Étude de la diversité chimique des lichens par LC-MS : acquisition et optimisation du traitement des données métabolomiques

Damien Olivier-Jimenez

▶ **To cite this version:**

Damien Olivier-Jimenez. Étude de la diversité chimique des lichens par LC-MS : acquisition et optimisation du traitement des données métabolomiques. Bio-informatique [q-bio.QM]. Université Rennes 1, 2021. Français. NNT : 2021REN1S030 . tel-03376459

HAL Id: tel-03376459
https://theses.hal.science/tel-03376459

Submitted on 13 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
Faculté des Sciences pharmaceutiques et biologiques

ECOLE DOCTORALE N° 596
Matière, Molécules, Matériaux
Spécialité : *Chimie Moléculaire et Macromoléculaire*

Par
Damien OLIVIER-JIMENEZ

Etude de la diversité chimique des lichens par LC-MS^n :

acquisition et optimisation du traitement des données métabolomiques

VOLUME 1

Thèse présentée et soutenue à Rennes, le 15/01/2021
Unité de recherche : UMR CNRS 6226, équipe COrInt

Rapporteurs avant soutenance :

Marion Millot	Maître de Conférences à l'Université de Limoges
Samuel Bertrand	Maître de Conférences à l'Université de Nantes

Composition du Jury :

Président :		
Examinateurs :	Carlos Afonso	Professeur à l'Université de Rouen
	Jean-Luc Wolfender	Professeur à l'Université de Genève
	Joëlle Quetin-Leclercq	Professeure à l'Université Catholique de Louvain
	Marion Millot	Maître de Conférences à l'Université de Limoges
	Samuel Bertrand	Maître de Conférences à l'Université de Nantes
Dir. de thèse :	Joël Boustie	Professeur à l'Université de Rennes 1
Co-dir. de thèse :	David Rondeau	Professeur à l'Université de Rennes 1
Co-enc. de thèse :	Marylène Chollet-Krugler	Maitre de conférences à l'Université de Rennes 1

Etude de la diversité chimique des lichens par LC-MS^n

Sommaire

Productions scientifiques iii

Liste des Figures & Tableaux iv

Introduction 1

Glossaire & recontextualisation 14

Chapitre I – LDB-Lit 47

Chapitre II – LDB 70

Chapitre III – Extension de la LDB 91

Chapitre IV – LDB-MotifDB & *Classnotator* 110

Chapitre V – *Molnotator* 128

Chapitre VI – Diversité chimique des lichens 157

Conclusion 188

Bibliographie 201

Production scientifique

Publications :

Damien Olivier-Jimenez, Marylène Chollet-Krugler, David Rondeau, Mehdi A. Beniddir, Solenn Ferron, Thomas Delhaye, Pierre-Marie Allard, Jean-Luc Wolfender, Harrie J. M. Sipman, Robert Lücking, Joël Boustie, Pierre Le Pogam. A database of high-resolution MS/MS spectra for lichen metabolites. *Scientific Data* (2019) 6:294

Solenn Ferron, Olivier Berry, Damien Olivier-Jimenez, Isabelle Rouaud, Joël Boustie, Françoise le Dévéhat, Rémy Poncet, Chemical diversity of five coastal *Roccella* species from mainland France, the Scattered Islands, and São Tomé and Príncipe. *Plant and Fungal Systematics* (**Soumise & acceptée**).

Communications orales :

Method setup and optimisation for the study of lichen secondary metabolites
Journée des Doctorants 2018 (Rennes, France, 18 Juin 2018).

Method setup and optimisation for the study of lichen secondary metabolites
3ème Symposium international AFERP & STOLON (Rennes, France, 18-20 Juillet 2018).

Lichen Metabolites Database, A modern MS/MS library for lichen metabolites
Journée Pharmacie Recherche (Rennes, France, 16 Novembre 2019).

Lichen Database, an MS/MS library for lichen metabolites
67th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Product Research (GA), (Innsbruck, Autriche, 1-5 Septembre 2019).

Lichen Database, an MS/MS library for lichen metabolites
Indo-French Joint Laboratory for Natural Products and Synthesis towards Affordable Health (NPSAH), (Rennes, France, 30 Octobre 2019).

Liste des Figures & Tableaux

Liste des Figures :

Fig.	Description	Page	Fig.	Descrpition	Page
1	*Diversité morphologique des lichens*	2	2	*Diversité morphologique des lichens*	3
3	*Diversité morphologique des lichens*	4	4	*Diversité morphologique des lichens*	5
5	*Couleurs obtenues à partir de lichens, extrait de Svenska Lafvarnas Färghistoria (1805, Johan Peter Westring).*	7	6	*Origine biosynthétique des groupes structuraux répertoriés dans les lichens.*	9
7	*Classification produite par ClassyFire sur deux molécules lichéniques : l'acide polyporique et l'atranorine.*	18	8	*Codes InChI et InChIKey pour la palytoxine ($C_{129}H_{223}N_3O_{54}$).*	22
9	*Spectre MS/MS individuel pour l'acide olivétolique au format MGF.*	27	10	*Extrait de données LC-MS/MS par DDA au format mzXML.*	29
11	*Exemple de réseau moléculaire avec des attributs.*	32	12	*Code SMILES pour la palytoxine ($C_{129}H_{223}N_3O_{54}$).*	33
13	*Un graphe et de réseau.*	34	14	*Exemple d'un Feature-Based Molecular Networking (FBMN) utilisant MZmine, le GNPS et Cytoscape.*	38
15	*Réseau représentant une classification taxonomique allant de la classe au genre, avec pour chaque taxon, le nombre de molécules qui y ont été décrites dans la littérature.*	39	16	*Réseau représentant les molécules d'une base de données classée par ClassyFire avec pour chaque classe / nœud, le nombre de molécules qui y ont été répertoriées.*	40
17	*Réseau de fragmentation, formant des blocs à partir d'ions précurseurs et leurs fragments de source.*	40	18	*Réseau produit par Adnotator, regroupant tous les ions produits par une molécule neutre hypothétique.*	41
19	*Bloc de fragmentation généré par Fragnotator.*	42	20	*Création d'une Hypothèse de relation.*	44
21	*Regroupement des Hypothèses de relation dépendantes en Cohortes.*	45	22	*Croissance attendue du nombre de molécules d'origine lichénique telle que décrite par Culberson* (C. F. Culberson and Culberson 2001).	50
23	*Echantillon du tableau contenant les données sur LIAS.*	52	24	*Echantillon du tableau produit à partir des données du LIAS avec, pour chaque métabolite lichénique, la liste des organismes producteurs.*	52
25	*Classes chimiques présentées dans Hun&Yosh96 avec le nombre de molécules présentes. FDD : Fragments de Depsides et de Depsidones.*	56	26	*Représentation sous forme de réseaux de la diversité chimique des lichens en utilisant des données du LIAS, de Hun&Yosh96 et de la LDB-Lit.*	59
27	*Détails des quatre principaux blocs de la LDB-Lit pour lesquels des structures représentatives de certains nœuds ont été représentées.*	61	28	*Réseaux présentant le nombre de molécules à chaque niveau taxonomique pour les données de la LDB-Lit (A) et du LIAS (B).*	64
29	*Réseaux représentant le nombre de métabolites décrits par taxon dans l'ordre des Lécanoromycètes pour les données de la LDB-Lit (A) et du LIAS (B).*	65	30	*Détails des principaux ordres pour lesquels des données ont été rajoutées par la LDB-Lit.*	66
31	*Déréplication à l'aide de la LDB dans le cadre d'un FBMN.*	73	32	*Métriques et caractéristiques de la LDB.*	79
33	*Réseau moléculaire généré à partir des extraits acétone d'Ophioparma ventosa, Evernia prunastri et Hypogymnia physodes par FBMN.*	82	34	*Résultat de la déréplication des extraits d'Ophioparma ventosa, Evernia prunastri et Hypogymnia physodes.*	83
35	*Représentation en miroir de différents spectres MS/MS d'Ophioparma ventosa.*	85	36	*Représentation en miroir de spectres MS/MS d'Evernia prunastri vis-à-vis de composés proches dans la LDB.*	86
37	*Réprésentation en miroir de l'acide physodalique de la LDB contre un ion à m/z 417 d'Hypogymnia physodes, soupçonné être l'acide conphysodalique.*	88	38	*Traitement des fichiers bruts mzXML sur R.*	96
39	*Création du spectre consensus de l'acide glomellique sous la forme $[M-H]^-$ (Xevo G2-XS).*	97	40	*Vérification manuelle des spectres sur la base de leur similarité d'un appareil à l'autre.*	97

41	*Nombre de spectres détectés pour chaque catégorie d'adduit pour chaque instrument analytique utilisé en mode positif (A) et négatif (B).*	99	42	*Pourcentage de spectres identifiés en les comparant d'un instrument à l'autre avec un seuil de similarité cosinus à 0.7.*	101
43	*Efficacité d'identification des spectres de la LDB par similarité cosinus en fonction du seuil fixé.*	102	44	*Pourcentage de spectres reconnus ou non au sein de la LDB.*	103
45	*Représentation de la diversité en adduits de la LDB et des librairies du GNPS suivant le mode d'ionisation.*	105	46	*Impact de l'inadéquation de diversité d'adduits lors de la déréplication, exemple de la LDB.*	108
47	*Analogie entre LDA pour le texte et MS2LDA (extrait de J. J. J. van der Hooft et al., 2016).*	113	48	*Exemple d'annotation MS2LDA sur un réseau moléculaire.*	114
49	*Obtention des motifs purs à partir des motifs de la LDB.*	118	50	*Algorithmes utilisés pour évaluer le partage des motifs entre adduits (A) et entre instruments (B).*	119
51	*Couverture de la LDB par les motifs purs dans les deux modes d'ionisation.*	123	52	*Exemples de blocs moléculaires générés par Molnotator.*	131
53	*Fonctionnement général de Molnotator.*	133	54	*Fonctionnement de Fragnotator.*	136
55	*Fonctionnement général d'Adnotator.*	139	56	*Exemple d'une Hypothèse de relation.*	140
57	*Regroupement des Hypothèses de relation dépendantes en Cohortes.*	141	58	*Annotation des ions/nœuds par Classnotator.*	143
59	*Déréplication orientée par la LDB de données LC-MS/MS annotées.*	144	60	*Blocs de fragmentation correspondant aux ions de l'acide glomellique.*	147
61	*Blocs moléculaires de l'acide glomellique après traitement par Adnotator.*	149	62	*Bloc moléculaire de l'acide glomellique après traitement par Mixmoder.*	150
63	*Comparaison des réseaux générés par similarité cosinus au réseau de polarité mixte généré par Molnotator.*	151	64	*Résultats de Molnotator sur la prédiction des nœuds neutres et les identifications par Classnotator.*	152
65	*Blocs moléculaires du diacétylpyxinol à partir de l'analyse de son standard puis du standard d'une autre molécule, mettant en évidence d'autres formes d'ionisation.*	154	66	*Fonctionnement général de File Merger.*	164
67	*Réseau final sous une forme simplifiée ne représentant que les neutres et les adduits.*	165	68	*Distribution des adduits en mode négatif.*	166
69	*Distribution des adduits en mode positif.*	166	70	*Données sur le nombre d'adduits et la déréplication des neutres prédits.*	167
71	*Distribution des neutres dans les classes prédites par Classnotator.*	168	72	*Résultats bruts de l'analyse d'Evernia prunastri (S025).*	169
73	*Neutres remarquables du bloc 11 d'Evernia prunastri.*	171	74	*Autres nœuds remarquables d'Evernia prunastri.*	172
75	*Neutres remarquables du bloc 1 d'Evernia prunastri.*	173	76	*Neutres remarquables du bloc 4 d'Evernia prunastri.*	174
77	*Neutres remarquables du bloc 9 d'Evernia prunastri.*	175	78	*Neutres remarquables du bloc 12 d'Evernia prunastri.*	176
79	*Neutres remarquables du bloc 15 d'Evernia prunastri.*	176	80	*Neutres remarquables du bloc 18 d'Evernia prunastri.*	177
81	*Résultats bruts de l'analyse de Cladonia gracilis (S015).*	178	82	*Neutres remarquables du bloc 1 de Cladonia gracilis.*	180
83	*Neutres remarquables du bloc 3 de Cladonia gracilis.*	181	84	*Neutres remarquables du bloc 4 de Cladonia gracilis.*	182
85	*Neutres remarquables du bloc 5 de Cladonia gracilis.*	183	86	*Neutres remarquables du bloc 55 et 59 de Cladonia gracilis.*	183
87	*Neutres remarquables du bloc 141,190 et 229 de Cladonia gracilis.*	184	88	*Autres neutres notables de Cladonia gracilis non interprétés (blocs 6, 7 et 8).*	184
89	*Autres neutres notables de Cladonia gracilis non interprétés (blocs 11, 14 et 20).*	185	90	*Autres neutres notables de Cladonia gracilis non interprétés (blocs 30, 31 et 46).*	185
91	*Requêtes à développer pour la LDB-Lit*	190	92	*Impact des énergies de collision de différents instruments LC-MS sur la fragmentation de la leucine-enképhaline, représenté par le rendement de survie SY.*	191
93	*Exemples d'applications de données LC-MS/MS à la taxonomie.*	195			

Liste des Tableaux :

Tab.	Description	Page	Tab.	Description	Page
1	*Quelques classes utilisées par ClassyFire et leurs définitions, traduites de l'anglais.*	18	2	*Exemple d'un tableau d'adduits.*	43
3	*Paramètres utilisés pour générer le réseau moléculaire à partir des spectres de la LDB.*	75	4	*Lichens utilisés pour la validation technique dans l'herbier de Rennes.*	76
5	*Paramètres MZmine utilisés pour traiter les fichiers LC-MS/MS des trois lichens.*	76	6	*Paramètres utilisés pour générer le réseau moléculaire à partir des trois lichens.*	77
7	*Métabolites annotés sur le réseau moléculaire.*	84	8	*Adduits recherchés avec les paramètres nécessaires au calcul de leur rapport m/z.*	95
9	*Adduits recherchés suivant le mode d'ionisation et le nombre de spectres trouvés pour chacun.*	100	10	*Adduits considérés pour la LDB et les librairies du GNPS.*	106
11	*Exemple d'un tableau molmotif.*	116	12	*Exemple de tableau Motifs x Classes répertoriant le nombre de fois que chaque motif a été détecté dans chaque classe.*	117
13	*Motifs purs retrouvés pour chaque classe chimique en fonction du mode d'ionisation.*	120	14	*Répartition des motifs purs du mode négatif en fonction du nombre de combinaisons successives et du nombre de molécules dans lesquelles elles ont été décrites.*	121
15	*Répartition des motifs purs du mode positif en fonction du nombre de combinaisons successives et du nombre de molécules dans lesquelles elles ont été décrites.*	121	16	*Classes structurales et leurs motifs pur, à l'exception de ceux pour lesquels n_combi = 2 et n_mol = 1.*	122
17	*Motifs retrouvés entre adduits en mode négatif.*	124	18	*Motifs retrouvés entre adduits en mode positif.*	124
19	*Motifs retrouvés entre instruments LC-MS.*	125	20	*Paramètres MZmine.*	134
21	*Paramètres utilisés pour l'algorithme combinatoire.*	138	22	*Adduits produits en mode positif par l'algorithme combinatoire.*	138
23	*Adduits produits en mode négatif par l'algorithme combinatoire.*	139	24	*Adduits retenus en mode négatif et positif.*	145
25	*Liste d'adduits réutilisés par Adnotator en mode négatif.*	146	26	*Liste d'adduits réutilisés par Adnotator en mode positif.*	146
27	*Récapitulatif du traitement de la LDB-Orbitrap par Molnotator.*	151	28	*Paramètres MZmine.*	161
29	*Liste d'adduits utilisés par Adnotator en mode négatif.*	163	30	*Liste d'adduits utilisés par Adnotator en mode positif.*	163
31	*Déréplication guidée (automatique) des molécules prédites dans Evernia prunastri.*	169	32	*Déréplication guidée (automatique) des molécules prédites dans Cladonia gracilis.*	178
33	*Composition chimique de quatre chimiotypes de Pseudevernia et la détection de ces molécules dans quatre échantillons analysés ici.*	194			

Introduction

Sommaire

1 – Lichens, naturalisme et chimie aux XVIIIe et XIXe siècles 6

2 – L'étude moderne des lichens aux XXe 7

3 – L'impact de la génomique 10

4 – Avènement de la métabolomique 10

5 – Les lichens et la métabolomique & la contribution de ces travaux 11

Les lichens sont des formes de vie ubiquitaires, se développant sur divers substrats, comme la roche (saxicoles), le sol (terricoles) ou le bois (lignicoles), dont l'existence est traçable jusqu'à il y a 600 millions d'années (Yuan, Xiao, and Taylor 2005). De par les spécificités de certains vis-à-vis de leur environnement, ils sont utilisés pour évaluer la qualité de l'air. Ils correspondent à un mode de vie adopté par certains champignons : une symbiose comprenant un partenaire fongique, hétérotrophe par nature, et un ou plusieurs organismes autotrophes (algue unicellulaire ou cyanobactérie). La classification des lichens repose sur celle du partenaire fongique ou mycobiote, les photobiotes n'étant pas capables de reproduction sexuée dans cette symbiose. Comme souvent pour les relations symbiotiques, ce mode de vie est apparu plusieurs fois dans l'histoire évolutive des champignons. Les champignons lichénisés forment donc un groupe hétérophylétique, représenté notamment à 99.1% par des Ascomycètes, les Basidiomycètes ne comptant que 0.9% des 19 387 espèces actuellement recensées (Lücking, Hodkinson, and Leavitt 2017). Ces critères confèrent aux lichens une importante variabilité morphologique.

A : *Lichenomphalia umbellifera* (L.) Redhead, Lutzoni, Moncalvo & Vilgalys, Lille, France. Basidiolichen terricole, thalle crustacé à chlorococcales, produisant des sporophores classiques de basidiomycète.[1]

B : *Rhizocarpon geographicum (L.) DC.*, presqu'île de Crozon, France. Lichen crustacé saxicole, apothécies noires incrustées dans le thalle.[2]

C : *Ochrolechia frigida (Sw.) Lynge,* Hraunhafnartangi, Islande. Lichen à thalle crustacé verruqueux, présentant de nombreuses apothécies orange et produisant de l'acide gyrophorique et lécanorique.[4]

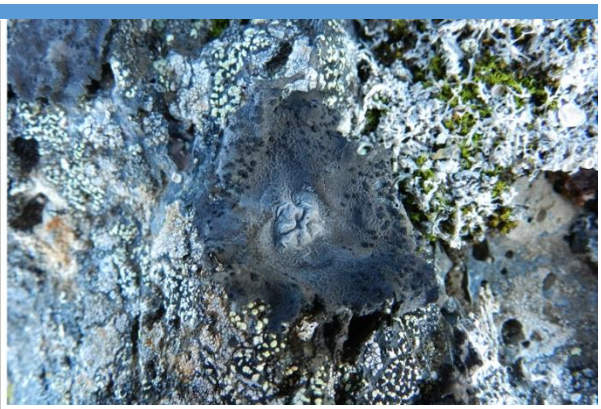
D : *Teloschistes chrysophthalmus (L.) Beltr.* Lichen à thalle fruticuleux, lignicole produisant de nombreuses apothécies.[3]

Figure 1 – Diversité morphologique des lichens. Références bibliographiques : 1 – Lichens des sols (Van Haluwyn et al. 2012), 2 – Lichens des roches (Asta et al. 2016), 3 – Lichens des arbres (Van Haluwyn, Asta, and Gavériaux 2013), 4 – Íslenskar fléttur (Kristinsson 2016).

Le thalle lichénique peut être divisé en deux parties : le thalle végétatif et les structures reproductrices. Le thalle adopte une certaine variété de formes, comme les exemples suivants :

- Thalle crustacé, formant une croûte très adhérente au substrat, comme *Rhizocarpon geographicum, Ochrolechia frigida* (**Fig. 1-B, C**) ou *Placopsis gelida* (**Fig. 4-B**).
- Thalle foliacé, formant des lames rappelant des feuilles, comme *Umbilicaria cylindrica*, *Umbilicaria proboscidea* (**Fig. 2-A, B**), *Lasallia pustulata*, *Peltigera membranacea* (**Fig. 3-B, D**) ou *Peltigera leucophlebia* (**Fig. 4-A**).
- Thalle fruticuleux, d'un aspect buissonnant, comme *Teloschistes chrysophthalmus* (**Fig. 1-D**), *Cladonia portentosa*, *Sphaerophorus globosus* (**Fig. 2-C, D**), *Roccella fuciformis*, *Ramalina cuspidata* (**Fig. 3-A, C**).

A : *Umbilicaria cylindrica* (L.) Delise, alentours de Mývatn, Islande. Lichen à thalle foliacé présentant de nombreuses apothécies noires avec des plis concentriques. Ce lichen présente des réactions variables à la potasse (K) et est décrit parfois avec de l'acide norstictique[2], ou sans, notamment dans les spécimens islandais.[4]



B: *Umbilicaria proboscidea* (L.) Schrad., alentours de Mývatn, Islande. Lichen foliacé avec le même type d'apothécies qu'*Umbilicaria cylindrica*, décrit le plus souvent avec de l'acide gyrophorique et norstictique, plus rarement qu'avec de l'acide gyrophorique.[4]

C : *Cladonia portentosa* (Dufour) Coem., Landes de Dreffeac, France. Thalle complexe à chlorococcales.[1]

D : *Sphaerophorus globosus* (Huds.) Vain., Hraunhafnartangi, Islande. Lichen fruticuleux contenant une algue du genre *Trebouxia*, ainsi que plusieurs composés : la sphérophorine, l'acide squamatique, l'acide hypothamnolique et/ou l'acide thamnolique.[4]

Figure 2 – Diversité morphologique des lichens (suite). Références bibliographiques : 1 – Lichens des sols (Van Haluwyn et al. 2012), 2 – Lichens des roches (Asta et al. 2016), 3 – Lichens des arbres (Van Haluwyn, Asta, and Gavériaux 2013), 4 – Íslenskar fléttur (Kristinsson 2016).

Les modes de reproduction sont aussi variables. Les organes de reproduction sexuée les plus communs sont les apothécies, en forme de coupe, permettant la libération de spores à maturité, bien observables ici sur *Ochrolechia frigida* et *Teloschistes chrysophthalmus*. Elles sont aussi présentes chez *Umbilicaria cylindrica* et *proboscidera* mais sous une forme plus atypique, présentant des plis plutôt qu'une surface plane. Dans le cas de *Lichenomphalia umbilifera*, la structure reproductrice est semblable à celle d'un Basidiomycète non lichénisé (**Fig. 1-A**).

A : *Roccella fuciformis* (L.) DC., presqu'île de Crozon, France. Lichen saxicole du littoral à thalle fruticuleux, avec une algue du genre *Trentepohlia*. Forme de reproduction végétative : soralies blanches aux bords des lanières.[2]

B : *Lasallia pustulata* (L.) Mérat, plateau du Landonnais, France. Lichen foliacé saxicole, dont la face supérieure est couverte de pustules et les bords d'isidies coralloïdes noires.[2]

C : *Ramalina cuspidata* (Ach.) Nyl., presqu'île de Crozon, France. Thalle fruticuleux à chlorococcales, littoral, apothécies, pycnides à ostiole noire.[2]

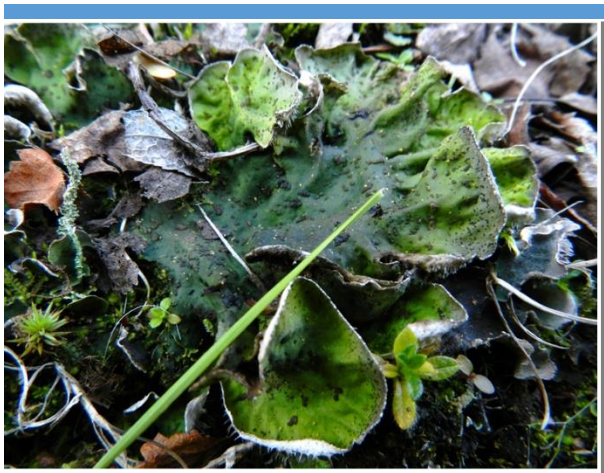
D : *Peltigera membranacea* (Ach.) Nyl., Þingvellir, Islande. Lichen foliacé terricole contenant des cyanobactéries du genre *Nostoc*[1]. Ce lichen ne réagit pas aux réactifs utilisés en lichénologie, et par conséquent, comme souvent pour les cyanolichens, n'a pas de molécules décrites.[4]

Figure 3 – Diversité morphologique des lichens (suite). Références bibliographiques : 1 – Lichens des sols (Van Haluwyn et al. 2012), 2 – Lichens des roches (Asta et al. 2016), 3 – Lichens des arbres (Van Haluwyn, Asta, and Gavériaux 2013), 4 – Íslenskar fléttur (Kristinsson 2016).

Les modes de reproduction asexuée sont communs chez les lichens. Certains développent des soralies, zones où le thalle s'ouvre pour exposer des filaments mycéliens entourant des cellules du photosymbiote. Ces soralies peuvent alors être disséminées par anémochorie, ou dissémination par le vent, et reformer un thalle entier. Elles peuvent ici être observées sur les bords des lanières de *Roccella fuciformis*. Les isidies sont une autre

forme de reproduction végétative, cette fois sous forme d'aiguilles qui auraient également un rôle métabolique en élargissant les surfaces d'échange gazeux. Elles sont observables ici sur *Lasallia pustulata* sous la forme de cils noirs sur les bords du thalle. D'autres, comme *Ramalina cuspidata*, produisent des spores asexuées contenues dans des pycnides intégrées à l'intérieur du thalle. Elles sont parfois visibles comme ici par leur ostiole de couleur noire, par laquelle les spores seront libérées à maturité.

La composition de la symbiose peut également changer. La plupart des lichens cités précédemment étant associés à des algues vertes (lichens à *chlorococcales*). D'autres comme *Peltigera membranacea* sont associés à des cyanobactéries plutôt qu'à des algues, capables de fixer l'azote atmosphérique. Dans certains cas, le lichen est associé à plusieurs photobiotes en même temps. *Peltigera leucophlebia* contient des algues vertes sur les lames de son thalle et concentre les cyanobactéries dans des verrues noires, des céphalodies. Ces céphalodies n'ont pas la même forme dans *Placopsis gelida*, où elles sont concentrées au centre de thalle avec une couleur brune, laissant le reste du thalle aux algues vertes.



A *: Peltigera leucophlebia* (Nyl.) Gyeln, alentours de Mývatn, Islande. Lichen foliacé tripartite, contenant une algue du genre *Coccomyxa* dans les lames du thalle et des cyanobactéries du genre Nostoc dans les céphalodies noires bien visibles ici.[1] Les données chimiques concernant cette espèce lui attribuent la ténuiorine, l'acide gyrophorique et le méthylgyrophorate.[4]

B *: Placopsis gelida* (L.) Linds., côte Sud-Ouest de l'Islande. Lichen à thalle crustacé placodiomorphe, saxicole, décrit avec de l'acide gyrophorique et des traces d'acide lécanorique. Il s'agit d'une symbiose tripartite, avec des algues du genre *Trebouxia* dans le thalle blanc et des cyanobactéries du genre *Stigonema* dans les céphalodies brunes.[3,4]

Figure 4 – Diversité morphologique des lichens (suite). Références bibliographiques : 1 – Lichens des sols (Van Haluwyn et al. 2012), 2 – Lichens des roches (Asta et al. 2016), 3 – Lichens des arbres (Van Haluwyn, Asta, and Gavériaux 2013), 4 – Íslenskar fléttur (Kristinsson 2016).

En plus de l'association mycobiote – photobiote, un lichen est composé d'une multitude d'autres organismes (Grube and Berg 2009; Millot and Mambu 2019), la présence de certains pouvant influencer grandement l'aspect de la symbiose au sein d'une même espèce, comme les levures basidiomycètes mises en évidence dans le genre *Bryoria* (Spribille et al. 2016). Ces variations morphologiques et la difficulté d'identification a très tôt amené à l'intégration de données chimiques dans l'identification des lichens. La plupart des lichens illustrés précédemment sont associés à une liste de composés utilisés lors des diagnoses, et ce, même dans les ouvrages dédiés au grand public. Bien que cette utilisation de leur chimie puisse paraitre étonnante, elle trouve ses origines dans des temps plus anciens, grâce à une histoire commune avec l'Homme qui le poussera à étudier les métabolites lichéniques jusqu'à l'époque moderne. Depuis l'antiquité, ils ont été utilisés dans la production de teintures pourpres, notamment dans la Grèce antique et à l'ère Romaine. Cette pratique aurait survécu dans la région méditerranéenne jusqu'au Moyen Âge tardif, période à laquelle ces lichens ont été commercialisés plus largement, comme à Florence, en Espagne et au Portugal. Ils étaient alors désignés comme étant des algues, malherbes ou plantes des côtes (Kok 1966).

