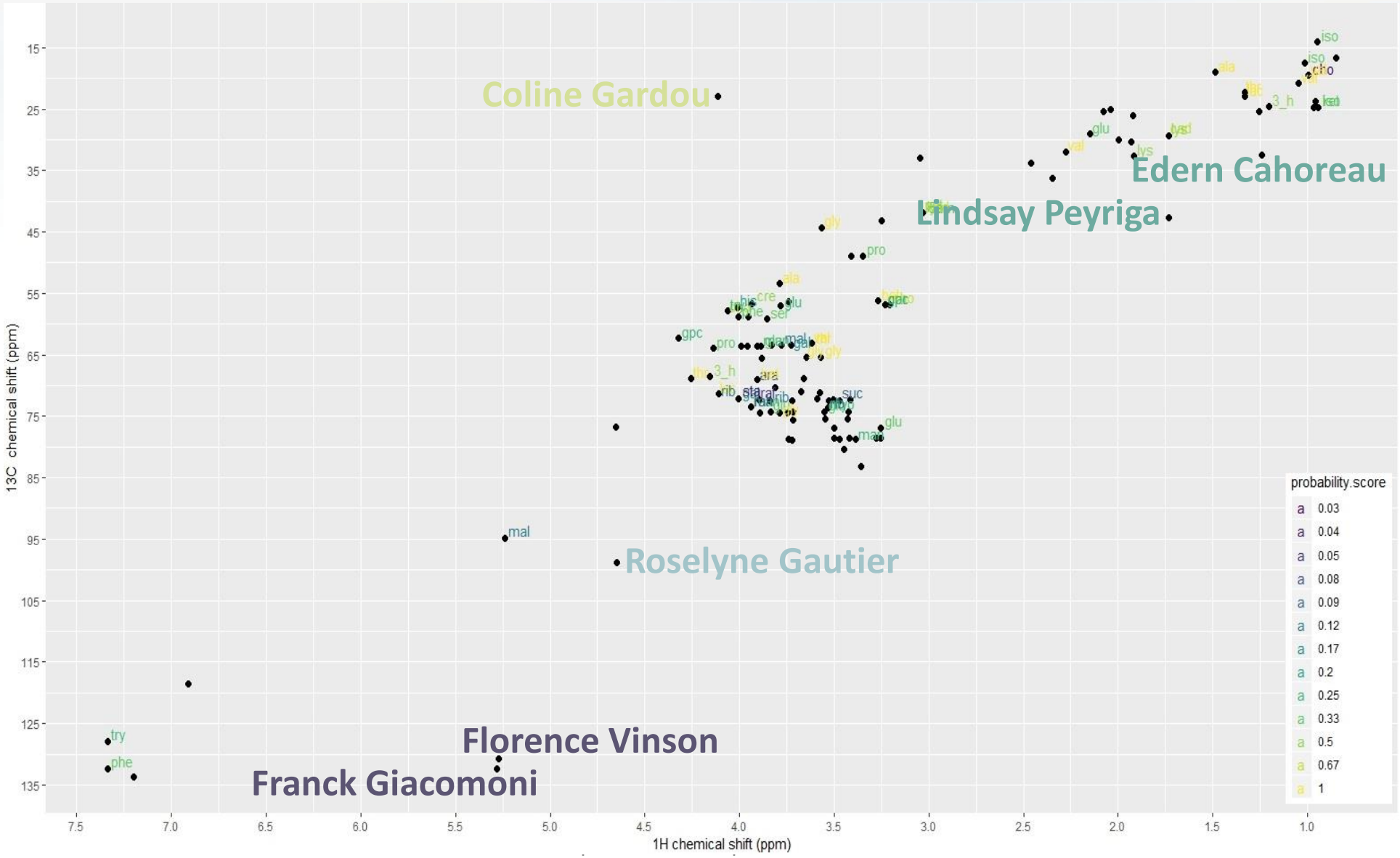
PERSPECTIVES

- Autres matrices à tester
  - Plasma NIST: vérification annotation par experts
  - Extraits de tissus
  - E. coli

- Lien avec MetExplore

- Test de l'algorithme avec la séquence SAPPHIRE-PSYCHE, dans laquelle les multiplicité des pics sont transformées en raies uniques pour améliorer l'annotation des signaux en diminuant la superposition des signaux

REMERCIEMENTS







MERCI DE VOTRE ATTENTION
Des questions?