1. Lichens, naturalisme et chimie aux XVIIIe et XIXe siècles.

Au cours sur XVIIIe siècle, Carl von Linné (1707-1778) propose une classification du vivant dans son *Systema naturæ* (1735). Plus spécifiquement dans *Species plantarum* (1753) (Linné 1753), il introduit parmi les algues cryptogames les lichens pour lesquels il décrit déjà plusieurs espèces. Ce sont les travaux d'Erik Acharius (1757-1819) qui mettront l'accent sur ces organismes, auxquels seront attribués la plupart des noms connus aujourd'hui (Acharius 1798, 1803, 1810, 1814). C'est également au cours du XVIIIe siècle qu'une collaboration s'établit entre les botanistes et les chimistes pour développer de nouvelles teintures et techniques de coloration. Par exemple, Johan Peter Westring (1753-1833) publie à partir de 1805 plusieurs tomes de son ouvrage *Svenska Lafvarnas Färghistoria* où sont détaillées les méthodes d'extractions et de préparation de teintures à partir de différents lichens ainsi que des échantillons des couleurs obtenues (Westring 1805) (**Figure 5**). Il est intéressant de noter que l'auteur a utilisé la classification binomiale avec les noms originaux donnés par Linné et les noms actualisés par Acharius. L'étude et le développement des teintures a connu un certain essor en France qui se prolongera au XIXe siècle. Ceci contribuera à solidifier dans le pays une industrie des teintures, comme en témoigne l'*Essai analytique des lichens de l'orseille* publié en 1829 par Pierre-Jean Robiquet (1780-1840) (Wisniak 2013). Cette capacité des substances lichéniques à changer de couleur n'aura pas échappé aux naturalistes : en 1866, Wilhelm Nylander (1822-1899) utilise de l'hypochlorite de soude et de l'hydroxyde de potassium pour provoquer une réaction colorée directement sur le thalle de lichens (Vitikainen 2001; Nylander 1866). Cette opération portera le nom de *réaction thalline* et sera utilisée jusqu'aujourd'hui pour contribuer aux diagnoses de lichens (Ekman and Tønsberg 2019; K. Kalb and Aptroot 2018; J. Kalb, Lücking, and Kalb 2018; Andre Aptroot and da Silva Caceres 2018).

Figure 5 – *Couleurs obtenues à partir de lichens, extrait de Svenska Lafvarnas Färghistoria (1805, Johan Peter Westring). Lichen prunastri correspond à l'actuel Evernia prunastri et Lichen pustulatus à Lasallia pustulata.*

Bien que les lichens aient eu jusque-là une place à part parmi les plantes, leur nature symbiotique n'avait pas encore été mise en évidence. Ceci est sur le point de changer à la fin du XIXe siècle avec les travaux de Simon Schwendener (1829-1919) qui propose que ces organismes ne soient pas des plantes mais une symbiose entre un champignon et une algue. Le concept de symbiose était encore nouveau, proposé notamment par Edouard Van Beneden (1846-1910) et Anton de Bary (1831-1888) et cette double nature des lichens a été mal accueillie par les principaux lichénologues de l'époque (Honegger 2000; Perru and Colin 2006). Il sera nécessaire d'attendre le XXe siècle pour que son hypothèse soit plus largement acceptée.

2. L'étude moderne des lichens aux XXe.

C'est au tournant du siècle que les composés lichéniques commencent à être étudiés de façon plus exhaustive. Les méthodes d'élucidation structurale étaient basées sur la dégradation des molécules en fragments connus, ce qui rendait la détermination de structures difficile. Des travaux de recensement de ces molécules lichéniques ont néanmoins été entrepris par Friedrich Wilhelm Zopf (1846-1909) et par Oswald Hesse (1835-1917). 150 molécules ont ainsi pu être cataloguées, une petite partie d'entre elles

avec une structure identifiée (Zopf 1895b, 1895a, 1897, 1907; Hesse-Feuerbach 1912). Des avancements majeurs ont été apportés par les travaux de Yasuhika Asahina (1881-1975) et de Shoji Shibata (1915-2016) grâce aux techniques de microcristallisation, permettant d'identifier les molécules sur la base de la forme de leurs cristaux visualisés au microscope (Shibata 2000; Asahina 1951; Asahina and Shibata 1954).

La recherche sur les lichens prend un nouvel essor dans l'après-guerre. Leur culture en laboratoire a permis la publication de nombreux travaux sur leur physiologie dans les années 60, notamment les relations entre de mycobiote et le photobiote sur les échanges de carbone au sein de la symbiose, leur reviviscence de façon générale, leur métabolisme (D. H. S. Richardson, Hill, and Smith 1968; D. H. S. Richardson and Smith 1968a, 1968b; Hill and Smith 1972; D. C. Smith and Molesworth 1973; Green and Smith 1974; S. Chambers, Morris, and Smith 1976; D. C. Smith and Drew 1965; D. H. S. Richardson, Smith, and Lewis 1967; Corbett and Smith 1969; Ahmadjian 1967; D. C. Smith and Seaward 2013). L'étude de la chimie des lichens n'est pas en reste : les travaux de Chicita Frances Culberson et William Louis Culberson soulignent l'intérêt des substances lichéniques pour différencier des espèces et comprendre l'évolution des lichens à travers leurs voies de biosynthèse (W. L. Culberson 1969; C. F. Culberson 1963, 1969; C. F. Culberson and Kristinsson 1969; W. L. Culberson and Culberson 1970). En collaboration avec le lichénologue islandais Hörður Kristinsson sera publiée une méthode standardisée pour l'identification par CCM des substances lichéniques (C. F. Culberson and Kristinsson 1970). Cette méthode restera jusqu'aujourd'hui la technique privilégiée par les lichénologues pour le profilage des lichens. Plusieurs études se pencheront sur la diversité chimique des lichens et des variations de celle-ci au sein d'un taxon : les chémosyndromes (C. F. Culberson, Culberson, and Esslinger 1977; C. F. Culberson and Culberson 1976, 1978; Stocker-Wörgötter 2004). C'est à cette période que seront identifiées la plupart des composés lichéniques, repérés à l'aide de la CCM puis isolés et caractérisés. La LC-DAD commence également à être utilisée notamment par Guido Benno Feige (Feige et al. 1993) et plusieurs travaux d'isolement et d'identification seront conduits par Isao Yoshimura (Yoshimura et al. 1994) et John Elix (Elix and Crook 1992; Elix and Gaul 1986; Elix, Jenie, and Parker 1987). Toutes les données accumulées durant cette période seront résumées dans *Identification of Lichen Substances* par Siegried Huneck et Isao Yoshimura en 1996 (Huneck and Yoshimura 1996). A ce stade, il est évident que les lichens produisent des molécules qui leur sont propres, différentes de celles des plantes. La plupart de ces molécules proviennent de la voie des polycétides, produites par le mycobiotes, notamment des depsides et des depsidones. Les molécules de la voie du shikimate et du mévalonate sont moins nombreuses ou du moins, sont moins bien étudiées (**Figure 6**) (Le Pogam, Herbette, and Boustie 2015; Elix 1996; Stocker-Wörgötter 2008).

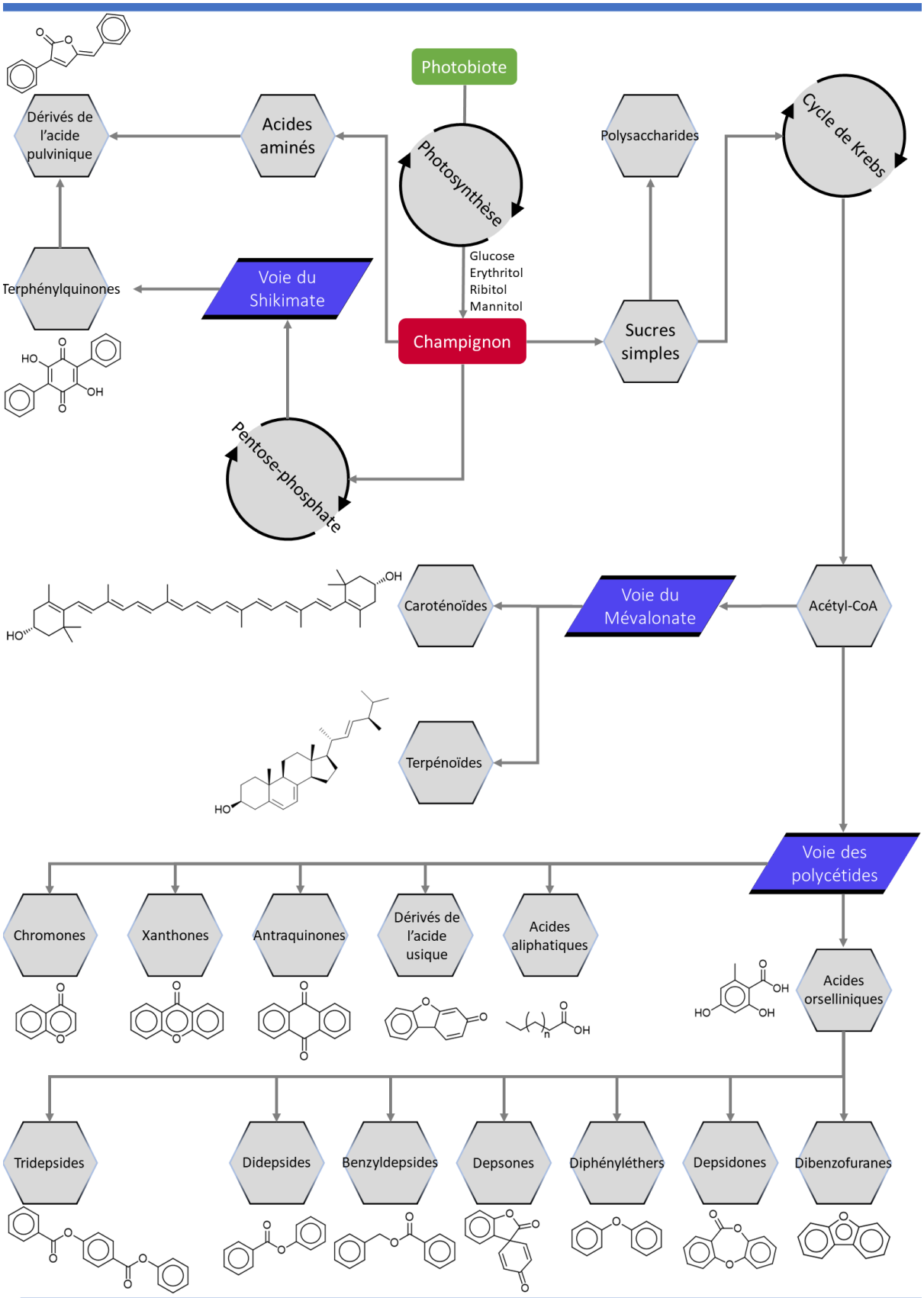


Figure 6 – Origine biosynthétique des groupes structuraux répertoriés dans les lichens. Figure produite à partir d'une autre dans un chapitre de livre (Le Pogam, Herbette, and Boustie 2015).

3. L'impact de la génomique.

Au terme de près d'un siècle de recherche, la structure de l'ADN fut identifiée au milieu du XXe siècle (Franklin and Gosling 1953b, 1953a; Watson and Crick 1953). Des techniques d'amplification se sont développées (Mullis 1990) et il allait devenir possible d'étudier les organismes à travers des séquences de leur ADN. Ceci semblait cependant contre-intuitif pour un lichen, n'étant pas à proprement parler un organisme mais une symbiose complexe composée d'un mycobionte, d'un phycobionte et comme il sera découvert plus tard, de nombreux autres organismes (Lawrey 2003; Arnold et al. 2009; Grube and Berg 2009; Grube et al. 2009; Bates et al. 2011; Spribille et al. 2016). Il a été établi par la suite qu'il serait nécessaire de se focaliser sur le champignon et des techniques spécifiques ont été développées pour isoler son ADN de celui des autres organismes (Grube et al. 1995).

Des différends se créent entre les lichénologues prônant la prédominance des caractères morpho-anatomiques, ceux qui veulent les coupler aux profils chimiques, et ceux qui veulent intégrer des données génétiques. La rupture entre les taxonomistes classiques et ceux prenant en compte la biologie moléculaire avait déjà été observé pour d'autres taxons. L'intérêt de la chimie dans la taxonomie est plus propre aux lichens : la corrélation des profils chimiques aux délimitations des espèces, la variation de ces profils en fonction des individus, de la population et de la géographie était déjà sujet à débat avant même de considérer l'impact de la génétique (Lumbsch 1998a; Brodo 1978, 1986; Egan 1986; R. W. Rogers 1989; Lumbsch 1998b; LaGreca 1999; Lumbsch and Leavitt 2011).

4. Avènement de la métabolomique.

En parallèle et à l'écart des débats tenus dans le milieu de la lichénologie, la quantité de données générée par l'étude du génome, du transcriptome et du protéome a promu dans les années 90 l'utilisation de l'informatique dans ces domaines. Les sciences de l'« omic » se sont ainsi développées : la génomique, la transcriptomique, la protéomique et la métabolomique. La métabolomique représentait la portion des constituants d'une cellule occupée par les « métabolites » ou petites molécules (<1500 Da) non couvertes par la protéomique. A la différence de l'ADN, de l'ARNm et des protéines, les métabolites ne sont pas composés de blocs prévisibles et il n'est pas facile d'établir leur structure directement lors de leur analyse. La métabolomique était cependant le dernier élément permettant de faire le lien entre le génome et le phénotype d'un organisme, et il a été nécessaire de développer cette science. Dès le début des années 2000, les bases nécessaires pour son développement ont été établies :

- Une approche holistique et non réductionniste, impliquant l'étude de toutes les molécules indépendamment de leurs classes structurales.
- L'usage d'outils informatiques performants pour traiter et « miner » les données d'acquisition. Les algorithmes évolutifs sont déjà recommandés.
- La création de bases de données pour identifier les signaux mesurés.
- Des outils de visualisation adaptés à la taille de l'information.

- Le développement d'outils permettant de transformer les données en savoir.

Compte tenu du manque cruel d'informations sur le métabolome, il était déjà suggéré que les voies de biosynthèse soient re-étudiées à la lumière des résultats de la métabolomique. La communauté des chimistes des produits naturels (notamment des plantes) s'est rapidement emparée de la métabolomique non seulement pour étudier leur métabolisme dans le contexte général des *omics*, mais également pour la recherche de métabolites secondaires bioactifs.

Bien qu'initialement très axée sur la GC-MS, la métabolomique utilisera également l'HPLC puis l'HPLC-MS après son interfaçage avec des spectromètres de masse grâce aux sources électrospray (Dole et al. 1968; Mack et al. 1970; Whitehouse et al. 1985; Wong, Meng, and Fenn 1988). Les techniques mélangeant chromatographie liquide et spectrométrie de masse se multiplieront, avec des méthodes séparatives devenant de plus en plus performantes (chromatographie capillaire, UPLC de Waters (Swartz 2005), UHPLC) couplées à différents spectromètres de masse (ToF, FTICR, Orbitrap de Thermo Fisher Scientific), grâce à des sources d'ionisations diverses (ESI, APCI, PI). La complexité des données générées par ces instruments a entrainé le développement de nombreux outils pour leur traitement : des packages sont développés sur R (R statistical programming language), des librairies sur Python et Matlab, des logiciels comme MZmine (Katajamaa, Miettinen, and Orešič 2006; Pluskal et al. 2010), XCMS (C. A. Smith et al. 2006; Tautenhahn et al. 2012; Gowda et al. 2014), MetaboAnalyst (Xia et al. 2009, 2012, 2015; Chong et al. 2018), Galaxy et Workflow4Metabolomics (Davidson et al. 2016; Guitton et al. 2017) et openMS (Sturm et al. 2008; Röst et al. 2016) pour ne citer que quelques exemples de logiciels en libre accès. Des bases de données ont été créées pour accompagner les études de métabolomique, comme le Human Metabolome DataBase (HMDB) (Wishart et al. 2007), MassBank (Horai et al. 2010), MetaboLights (Haug et al. 2013), la base de données du NIST, celles du GNPS (Wang et al. 2016) et le KEGG pour remettre des données dans leur contexte. Plus récemment, le développement des réseaux moléculaires a permis de visualiser l'intégralité des données d'une expérience de métabolomique sous la forme d'un réseau. La métabolomique, à l'origine une branche peu étudiée des « omics » est devenue un domaine de recherche hautement multidisciplinaire, encore en pleine expansion, qui pourra à terme faire le pont entre la génétique et le phénotype dans des études multi-omiques (Fiehn 2001, 2002; Mendes 2002; Goodacre 2005; Hall 2006; Rabinowitz et al. 2011; Misra, Fahrmann, and Grapov 2017).

5. Les lichens et la métabolomique & la contribution de ces travaux.

Comme pour les plantes plus tôt, des études de métabolomique commencent à être conduites sur les lichens, connus pour leurs polycétides spécifiques et bioactifs (Crittenden and Porter 1991; Grube 2019; Molnár and Farkas 2010; Goga et al. 2018; Schinkovitz et al. 2018; Varol 2018; Boustie, Tomasi, and Grube 2011; Huneck 1999). Les bases de données de métabolomique ne couvrent que très peu ces composés lichéniques, mettant en difficulté les chercheurs qui se retrouvent face à trop d'inconnues

accompagnées de faux-positifs issus de molécules végétales ou des médicaments dérépliqués dans leurs données. L'objectif de cette thèse est de développer les outils nécessaires à l'étude des lichens dans le domaine de la métabolomique, et d'explorer cette diversité chimique que les lichens devraient contenir à l'aide de ces outils innovants.

Dans le *Chapitre I*, un outil essentiel pour étudier les lichens est développé : une base de données pour les métabolites lichéniques produite à partir de la littérature (LDB-Lit / Lichen DataBase – Littérature). Les composés lichéniques sont absents de la plupart des bases de données, étant souvent liés au métabolisme du champignon modifié par la présence du photobiote. Cette base de données contient les données structurales de 1662 molécules recensées ainsi que leurs organismes producteurs. Bien qu'elle ne contienne pas de données spectrales, elle permet une déréplication sur la base de la masse exacte des composés, de leur origine biologique et des caractéristiques structurales. La taille du document étant déjà assez importante et compte tenu du caractère questionnable d'une base de données au format papier, cette LDB-Lit ne sera consultable publiquement que lorsqu'elle sera publiée.

Dans le *Chapitre II*, une deuxième base de données plus réduite est créée, cette fois avec des données spectrales permettant une meilleure identification des composés dans une analyse LC-MS classique de métabolomique (LDB / Lichen DataBase). Ces données correspondent aux spectres de fragmentation (MS/MS) de plusieurs standards lichéniques issus de la collection Siegried Huneck conservée à Berlin. Cependant, même avec des bases de données spectrales, il est couramment admis que la plupart des signaux de LC-MS restent non identifiés. La liste de ces composés est consultable en **Annexe** (**Tableau S-1**)

Dans le *Chapitre III*, les raisons pouvant être à l'origine de ce faible taux d'identification sont explorées, notamment par l'impact des adduits produits en LC-MS et par les différences entre spectres produits par différents instruments.

Comme la déréplication des spectres individuels par comparaison à des standards ne permet la déréplication que de quelques signaux, une base de données de mass2motifs est créée sur MS2LDA à l'aide de la LDB dans le *Chapitre IV*. Elle permet de guider l'utilisateur lors de l'interprétation des données LC-MS/MS en détectant les motifs communs entre les ions, notamment les ions produits par des molécules connues. Un outil pour automatiser cette interprétation a été développé dans ce chapitre : *Classnotator*.

Dans le *Chapitre V*, *Classnotator* est intégré en tant que module dans un plus grand outil permettant de résoudre certains problèmes soulevés dans le *Chapitre III* : *Molnotator*. Il a pour objectif principal d'identifier les espèces ioniques dans les analyses LC-MS, en tant que fragment de source ou en tant qu'adduit d'une molécule, de prédire la masse de la molécule à laquelle ces entités appartiennent et de guider la déréplication automatisée pour réduire les taux de faux-positifs. *Molnotator* est validé sur les données de la LDB pour permettre de correctement prédire les espèces ioniques.

Dans le *Chapitre VI*, 300 échantillons de lichens ont été analysés par LC-MS pour tenter d'explorer leur diversité chimique. Les données ont été traitées sur MZmine puis par *Molnotator* pour prédire les molécules qui sont produites par ces 300 lichens. Bien que cette méthode reste à améliorer, ce genre d'approche commence à se développer (Schmid et al. 2020) et constitue une étape décisive pour l'étude d'organismes dans un contexte multi-omique (Mark et al. 2019; Bertrand, Abdel-hameed, and Sorensen 2018).

Pour des raisons pratiques, un volume important de données est présenté en **Annexe** en suivant les chapitres du premier volume.

Un glossaire a été positionné à la suite de cette introduction pour couvrir les termes et abréviations utilisés dans le document et donner des précisions sur des points spécifiques. Le caractère hautement multidisciplinaire de la métabolomique le rend assez conséquent, et une contextualisation y est également intégrée pour permettre de mieux comprendre ce qui a été développé ici.

Glossaire & Recontextualisation

Résumé

Cette section a été divisée en deux : un glossaire conséquent du fait du caractère multidisciplinaire de la métabolomique, et une remise en contexte nécessaire pour mieux comprendre les outils utilisés ou développés dans le cadre d'une application de la métabolomique aux lichens. Cette dernière partie explique notamment ce qu'est un réseau moléculaire et les variantes créées ici.

Sommaire

1 – Définitions, Abréviations et Sites utilisés 16

2 – Recontextualisation dans le cadre de la thèse 36

2.1 Les réseaux moléculaires 36

2.2 Autres réseaux utilisés 39

2.3 Termes spécifiques à *Fragnotator* 41

2.4 Termes spécifiques à *Adnotator* 42

2.5 Termes spécifiques à *Classnotator* 45

2.6 Remarques supplémentaires 46

1

Définitions, Abréviations et Sites utilisés

ACN : acétonitrile.

Adnotator : algorithme développé dans le cadre de cette thèse permettant de regrouper tous les ions produits par une molécule dans une analyse LC-MS/MS et de les connecter à une molécule neutre hypothétique, dont la masse est calculée par triangulation après l'annotation de tous ses adduits. Le résultat peut être visualisé sous la forme d'un réseau (*voir section 2.4*).

AF, FA : acide formique, formic cid.

Analyseur quadripolaire : spectromètre de masse constitué d'un quadrupôle, ou quatre électrodes parallèles disposées en carré, les deux électrodes opposées étant soumises au même potentiel. Ceci crée un champ électrique qui permet à un ion avec un rapport *m/z* donné dans un plan x-y orthogonal à la longueur du quadrupôle de le traverser selon l'axe z.

Analyseur à temps de vol, TOF : spectromètre de masse séparant les ions par leurs rapports *m/z* dans une zone de vol sans champ électrique après accélération grâce à un potentiel donné. Les ions avec la même accélération et avec des rapports *m/z* différents ne mettront pas le même temps à traverser la zone de vol.

APCI, Ambient Pressure Chemical Ionisation : ionisation chimique à pression ambiante, ionisation chimique d'un échantillon, gaz ou liquide nébulisé, à l'aide d'une décharge corona ou d'un émetteur bêta comme le ^{63}Ni (Murray et al. 2013).

Apothécie : ascome des discolichens en forme de coupe contenant l'hyménium exposé à l'air libre (Association Française de Lichénologie 2016).

Arête (edge) : relie deux *nœuds* dans un *graphe* ou un *réseau* (voir *Graphe*, *Théorie des graphes*, *Réseaux moléculaires*).

Batch mode : enchaînement automatique de plusieurs opérations à l'aide d'un *batch file* fourni par l'utilisateur.

Binning : groupement des données par classes. Ces classes sont caractérisées par une fenêtre de valeurs (*bin size*) permettant d'y intégrer les données

Blocs (cluster) : Ensemble de *nœuds* reliables par une ou une suite continue *arêtes*.

CCM – Chromatographie sur Couche Mince : technique de chromatographie pour la séparation des analytes à l'aide d'une phase stationnaire (généralement du gel de silice) déposée sur une plaque, et d'une phase mobile qui va migrer le long de la phase stationnaire en entrainant plus ou moins vite les analytes. Cette technique a été et est toujours largement utilisée pour l'étude des lichens, notamment depuis la standardisation des mélanges de solvants d'élution et depuis que les données de

migration des métabolites lichéniques dans chacun de ces solvants sont rapportées (Huneck and Yoshimura 1996; C. F. Culberson and Kristinsson 1970).

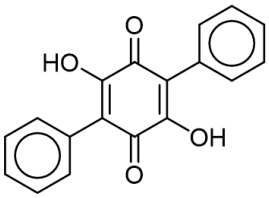
CFM-ID – Competitive Fragmentation Modeling for Metabolite Identification (https://cfmid.wishartlab.com/) : outil permettant l'annotation des pics d'un spectre MS/MS à une structure connue, la prédiction de spectre MS/MS pour une structure donnée et l'annotation putative de métabolite. Les algorithmes utilisés sont basés sur des techniques de *machine learning* pour modéliser les processus de fragmentation MS/MS (Allen et al. 2014, 2016; Djoumbou-feunang et al. 2019).

Chromatographie en phase liquide : technique séparative pour des mélanges de composés migrant plus ou moins vite à travers une phase stationnaire par leur entrainement grâce à une phase mobile liquide. Dans le contexte de cette thèse, elle sera abrégée LC (Liquid Chromatography) et désignera la séparation des mélanges complexes d'extraits lichéniques ou de standards à travers une colonne. Les systèmes de chromatographie utilisés ici sont de type UHPLC (Ultra High Performance LC, colonnes avec des particules de diamètres compris entre 2 et 5 µm) ou UPLC dans le cas des instruments produits par la compagnie Waters (colonnes avec des particules de diamètres inférieurs à 2 µm).

Chromatogramme (en LC-MS) : par défaut dans cette thèse, chromatogramme de courant ionique total. Chromatogramme créé en traçant le *courant ionique total* dans une série de *spectres de masse* enregistrés en fonction du temps de rétention (Murray et al. 2013). Il existe aussi des chromatogrammes produits à partir de l'intensité du pic de base d'un spectre (BPI ou *Base Peak Intensity*), appelés BPI chromatograms.

Classnotator : algorithme développé dans le cadre de cette thèse permettant d'annoter les ions d'une analyse LC-MS/MS à partir de motifs produits sur MS2LDA et d'une classification structurale fournie par l'utilisateur.

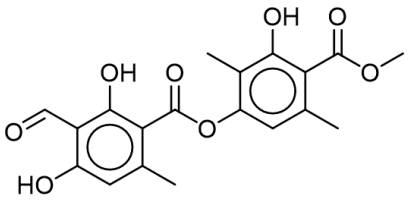
ClassyFire (http://classyfire.wishartlab.com/) : Logiciel en ligne pour la classification automatisée d'entités chimiques. ClassyFire fournit une classification chimique hiérarchisée comptant 4825 classes chimiques de composés organiques et inorganiques. Ces classes sont organisées suivant un modèle s'inspirant de la classification du vivant avec les niveaux *Kingdom*, *Superclass*, *Class*, *Subclass*, et de quatre niveaux (*levels*) allant de 5 à 9. ClassyFire est également accessible à l'aide d'un *package* R (https://cran.r-project.org/web/packages/classyfireR/index.html). La liste des classes chimiques est décrite sur le site de ClassyFire : http://classyfire.wishartlab.com/tax_nodes. Un exemple pour la classification de l'acide polyporique et de l'atranorine est présenté en **Figure 7** et la description ClassyFire des classes détectées est présentée dans le **Tableau 1**.



Acide polyporique
Formule brute : $C_{18}H_{12}O_4$
SMILES :
c1ccc(cc1)C1=C(C(=O)C(=C(C1=O)O)c1ccccc1)O

Classification Hun&Yosh96 :
Benzoquinones

Classification ClassyFire :
Kingdom : Composés organiques
Superclass : Composés organiques oxygénés
Class : Composés organooxygénés
Subclass : Composés carbonylés
Level 5 : Cétones
Level 6 : Cétones cycliques
Level 7 : Quinones
Level 8 : Benzoquinones
Level 9 : P-benzoquinones



Atranorine
Formule brute : $C_{19}H_{18}O_8$
SMILES:
Cc1cc(c(c(c1C(=O)Oc1c(c(c(c(c1)C)C(=O)OC)O)C)O)C=O)O

Classification Hun&Yosh96 : Depsides (Didepsides)

Classification ClassyFire :
Kingdom : Composés organiques
Superclass : Phénylpropanoïdes et polycétides
Class : Depsides et depsidones
Subclass : NA
Level 5 : NA
Level 6 : NA
Level 7 : NA
Level 8 : NA
Level 9 : NA

Figure 7 – Classification produite par ClassyFire sur deux molécules lichéniques : l'acide polyporique et l'atranorine. Leur classification selon Hun&Yosh96 est également rapportée. Les noms de chaque classe ont été traduits en français et leurs descriptions sont décrites dans le tableau ci-dessous.

Tableau 1 – Quelques classes utilisées par ClassyFire et leurs définitions, traduites de l'anglais.

Nom & Identifiant	Description
COMPOSÉS ORGANIQUES *CHEMONTID:0000000*	Composés qui contiennent au moins un atome de carbone, à l'exclusion des isocyanures/cyanures et de leurs dérivés non-hydrocarbylés, du thiophosgène, du diséléniure de carbone, du monosulfure de carbone, du disulfure de carbone, du sous-sulfure de carbone, du monoxyde de carbone, du dioxyde de carbone, du sous-oxyde de carbone et du monoxyde de dicarbone.
COMPOSÉS ORGANIQUES OXYGÉNÉS *CHEMONTID:0004603*	Composés organiques qui contiennent un ou plusieurs atomes d'oxygène.
PHÉNYLPROPANOÏDES ET POLYCÉTIDES *CHEMONTID:0000261*	Composés organiques qui sont synthétisés soit à partir de l'acide aminé phénylalanine (phénylpropanoïdes), soit par la condensation décarboxylative de malonyl-CoA (polycétides). Les phénylpropanoïdes sont des composés aromatiques basés sur le squelette du phénylpropane. Les polycétides sont généralement constitués de groupes carbonyles et méthylènes alternés (bêta-polycétones), dérivés biogénétiquement de la condensation répétée de l'acétylcoenzyme A (via le malonyl-coenzyme A), et généralement des composés qui en sont dérivés par d'autres condensations.
COMPOSÉS ORGANOOXYGÉNÉS *CHEMONTID:0000323*	Composés organiques contenant une liaison entre un atome de carbone et un atome d'oxygène.
DEPSIDES ET DEPSIDONES *CHEMONTID:0001645*	Composés polycycliques qui est soit un composé polyphénolique composé de deux ou plusieurs unités aromatiques monocycliques liées par une liaison ester (depside), soit un composé contenant la structure depsidone (depsidone).
COMPOSÉS CARBONYLÉS *CHEMONTID:0001831*	Composés organiques contenant un groupe carbonyle, de structure générale RC(=O)R', où R=organyle, R'=H, N, O, groupe organyle ou groupe halogénure.

Tableau 1 – Suite.

Nom & Identifiant	Description
CÉTONES *CHEMONTID:0000118*	Composés organiques dans lesquels un groupe carbonyle est lié à deux atomes de carbone $R_2C=O$ (aucun R ne peut être un atome d'hydrogène). Les cétones qui ont un ou plusieurs atomes d'hydrogène alpha subissent une tautomérisation céto-énolique, le tautomère étant un énol.
CÉTONES CYCLIQUES CHEMONTID:0003487	Composés organiques contenant une cétone qui est conjuguée à un groupement cyclique.
QUINONES *CHEMONTID:0002495*	Composés ayant une structure de dione cyclique entièrement conjuguée, telle que celle des benzoquinones, dérivés de composés aromatiques par conversion d'un nombre pair de groupes -CH= en groupes -C(=O)- avec tout réarrangement nécessaire de doubles liaisons (les analogues polycycliques et hétérocycliques sont inclus).
BENZOQUINONES *CHEMONTID:0002384*	Composés organiques contenant un cycle benzénique qui porte deux groupements cétoniques en positions 1 et 4 (cyclohexa-2,5-diène-1,4-dione).
P-BENZOQUINONES *CHEMONTID:0002494*	Benzoquinones où les deux groupes C=O sont attachés aux positions 1 et 4, respectivement.

Complexe (complexe ion/neutre) : espèce formée par la combinaison entre un ion et une molécule neutre via des interactions faibles (Murray et al. 2013).

Contexte d'ionisation : notion abordée dans ces travaux pour faire référence à l'ensemble des ions créés et détectés au cours d'une analyse LC-MS dans une fenêtre de temps ΔTR. Ainsi, tous les ions produits par l'ionisation d'une molécule inconnue devraient pouvoir être retrouvés dans ce contexte et ainsi, permettre par triangulation de retrouver les caractéristiques de la molécule qui les a produits (masse exacte, données structurelles…).

Courant Ionique Total (Total Ion Current, TIC) : Somme de toutes les intensités des ions composant un *spectre de masse* (Murray et al. 2013).

CSV – Comma-Separated Values : média de texte séparant les valeurs par des virgules. Ainsi un tableau peut être enregistré en séparant les éléments de chacune de ses lignes par une virgule (« , ») et en insérant une *fin de ligne* à la fin pour passer à la ligne suivante. Une première ligne peut être insérée en tant que « *header* » pour donner un nom à chaque colonne (Shafranovich and Network Working Group 2005). Le format CSV peut changer en fonction des pays. En France, les décimales sont séparées par une virgule « , » au lieu d'un point « . » comme dans d'autres systèmes, rendant le CSV en tant que tel inutilisable. Le CSV en France utilise donc souvent le point-virgule « ; » pour palier à ce problème (Wikipedia 2020a).

Cytoscape (https://cytoscape.org/) : logiciel open-source pour la visualisation de *réseaux* et l'intégration de leurs attributs.

Data-Dependent Acquisition (DDA) : mode d'acquisition de données de *spectrométrie de masse en tandem* permettant dans un premier temps de sélectionner un nombre donné d'*ions précurseurs* à l'aide de leurs rapports *m/z* pour les soumettre dans un deuxième temps à une analyse MS/MS (Murray et al. 2013).

Data mining : exploration ou fouille des données, extraction des connaissances à partir de grandes quantités de données par des méthodes automatiques. Quand il est appliqué à du texte, on parle de *Text mining*.

DDA : Data Dependent Acquisition.

Déconvolution : dans le contexte du traitement de données LC-MS, se réfère à la filtration des données pour éliminer les signaux assimilables à du bruit et les signaux redondants d'une même molécule.

Déréplication : Mot utilisé pour la première fois en 1980 pour désigner les méthodes permettant de reconnaître et d'éliminer les substances actives déjà étudiées dans les premières étapes du criblage. La définition actuelle de la déréplication s'est élargie au criblage des produits naturels, notamment par l'identification de signaux LC-MS/MS déjà répertoriés dans les bases de données (Ito and Masubuchi 2014).

Edge table (« Tableau d'arêtes ») : tableau utilisé pour créer un réseau avec un logiciel approprié, contenant toutes les paires / nœuds reliés deux à deux.

ESI, Electrospray Ionisation : Ionisation par électrospray, procédé par lequel des cations ou des anions en solution sont transférés en phase gazeuse par la formation et la désolvatation à pression atmosphérique d'un faisceau de gouttelettes hautement chargées résultant d'une différence de potentiel appliquée entre le bout de l'aiguille d'électrospray et une électrode (Murray et al. 2013).

Expression régulière (regex) : séquence de caractères qui définit un motif à chercher, utilisé par des algorithmes de recherche de chaînes de caractères (Wikipedia 2020b).

FBMN : Feature-Based Molecular Networking.

FDD : Fragments de Depsides et Depsidones, traduction simplifiée de la classe Hun&Yosh96 « Cleavage products of depsides and depsidones ».

Feature-Based Molecular Networking (FBMN) : méthodes consistant à combiner les traitements de données LC-MS/MS et les réseaux moléculaires. Les données LC-MS/MS sont traitées de façon à les exprimer sous forme de liste de pics, ou *feature list*. Une *feature* dans ce contexte est un signal de masse accompagné d'un spectre MS/MS. Il peut être issu d'une molécule (*dé*)*protonée*, d'un *adduit*, d'un *fragment de source* ou d'un pic du *massif isotopique*. Le FBMN est opposé aux réseaux moléculaires produits auparavant à partir de données LC-MS/MS brutes, où une détection de pics rudimentaire était effectuée sans être aussi efficace que les logiciels dédiés. L'un des avantages notables du FBMN est la différenciation des pics isobares et des molécules isomères lorsque ceux-ci peuvent l'être grâce à des temps de rétention différents. Ils étaient avant regroupés sous la forme d'un seul pic étant donné que le temps de rétention n'était pas pris en compte. La méthode la plus courante actuellement est la combinaison de MZmine et du GNPS (Nothias et al. 2019; Olivon et al. 2017). D'autres combinaisons sont possibles, comme l'usage de XCMS pour le traitement LC-MS et de MetGem pour les réseaux moléculaires

Format ouvert (Open format) : format de fichier librement accessible par n'importe quel individu ou organisation (Wikipedia 2020d).

Format propriétaire (Proprietary format) : format de fichier d'une compagnie contenant des données organisées à l'aide d'un encodage particulier produit par la compagnie pour garder les données secrètes. Ainsi, le décodage et l'interprétation des données n'est possible qu'avec un logiciel fourni par la compagnie elle-même (Wikipedia 2020d). Dans le contexte de la spectrométrie de masse, les formats propriétaires évoqués lors de cette thèse sont le .D (Agilent), et les .RAW (Thermo & Waters).

Forme d'ionisation / d'ion : désigne ici les molécules déprotonées et tous les adduits résultant de l'ionisation d'une molécule. Comme il est parfois difficile de conserver la simplicité d'une phrase avec « forme d'ionisation », le mot « adduit » lui sera occasionnellement substitué.

Fragment de source : Ion formé par la dissociation d'un ion précurseur au niveau de la source d'ions. Les fragments de source peuvent se former lorsque les voltages utilisés au niveau de la source sont trop élevés, ou que l'ion est trop fragile (Murray et al. 2013).

Fragnotator : algorithme développé dans le cadre de cette thèse permettant de regrouper les fragments de source d'une analyse LC-MS/MS à leur ion précurseur. Le résultat du traitement peut être visualisé sous la forme d'un réseau (*voir section 2.3*).

GenOuest Bioinformatics : Plate-forme bioinformatique hébergée par l'INRIA Rennes-Bretagne Atlantique et l'IRISA, donnant accès ici à un *cluster* de calculs (https://www.genouest.org/).

GNPS – Global Natural Products Social Molecular Networking (https://gnps.ucsd.edu) : site pour le partage de données LC-MS brutes, traitées et de spectres MS/MS identifiés. Les spectres identifiés par la communauté scientifique et déposés dans des bases de données participatives sur le GNPS sont en libre accès. Les spectres expérimentaux de l'utilisateur peuvent y être *dérépliqués* contre les spectres de référence de ces bases de données à l'aide des calculs de similarité cosinus. Le GNPS offre aussi à l'utilisateur la possibilité de produire des réseaux moléculaires à l'aide de ses données LC-MS/MS, et plus récemment par FBMN, permettant la combinaison de traitement de données LC-MS et les réseaux moléculaires (Wang et al. 2016; Olivon et al. 2017) (Voir aussi : Réseaux moléculaires).

Graphe : Un graphe est un dessin géométrique défini par un ensemble de points (appelés sommets ou nœuds), reliés entre eux par un ensemble de lignes ou de flèches (appelées arêtes ou arcs). Chaque arête a pour extrémités deux points, éventuellement confondus (Encyclopédie Larousse en ligne n.d.).

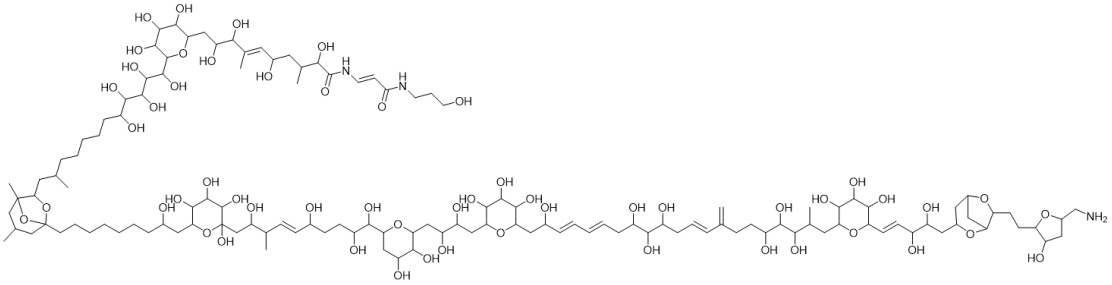
HPLC – High Performance Liquid Chromatography : Chromatographie en phase liquide à haute performance (voir Chromatographie en phase liquide).

Hun&Yosh96 : « Identification of Lichen Substances » par S. Huneck et I. Yoshimura, publié en 1996.

HRMS (High Resolution Mass Spectrometry) : Spectrométrie de Masse Haute Résolution. Régulièrement simplifié en MS, toutes les données évoquées dans cette thèse étant issues d'analyses de spectrométrie de masse haute résolution.

ICN – International Code of Nomenclature for algae, fungi, and plants : anciennement l'ICBN (International Code of Botanical Nomenclature), ensemble de règles et recommandations au sujet des noms donnés à différents groupes d'organismes. *MycoBank* et *Index Fungorum* sont les bases de données acceptées pour le dépôt d'identifiants d'espèces fongiques, bien que *MycoBank* soit devenu le site de référence.

InChI et InChIKey – International Chemical Identifier : Système de notation chimique non-propriétaire, dont la première version a été publiée en 2005. L'InChIKey a été construit pour être un substitut presque unique de l'InChI (McNaught 2006; Pletnev et al. 2012; S. Heller et al. 2013; S. R. Heller et al. 2015). Un exemple est donné en **Figure 8**, en comparant le nom IUPAC de la palytoxine à son code InChI et InChIKey.



Nom IUPAC : 10-[6-[12-[5-[9-[6-[10-[6-[4-[6-[21-[6-[5-[7-[2-[5-(aminomethyl)-3-hydroxyoxolan-2-yl]ethyl]-2,6-dioxabicyclo[3.2.1]octan-3-yl]-3,4-dihydroxypent-1-enyl]-3,4,5-trihydroxyoxan-2-yl]-2,8,9,10,17,18,19-heptahydroxy-20-methyl-14-methylidenehenicosa-3,5,12-trienyl]-3,4,5-trihydroxyoxan-2-yl]-2,3-dihydroxybutyl]-4,5-dihydroxyoxan-2-yl]-2,6,9,10-tetrahydroxy-3-methyldec-4-enyl]-3,4,5,6-tetrahydroxyoxan-2-yl]-8-hydroxynonyl]-1,3-dimethyl-6,8-dioxabicyclo[3.2.1]octan-7-yl]-1,2,3,4,5-pentahydroxy-11-methyldodecyl]-3,4,5-trihydroxyoxan-2-yl]-2,5,8,9-tetrahydroxy-*N*-[3-(3-hydroxypropylamino)-3-oxoprop-1-enyl]-3,7-dimethyldec-6-enamide

InChI : InChI=1S/C129H223N3O54/c1-62(29-33-81(143)108(158)103(153)68(7)47-93-111(161)117(167)110(160)91(180-93)36-35-76(138)82(144)51-73-50-74-53-92(178-73)90(177-74)38-37-89-85(147)52-75(61-130)179-89)23-20-28-78(140)105(155)77(139)26-18-13-16-25-70(135)48-94-112(162)118(168)113(163)97(181-94)55-84(146)83(145)54-95-107(157)87(149)57-96(182-95)106(156)80(142)34-32-69(134)31-30-65(4)88(150)60-129(176)125(174)123(173)115(165)99(184-129)49-71(136)24-15-10-9-11-19-40-128-59-64(3)58-127(8,186-128)100(185-128)44-63(2)22-14-12-17-27-79(141)109(159)116(166)120(170)122(172)124-121(171)119(169)114(164)98(183-124)56-86(148)102(152)66(5)45-72(137)46-67(6)104(154)126(175)132-42-39-101(151)131-41-21-43-133/h13,16,18,20,23,25,30-31,35-36,39,42,45,63-65,67-100,102-125,133-150,152-174,176H,1,9-12,14-15,17,19,21-22,24,26-29,32-34,37-38,40-41,43-44,46-61,130H2,2-8H3,(H,131,151)(H,132,175)

InChIKey : CWODDUGJZSCNGB-UHFFFAOYSA-N

Figure 8 – Codes InChI et InChIKey pour la palytoxine ($C_{129}H_{223}N_3O_{54}$).

Index Fungorum (http://www.indexfungorum.org/) : base de données taxonomique pour les champignons, en partenariat avec Royal Botanic Gardens Kew et sa section mycologique, Landcare Research-NZ et son groupe mycologique et l'Institut de Microbiologie de l'Académie des Sciences Chinoise.

Ion : Espèces atomiques, moléculaires ou radicalaires ayant une charge électrique nette non nulle (Murray et al. 2013).

Ion adduit : ion formé par l'interaction d'un *ion précurseur* avec un ou plusieurs atomes ou molécules pour former un ion contenant tous les atomes constitutifs de l'ion précurseur ainsi que les atomes supplémentaires des atomes ou molécules associés (Murray et al. 2013).

Ion fragment : *Ion produit* qui résulte de la dissociation d'un *ion précurseur* (Murray et al. 2013).

Ion précurseur : Ion qui réagit pour former des *ions produits* particuliers ou qui subit des *pertes neutres* pour former des *ions fragments* (Murray et al. 2013).

Ion produit : Ion formé comme le produit d'une réaction impliquant un *ion précurseur* particulier (Murray et al. 2013).

ITIS – Integrated Taxonomic Information System (https://www.itis.gov/) : site fournissant des données taxonomiques sur les plantes, animaux, champignons et microbes. L'ITIS est en partenariat avec l'USGS (Institut d'études géologiques des Etats-Unis) et le Smithsonian Institute (Complexe regroupant des muséums, des structures de l'éducation et de la recherche).

Known unknowns : dans le domaine des produits naturels, désigne les molécules qui n'ont pas pu être dérépliquées mais qui sont déjà connues, par opposition aux *unknown unknowns* qui sont des molécules qui n'ont pas pu être dérépliquées et qui n'ont jamais été identifiées.

LC, LC-MS, LC-MS/MS : voir *Liquid Chromatography – Mass spectrometry*.

LDB – Lichen DataBase : base de données développée au cours de cette thèse. Par défaut, LDB se réfère à une base spectrale regroupant les spectres MS/MS de molécules lichéniques. La LDB admet plusieurs variantes : la LDB-Orbitrap, -Agilent, -Waters, se référence aux spectres de la LDB acquis respectivement sur une Orbitrap Q-Exactive Focus de Thermo Fisher, sur un qToF 6530 d'Agilent et sur un qToF Xevo G2-XS de Waters. La LDB-étendue se réfère à la LDB enrichie avec des spectres autres que ceux des molécules (dé)protonées. La LDB-Lit quant à elle se réfère à une variante de la LDB sans spectres mais avec les données structurales des molécules lichéniques répertoriées dans la littérature, ainsi que leurs masses exactes et sources biologiques.

LIAS – Lichen Information System, A Global Information System for Lichenized and Non-Lichenized Ascomycetes (http://www.lias.net/) : Système d'information pour la collecte et la distribution de données descriptives sur la biodiversité des lichens et ascomycètes

non-lichénisés (Rambold 1996). Le LIAS est divisé en plusieurs parties fournissant chacune des informations complémentaires. *LIAS names* permet la recherche d'un taxon par son nom. *LIAS glossary* fournit un glossaire pour les termes utilisés dans l'identification des lichens. *LIAS Light* (Rambold et al. 2014) donne accès à la liste des lichens de leur base de données et leurs descriptions. *LIAS Metabolites* (Elix et al. 2012) est un outil d'aide à l'interprétation pour l'identification les lichens et leurs métabolites par CCM. *LIAS gtm* (Rambold et al. 2016) permet la visualisation de traits phénotypiques lichéniques à l'aide des données de LIAS Light et des répartitions fournies par le GBIF (https://www.gbif.org/).

Libmetgem (https://github.com/metgem/libmetgem) : *Librairie* pour la production de réseaux moléculaires disponible sur *Python*. Cette *librairie* a été créée pour servir dans le logiciel MetGem (Olivon et al. 2018). (Voir *Réseaux moléculaires*).

Librairie, package, bibliothèque logicielle : collection de codes prêts à être utilisés par des programmes pour des traitements spécifiques bien identifiés (Wikipedia 2019).

Lichen : association symbiotique entre un champignon (mycosymbiote) et un (parfois plusieurs) symbiote photosynthétisant (photosymbiote) pour réaliser une structure spécifique, le thalle lichénique dans lequel le mycosymbiote, qui représente 90% de la structure enserre le photosymbiote. Le photosymbiote est une algue ou/et une cyanobactérie, sur ou dans lequel le mycosymbiote peut émettre des haustoria pour assurer sa nutrition par absorption. Les lichens ne constituent pas une unité systématique, mais un groupe biologique, tous sont des champignons, dits lichénisés. Ce champignon, représentant le seul partenaire qui présente une reproduction sexuée, est utilisé pour donner un nom au lichen (Association Française de Lichénologie 2016).

Lichen à thalle crustacé : thalle formant une croûte fortement adhérente au substrat. Plus de 4/5^{e} des lichens ont des thalles crustacés (Association Française de Lichénologie 2016).

Lichen à thalle foliacé : thalle en forme de lames ayant plus ou moins l'apparence de feuilles constituées de lobes diversement orientées (thalle foliacé à rosettes) ou de squamules ombiliquées (thalle foliacé ombiliqué) (Association Française de Lichénologie 2016).

Lichen à thalle fruticuleux : thalle plus ou moins buissonnant, dressé, pendant ou prostré, n'adhérant au substrat que par une surface très réduite. On peut souvent discerner un tronc principal et des rameaux (primaires et secondaires) ainsi qu'une symétrie plus ou moins radiaire avec un cortex périphérique, une couche algale autour d'une médulle centrale (Association Française de Lichénologie 2016).

Liquid Chromatography – Mass Spectrometry (LC-MS, LC/MS) : Technique par laquelle un mélange d'analytes est séparé en composants individuels par chromatographie liquide (généralement une chromatographie liquide à haute performance), suivie d'une détection avec un spectromètre de masse (Murray et al.

2013). Par défaut dans cette thèse, l'analyse en spectrométrie de masse est faite en tandem (LC-MS/MS) en DDA.

Liste d'exclusion dynamique : dans le contexte des analyses LC-MS/MS en DDA, liste de rapports *m/z* déjà analysés en MS/MS, exclus de toute nouvelle analyse pendant un certain temps fixé par l'utilisateur. Ceci permet d'analyser en MS/MS d'autres ions que les quelques-uns les plus intenses détectés à un moment donné.

Lock mass : masse d'un ion de rapport m/z connu, dérivé d'un standard adapté introduit dans une source d'ions avec l'échantillon analysé, permettant une recalibration en temps réel en corrigeant les décalages de m/z résultant de la dérive instrumentale (Murray et al. 2013).

Machine learning (Apprentissage automatique, apprentissage machine) : étude des algorithmes informatiques qui s'améliorent automatiquement par l'expérience (Mitchell 1997). Les algorithmes de *machine learning* permettent par un grand nombre de répétitions, d'automatiser de façon empirique certaines tâches, comme la production d'un spectre MS/MS à partir d'une structure donnée.

Massif isotopique : pics représentant des ions de la même composition élémentaire, mais de composition isotopique différente (Murray et al. 2013).

Métabolomique : Étude complexe et complète du métabolome ; identification et quantification des petites molécules / produits métaboliques d'un système biologique (cellule, tissu, organe, fluide biologique ou organisme) à un moment précis dans le temps (Labuda et al. 2018). Le terme « métabolome » est apparu pour la première fois en 1998 pour désigner la part des métabolites dans le contenu des tissus (Oliver et al. 1998).

MetaboLights (https://www.ebi.ac.uk/metabolights/) : Base de données pour les expériences de *métabolomique* et des données dérivées. MetaboLights est le site de dépôt recommandé par plusieurs journaux (Haug et al. 2013).

Métabolite : Tout composé chimique du système biologique qui n'est pas codé génétiquement et qui est un substrat, un intermédiaire ou un produit du métabolisme ; qui est consommé de l'environnement extérieur ; ou qui provient de micro-organismes endogènes, tels que la microflore intestinale (Labuda et al. 2018).

Métabolite primaire : Métabolites directement impliqués dans les processus normaux de la vie de chaque cellule (Labuda et al. 2018). Réflexion des auteurs : Cette définition, bien que communément admise, est particulièrement floue du fait de son recours au concept de normalité dans la vie d'une cellule. Intuitivement, les métabolites primaires renvoient aux sucres, aux acides gras, aux acides aminés et aux bases nucléiques, composés associés au fonctionnement « vital » d'une cellule, sans lesquels, la cellule mourrait. Cependant, aucune frontière nette ne différencie un métabolite primaire d'un métabolite secondaire. Les métabolites résultant d'une combinaison entre métabolite primaire et secondaire sont *de facto* considérés comme des métabolites secondaires. Par

ailleurs, certains de ces métabolites primaires peuvent être absents de la cellule sans pour autant entrainer sa mort.

Métabolites secondaires : Métabolites produits par des voies autres que les voies métaboliques normales, principalement après la phase de croissance active et dans des conditions de carence. La signification biologique de nombreux métabolites secondaires n'est pas exactement connue (McNaught, Wilkinson, and Chalk 2019). Réflexion des auteurs : la définition des métabolites secondaires est encore plus floue que celle des métabolites primaires. Elle peut ici se résumer aux petites molécules qui ne sont pas des métabolites primaires et qui sont produites dans des conditions « anormales » pour la cellule. Ils sont régulièrement considérés comme des mécanismes non essentiels à la vie, bien que souvent associés à des mécanismes de défense contre les prédateurs et pathogènes, fonction pourtant indispensable à la vie d'un organisme. Du fait de cette fonction défensive, les métabolites secondaires ont été recherchés pour leurs propriétés bioactives, bien que leurs fonctions dans l'organisme producteur restent largement inconnues. Ils ont aussi été rebaptisés *métabolites spécialisés* pour faire allusion à la production limitée à certains taxons, par opposition aux *métabolites non-spécialisés* qui sont partagés par la majorité des organismes. Bien que ces définitions restent floues, elles sont meilleures que celles communément admises des métabolites primaires et secondaires. Elles sont basées sur un critère qui pourrait être mesuré, en l'occurrence l'abondance à travers le vivant, et non sur la base de voies de biosynthèse choisies arbitrairement et le fonctionnement « normal » d'une cellule.

Métabolome : Ensemble quantitatif complet de composés organiques ou inorganiques de masse moléculaire relativement faible (de 50 à 1000) - métabolites - présents dans la cellule ou l'organisme et qui participent aux réactions métaboliques nécessaires à la croissance, au maintien et au fonctionnement normal (Labuda et al. 2018).

MGF – Mascot Generic Format : format standard pour le stockage de spectres MS/MS en protéomique, à présent communément utilisé dans le domaine de la métabolomique. Les champs essentiels incluent la masse du précurseur (PEPMASS), la charge (CHARGE) et les paires *m/z* – intensité (Matrixscience 2019; Fiehn Lab 2016). D'autres champs peuvent être ajoutés pour compléter les métadonnées du spectre. Un exemple de la structure d'un fichier MGF est donné dans la **Figure 9**.

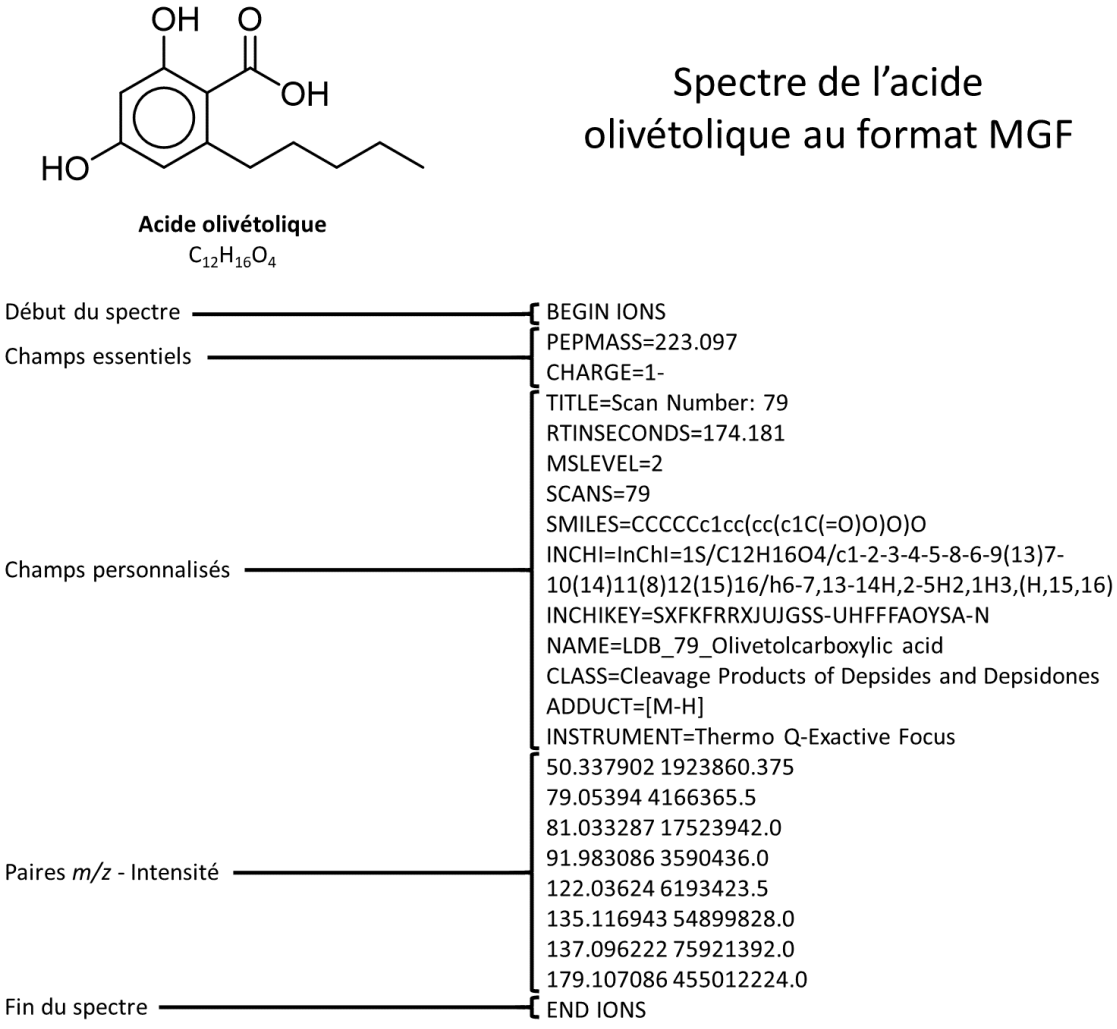


Figure 9 – Spectre MS/MS individuel pour l'acide olivétolique au format MGF. Des champs personnalisés ont été rajoutés pour rapporter le nom de la molécule, sa structure, sa classe chimique, l'instrument d'acquisition et la forme sous laquelle elle a été ionisée.

Molécule déprotonée : Ion formé par l'élimination d'un proton d'une molécule M pour produire un anion représenté par $[M-H]^-$ (Murray et al. 2013).

Molécule protonée : *Ion adduit*, représenté par $[M+H]^+$, formé par l'interaction d'une molécule avec un proton (Murray et al. 2013).

MS^1 : Mass Spectrometry[1], voir *spectre de masse*.

MS^2 : Mass Spectrometry[2], voir *spectre MS/MS* et *Spectrométrie de masse en tandem*.

MS2LDA – Mass Spectrometry[2] Latent Dirichlet Allocation (http://ms2lda.org/) : Logiciel en ligne permettant d'extraire à partir des spectres de fragmentation (MS/MS) des motifs récurrents de façon non-supervisée. Ces motifs, appelés *Mass2Motif*, représentent des fragments ou des pertes de neutres assimilables à des marqueurs biochimiques et permettent de regrouper les molécules à partir de sous-structures communes (van der Hooft et al. 2016).

MS/MS : Mass Spectrometry / Mass Spectrometry, voir *spectre MS/MS* et *Spectrométrie de masse en tandem*.

MSConvert : outil disponible sur *ProteoWizard*, en ligne de commande ou avec interface graphique, permettant de convertir plusieurs formats propriétaires de données *LC-MS* en formats ouverts, tels que le *MZXML* et le *MGF* (voir *ProteoWizard*) (Adusumilli and Mallick 2017).

MycoBank (http://www.mycobank.org/) : base de données taxonomique en ligne sur les champignons au service de la communauté mycologique et scientifique, permettant également des alignements de séquences. *MycoBank* est en partenariat avec l'IMA (International Mycological Association), le Westerdijk Fungal Biodiversity Institute (Pays-Bas), et le DGfM (Societé Allemande de Mycologie).

MZmine (http://mzmine.github.io/) : logiciel open-source pour le traitement de données de spectrométrie de masse, avec une focalisation sur les données *LC-MS*. Le but de MZmine est d'offrir un outil facile à prendre en main, flexible et améliorable avec un ensemble de modules couvrant l'intégralité du traitement des données *LC-MS* (Pluskal et al. 2010).

MZXML – MZ eXtensible Markup Language : format *XML* dédié aux données de spectrométrie de masse, chaque champ de données étant encadré de balises contenues entre un chevron « < » et « > » pour une balise d'ouverture de champ de donnés, « </ » et « > » pour une balise de fermeture de champ (Wikipedia 2020c). Le mzXML a été créé en 2004 dans le contexte de la protéomique pour résoudre les problèmes liés aux formats propriétaires des différents spectromètres de masse. Ces problèmes sont notamment : l'impossibilité de l'échange des données d'un appareil à l'autre et donc l'impossibilité de la comparaison des données entre plusieurs laboratoires ; et l'impossibilité pour la communauté bioinformatique d'accéder aux données pour le développement de logiciels (Pedrioli et al. 2004; Lin et al. 2005). Bien que ce format soit fréquemment utilisé, d'autres formats plus récents comme le mzML (combinaison du mzData et du mzXML) sont jugés plus appropriés (Deutsch 2008). Un exemple de la structure d'un mzXML est donné en **Figure 10**.

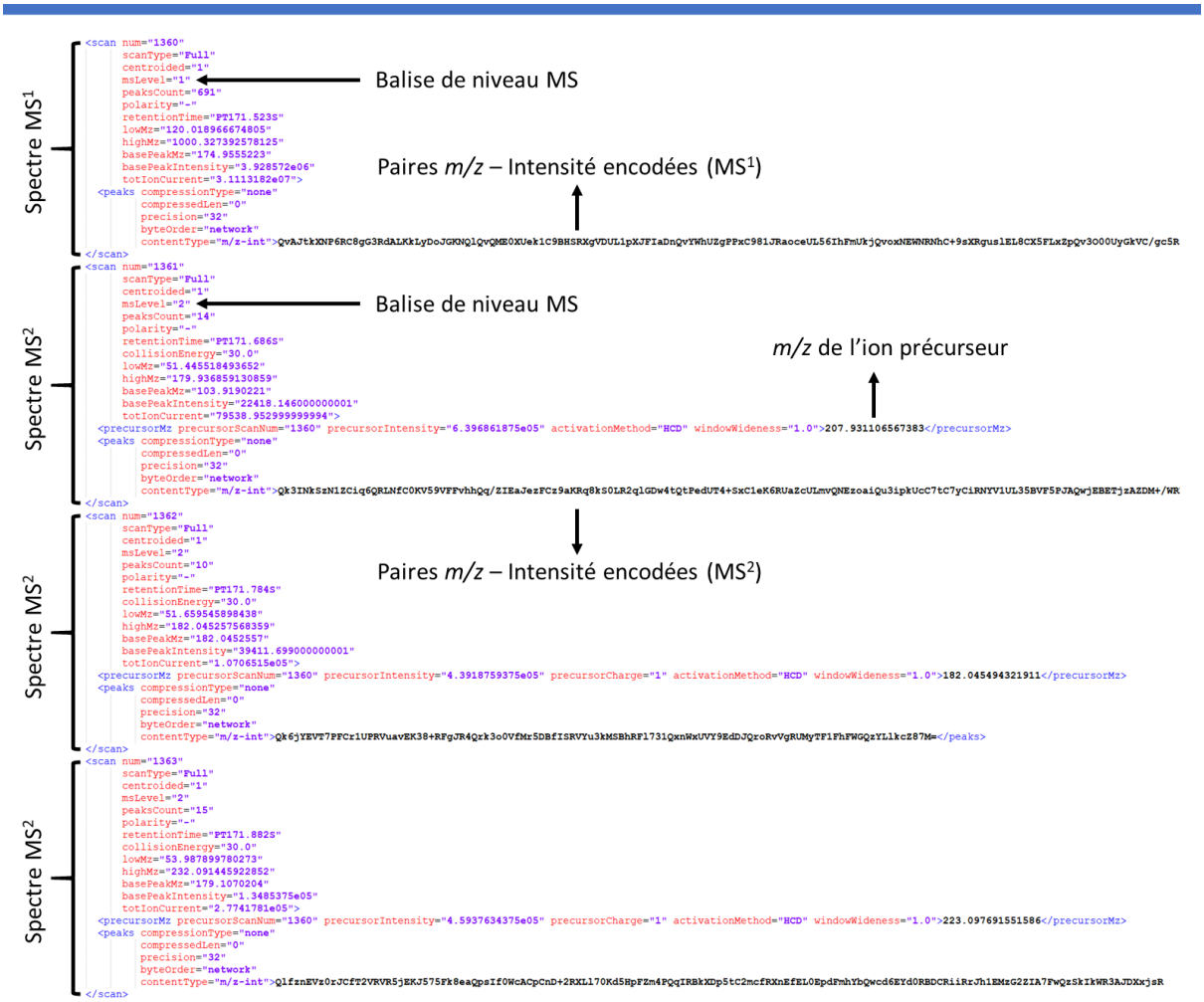


Figure 10 – Extrait de données LC-MS/MS par DDA au format mzXML. Sont représentés un spectre MS[1] et tous les ions détectés, suivi des spectres MS[2] des trois ions les plus intenses détectés dans le spectre MS[1]. Les champs relatifs aux conditions expérimentales ont été omis. Les éléments qui composent cet exemple sont « scan » (données relatives à un spectre de masse donné), composé des sous-éléments « precursorMz » (données de l'ion précurseur) et « peaks » (données sur le spectre de masse). Chaque élément est accompagné d'un certain nombre d'attributs qui rapportent les données de chaque élément.

m/z : Rapport *m/z* (voir *Rapport masse sur charge*).

NCBI – National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) : structure établie le 4 novembre 1988 dans le but de créer des systèmes automatisés pour stocker et analyser les connaissances sur la biologie moléculaire, la biochimie et la génétique. Le NCBI donne l'accès à diverses ressources, notamment à des données de taxonomie dans le cadre de cette thèse.

Node table (« Tableau de nœuds ») : tableau utilisé dans un réseau, contenant tous les attributs de chaque nœud.

Nœud (node) : objet dans un *graphe* ou un *réseau,* éventuellement relié à d'autre nœuds par le biais d'une *arête*. Dans le contexte des réseaux moléculaires, il représente généralement un ion associé à un spectre MS/MS. Dans ce contexte, les termes *ion* et

nœud sont utilisés de façon interchangeable (voir *Graphe*, *Théorie des graphes*, *Réseaux moléculaires*).

OpenBabel (http://openbabel.org/wiki/Main_Page) : Logiciel open-source permettant de convertir, analyser et stocker des données de modélisation moléculaire, de chimie, des matériaux solides, de biochimie et d'autres secteurs (O'Boyle et al. 2011).

Open-source (code-source ouvert, logiciel libre) : Se dit d'un logiciel dont le code source est libre d'accès, réutilisable et modifiable (Dictionnaire Larousse en Ligne n.d.).

Orbitrap : marque déposée par la compagnie Thermo Scientific pour désigner un analyseur à trappe orbitale. Ils sont composés d'une électrode externe en forme de tonneau et d'une électrode interne coaxiale en forme de tige, formant un champ électrique. Les ions vont emprunter des trajectoires orbitalaires autour de l'électrode interne et la fréquence de leurs oscillations est inversement proportionnelle à leurs rapports *m/z*, ce qui permet de les mesurer par transformée de Fourrier.

Pic de base (Base Peak, BP) : Pic dans un *spectre de masse* avec l'intensité la plus élevée (Murray et al. 2013).

Perte neutre : Perte d'une espèce non chargée d'un *ion* lors de sa dissociation (Murray et al. 2013).

ProteoWizard (http://proteowizard.sourceforge.net/) : Logiciel open-source rassemblant plusieurs librairies et outils pour faciliter l'analyse en protéomique (M. C. Chambers et al. 2012; Kessner et al. 2008). Ces outils sont à présent également utilisés en métabolomique.

Python : langage de programmation, version utilisée : 3.7.3 (Van Rossum and Drake 2009).

Q-ToF : spectromètre de masse hybride consistant d'un analyseur quadripolaire (Q) couplé à un analyseur de temps de vol (TOF : Time of Flight).

R : langage de programmation, version utilisée : 3.6.1 (R Development Core Team 2008).

Rapport masse sur charge, *m/z* : valeur sans unité, calculée par la division de la masse m (kg) d'un ion par sa charge q (C). Le « *rapport m/z* » sera régulièrement simplifié en « *m/z* » dans la thèse.

$$Rapport\ masse\ sur\ charge\ = \frac{m}{q}$$

Réaction de fragmentation : Réaction d'un ion qui produit deux ou plusieurs fragments dont au moins un est un *ion* (Murray et al. 2013).

Réaction thalline : usage de réactifs chimiques sur le thalle, ou autre partie du lichen, pour provoquer une coloration mettant en évidence la présence de certains composés, permettant la diagnose du lichen.

Regroupement : ici, traduction du mot « clustering » en anglais, en l'occurrence le regroupement des nœuds dans un réseau sous forme de blocs, soit par similarité cosinus, soit par d'autres méthodes (*Fragnotator*, *Adnotator*…).

Réseaux moléculaires : présentation sous forme de réseaux de l'espace chimique dans des expériences LC-MS/MS. Chaque nœud (spectre MS/MS d'un ion) peut être relié à un autre nœud par une arête si le *score de similarité cosinus* de la paire est supérieur à un seuil fixé par l'utilisateur (généralement 0.7). Le postulat de base est que si deux spectres de fragmentation sont similaires, les structures des molécules desquelles ils sont issus sont similaires. Un réseau moléculaire est ainsi composé de plusieurs *blocs* (clusters) de spectres proches, qui peuvent être assimilés à des familles moléculaires. Certains nœuds ne seront jamais assez similaires à d'autres pour dépasser le seuil fixé, et formeront des *blocs* à un nœud relié à lui-même : des *self-looped nodes*. Certaines molécules structurellement différentes peuvent en revanche être regroupées dans le même bloc, notamment des molécules présentant peu de fragments en dehors de certains groupements chimiques en commun et fréquents comme les acides carboxyliques. Un autre intérêt des réseaux moléculaires est l'association d'attributs aux nœuds et aux arêtes de ces derniers, comme le $\Delta m/z$ et le score de similarité entre deux ions, le temps de rétention, le rapport m/z de l'ion précurseur du nœud, la structure associée au nœud si celui-ci a été dérépliqué, l'intensité de l'ion précurseur etc... (**Figure 11**). Ainsi, un réseau peut fournir de nombreuses informations permettant de faciliter l'interprétation des données LC-MS. Un nœud dérépliqué dans un bloc peut impliquer que les autres molécules du même bloc soient des dérivés structuraux. Le temps de rétention permet de supposer la polarité du composé étudié et les $\Delta m/z$ permettent de vérifier si l'ion voisin est un fragment de source, un adduit ou un dérivé potentiel. La plateforme du *GNPS* a permis de démocratiser l'usage des réseaux moléculaires (Wang et al. 2016; Yang et al. 2013; Watrous et al. 2012; Nguyen et al. 2013). Des solution locales et faciles à prendre en main sont à présent disponibles depuis la mise à disposition de la librairie *libmetgem* sur *Python* et du logiciel MetGem (Olivon et al. 2018).

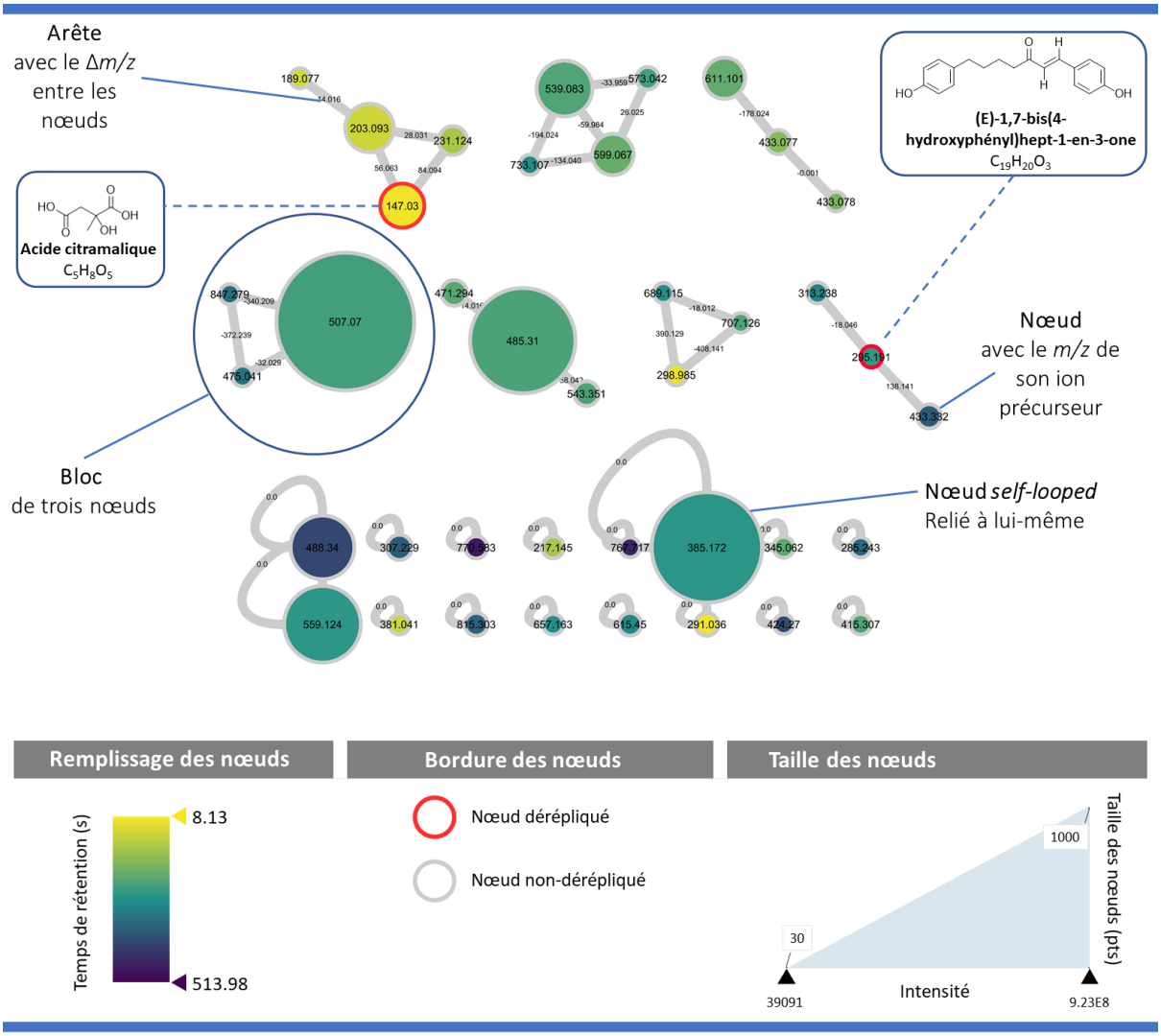


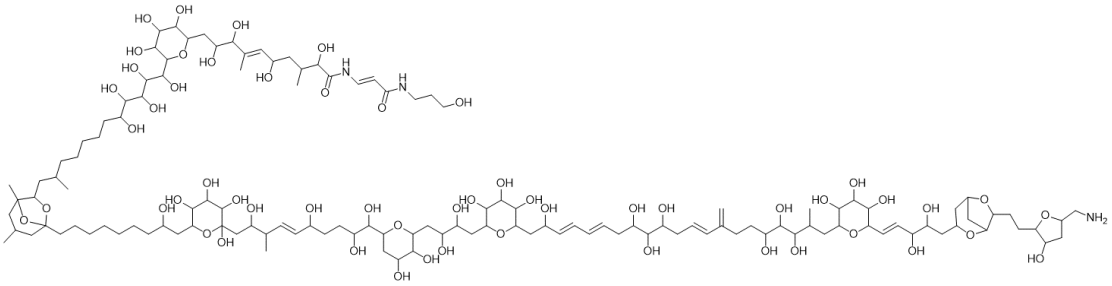
Figure 11 – Exemple de réseau moléculaire avec des attributs.

Similarité cosinus appliquée à la spectrométrie de masse : Méthode pour évaluer la similarité entre deux spectres MS/MS. Les spectres sont dans un premier temps filtrés pour ne conserver que les 6 pics les plus intenses par fenêtre de 50 Da, ce qui permet de maximiser les chances de conserver des pics signaux plutôt que des pics de bruit. Les spectres sont ensuite divisés en fenêtres de 2 Da et la présence d'un pic dans la fenêtre est notée 1, l'absence est notée 0. Chaque spectre est ainsi associé à une chaîne de bits et peut-être défini comme un vecteur dans une sphère à n-dimensions. La distance spectrale peut alors être définie comme la distance Euclidienne des points sur la sphère. Ainsi, la similarité entre un spectre *x* et *y* séparés d'un angle θ peut être calculée par cos(θ), où des spectres similaires auront un score proche de 1, et des spectres dissimilaires auront un score proche de 0 (Dutta and Chen 2007).

$$\cos\theta = \frac{x \cdot y}{\|x\| \, \|y\|}$$

SMILES – Simplified Molecular Input Line Entry Specification : Système de notation chimique utilisable dans le traitement de l'information chimique. Basé sur les principes de la *théorie des graphes*, le SMILES permet une spécification rigoureuse des structures

par l'utilisation d'une grammaire réduite et naturelle. La notation en SMILES est par ailleurs bien adaptée aux traitements informatiques intensifs. Ceci permet à l'utilisateur de générer une notation unique, de récupérer des molécules dans des bases de données, la recherche de sous-structure et la prédiction de propriétés chimiques (Weininger 1988; Weininger, Weininger, and Weininger 1989; Weininger 1990). Un exemple de code SMILES est donné en **Figure 12** pour la palytoxine en comparaison de son nom IUPAC.



Nom IUPAC : 10-[6-[12-[5-[9-[6-[10-[6-[4-[6-[21-[6-[5-[7-[2-[5-(aminomethyl)-3-hydroxyoxolan-2-yl]ethyl]-2,6-dioxabicyclo[3.2.1]octan-3-yl]-3,4-dihydroxypent-1-enyl]-3,4,5-trihydroxyoxan-2-yl]-2,8,9,10,17,18,19-heptahydroxy-20-methyl-14-methylidenehenicosa-3,5,12-trienyl]-3,4,5-trihydroxyoxan-2-yl]-2,3-dihydroxybutyl]-4,5-dihydroxyoxan-2-yl]-2,6,9,10-tetrahydroxy-3-methyldec-4-enyl]-3,4,5,6-tetrahydroxyoxan-2-yl]-8-hydroxynonyl]-1,3-dimethyl-6,8-dioxabicyclo[3.2.1]octan-7-yl]-1,2,3,4,5-pentahydroxy-11-methyldodecyl]-3,4,5-trihydroxyoxan-2-yl]-2,5,8,9-tetrahydroxy-*N*-[3-(3-hydroxypropylamino)-3-oxoprop-1-enyl]-3,7-dimethyldec-6-enamide

SMILES : CC1CC2(C(OC(C1)(O2)CCCCCCCC(CC3C(C(C(C(O3)(CC(C(C)C=CC(CCC(C(C4CC(C(C(O4)CC(C(CC5C(C(C(C(O5)CC(C=CC=CCC(C(C(CC=CC(=C)CCC(C(C(C(C)CC6C(C(C(C(O6)C=CC(C(CC7CC8CC(O7)C(O8)CCC9C(CC(O9)CN)O)O)O

Taxonomie : branche des sciences naturelles qui a pour objet de décrire la diversité des organismes vivants et de les regrouper en entités appelées taxons afin de les identifier (notamment grâce aux clés de détermination), les décrire, les nommer et les classer (Wikipedia 2020e).

Temps de rétention (TR, RT) : dans le contexte de la chromatographie, temps de migration d'un analyte à travers une phase stationnaire entre son injection et sa détection par le détecteur. Abrégé TR (anglais : RT, Retention Time).

Thalle lichénique : terme créé en 1810 par Acharius pour désigner l'appareil végétatif des lichens, ce terme a ensuite été adopté pour désigner, dans la classification deux règnes (animaux et végétaux), l'appareil végétatif des végétaux dépourvus de racines, de tiges, de feuilles et de système de vascularisation (algues, champignons, lichens regroupés au sein des thallophytes) (Association Française de Lichénologie 2016).

Théorie des graphes : Etude des *graphes*, structures qui modélisent les relations entre objets. La théorie des graphes permet de représenter et organiser des tâches de façon optimale : après avoir traduit un problème sous forme de *graphe*, on cherche des méthodes systématiques qui permettent de trouver la succession la plus rapide ou la moins coûteuse pour effectuer toutes les tâches (Wikipedia 2020f; Encyclopédie Larousse en ligne n.d.). Un réseau est un graphe auquel des attributs ont été rajoutés aux nœuds et aux arêtes (**Figure 13**), comme les réseaux moléculaires dans le cadre de cette thèse (Voir *Réseaux moléculaires*) (Wikipedia 2020g).

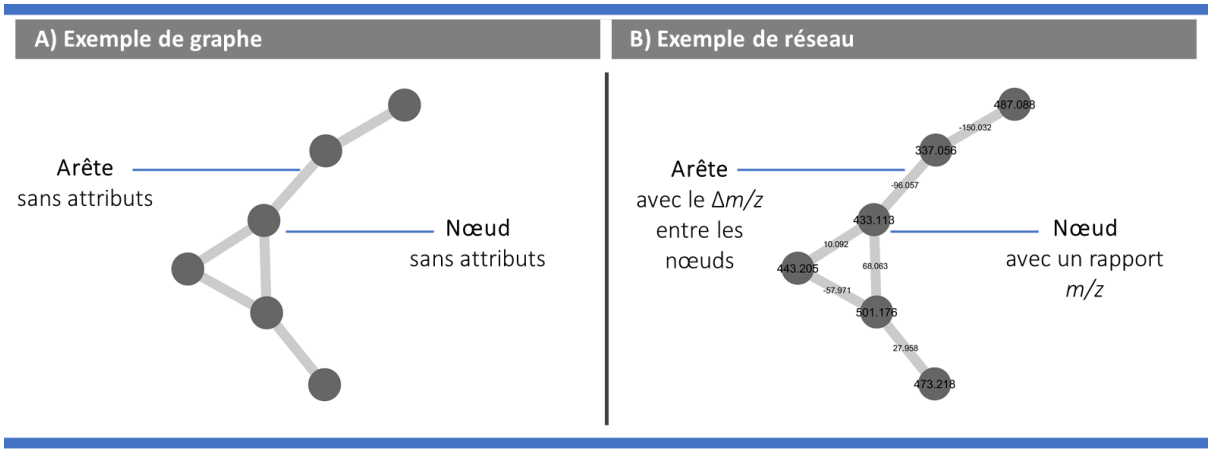


Figure 13 – Un graphe et de réseau. Le réseau est ici représenté avec des attributs utilisés dans les réseaux moléculaires, notamment le rapport m/z de chaque nœud qui représente un ion, et les valeurs sur les arêtes représentent la différence de rapport m/z entre les nœuds reliés.

TIC : Total Ion Current (voir *Courant Ionique Total*).

TR : Temps de rétention (voir *Temps de rétention*).

TSV – Tab-Separated Values : média de texte séparant les valeurs par des tabulations. Le fonctionnement est le même que celui du *CSV*, avec les éléments de chaque ligne séparés par un tabulation (IANA 2020).

UHPLC – Ultra High Performance Liquid Chromatography : Chromatographie liquide à ultra haute performance (voir Chromatographie en phase liquide).

UPLC – Ultra Performance Liquid Chromatography : marque déposée par la compagnie Waters pour désigner des systèmes chromatographiques (voir Chromatographie en phase liquide).

Unknown unknowns : dans le domaine des produits naturels, désigne les molécules qui n'ont pas pu être dérépliquées et qui restent inconnues à ce jour, par opposition aux *known unknowns* qui sont des molécules connues qui n'ont pas pu être dérépliquées.

2

Recontextualisation dans le cadre de la thèse

Compte tenu des sujets abordés, il est nécessaire d'aller au-delà des définitions et de démontrer davantage le rôle que les réseaux et les outils de visualisation des données LC-MS/MS vont occuper.

2.1 Les réseaux moléculaires :

Les réseaux moléculaires se sont démocratisés dans le domaine de la métabolomique et des produits naturels depuis la mise en place du GNPS en 2016 pour permettre aux utilisateurs de créer leurs propres réseaux (Wang et al. 2016). Bien que cette plateforme permette de créer un réseau sans avoir à coder, la logique derrière ces derniers et les subtilités de l'acquisition de données échappent encore largement à la plupart des chercheurs. La production d'un réseau moléculaire est détaillée ici en suivant un *Feature-Based Molecular Networking* (FBMN) passant par MZmine et le GNPS (Olivon et al. 2017; Nothias et al. 2019), bien qu'il existe d'autres méthodes (**Figure 14**).

Acquisition des données : Un réseau moléculaire provient nécessairement d'une analyse LC-MS/MS, idéalement en DDA avec une liste d'exclusion dynamique et une vitesse de balayage assez élevée pour fragmenter autant d'ions que possible à un temps de rétention donné. Pour pouvoir traiter les données LC-MS en amont de la création du réseau moléculaire, elles sont converties de leur format propriétaire à un format ouvert (ici en mzXML). Si le logiciel fournisseur ne permet pas cette conversion, MSConvert (Adusumilli and Mallick 2017) est utilisé. Les fichiers convertis contiendront les enchainements de spectres MS^1 et MS^2 à chaque temps de rétention, avec pour les spectres MS^2, le rapport *m/z* de l'ion précurseur. Ce lien entre un spectre MS/MS et l'ion précurseur est indispensable à la création des réseaux moléculaires. Par conséquent, les analyses dites *données indépendantes* telles que le mode MS^E (Waters) ne sont pas adaptées. Des algorithmes de protéomique existent pour reformer ce lien mais ce mode d'analyse reste marginal dans le domaine de la métabolomique.

Traitement des données LC-MS : Il existe plusieurs logiciels open-source pour le traitement de données LC-MS et tous suivent plus ou moins les mêmes procédés. MZmine 2 (Katajamaa, Miettinen, and Orešič 2006; Pluskal et al. 2010) est utilisé ici. Un seuil de bruit est d'abord fixé pour les spectres MS^1 et MS^2, puis une détection de pics est réalisée, produisant une liste de pics (les *Features* du FBMN). Cette liste de pics sera déconvoluée pour éliminer les signaux de « bruit » et séparer les ions isobares. Les massifs isotopiques sont éliminés et les listes de pics de tous les échantillons sont combinées en une seule liste par l'alignement de leurs pics. Les ions auxquels aucun spectre MS/MS n'est associé sont éliminés avant d'exporter cette liste finale de spectres MS/MS au format MGF. Il est important de signaler à ce stade qu'un ion ne représente pas une molécule unique dans l'analyse : il peut s'agir de la molécule (dé)protonée, d'un adduit, d'un fragment de source

ou d'un complexe ion-neutre. Les logiciels, dont MZmine, fournissent généralement des algorithmes pour annoter ces pics.

Génération du réseau moléculaire : Les ions de la liste sont comparés deux à deux en calculant leur score similarité cosinus (Dutta and Chen 2007). Si le score de la paire est supérieur au seuil fixé par l'utilisateur (habituellement 0.7), les deux ions sont reliés par une arête. Certains ions ne présentent jamais de similarité suffisante avec un autre ion pour dépasser le seuil : une paire est alors formée en reliant l'ion à lui-même (*self-looped*). Une fois les comparaisons terminées, les paires sont soumises à un logiciel de visualisation de réseaux (Cytoscape par exemple (Shannon et al. 2003)) qui va construire le réseau à partir du fichier de paires de nœuds (*edge table*). Chaque ion est représenté sous forme de nœud, plusieurs nœuds regroupés formant un bloc (*cluster*). Deux blocs sont formés lorsqu'aucun des ions entre les deux blocs ne sont assez similaires pour être reliés par une arête. Les données relatives à chaque nœud sont importées séparément : le GNPS fournit notamment des données extraites du fichier MGF comme le temps de rétention, le rapport *m/z* de l'ion précurseur et de nombreuses informations relatives à la déréplication de l'ion si elle a eu lieu (structure sous forme de SMILES ou d'InChI, la base de données spectrale utilisée, le score de similarité avec le spectre de référence etc…). D'autres fichiers d'informations complémentaires sur les nœuds peuvent être importés, notamment les fichiers exportés depuis MZmine qui comportent l'intensité de l'ion précurseur pour chaque échantillon dans lequel il a été détecté. Ces fichiers, issus du GNPS ou de MZmine, contiennent des données relatives à chaque nœud : ce sont des fichiers attributs ou *node tables*.

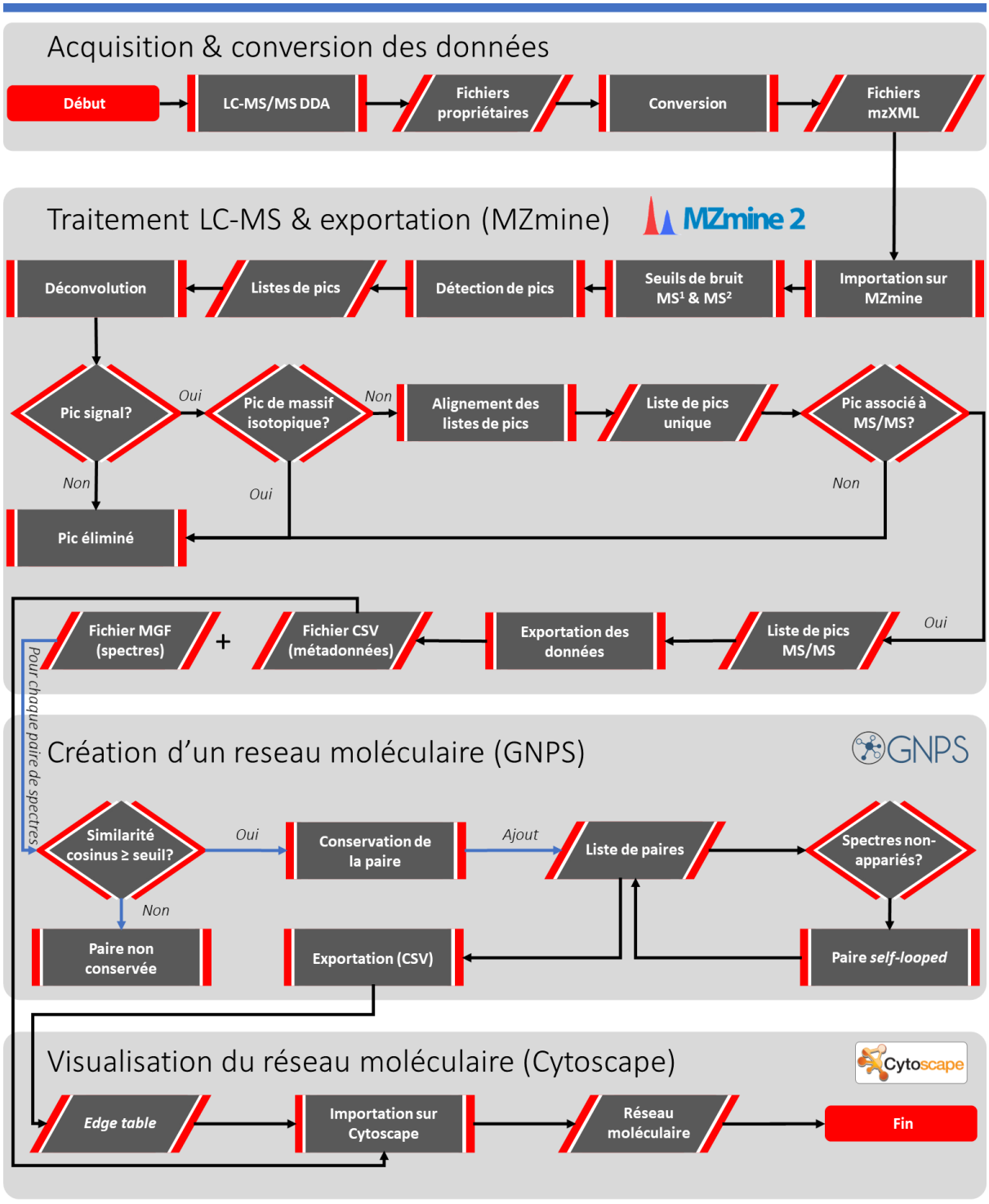


Figure 14 – Exemple d'un Feature-Based Molecular Networking (FBMN) utilisant MZmine, le GNPS et Cytoscape.

2.2 Autres réseaux utilisés :

Un réseau n'étant qu'une représentation graphique de données, il est possible de les utiliser dans d'autres contextes que les données LC-MS. Certains réseaux présentés ici ne sont pas des réseaux moléculaires et n'ont pas nécessairement recours aux données MS/MS et aux regroupements par similarité cosinus.

Dans le *Chapitre I*, des réseaux ont été produits pour visualiser la diversité chimique des lichens sur la base de la bibliographie. Ils représentent sous forme de nœuds plusieurs taxons de champignons lichénisés allant du genre à la classe. Ces nœuds sont reliés en suivant la logique de la taxonomie : les genres d'une même famille sont reliés au nœud de cette famille, les familles à leurs ordres et les ordres à leurs classes.

Comme il s'agit d'étudier la diversité chimique décrite dans cette taxonomie, les nœuds sont associés à des attributs reflétant cette diversité. Les nœuds sont d'abords colorés en fonction de leur niveau taxonomique. Leur diamètre est proportionnel à la quantité de molécules qui y sont décrites. Le nombre de molécules décrites à chaque niveau taxonomique est par ailleurs représenté sur chaque nœud.

L'exploration d'un tel réseau permet d'observer rapidement le degré d'étude de chaque taxon : ceux qui sont sous-étudiés ou sur-étudiés (**Figure 15**).

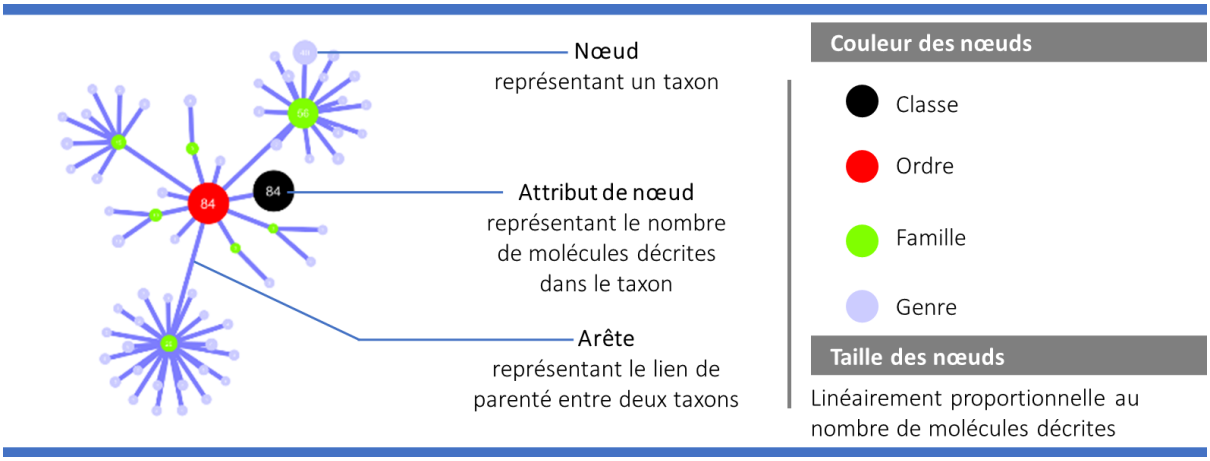


Figure 15 – Réseau représentant une classification taxonomique allant de la classe au genre, avec pour chaque taxon, le nombre de molécules qui y ont été décrites dans la littérature.

Dans le même esprit, puisque qu'il est possible de classer les composés d'une base de données avec la même logique que les classifications taxonomiques (ClassyFire (Djoumbou Feunang et al. 2016), NPClassifier (Kim et al. 2020)), la même représentation en réseau est envisageable. Dans ce même *Chapitre I*, une base de données est soumise à ClassyFire sur la base de codes SMILES pour associer à chacune de ses molécule une classification ontologique. Comme précédemment, les nœuds « parents » seront reliés à leurs « enfants », représentant les catégories chimiques. Pour chaque catégorie, le nombre de molécules qui y sont décrites dans la base de données est reporté numériquement ainsi qu'indirectement par le diamètre des nœuds. Ceux-ci sont là encore colorés en fonction de leur niveau taxonomique (**Figure 16**).

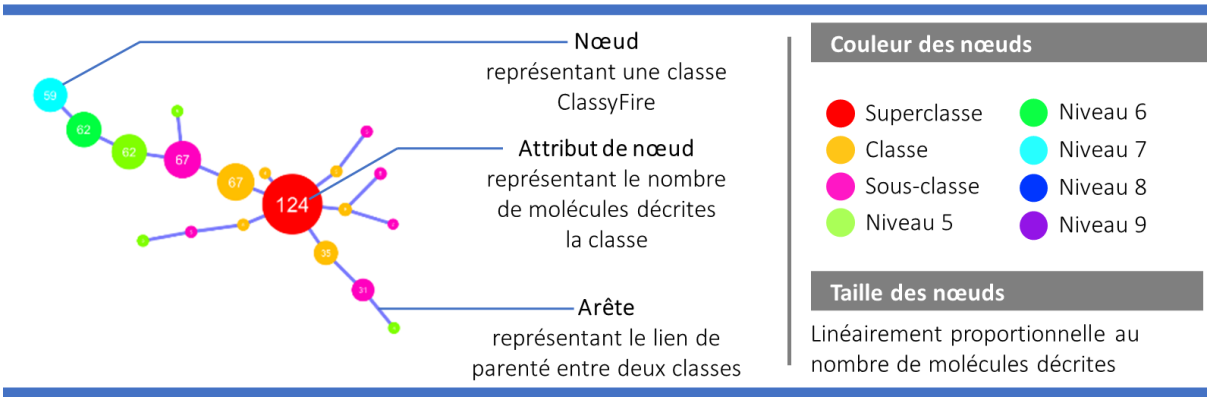


Figure 16 – Réseau représentant les molécules d'une base de données classée par ClassyFire avec pour chaque classe / nœud, le nombre de molécules qui y ont été répertoriées.

Ces réseaux n'impliquent pas de données LC-MS, mais des données issues de la littérature et de bases de données. Ci-dessous sont abordés des variantes de réseaux moléculaires qui sont eux, basés sur la LC-MS. Ils n'utilisent cependant pas de regroupement par similarité cosinus, une méthode presque indissociable des réseaux moléculaires. Il convient alors de les désigner comme des variantes de réseaux moléculaires.

Un outil développé dans cette thèse permet un regroupement sur la base des relations de fragmentation entre les ions lors d'une analyse LC-MS : *Fragnotator*. La portion occupée par les fragments de source dans les analyses LC-MS n'est que rarement abordée et cet outil a pour but d'y apporter une réponse. Sur la base de leurs spectres MS^2, un ion précurseur est relié à son fragment de source qui présente en toute logique le même temps de rétention (**Figure 17**). Une section est dédiée à l'explication détaillée de *Fragnotator* plus bas.

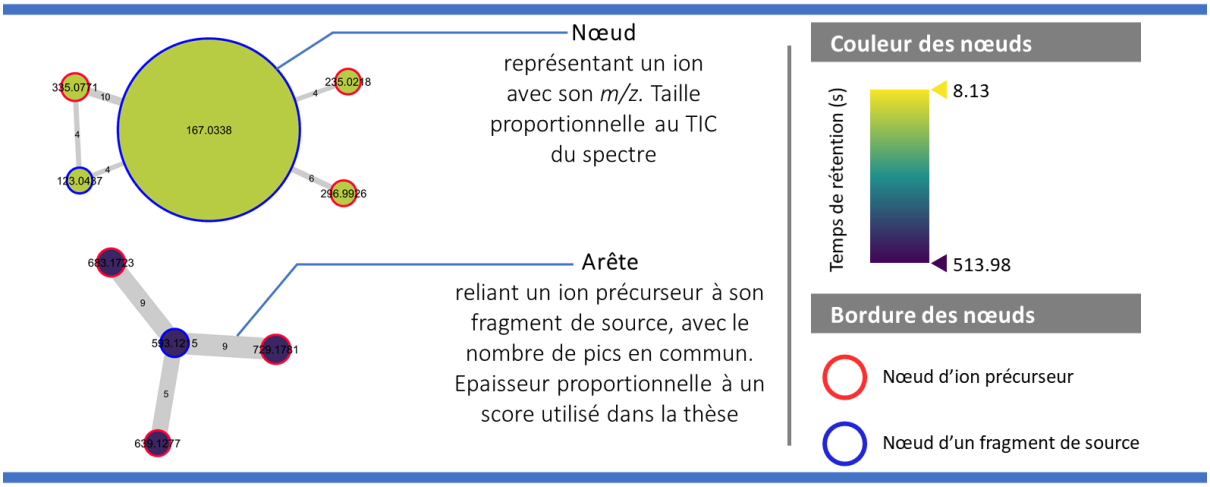


Figure 17 – Réseau de fragmentation, formant des blocs à partir d'ions précurseurs et leurs fragments de source.

Une autre variante de réseau moléculaire a été développée dans cette thèse avec la création d'*Adnotator* (voir le paragraphe dédié, plus bas). Ces réseaux sont basés sur le regroupement des différents adduits pointant vers une même molécule neutre

hypothétique. Comme pour les fragments de source, la proportion des différents adduits dans une analyse LC-MS est trop rarement prise en compte.

Ces réseaux peuvent être complétés par les données d'un réseau de fragmentation comme celui décrit au-dessus, ce qui permet de regrouper autour d'une molécules ses ions, adduits ou fragments de source. Du fait des fragmentations parfois similaires des différents adduits d'une même molécule, certains d'entre eux peuvent être considérés comme fragments d'un ion et en même temps comme une variante d'adduit de la molécule neutre calculée (**Figure 18**). Par ailleurs, la prédiction des molécules dans un mode d'ionisation permet de faire le lien avec la même molécule dans un autre mode, comme il sera démontré plus tard.

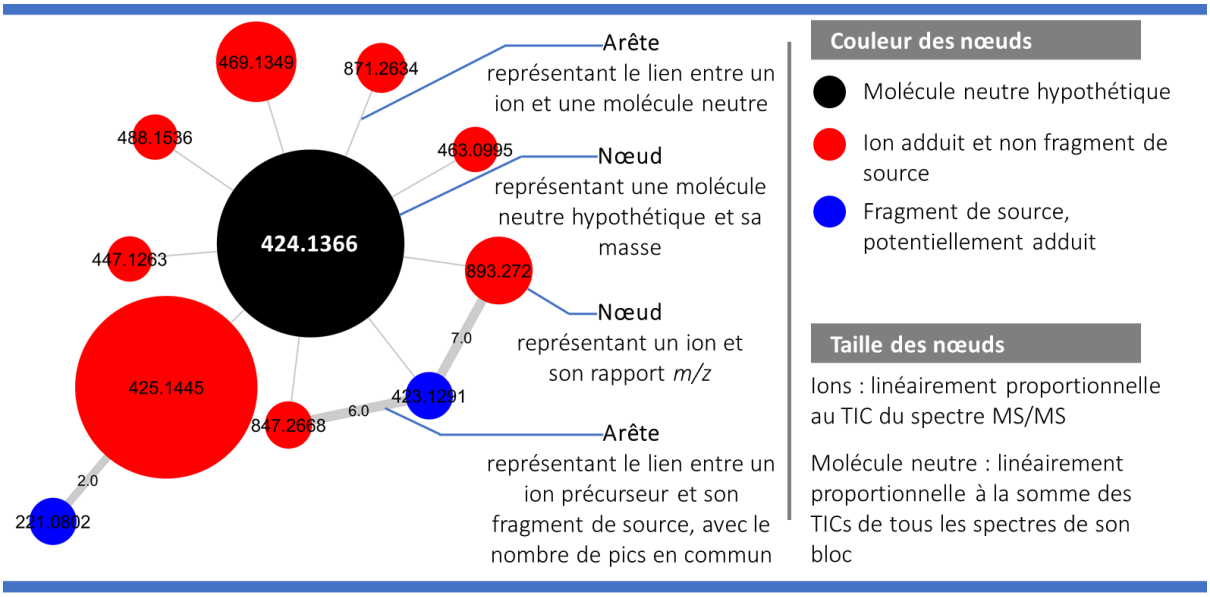


Figure 18 – Réseau produit par Adnotator, regroupant tous les ions produits par une molécule neutre hypothétique.

Pour finir, il convient de signaler que l'esthétique d'un réseau se fait à la discrétion de l'utilisateur. Ainsi, il est possible de représenter les nœuds sous forme de carrés, de triangles, d'étoiles ou n'importe quelle autre forme que le rond, les arêtes peuvent être droites, recourbées, orientées ou non. Les couleurs sont arbitraires : si dans la **Figure 17** les nœuds sont colorés en fonction d'un gradient de temps de rétention, ils sont colorés en noir rouge ou bleu dans la **Figure 18** selon qu'ils représentent des neutres, des adduits ou des fragments de source. Plus tard, des variantes de rouge représenteront des cations et les bleus des anions. Le diamètre des nœuds dans les **Figures 15** et **16** représente le nombre de molécules décrites pour la classe en question, alors qu'il représente le TIC du spectre dans les **Figures 17** et **18**. Les seules légendes valables sont celles fournies avec la figure du réseau.

2.3 Termes spécifiques à *Fragnotator* :

Fragnotator : outil développé dans le cadre de cette thèse sur Python 3.7 permettant de regrouper les fragments de source (*Ions 2*) d'une analyse LC-MS/MS à leur ion précurseur

coélué (*Ions 1*). Ceci est fait par le biais d'un *matching score* et d'un seuil de pics partagés. Le résultat du traitement peut être visualisé sous la forme d'un réseau (**Figure 19**).

Bloc de fragmentation : ensemble de nœuds reliables par des arêtes produites par *Fragnotator*, donc par des relations précurseur-fragment de source.

Ion 1 : ion précurseur d'un *Ion 2* qui est, dans la majorité des cas, son fragment de source. L'*Ion 1* doit contenir dans son spectre MS/MS le rapport *m/z* de l'*Ion 2*.

Ion 2 : cet ion est en principe un fragment de source de l'*Ion 1*, mais il peut également s'agir d'un autre ion / adduit de la molécule du fait de leur fragmentation proche et de la coélution. Ceci a par ailleurs l'avantage de capter des adduits insoupçonnés par l'utilisateur. Un *Ion 2* est toujours considéré comme « Fragment », même s'il est lui-même précurseur d'un autre fragment de source (**Figure 19**).

Matching score : score utilisé par *Fragnotator* calculé en divisant le nombre pics partagés par l'*Ion 2* avec l'*Ion 1*, par le nombre total de pics de l'*Ion 2*.

Seuil de pics partagés : nombre de pics minimal que doivent partager les spectres de l'*Ion 1* et de l'*Ion 2*.

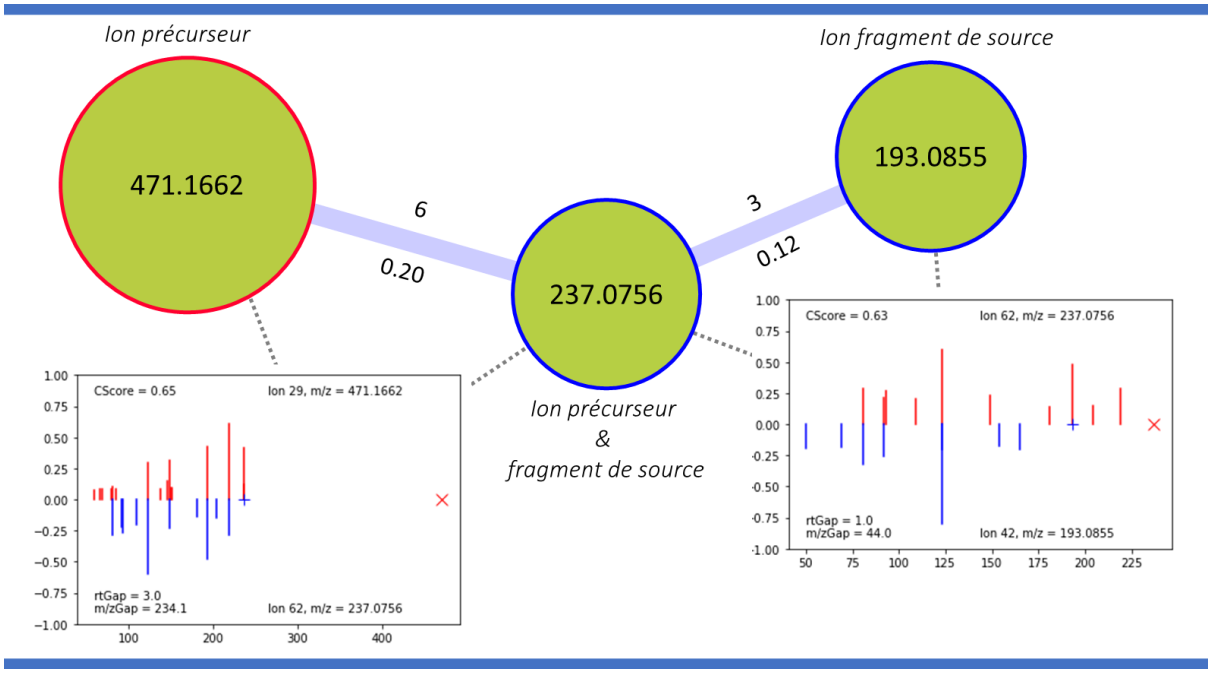


Figure 19 – Bloc de fragmentation généré par Fragnotator. Il est constitué d'un ion précurseur (bords rouges) partageant avec un fragment de source (bords bleus) 6 pics (matching score de 0.20). Ce fragment de source est lui-même le précurseur d'un autre fragment de source avec lequel il partage 3 pics (score de 0.12).

2.4 Termes spécifiques à *Adnotator* :

Adnotator : outil développé dans le cadre de cette thèse sur Python 3.7 permettant de regrouper à une molécule neutre hypothétique tous ses ions. La masse de cette molécule est calculée par triangulation en choisissant les ensembles de combinaisons adduit –

neutre – adduit qui interprètent mieux des données LC-MS. Le résultat peut être visualisé sous la forme d'un réseau.

Bloc moléculaire : un neutre et les ions qui lui sont directement reliés (ses adduits). Ils sont habituellement contenus dans des blocs plus grands qui comprennent également les fragments de source ainsi qu'éventuellement d'autres blocs moléculaires.

Tableau d'adduits : tableau contenant tous les adduits à rechercher dans les données LC-MS/MS (**Tableau 2**). Il peut être fourni par l'utilisateur ou produit par *Adnotator*.

Tableau 2 – Exemple d'un tableau d'adduits.

Code adduit	*Adduit*	*q*	*Δm*	*x*	*C*
M1|m1H|pHCOOH	$[M-H+HCOOH]^-$	-1	44.997655	1	3
M1|m1H|	$[M-H]^-$	-1	-1.007825	1	1
M1|p1Cl|	$[M+Cl]^-$	-1	34.968853	1	2
M1|m2Hp1Na|pHCOOH	$[M-2H+Na+HCOOH]^-$	-1	66.9796	1	5
M1|m2Hp1Na|	$[M-2H+Na]^-$	-1	20.97412	1	4
M2|m1H|pHCOOH	$[2M-H+HCOOH]^-$	-1	44.997655	2	4
M2|m1H|	$[2M-H]^-$	-1	-1.007825	2	3
M2|p1Cl|	$[2M+Cl]^-$	-1	34.968853	2	3
M2|m2Hp1Na|pHCOOH	$[2M-2H+Na+HCOOH]^-$	-1	66.9796	2	6
M2|m2Hp1Na|	$[2M-2H+Na]^-$	-1	20.97412	2	5
M2|m2Hp1K|	$[2M-2H+K]^-$	-1	36.948058	2	5
M3|m1H|	$[3M-H]^-$	-1	-1.007825	3	4

Le *Code adduit* représente une chaîne de caractères permettant à *Adnotator* d'explorer les différents composants d'un ion en conservant son écriture conventionnelle dans la colonne *Adduit*. Il est créé à partir de trois éléments fondamentaux pouvant composer un ion, à savoir une partie « molécule » composée de x molécules M, une partie « ion » composée de y espèces chargées I et d'une partie « neutre » composée de z espèces neutres N. Une charge q résulte de la combinaison de l'ensemble des parties :

$$Ion = \left[\binom{M}{x} + \binom{I}{y} + \binom{N}{z}\right]^q$$

Ces *Codes adduits* sont accompagnés de paramètres permettant les calculs d'*Adnotator* : la charge q, une masse Δm (masses totales de $I + N$), x et un score de complexité C ($x + y + z$, sauf pour les molécules (dé)protonées pour lesquelles $C = 1$). Ce tableau est utilisé comme base pour calculer toutes les molécules neutres possibles pour chaque ion d'une analyse LC-MS, et également tous les ions pouvant être générés à partir d'une molécule neutre.

Ion 1 : ion sélectionné dans un fichier MGF dont l'identifiant, le rapport *m/z*, la charge et le temps de rétention sont utilisés pour générer ses molécules neutres hypothétiques et l'associer à un *Ion 2* avec lequel il est coélué qui pourrait correspondre à un autre ion d'une molécule neutre générée (*voir* **Figure 20**).

Hypothèse d'adduit 1 : forme hypothétique que revêt un *Ion 1*, sélectionnée à partir du tableau d'adduits. Elle permet de calculer la *molécule neutre hypothétique* pour l'*Ion 1*.

Hypothèse de neutre / Molécule neutre hypothétique : masse hypothétique correspondant à la molécule neutre de l'*Ion 1* calculée à partir de l'*Hypothèse d'adduit 1*.

Hypothèse d'adduit 2 : forme hypothétique que revêt un *Ion 2*, sélectionnée à partir du tableau d'adduits. Ces hypothèses ne se font que sur la base d'une *Hypothèse d'adduit 1* et d'une *Hypothèse de neutre* spécifiques, renvoyant vers un potentiel *Ion 2* coélué présentant le bon rapport *m/z* pour expliquer l'ensemble des hypothèses.

Ion 2 : ion coélué avec l'*Ion 1* à ΔTR près, dont le rapport *m/z* peut correspondre à un ion produit par la *molécule neutre hypothétique* via une *Hypothèse d'adduit 2*.

Hypothèse de relation : paire d'ions reliés par une molécule neutre et les *Hypothèses d'Adduit 1* et *2* leur attribuant une annotation (**Figure 20**). Chaque *Hypothèse de relation* est accompagnée d'un score *HRS* calculé de la façon suivante :

$$HRS = \frac{1 + F + N_c}{C}$$

Les éléments permettant de calculer HRS sont les suivants :

- *F* (fragmentation) : par défaut 0, ou 1 si l'*Ion 1* et *2* sont déjà reliés par *Fragnotator* (souvent le cas pour deux ions de la même molécule)
- N_c (confirmation de la fraction neutre) : par défaut 0, ou à 1 si l'*Ion 2* possède une partie neutre (N) et qu'un autre ion sans cette partie neutre est détecté parmi les ions coélués.
- *C* (complexité) : le score de complexité de l'Ion 2 (*voir plus haut,* ***Tableau 2***).

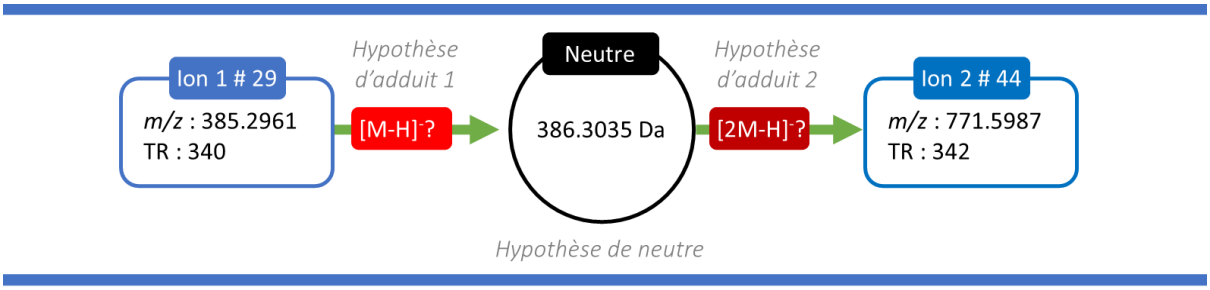


Figure 20 – Création d'une Hypothèse de relation. Elle est constituée d'un Ion 1, sa molécule neutre produite à partir de son Hypothèse d'adduit 1, et un Ion 2 détecté pouvant correspondre à un autre ion de la même molécule neutre à partir de l'Hypothèse d'adduit 2.

Cohorte : ensemble d'*Hypothèses de relation dépendantes*, c'est-à-dire des *Hypothèses de relation* partageant au moins un ion (**Figure 21**). L'annotation de ces ions peut être consensuelle (même annotation pour un ion donné) ou conflictuelle (différentes annotations pour le même ion). Les *Hypothèses de relation* présentant des annotations conflictuelles sont dites *incompatibles* et celles ne présentant que des annotations consensuelles et/ou indépendantes sont dites *compatibles*.

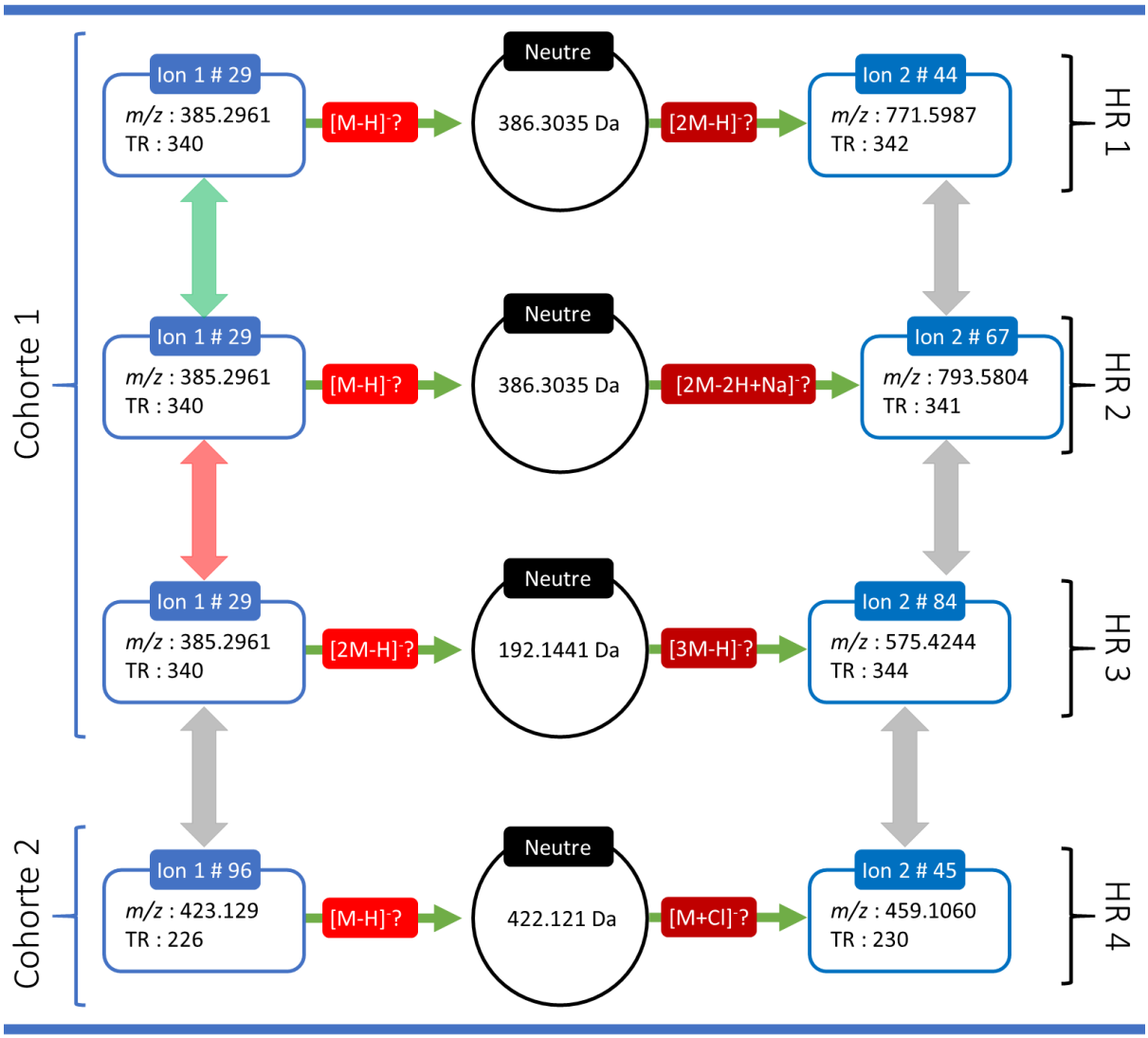


Figure 21 – Regroupement des Hypothèses de relation dépendantes en Cohortes. Les flèches entre les ions représentent des annotations consensuelles (vert), conflictuelles (rouge) ou indépendantes (gris). HR : Hypothèse de relation.

Cour : ensemble d'*Hypothèses de relation compatibles* au sein d'une *Cohorte*. Les *cours* avec le meilleur score sont retenues pour annoter les données LC-MS. Le score d'une *cour* est calculé avec la somme des scores des *Hypothèses de relation* qui la composent.

2.5 Termes spécifiques à *Classnotator* :

Classnotator : outil développé dans le cadre de cette thèse sur Python 3.7 permettant d'annoter les ions d'une analyse LC-MS/MS à partir de motifs produits sur MS2LDA (van der Hooft et al. 2016) et d'une classification structurale fournie par l'utilisateur. *Classnotator* nécessite une base de données MS/MS annotée avec une classe structurale pour chaque ion. La base de données est ensuite soumise à MS2LDA pour rechercher les mass2motifs (motifs ou ensemble de pics récurrents entre spectres MS^2). Les données MS2LDA sont ensuite utilisées pour trouver des motifs ou combinaisons de motifs spécifiques à certaines classes structurales de la base de données (« *motifs purs* »). Ces *motifs purs* sont accompagnés d'un score *n_mol* représentant le nombre de molécules de

la base de données dans lesquelles ils ont été détectés. Plus *n_mol* est élevé, plus le *motif pur* a des chances d'être retrouvé dans d'autres molécules de la même classe chimique, un *n_mol* faible (par exemple : 1) signifiant que le motif est spécifique d'une molécule et de ses dérivés éventuels plutôt que de sa classe structurale. Ces *motifs purs* peuvent ensuite être utilisés pour annoter des ions dans une analyse LC-MS après détection de ses motifs sur MS2LDA en utilisant la même base de données de motifs. Si une annotation résulte en l'attribution de plusieurs classes pour un ion, celle avec le *n_mol* cumulé le plus élevé sera sélectionnée.

Tableau molmotif : tableau reliant des molécules (« mol », lignes) aux motifs détectés sur MS2LDA (« motif », colonnes). Les paires molécules-motif sont notées 1 si le motif a été détecté dans la molécule, un 0 qu'il n'a pas été détecté.

Motif pur : motif ou combinaison de motifs spécifiques à une classe structurale pour un jeu de données.

2.6 Remarques supplémentaires :

Fragnotator, *Adnotator* et *Classnotator* sont les principaux modules d'un plus grand outil : *Molnotator* (Python 3.7). La compréhension du fonctionnement de ces modules permet d'avoir une idée globale du fonctionnement de *Molnotator* qui lui, ne sera abordé que dans le *Chapitre V*. De façon sommaire, *Molnotator* est un outil développé ici comme une alternative aux réseaux moléculaires classiques pour représenter directement des molécules et non des ions sur les réseaux. Il se base sur un haut degré de triangulation sur les modes d'ionisation positive et négative pour présenter chaque molécule dans son contexte d'ionisation et réduire au maximum les données redondantes. Le but ultime de *Molnotator* est ici d'explorer la diversité chimique des lichens, chose qui n'est pas possible avec les outils actuellement disponibles dans le milieu de la métabolomique. Bien que *Molnotator* puisse être utilisé dans d'autres contextes, les méthodes de déréplication qu'il emploie utilisent des bases de données spécifiques aux lichens. Une grande partie de ces travaux se focalisent sur le développement de ces bases de données.

Chapitre I

– LDB-Lit –

Ou la création d'une base de données répertoriant les métabolites lichéniques décrits dans la littérature, utilisée ici pour l'étude de la diversité chimique des lichens

Intervenants extérieurs : aucun.

Résumé

Dans ce premier chapitre, une base de données est créée à partir de données bibliographiques pour permettre un état des lieux de ce qui est connu de la chimie des lichens. Elle s'intitule LDB-Lit (Lichen DataBase – Literature) et combine les données issues de 283 publications, d'ouvrages comme Hun&Yosh96 ainsi que de bases de données comme le LIAS Light. Son but premier est de fournir des informations sur les substances lichéniques, chaque composé étant conservé sous la forme de codes SMILES, InChI et InChIKey. Ce sont pour l'instant 1662 métabolites qui y sont représentés, associés à leurs lichens producteurs notamment grâce aux données du LIAS. La LDB-Lit est utilisée dans ce chapitre pour représenter la diversité chimique des composés lichéniques ainsi que leur répartition au sein des différents taxons de champignons lichénisés. La LDB-Lit n'est pas encore rendue disponible et les possibilités d'hébergement sont encore discutées.

Sommaire

1 – Introduction 49

2 – Méthodes 51

2.1 Numérisation des données Hun&Yosh96 51

2.2 Extraction des données de la littérature 51

2.3 Web scraping du LIAS 51

2.4 Classification & mise-à-jour des espèces 52

2.5 Classification des molécules par ClassyFire 53

2.6 Représentations sous forme de réseaux 54

3 – Résultats 55

3.1 Numérisation des données Hun&Yosh96 55

3.2 Extraction des données de la littérature 56

3.3 Web scraping du LIAS 56

3.4 Diversité des produits lichéniques étudiée avec *ClassyFire* 57

3.5 Répartition taxonomique des métabolites lichéniques 61

4 – Conclusion 68

4.1 De la création de la LDB-Lit 68

4.2 De la diversité chimique des lichens et sa distribution 68

4.3 De l'étude du métabolome des lichens 68

1 Introduction

Les sources d'information sur les métabolites lichéniques restent à ce jour limitées. Une étape majeure dans l'évolution de nos connaissances sur les métabolites lichéniques a été la publication du livre « Identification of Lichen Substances » par S. Huneck et I. Yoshimura en 1996 (Huneck and Yoshimura 1996), abrégé « Hun&Yosh96 » pour la suite. Ce livre regroupe les résultats de la majorité des travaux précédents sur le sujet en présentant pour chaque molécule des données physicochimiques. Parmi ces études se trouvent les travaux d'isolement de W. Zopf et de O. Hesse à la du fin XIXe et au début du XXe siècle (Zopf 1897; Hesse-Feuerbach 1912; Zopf 1895b, 1907, 1895a), les expériences de microcristallisation de Y. Asahina et S. Shibata (Shibata 2000; Asahina 1951; Asahina and Shibata 1954), les travaux de déréplication des Culberson (C. F. Culberson and Kristinsson 1970; C. F. Culberson and Culberson 1978; C. F. Culberson, Culberson, and Esslinger 1977), les études phytochimiques de S. Huneck, ainsi que les données d'analyse par LC-DAD et d'isolement de G. Feige (Feige et al. 1993), I. Yoshimura (Yoshimura et al. 1994) et J. Elix (Elix and Crook 1992; Elix and Gaul 1986; Elix, Jenie, and Parker 1987). Ce compendium donne un aperçu de ce qui était globalement connu dans la chimie des lichens à sa parution. Les pronostics de l'époque envisageaient une évolution exponentielle des connaissances sur les produits lichéniques qui irait de pair avec l'amélioration des techniques d'analyse (C. F. Culberson and Culberson 2001) (**Figure 22**).

Des travaux complémentaires ont été entrepris, parmi ceux-là, les mises-à-jour par S. Huneck (Huneck 1999, 2001, 2006) et les travaux d'isolement ainsi que les contributions aux données de migration sur CCM par Elix (Elix 2014; Schumm and Elix 2015) ainsi qu'une mise à jour sur les xanthones lichéniques (Le Pogam and Boustie 2016). En ce qui concerne les bases de données, les composés lichéniques y sont rarement présents, étant relayés à sites spécialisés. Parmi ces sites spécifiques, le LIAS, qui traite en détail des ascomycètes lichénisés et leurs molécules. Ce projet a été initié durant le symposium de l'IAL 5 en 1995 (Rambold 1996) et la base de donnés est disponible en ligne depuis 2007 (Rambold and Triebel 2007). Bien que constamment mises à jour, ses données sont dédiées à des buts taxonomiques et naturalistes. Par conséquent les informations sur les molécules se limitent à leur nom et aux données de migration sur CCM et HPLC disponibles. Ces données sont regroupées dans un outil complémentaire : LIAS metabolites (Elix et al. 2012; Rambold et al. 2014, 2016; Mietzsch, Lumbsch, and Elix 1993). Wintabolites est un autre outil pour l'identification de métabolites lichéniques à partir d'analyses CCM, similaire à ce qui est disponible dans le LIAS (Mietzsch, Lumbsch, and Elix 1992; Lafferty and Bungartz 2018). Comme ce dernier, il utilise le catalogue de composés produit par J. Elix, en plus des données du CNALH (Consortium of North American Lichen Herbaria, https://lichenportal.org/cnalh/index.php). Le LIAS et Wintabolites contiennent un nombre de molécules similaires, respectivement 844 et 833.

Cependant, les techniques permettant d'étudier la chimie des lichens incluent de plus en plus de spectrométrie de masse couplée à de la chromatographie en phase liquide. Les bases de données évoquées précédemment ne donnent pas les informations nécessaires à la déréplication par LC-MS (structure, masse exacte). Il a par conséquent été entrepris de développer ici une base de données, la LDB-Lit : Lichen DataBase – Literature. Elle a été réalisée grâce à l'étude d'articles de la littérature, de livres et de bases de données disponibles. Pour chaque molécule, la structure a été enregistrée sous forme de codes SMILES, InChI et InChIKey. Les espèces dans lesquelles ces métabolites ont été décrits ont également été rapportées, avec leur nom mis à jour et associé à leur classification taxonomique. La LDB-Lit permettra de guider l'utilisateur lors d'une analyse LC-MS et peut notamment être utilisée pour l'étude de la diversité chimique des lichens, comme il sera démontré dans ce chapitre. Les options d'hébergement de cette base de données sont encore discutées.

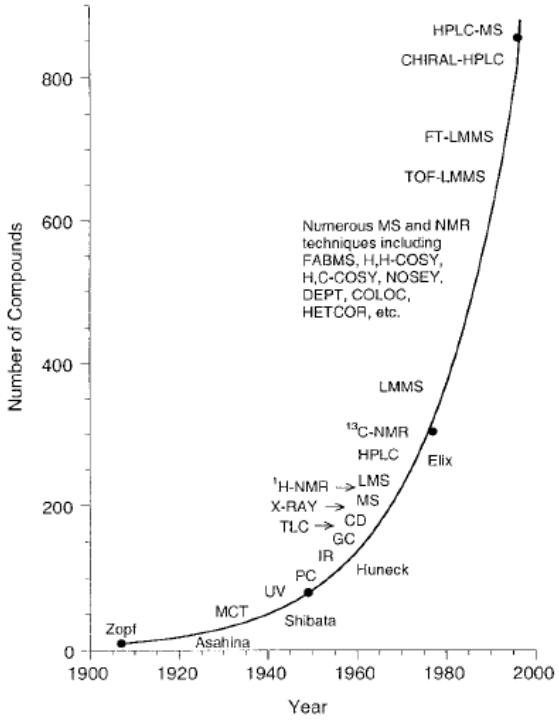


Figure 22 – Croissance attendue du nombre de molécules d'origine lichénique telle que décrite par Culberson (C. F. Culberson and Culberson 2001).

2 Méthodes

2.1 Numérisation des données de Hun&Yosh96.

La première étape entreprise pour produire la LDB-Lit a été de numériser les données contenues dans Hun&Yosh96, c'est-à-dire de reporter dans un format CSV les noms des molécules, leurs sources biologiques et leurs structures. Ceci a été fait manuellement et les structures ont été enregistrées en format SMILES, InChI et InChIKey. L'intégration des données sous cette forme a permis l'usage d'outils chémoinformatiques pour rajouter leurs formules brutes et leurs masses exactes.

2.2 Extraction de données de la littérature.

Des rajouts ont été faits à l'aide de 283 articles traitant de sujets allant de la détection à l'isolement de molécules lichéniques. Aucun jugement n'a été fait sur la qualité des données recueillies et l'utilisateur est libre de choisir par lui-même celles qui lui paraissent fiables en étudiant les sources de chaque donnée. Ces articles ont été choisis en ciblant arbitrairement certains genres de lichens, tant que le nombre de publications les concernant n'était pas trop important. Ces genres étaient *Caloplaca*, *Evernia*, *Hypogymnia*, *Lichina*, *Lobaria*, *Ophioparma*, *Porpidia*, *Pseudevernia*, *Ramalina*, *Roccella* et *Stereocaulon*. Pour les raisons d'exhaustivité et pour ne pas avoir à étudier des articles déjà abordés par la suite, l'intégralité des informations qui y sont contenues ont été extraites, même si elles traitaient d'autres genres de lichens. Les articles déjà étudiés sont alors mis sur une liste d'exclusion et réduire progressivement la quantité de travail.

2.3 Web scraping du LIAS.

Les données contenues dans le LIAS ont été collectées automatiquement par *Web Scraping* à l'aide de Python 3.7 (Van Rossum and Drake 2009) et de la librairie *Beautiful Soup* 4 (L. Richardson 2007). La page des espèces (http://liaslight.lias.net/) a été ciblée et les noms des lichens ainsi que la liste des molécules qu'ils contiennent ont été rapportés dans un tableau (exemple en **Figure 23**).

Index	Link	Species	Metabolites
0	http://liaslight.lias.net/Descriptions/ItemID_11827.html	Absconditella amabilis	NA
1	http://liaslight.lias.net/Descriptions/ItemID_14231.html	Absconditella antarctica	NA
2	http://liaslight.lias.net/Descriptions/ItemID_2.html	Absconditella celata	NA
3	http://liaslight.lias.net/Descriptions/ItemID_3.html	Absconditella delutula	NA
4	http://liaslight.lias.net/Descriptions/ItemID_4.html	Absconditella lignicola	NA
5	http://liaslight.lias.net/Descriptions/ItemID_5.html	Absconditella pauxilla	NA
6	http://liaslight.lias.net/Descriptions/ItemID_6.html	Absconditella sphagnorum	NA
7	http://liaslight.lias.net/Descriptions/ItemID_7.html	Absconditella trivialis	NA
8	http://liaslight.lias.net/Descriptions/ItemID_8126.html	Acanthothecis abaphoides	protocetraric acid
9	http://liaslight.lias.net/Descriptions/ItemID_12355.html	Acanthothecis africana	norstictic acid
10	http://liaslight.lias.net/Descriptions/ItemID_9362.html	Acanthothecis aquilonia	NA
11	http://liaslight.lias.net/Descriptions/ItemID_13727.html	Acanthothecis archeri	NA
12	http://liaslight.lias.net/Descriptions/ItemID_12356.html	Acanthothecis aurantiaca	connorstictic acid|norstictic acid|stictic acid

Figure 23 – Echantillon du tableau contenant les données sur LIAS. « NA » signifie l'absence de données. Les caractères spéciaux (α, β, ë, ö…) ont été remplacés par des lettres ou des mots compatibles avec la plupart des logiciels (alpha, beta, e, o…).

Un autre tableau a été produit en créant une liste de tous les métabolites uniques présents sur le LIAS (en se basant sur leurs noms) et les espèces qui les produisent leur ont été associées (**Figure 24**).

Metabolite	Species
(+)-dechlorogriseofulvin	Lecanora griseofulva
(+)-griseofulvin	Lecanora griseofulva
19-acetoxylichesterinic acid	Remototrachyna flexilis
19-acetoxyprotolichesterinic acid	Remototrachyna flexilis
1'-chloronephroarctin	Pseudocyphellaria nermula|Pseudocyphellaria pickeringii
1'-methyl hypothamnolate	Pertusaria tropica
2- methylene-3-carboxy-18-hydroxynonadecanoic acid	Hypogymnia wilfiana
2"-O-methyltenuiorin	Pseudocyphellaria billardierei|Pseudocyphellaria glaucescens
2,2'-di-O-methyldivaricatic acid	Pertusaria subplanaica
2,2'-di-O-methylstenosporic acid	Pertusaria alboaspera|Pertusaria alboaspera|Pertusaria subplanaica|Pertusaria trimera|Pertusaria verruculifera

Figure 24 – Echantillon du tableau produit à partir des données du LIAS avec, pour chaque métabolite lichénique, la liste des organismes producteurs.

Puisque le LIAS n'associe pas de structures à ses noms de molécules, elles ont été générées automatiquement quand l'opportunité se présentait, en utilisant un algorithme de conversion nom-structure ou alors en faisant le lien avec un nom déjà présent dans la LDB-Lit. Quand un traitement automatique n'était pas envisageable, les structures ont été recueillies à partir de la littérature. La LDB-Lit intègre à ce moment les données de Hun&Yosh96, de quelques centaines d'articles ainsi que celles du LIAS. Les sources pour chaque information ont été conservées dans une colonne dédiée.

2.4 Classification & mise-à-jour des espèces.

Les noms d'espèces ont été conservés tels qu'ils étaient décrits dans les sources, bien que certaines espèces aient changé de nom (comme par exemple *Cryptothelium cecidiogenum* qui est devenu *Astrothelium cecidiogenum*). Conserver les noms d'espèce tels qu'ils ont été reportés engendre cependant des problèmes : une même espèce pourrait être présente sous plusieurs entrées et son contenu chimique serait divisé. Une deuxième colonne d'espèce a été produite avec les noms actualisés pour chaque espèce, en utilisant le package *Taxize* (Chamberlain and Szocs 2013; Chamberlain et al. 2020) sur R 3.6.1 (R

Development Core Team 2008). Pour permettre cette actualisation, le nom de chaque espèce a été soumis à *Index Fungorum* (CBS and Landcare Research (custodians) 2019), une base de données spécialisée dans la taxonomie des champignons. Chaque nom changé a été revu manuellement pour corriger des problèmes d'expression régulière qui pourraient survenir. La classification taxonomique de chaque espèce a été téléchargée à partir de la base de données taxonomique du NCBI (Federhen 2012, 2015) en utilisant le package *myTAI* (Drost et al. 2018) sur R. Ainsi, chaque espèce s'est vue associée à différents niveaux de taxonomie désignés sur NCBI par les rangs *superkingdom*, *kingdom*, *phylum*, *subphylum*, *class*, *subclass*, *infraclass*, *cohort*, *order*, *suborder*, *infraorder*, *superfamily*, *family* et *subfamily*, rangs dont les noms seront écrits en français pour la suite. Pour simplifier les données, seuls les rangs « classe », « ordre » et « famille » ont été conservés puisqu'ils sont décrits pour la plupart des organismes. Les rangs au-dessus de classe n'ont pas été conservés car les lichens sont tous des eucaryotes (super règne), des champignons (règne) et presque exclusivement des ascomycètes (phylum). L'appartenance à un autre phylum pour le cas des quelques basidiolichens a été précisée dans les résultats. Malheureusement, myTAI ne permet l'accès qu'à NCBI et ITIS (https://www.itis.gov/), des bases de données qui ne sont pas spécialisées sur les champignons à la différence d'*Index Fungorum* et *Mycobank* (Crous et al. 2004; Robert et al. 2013). Les classifications produites ont donc été inspectées manuellement pour repérer les problèmes d'homonymie et les espèces absentes des bases de données du NCBI. Par exemple, les genres de lichen *Crypthothele*, *Epiphloea* et *Xyleborus* ont tous des homonymes, respectivement dans les araignées, les algues rouges et les coléoptères. Certains genres de lichens comme *Amazonomyces*, *Ameliella*, *Biatorella*, *Brasilicia* et *Loflammia* ne sont pas présents sur NCBI et leur taxonomie a été corrigée manuellement à l'aide de *MycoBank*. Par ailleurs, certaines espèces n'étaient présentes dans aucune des bases de données précédentes, n'existant que dans le LIAS (*Austroparmelina elixiana*, *Cladonia insularis*, *Fissurina saxicola*, *Flavoparmelia pseudosorediosa*, *Hypotrachyna palniensis*, *Loxospora macrosperma*, *Parmotrema olivarium*, *Platygramme norstictica*, *Ramalina azorica* and *Ramalina neopacifica*). Ces noms ont été conservés dans la LDB-Lit compte tenu de la qualité du LIAS en tant que base de données taxonomique.

2.5 Classification des molécules par ClassyFire.

Les molécules de la LDB-Lit ont été classifiées à l'aide de *ClassyFire* (Djoumbou Feunang et al. 2016) et son package dédié sur R. Cette méthode est adaptée à la situation car la LDB-Lit contient déjà les molécules sous forme de SMILES, ce qui permet de les classer automatiquement. La classification proposée par ClassyFire est par ailleurs basée sur la systématique classique avec plusieurs rangs organisés de façon hiérarchique désignés par les noms *kingdom*, *superclass*, *class*, *subclass*, et quatre niveaux additionnels (« *levels* ») numérotés 5 à 9.

2.6 Représentations sous forme de réseaux.

Deux types de réseaux sont produits dans ce chapitre :

- La répartition des molécules au sein de la classification taxonomique des lichens.
- La répartition des molécules de la LDB-Lit au sein des classes structurales de *ClassyFire*.

Ils se basent sur les ontologies de la classification taxonomique et de la classification par *ClassyFire*, en reliant chaque « taxon parent » à ses « taxons enfants » par des arêtes. Les attributs du premier type de réseau seront le nombre de molécules décrites dans chaque taxon lichénique, ceux du second seront le nombre de molécules décrites dans chaque classe *ClassyFire*. La classification taxonomique ne comprendra que le genre, la famille, la classe et l'ordre. Les groupes *ClassyFire* ont été conservés tels quels. Les *edge tables* (relations taxonomiques) et les *node tables* (attributs) ont été générés à partir des données de la LDB-Lit sur Python 3.7 (Van Rossum and Drake 2009) et les réseaux ont été créés et visualisés sur *Cytoscape* 3.7.1 (Shannon et al. 2003). A titre de comparaison, les mêmes réseaux ont été générés avec les données du LIAS et de Hun&Yosh96. La comparaison entre ces trois sources permettra d'étudier l'évolution de nos connaissances sur les métabolites lichéniques.

3 Résultats

3.1 Numérisation des données de Hun&Yosh96.

L'extraction des données de Hun&Yosh96 a été le point de départ de LDB-Lit, permettant de récupérer les données de 1011 molécules. Ce nombre est étonnamment proche du nombre total de métabolites lichéniques (1050) souvent cité dans la littérature, qui provient d'une communication personnelle par J. Elix utilisée dans un article en 2008 (Stocker-Wörgötter 2008). Ceci s'explique probablement par l'exclusion de certaines molécules de Hun&Yosh96 qui ne sont pas associées à des organismes producteurs, notamment plusieurs fragments de depsides et depsidones pour lesquels les structures ont été déduites de leurs molécules entières. Bien que ces fragments n'aient pas été décrits dans des lichens, il serait logique de les observer lors d'analyses. D'autres, comme les nostoclides, ne sont pas produites directement par un lichen mais par un *Nostoc* isolé de *Peltigera canina*. Après retrait de ces molécules, la LDB-Lit contenait 854 entrées.

Ceci a permis d'étudier la diversité chimique telle que présentée dans Hun&Yosh96. Les molécules y sont réparties en 20 grandes classes chimiques, certaines étant presque uniques aux lichens comme les depsides et les depsidones. La plupart des molécules sont des polycétides (62% : depsides, depsidones, xanthones, quinones, dibenzofuranes, diphenyléthers, chromones, depsones et benzyldepsides) et les terpènes arrivent en deuxième position avec 17% de représentants (**Figure 25**). Le livre ne contient que peu d'informations sur les lichens producteurs de ces molécules, ne fournissant qu'un seul exemple à chaque fois.

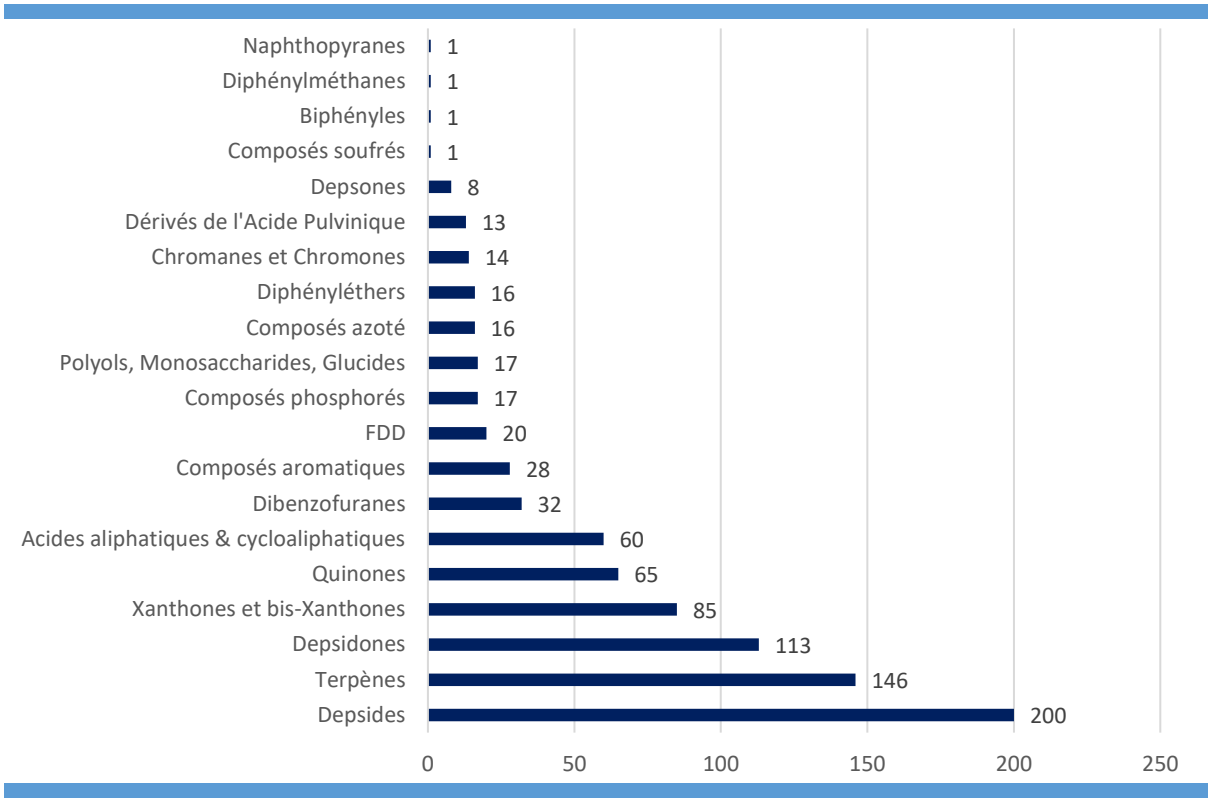


Figure 25 – Classes chimiques présentées dans Hun&Yosh96 avec le nombre de molécules présentes. FDD : Fragments de Depsides et de Depsidones.

3.2 Extraction de données de la littérature.

En rajoutant les données des 283 articles à celles de Hun&Yosh96, le nombre d'espèces dans la LDB-Lit a été augmenté à 1810 entrées (956 rajoutées) et le nombre total de composés à 1570 (716 rajoutés). Les genres recherchés ne représentent pas nécessairement les lichens les plus étudiés, comme *Cladonia*, *Usnea* et les différents genres de parmélias. L'étude de ces genres est prohibitive : les molécules du genre *Cladonia* par exemple a été abordées dans 415 articles (entre 1906 et 2018) en excluant les 283 publications déjà examinées. Au vu du nombre limité de taxons abordés, ces résultats sont tout de même significatifs : 1810 entrées pour seulement 283 articles, ce qui laisse présager bien plus de données si l'intégralité de la littérature pouvait être examinée.

Bien que le nombre de molécules et d'espèces aient été doublés, la quantité d'espèces est encore trop faible pour représenter la diversité des lichens. Pour y remédier, les données du LIAS ont été intégrées à la LDB-Lit.

3.3 Web scraping du LIAS.

Le *Web Scraping* du LIAS a permis de récupérer 10201 espèces et 844 noms de molécules. Les espèces pour lesquelles aucune molécule n'était décrite ont été éliminées des données (6103 espèces restantes). Parmi les molécules, 151 n'ont pas pu être associées

à des structures et ont été éliminées également. Ces molécules proviennent pour la plupart d'observations sur CCM : elles ont été nommées dans des publications pour permettre la description du profil chimique d'une espèce et l'identification de celle-ci par d'autres lichénologues. Leur structure n'a jamais été élucidée. Parmi ces molécules se trouvent les acides amphorothéciques A et B d'*Amphorothecium occultum* (McCarthy, Kantvilas, and Elix 2001), l'acide brialmontique et la butlerine G de *Xanthoparmelia kimberleyensis* (Elix 2003), l'acide cinnamomeique B de *Buellia thiopoliza* (van den Boom and Giralt 2011), l'eumitrine A3 d'*Usnea baileyi* (Laily et al. 2010) et d'*Hypotrachyna contradicta* (Sipman, Elix, and Nash 2009), les acides exuviatiques A, B et C issus d'espèces de *Nephroma* (Louwhoff 2005), les flavo-obscurines B1 et B2 d'*Heterodermia flabellata* (Laily et al. 2010), les isopigmentosines A et B d'espèces du genre *Menegazzia* (Kantivilas 2012), les leucomyeloconones 1 et 2, les myeloconones A1, B et C et les myelocoterpenes 1 à 3 dans des *Myeloconis* (McCarthy and Elix 1996), les acides paraensiques C et D dans le genre *Lecanora* (van den Boom and Brand 2008), la parvifolielline de *Phylopsora spp.* (Kistenich et al. 2019), les pigmentosines B, D, E et F (Sipman, Elix, and Nash 2009) et l'acide rélicinulinique B de *Relicina relicinula* (Bawingan, Lardizaval, and Elix 2019).

A l'issue de ces étapes de filtration, les données du LIAS comptaient 693 molécules réparties dans 5879 espèces de lichen. Toutes ont été réunies avec les données de la LDB-Lit qui a vu son nombre de molécules légèrement augmenté de 96 entrées (1662 au total) alors que son nombre d'espèces a été pratiquement quadruplé (6576 dont 4766 entrées rajoutées par le LIAS). Suite aux traitements des données sur R avec *Taxize* et *MyTAI*, chaque nom de lichen a été actualisé et associé à une classification. Il en a été de même pour les métabolites avec *ClassyFire*.

3.4 Diversité des produits lichéniques étudiée avec *ClassyFire*.

Pour représenter la diversité chimique des lichens, le nombre de représentants pour chaque classe ClassyFire a été compté pour les données du LIAS, de Hun&Yosh96 et la LDB-Lit. Un réseau a été créé avec ces données, représenté en **Figure 26** sur laquelle l'évolution de la diversité des classes et du nombre de molécules peut être constatée.

Le bloc majoritaire dans chacune des bases de données correspond à celui des « Phénylpropanoïdes et polycétides ». Ce bloc contient les depsides et les depsidones qui forment ensemble l'essentiel des molécules lichéniques comme établi auparavant (**Figure 25**). Ce bloc gagne en molécules avec la LDB-Lit et voit aussi l'apparition de nouvelles classes chimiques absentes du LIAS et de Hun&Yosh96. La **Figure 27-A** présente certaines de ces nouvelles molécules comme la roccanine (macrolactame) isolée de *Roccella canariensis* (Bohman-Lindgren 1972) et le rikuzénol isolé d'une culture du mycobiote de *Graphis rikuzensis* (Takenaka et al. 2003). La rutine a été détectée dans *Roccella montagnei* (Dixit et al. 2016), cependant la présence de flavonoïdes dans les lichens est souvent considérée comme une contamination chimique par des plantes, s'il ne s'agit pas d'une mauvaise interprétation des données dans l'article.

Le second bloc est celui des benzénoïdes, contenant majoritairement les produits de dégradation des depsides et depsidones comme le méthyldivarate et l'éthyldivaricatinate représentés en **Figure 27-B**. C'est également dans ce bloc que se trouvent les diphényléthers comme la buelline de *Diploicia canescens* (Millot et al. 2009) ainsi que différents groupes de quinones comme les perylènequinones dont l'unique représentant ici est l'isohypocrelline de *Phaeographis haematites* (Fazio et al. 2018), les naphthoquinones dont la cristazarine isolée à partir de cultures du mycobiote de *Cladonia cristatella* (Yamamoto et al. 1996), les hydroxyanthraquinones dont la 7-chlorocitréoroséine détectée dans de nombreuses espèces de Teloschistacées (Søchting 2016) au même titre que l'acide 7-chloroémodique (acides anthracéniques). Bien que le bloc de la LDB-Lit ait 5 fois plus de molécules que celui du LIAS, il n'est pas possible d'en supposer un accroissement de la découverte des molécules de type benzénoïde. Hun&Yosh96 contenant déjà deux fois plus de ces molécules que le LIAS. L'une des raisons expliquant la faible quantité de ces molécules dans le LIAS pourraient être que les produits de dégradation des depsides et depsidones soient d'un intérêt limité pour un lichénologue qui utilise des CCMs.

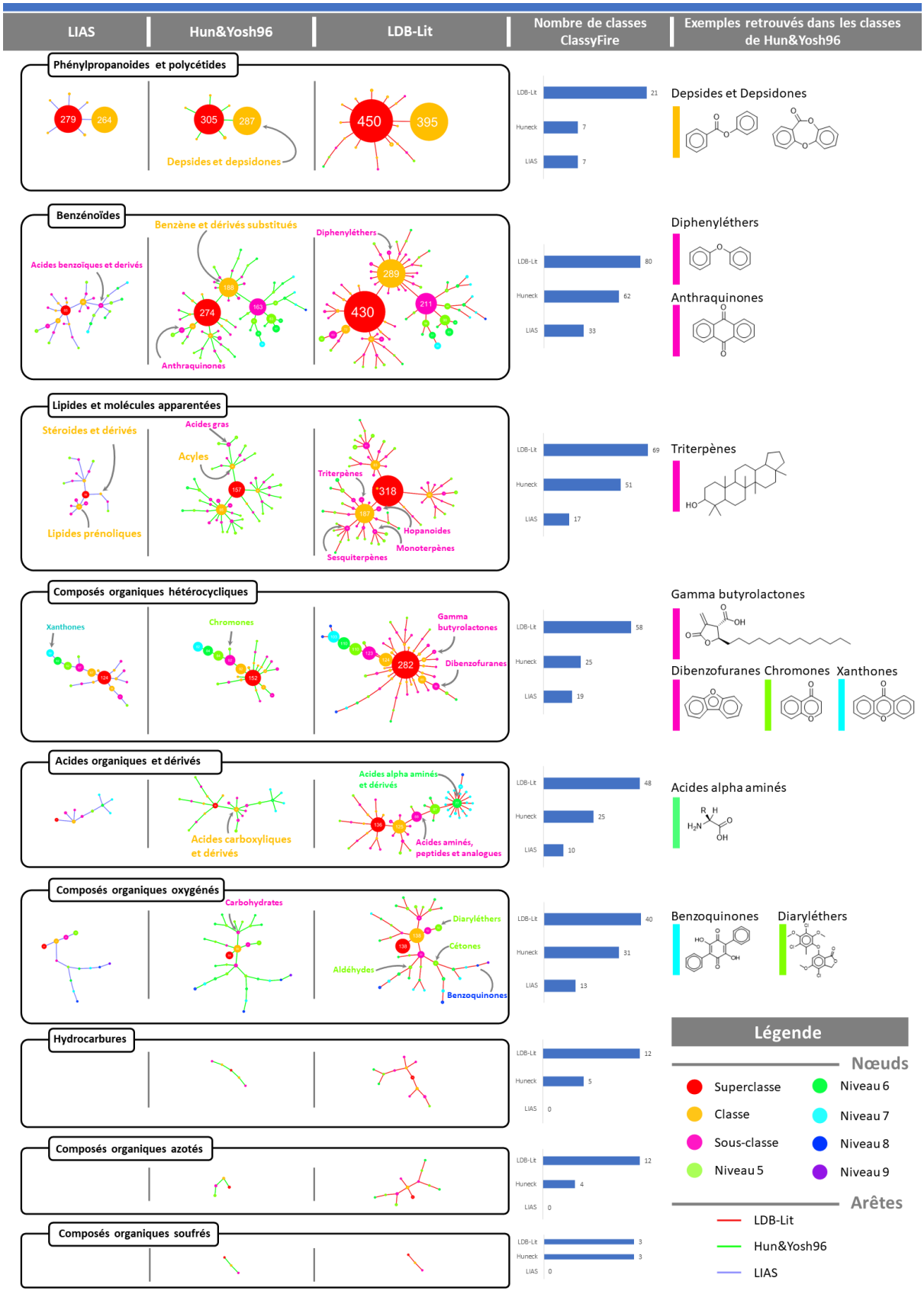


Figure 26 – *Représentation sous forme de réseaux de la diversité chimique des lichens en utilisant des données du LIAS, de Hun&Yosh96 et de la LDB-Lit. Le diamètre des nœuds est proportionnel au nombre de molécules décrites dans chaque classe ClassyFire. Le nombre de classes pour chaque base de données*

est reporté sous la forme d'histogrammes et des exemples de structures sont donnés pour certaines classes du Hun&Yosh96.

Le bloc des « Lipides et molécules apparentées » est particulièrement peu développé dans le LIAS qui ne compte que 56 molécules et 17 classes. Le Hun&Yosh96 en comptait déjà 157 et 51 classes et le nombre de molécules a doublé dans la LDB-Lit. Comme pour le bloc précédent, ceci peut s'expliquer par l'utilisation limitée des lipides de manière générale dans la diagnose des lichens. Parmi les molécules de ce bloc (**Figure 27-C**) se trouvent des dérivés de d'ergostérol comme l'épistérol de *Lobaria pulmonaria*, *L. scrobiculata* et *Usnea longissima* (Safe, Safe, and Maass 1975), des acides apparentés comme l'acide bourgeanique issu de *Ramalina bourgeana* (Stocker-Wörgötter 2008), des triterpènes comme le diacétylpyxinol isolé de *Pyxine endochrysina* (Yosioka, Yamauchi, and Kitagawa 1972), et des monoterpènes comme le (+)-alpha-fenchol détecté dans *Evernia prunastri*. La légitimité de l'identification de nombre de ces molécules dans les lichens est discutée par Joulain et Tabacchi (Joulain and Tabacchi 2009a, 2009b).

Le bloc des « Composés organiques hétérocycliques » est, avec les « Phénylpropanoïdes », celui qui connait le moins de variation entre les données du LIAS et de Hun&Yosh96. Il s'est considérablement développé avec la LDB-Lit qui y répertorie deux fois plus de molécules et de classes. Dans les molécules rajoutées dans la LDB-Lit se trouvent les scabrosines de *Xanthoparmelia scabrosa* (Ernst-Russell et al. 1999), des dibenzoxépines comme la graphisine B isolée de *Graphis tetralocularis* (Pittayakhajonwut et al. 2009), des azaphilones comme la sclérotiorine produite par le mycobionte cultivé de *Pyrenula japonica* (Takenaka et al. 2000), des quinones isoquinoliniques comme la 6-désoxy-8-méthylbostrycoïdine issue de cultures d'une espèce de *Haematomma* (Moriyasu et al. 2005) et des xanthones prénylées comme les umbilicaxanthosides d'*Umbilicaria proboscidea* (Řezanka, Jáchymová, and Dembitsky 2003).

Les blocs restants contiennent relativement peu de métabolites secondaires. Le bloc des « Acides organiques et dérivés » contient essentiellement des acides aminés et celui des « Composés organiques oxygénés », en dehors de quelques quinones et de diaryléthers, contient des glucides. Les « Composés organiques et azotés » ne contiennent que des métabolites primaires et le bloc des « Hydrocarbures » se compose de molécules considérées comme des contaminants observés en chromatographie phase gazeuse.

Cette analyse de la diversité chimique des lichens permet d'établir deux points :

Premièrement, bien que le LIAS soit postérieur à Hun&Yosh96, il présente moins de molécules que ce dernier. Ceci peut s'expliquer par une sélection des molécules présentant un intérêt pour les diagnoses en lichénologie comme en témoigne le nombre de molécules sans structure dans le LIAS. Ces différences sont moindres dans le bloc des phénylpropanoïdes et polycétides et dans celui des composés organiques hétérocycliques qui contiennent l'essentiel des molécules recherchées lors de l'identification des lichens.

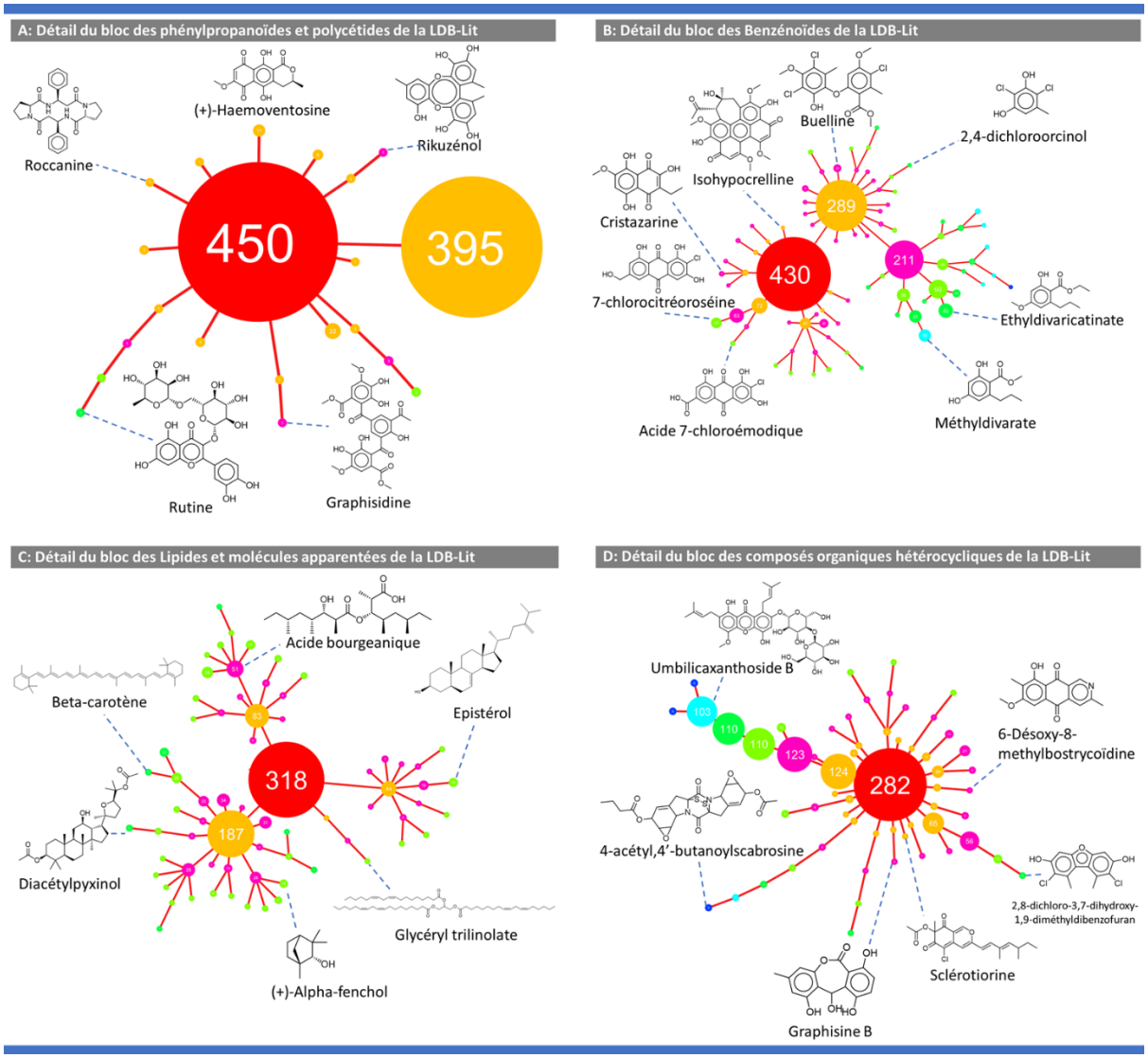


Figure 27 – Détails des quatre principaux blocs de la LDB-Lit pour lesquels des structures représentatives de certains nœuds ont été représentées.

Deuxièmement, le nombre de molécules est presque doublé pour la LDB-Lit par comparaison aux deux autres bases de données. Ceci s'accompagne systématiquement par des rajouts de classes *ClassyFire* dans la LDB-Lit ainsi que par une augmentation du nombre de molécules décrites dans les classes déjà existantes. La plupart de ces nouvelles classes ne contiennent que quelques molécules mais ces dernières présentent souvent des structures atypiques. Par ailleurs un certain nombre de ces molécules ont été obtenues à partir de cultures aposymbiotiques de mycobiontes, connues pour provoquer des changements dans l'expression de certaines voies de biosynthèse (Boustie, Tomasi, and Grube 2011).

3.5 Répartition taxonomique des métabolites lichéniques.

Tous les noms des organismes du LIAS et de la LDB-Lit ont été actualisés et reliés à leurs différents rangs taxonomiques. Le nombre de molécules pour chaque rang a été compté et des réseaux ont été construits à partir de ces données. La **Figure 28** représente les

taxons allant de la classe au genre et la **Figure 29** spécifiquement la classe des Lécanoromycètes. Les espèces n'ont pas été représentées dans les réseaux.

Le premier résultat constaté à partir de la **Figure 28** est que la plupart des molécules décrites sont produites par des Lécanoromycètes. Ceci est peu surprenant, étant donné que la plupart des lichens appartiennent à cette classe : dans les données de la LDB-Lit, 94% des espèces sont des Lécanoromycètes, 3.6% sont des Arthoniomycètes et les 2.4% restants sont des Eurotiomycètes, Dothidéomycètes, Candélariomycètes, Coniocybomycètes, Lichinomycètes et Agaricomycètes. Par ailleurs, c'est dans ces classes qui comptent le plus d'espèces que la plupart des molécules ont été décrites. En rajoutant les données de la LDB-Lit au LIAS, il a été possible de répertorier en moyenne 10% de taxons supplémentaires et le double de molécules.

La LDB-Lit apporte des données pour les Lichinomycètes décrits sans molécules dans le LIAS ainsi que pour les Agaricomycètes. Les Lichinomycètes ne sont composés que d'un ordre (Lichinales) et de trois familles : les Gloeoheppiacées, les Peltulacées et les Lichinacées. Seule la famille des Lichinacées est représentée ici avec les genres *Lichina* et *Synalissa*. Des espèces de ces genres produisent notamment des mycosporines (Roullier et al. 2009; Favre-Bonvin, Arpin, and Brevard 1976; Roullier et al. 2011; de la Coba et al. 2009) et également un arylhydrazide & des arylurées dans le cas de *Lichina pygmaea* : *la* pygméine et les pygmanilines (Roullier et al. 2010). Les Agaricomycètes ne sont pas des Ascomycètes mais des Basidiomycètes, ce qui explique leur absence du LIAS. Ils ne sont par ailleurs représentés ici que par l'espèce *Cora glabrata* (anciennement *Dictyonema glabratum* (Hawksworth 1988)), pour laquelle une seule molécule a été reportée : l'acide fumarique (Huneck and Yoshimura 1996).

Le bloc des Thélocarpales apparaît dans les deux bases de données. Il ne s'agit pas d'une classe mais d'un ordre, la position taxonomique de celui-ci entre l'ordre et la sous-division (Pezizomycotina) étant encore incertaine (*Incertae sedis*). Il ne comporte que trois genres dont *Thelocarpon* qui produit de l'acide pulvinique (Khodosovtsev et al. 2017; W. L. Culberson and Culberson 1970).

Les Coniocybomycètes restent inchangés avec comme seul représentant ici le genre *Chaenotheca* et ses 11 molécules. Ce sont de petites espèces à thalle poudreux, ce qui pourrait expliquer la faible quantité de données à propos de leur chimie.

Il en va de même pour les Candélariomycètes (7 molécules) dont les deux familles sont représentées (Candélariacées et Pycnoracées). Les données sur les genres *Candelariella*, *Candelaria* et *Candelina* permettent d'établir une dominance de molécules dérivées de l'acide pulvinique (calycine, acide rhizocarpique, acide vulpinique). Les espèces encroûtantes ou squamuleuses du genre *Pycnora* restent peu étudiées et seulement de l'acide alectorialique (benzyldepside) y a été détecté (Tsurykau, Khramchankova, and Motiejūnaitė 2012).

Les Dothidéomycètes n'ont qu'une famille de plus dans la LDB-Lit ainsi que 5 genres supplémentaires. Le nombre de molécules reste malgré tout de 18. Les molécules de cette classe sont variées, avec plusieurs xanthones, anthraquinones récurrentes ainsi que quelques naphthofuranes.

A partir des Eurotiomycètes, des différences importantes entre le LIAS et la LDB-Lit deviennent observables. Ce bloc passe de 11 à 22 taxons et le nombre de molécules est triplé en passant de 17 à 54. Cette classe contient de nombreux taxons dont des champignons non-lichénisés. Une famille y a notamment été rajoutée avec les données de la LDB-Lit : les Verrucariacées. Des genres tels que *Verrucaria* ou *Hydropunctaria* ont déjà été étudiés au laboratoire de Rennes notamment pour leurs mycosporines (mycosporines glutaminol et glutamicol). La plupart des molécules de famille des Pyrenulacées ont été isolées du genre *Pyrenula* (30 molécules sur 31). Parmi ces molécules se trouvent plusieurs xanthones dont la draculone (Mathey, Spiteller, and Steglich 2002) ainsi que plusieurs autres molécules issues de la culture du mycobionte de *P. japonica* (Takenaka et al. 2000). Les espèces de la famille des Mycocaliciacées présentent des molécules entièrement différentes, comme l'acide pinastrique et l'acide vulpinique du genre *Mycocalicium*, ce qui a déjà été relevé dans la littérature en comparaison des autres genres de cette famille (André Aptroot et al. 2016).

Les Arthoniomycetes subissent également de grands changements entre le LIAS et la LDB-Lit. Bien que le nombre de familles reste le même (7) et que le nombre de genres ne passe que de 54 à 61, le nombre de molécules décrites se voit plus que doublé (84 à 174). Ceci est dû au genre *Roccella* qui passe de 4 à 85 molécules, expliquant à lui seul les différences observées. Contrairement à la plupart des genres citées auparavant, *Roccella* est un genre de lichens fruticuleux et a attiré à lui de nombreuses études phytochimiques (Follmann 1987; Bohman-Lindgren 1972; T. H. Duong et al. 2017; Huneck et al. 1993; Tehler et al. 2010). La famille des Arthoniacées ne voit que très peu de changements (56 à 62 molécules). Le genre *Herpothallon* possède la majorité de ces métabolites : des naphthofuranes comme l'acide chiodectonique et l'acide rhodocladonique, des acides à longue chaîne comme les acides murolique et constipatique mais également plusieurs depsides et depsidones (acide perlatolique, lécanorique, confluentique, salazinique, stictique etc…).

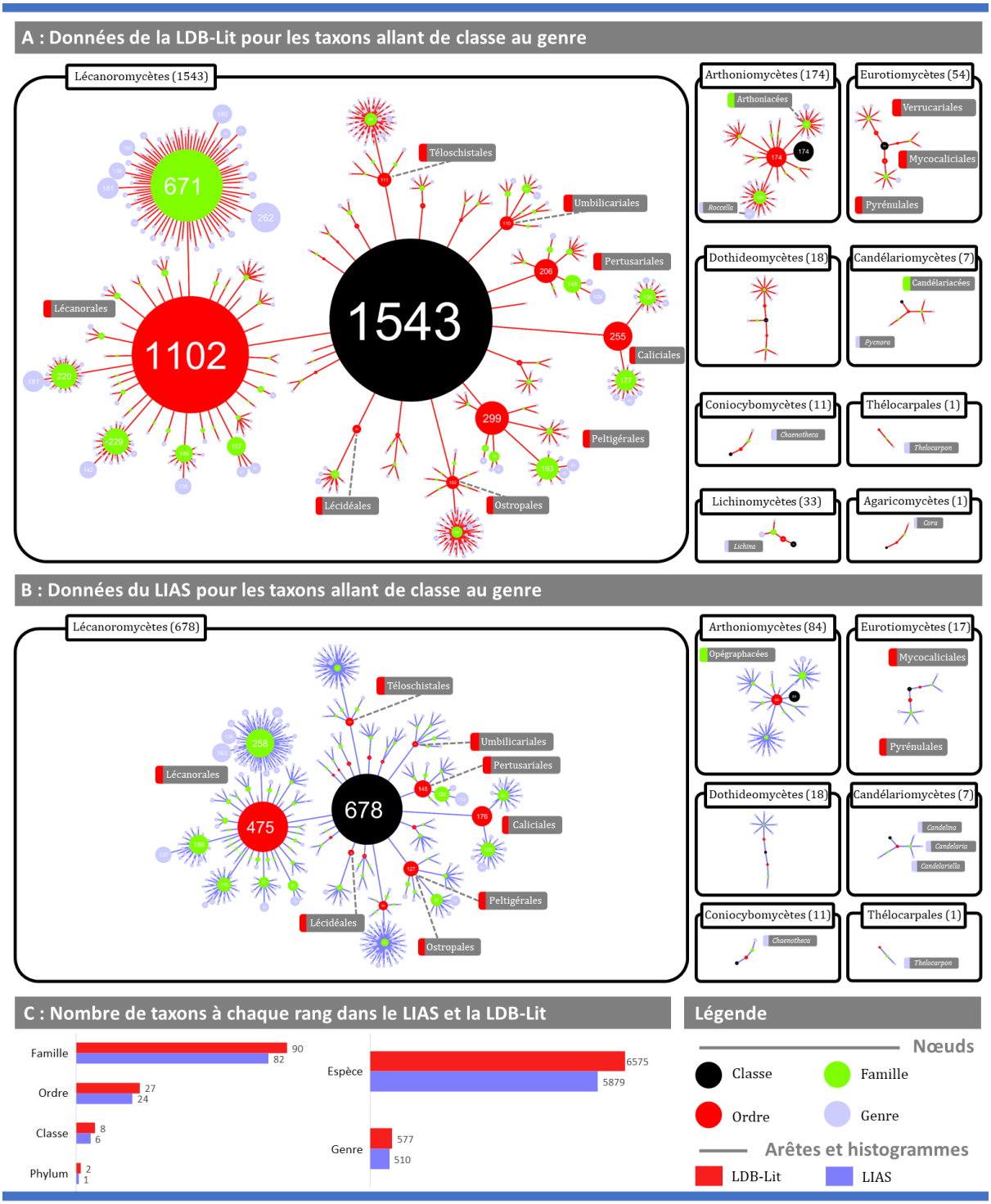


Figure 28 – Réseaux présentant le nombre de molécules à chaque niveau taxonomique pour les données de la LDB-Lit (A) et du LIAS (B). Le nombre total de taxons suivant leurs rangs a été représenté pour les deux bases de données (C).

Les Lécanoromycètes forment la classe comprenant le plus de taxons et le plus de molécules, que ce soit dans le LIAS ou la LDB-Lit. La **Figure 29** présente en détail les différents ordres formant les Lécanoromycètes. Compte tenu de la quantité d'information, les différents taxons ne seront abordés que superficiellement.

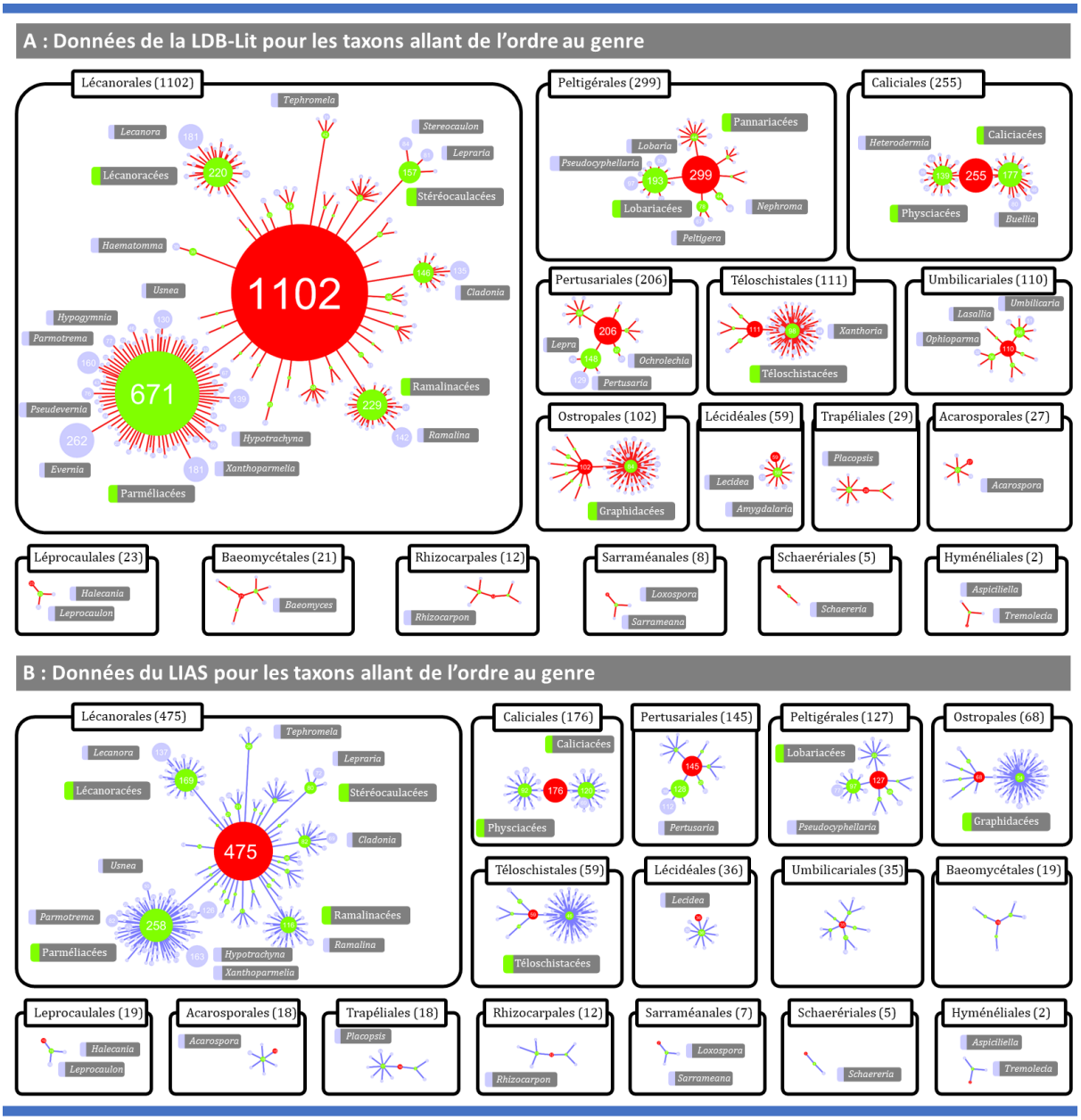


Figure 29 – Réseaux représentant le nombre de métabolites décrits par taxon dans l'ordre des Lécanoromycètes pour les données de la LDB-Lit (A) et du LIAS (B).

Le nombre total d'ordres reste inchangé entre le LIAS et la LDB-Lit. Parmi les plus petits d'entre eux, aucune ou presque aucune molécule ou taxon n'ont été rajoutés. Les Hyménéliales n'ont que très peu de données à leur sujet, au même titre que les Schaériales. Les Sarraméanales sont majoritairement représentées par le genre *Loxospora* qui produit notamment des depsides (Golubkov and Kukwa 2006; Tarasova et al. 2016; Lendemer 2013; Lumbsch, Archer, and Elix 2007). Les Rhizocarpales, essentiellement représentées par le genre *Rhizocarpon* ne subissent pas de changements non plus.

Les Baeomycétales et les Léprocaulales subissent des changements mineurs, avec l'ajout du genre *Phyllobaeis* pour le premier et le rajout de quatre molécules pour le second. Les

Acarosporales passent de 18 à 27 molécules notamment suite aux études sur *Pleopsidium gobiense* (=*Acarospora gobiensis*) (Řezanka and Guschina 1999). Il en va de même pour mes Trapéliales passent de 18 à 29 molécules grâce à des données sur les *Placopsis*.

Les ordres restants ont tous reçu des apports de la LDB-Lit, multipliant en moyenne le nombre de molécules par deux. Cette augmentation du nombre de molécules n'est pas reliée à l'éventuelle étude de nouveaux taxons absents du LIAS, les taxons rajoutés par la LDB-Lit étant peu nombreux (**Figures 30-B, C et D**). Ceci pourrait être dû à des études phytochimiques plus poussées permettant de découvrir de nombreux composés inédits.

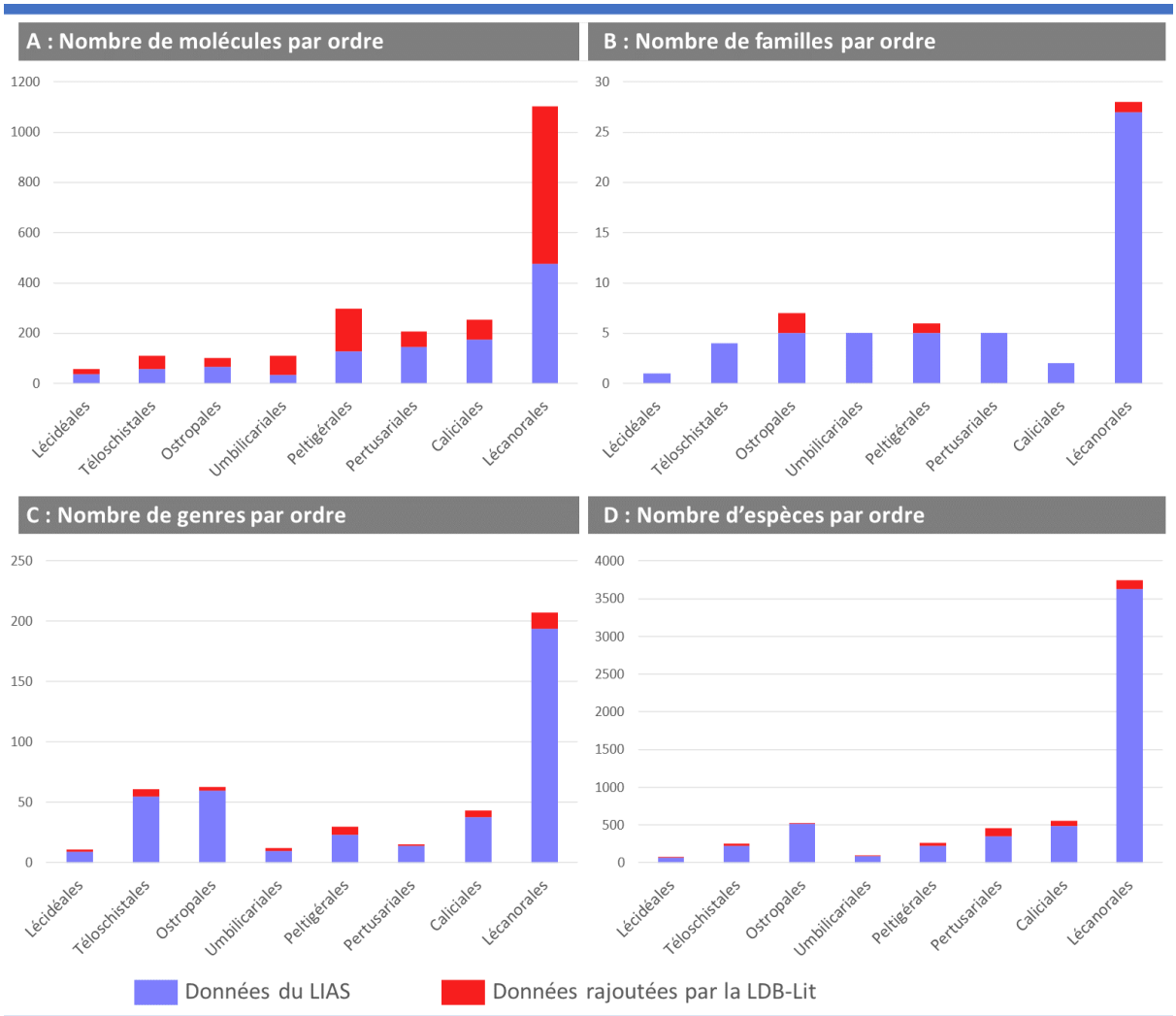


Figure 30 – Détails des principaux ordres pour lesquels des données ont été rajoutées par la LDB-Lit.

Les métabolites supplémentaires proviennent essentiellement d'organismes déjà étudiés : parmi les 618 molécules de la LDB-Lit n'étant répertoriées ni dans Hun&Yosh96 ni dans le LIAS, 575 (93%) proviennent de genres et 461 (75%) d'espèces déjà dans le LIAS.

Les Umbilicariales, les Peltigérales et les Lécanorales par exemple ont de deux à trois fois plus de molécules pour à peine 10% de taxons supplémentaires.

Dans la famille des Umbilicariacées, le genre *Umbilicaria* passe de 9 à 51 molécules. Parmi les molécules rajoutées se trouvent plusieurs glucides (Ranković, Mišić, and Sukdolak 2007; Yoshihiro, Kazuhiko, and Goichi 1973), des carotènes (Czeczuga and Yoshida 1991), des triterpènes déjà présents dans Hun&Yosh96, des didepsides, des tridepsides (Narui et al. 1998; Seriña et al. 1996) et les umbilicaxanthosides (Řezanka, Jáchymová, and Dembitsky 2003). Le genre *Lasallia* ne contenait que de l'acide gyrophorique dans le LIAS et 23 autres molécules du même type que celles décrites pour *Umbilicaria* ont été rajoutées (Briggs et al. 1972; Narui et al. 1996). La famille des Ophioparmacées porte également un genre qui à été davantage étudié : *Ophioparma*. Il ne contenait que de l'acide divaricatique, thamnolique et usnique, et avec l'ajout de la LDB-Lit, des quinones lui ont été associées (Le Pogam, Le Lamer, Siva, et al. 2016).

Pour les Peltigérales, le genre *Pseudocyphellaria* a été davantage étudié, avec la découverte des pseudocyphellarines A et B, l'isopseudocyphellarine A, de depsides comme la chlorogranulatine, de depsidones comme l'acide alpha-acétylconstictique, plusieurs dérivés de l'acide gyrophorique, ainsi que le rajout des mêmes carotènes signalés précédemment, de glucides et de lipides. Le genre *Lobaria* passe de 18 à 80 métabolites avec l'acide téléphorique, des depsidones comme l'acide salazinique et l'acide peristictique, des sesterpénoïdes comme les acides rétigéraniques A et B pour n'en citer que quelques-uns. Dans les Peltigeracées, le genre *Peltigera* passe de 13 à 67 molécules et le genre *Nephroma* passe de 26 à 44. La famille des Collématacées trouve ses premières molécules dans les genres *Leptogium*, *Lathagrium*.

Pour les Lécanorales, le nombre de molécules dans les genres principalement étudiés par les chercheurs a été généralement doublé : le genre *Cladonia* est passé de 69 à 135 molécules, *Stereocaulon* passe de 7 à 84, *Lecanora* de 137 à 181, *Ramalina* de 38 à 142 molécules, *Usnea* de 50 à 130 et *Parmotrema* de 82 à 160.

Sans aller plus loin dans l'interprétation de ces réseaux, la contribution de la LDB-Lit semble déjà évidente.

4 Conclusion

4.1 De la création de la LDB-Lit.

Les rajouts de molécules dans la LDB-Lit à partir de la bibliographie se basent sur un nombre limité de taxons. Bien qu'elle contienne 1662 entrées, ceci suggère que le décompte réel des métabolites lichéniques soit bien supérieur. Ce nombre remet déjà en cause l'idée couramment répandue que les lichens ne soient associés qu'à 1050 composés. Il pourrait être envisagé d'étudier l'intégralité de la bibliographie au sujet des métabolites lichéniques pour une mise à jour exhaustive. Cependant, un travail aussi conséquent expose la LDB-Lit à des erreurs d'entrées lors de l'examen laborieux des articles, mais également à l'ajout de données erronées dues à des interprétations douteuses faites dans des publications. Une stratégie plus efficace, serait d'étudier les lichens avec des méthodes modernes de métabolomique. Ceci permettrait non seulement de vérifier de façon expérimentale ces données de la littérature, mais également de révéler le métabolome des lichens à l'aide des méthodes plus précises. En comparaison du LIAS, où seules sembles conservées les données les plus fiables et utiles aux lichénologues, la LDB-Lit diverge dans son approche. Dans cette dernière, toutes les données sont prises en compte et leur fiabilité est laissée à la discrétion de l'utilisateur. La qualité des entrées sera inévitablement vérifiée par l'expérience et un filtre de fiabilité pourrait être mise en place.

4.2 De la diversité chimique des lichens et sa distribution.

Les nouvelles molécules rajoutées à la LDB-Lit ne proviennent pas, pour la plupart, de l'analyse de nouveaux taxons, mais de genres et d'espèces déjà étudiées. Ceci signifie qu'il est utile d'étudier à nouveau des lichens qui auraient déjà été profilés par le passé. En dehors de ceux-ci, la majorité des espèces de lichens reste encore à étudier : sur les 10201 espèces répertoriées dans le LIAS seulement la moitié est associée à au moins une molécule, et ce par profilage CCM. L'étude d'organismes en dehors des Lécanoromycètes pourrait également être à privilégier, bien qu'ils soient moins nombreux. Dans tous les cas, l'usage de méthodes analytiques avancées couplées à une déréplication rigoureuse serait à privilégier pour mettre en évidence l'étendue du métabolome lichénique, encore largement inexploré.

4.3 De l'étude du métabolome des lichens.

La LDB-Lit constitue un premier pas pour faciliter l'étude des lichens par LC-MS : bien qu'elle ne contienne aucune donnée expérimentale, elle renseigne sur la structure, la masse exacte, les sous-structures et l'origine biologique des métabolites. Elle pourrait être utilisée pour générer des spectres *in silico*, sur CFM-ID (Allen et al. 2014; Djoumbou-feunang et al. 2019) par exemple, et couplée à des outils de pondération des *hits* grâce au

contexte biologique de ces molécules. La LDB-Lit doit encore être revue une dernière fois avant publication et le choix de son hébergement est encore discuté. En plus de ces données bibliographiques, l'ajout de caractéristiques spectrales, comme les spectres de fragmentation, serait désirable pour permettre d'augmenter la fiabilité des déréplications. Ceci sera abordé dans le *Chapitre II* avec la création de la LDB, une base de données spectrale pour les substances lichéniques.

Chapitre II

– LDB –

Ou la création d'une base de données de spectres MS/MS haute résolution pour les métabolites lichéniques

Intervenants extérieurs : Mehdi A. Beniddir[(a)], Solenn Ferron[(b)], Thomas Delhaye[(c)], Pierre-Marie Allard[(d)], Jean-Luc Wolfender[(d)], Harrie J. M. Sipman[(e)], Robert Lücking[(e)], Pierre Le Pogam[(a)].

(a) : CNRS, BioCIS (Biomolécules : Conception Isolement et Synthèse)-UMR 8076, Univ Paris-Sud, Université Paris-Saclay F-92290 Châtenay-Malabry, France
(b) : CNRS, ISCR (Institut des Sciences Chimiques de Rennes)-UMR 6226, Univ Rennes, F-35000 Rennes, France
(c) : CNRS, IETR (Institut d'Électronique et Télécommunications de Rennes)-UMR 6164, Univ Rennes, F-35000 Rennes, France
(d) : EPGL, Université de Genève, Université de Lausanne, CMU, 1 Rue Michel Servet, 1211 Genève 4, Suisse
(e) : Botanischer Garten und Botanisches Museum, Freie Universität Berlin, Königin-Luise-Strasse 6–8, D-14195 Berlin, Allemagne

Contributions externes : PLP, TD – *Participation à la conception du projet*. PLP – *Récupération des standards à Berlin avec J. Boustie*. HJMS, RL – *Conservation et mise à disposition des standards de la chimiothèque Siegfried Huneck*. SF – *Pesée et préparation des échantillons*. MAB, PLP – *Accueil à BioCis et mise en place de la méthode d'analyse*. PMA, JLW – *Conseils pour le traitement MZmine et l'utilisation du FBMN*.

Résumé

Dans ce chapitre, la LDB, ou Lichen DataBase, est créée. A la différence de la LDB-Lit, la LDB est une base de données spectrale (MS^2), permettant de reconnaitre les ions générés par les composés lichéniques avec une fiabilité plus élevée. Les 250 composés utilisés proviennent essentiellement de la chimiothèque Siegfried Huneck conservée au Musée et Jardin Botanique de Berlin et constituent une sélection représentative de la diversité structurale connue pour les lichens. Ils ont été analysés avec un instrument LC-MS (Agilent 6530) et les fichiers ont été traités automatiquement de façon à créer la LDB sous forme de MGF. Une validation technique a ensuite été réalisée en dérépliquant les données issues de l'analyse de trois lichens : *Evernia prunastri* (L.) Ach., *Ophioparma ventosa* (L.) Norman et *Hypogymnia physodes* (L.) Nyl. Après cette validation qui a permis de retrouver la plupart des composés lichéniques attendus, la LDB a été rendue disponible sur le GNPS et MetaboLights.

Sommaire

1 – Introduction 72

2 – Méthodes 74

2.1 Obtention & préparation des standards 74

2.2 Acquisition des données 74

2.3 Constitution de la base de données 75

2.4 Création d'un réseau moléculaire avec les données de la LDB 75

2.5 Validation technique avec trois extraits de lichens 75

3 – Résultats 78

3.1 Constitution de la base de données 78

3.2 Création d'un réseau moléculaire avec les données de la LDB 79

3.3 Déréplications sur le réseau moléculaire des trois lichens 81

3.4 Déréplication de l'extrait d'*Ophioparma ventosa* 84

3.5 Déréplication de l'extrait d'*Evernia prunastri* 85

3.6 Déréplication de l'extrait d'*Hypogymnia physodes* 86

4 – Conclusion 89

4.1 Des constituants de la LDB 89

4.2 Des regroupements par similarité cosinus 89

4.3 De l'utilise de la LDB intégrée au FBMN 89

1 Introduction

Une idée encore largement repandue est que les champignons lichénisés sont associés à une faible diversité chimique en raison de l'origine commune de nombre de leurs métabolites : la voie des polycétides. Pourtant, cela peut être lié à une connaissance superficielle du métabolome spécialisé de ces organismes, basée, comme établi dans le chapitre précédent, sur des composés majoritaires profilés par CCM.

Cette technique manque de sensibilité et repose sur la coélution avec des molécules de référence. L'identification comparative peut être difficile puisqu'elle ne génère pas de données spectroscopiques.

Une grande gamme de techniques analytiques a pourtant été utilisée pour étudier les métabolites lichéniques (Le Pogam et al. 2015; Huovinen, Hiltunen, and Von Schantz 1982; Yoshimura et al. 1994; Le Pogam, Le Lamer, Legouin, et al. 2016; Le Corvec et al. 2016) et ces méthodes seraient à privilégier pour révéler la part d'inconnue dans leur métabolome. Des résultats récents ont par ailleurs démontré que certaines structures atypiques produites par les lichens restent à décrire, comme le montrent les squelettes originaux tsavoénones et sanctis, tous deux obtenus à partir d'espèces de *Parmotrema* collectées au Vietnam (T. Duong, Ha, et al. 2018; T. Duong, Beniddir, et al. 2018).

Pour cibler efficacement ces composés inconnus lors des analyses, les bases de données sont incontournables. La LDB-Lit permettrait éventuellement d'annoter certains ions d'analyses LC-MS mais seulement sur la base des masses mesurées et des origines biologiques. Une annotation solide doit passer par la comparaison de caractéristiques spectrales plus propres à la molécule et dans ce but, une base de données MS/MS spécialement dédiée aux substances lichéniques est créée dans ce chapitre : la LDB ou Lichen DataBase. Cette base de données pourra être employée dans des réseaux moléculaires avec des méthodes similaires à des projets précédents, comme la MIADB (Fox Ramos et al. 2019) (**Figure 31**).

La LDB se base sur des standards de métabolites lichéniques issus de la chimiothèque de Siegfried Huneck conservée à Berlin, ainsi que de la chimiothèque du laboratoire de Rennes. Un total de 250 composés ont été prélevés et analysés par LC-MS en ESI^-, ESI^+ et APCI. Une validation technique a été réalisée par la déréplication d'un réseau moléculaire de trois extraits de lichens, permettant l'annotation de onze molécules uniques qui n'auraient pas été reconnues sans la LDB. Bien qu'ils ne représentent que 15% des 1662 entrées de la LDB-Lit, les composés de la LDB représentent les classes structurales les plus communes au sein des lichens : depsides, depsidones, dibenzofuranes, diphényléthers, dérivés de l'acide pulvinique, quinones, xanthones et terpènes. Ceci permet d'établir un nouveau standard de déréplication dans la chimie des lichens, que ce soit sur le plan de la sensibilité de détection que sur la fiabilité de l'identification. L'usage

de telles bases de données permet également de contourner l'usage et de la conservation de standards purs. La LDB a été rendue disponible sur le GNPS (CCMSLIB00004751209 à CCMSLIB00004751517) (Wang et al. 2016) et MetabloLights (https://www.ebi.ac.uk/metabolights/MTBLS999) (Haug et al. 2013; Olivier-Jimenez, Chollet-Krugler, Rondeau, Beniddir, Ferron, Delhaye, Sipman, et al. 2019).

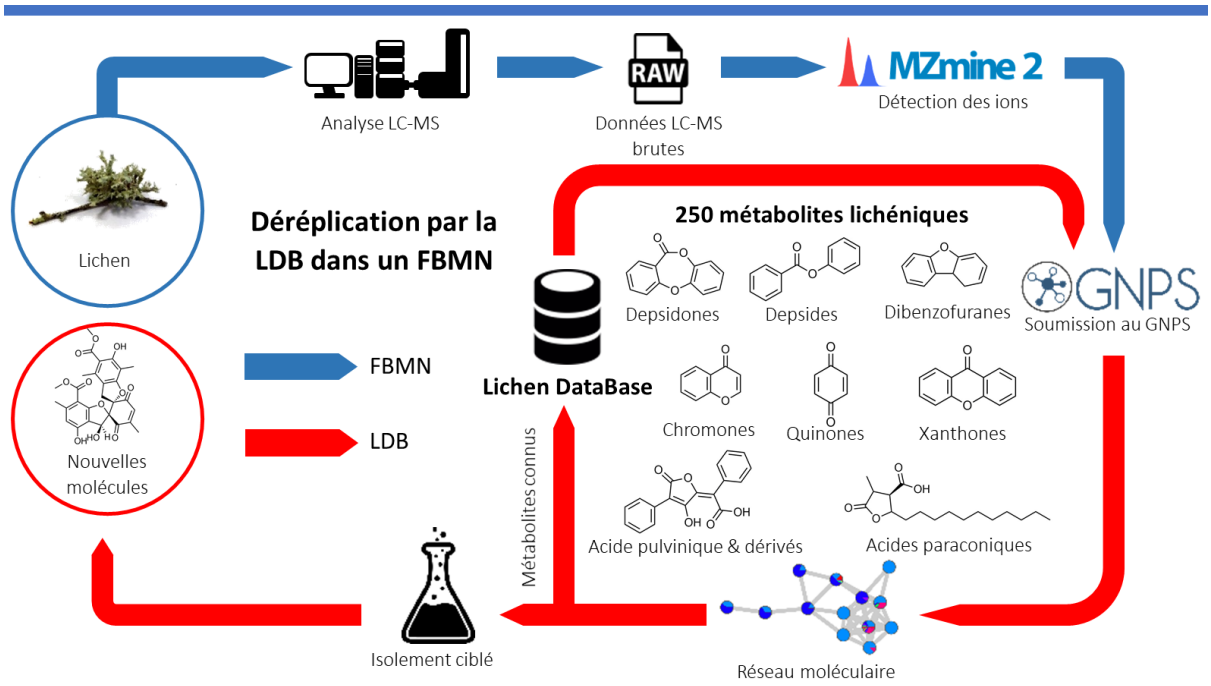


Figure 31 – Déréplication à l'aide de la LDB dans le cadre d'un FBMN.

2 Méthodes

2.1 Obtention & préparation des standards.

Les échantillons ont été obtenus à partir de deux sources : la chimiothèque Huneck du Jardin botanique et Musée botanique de Berlin (B), et la chimiothèque du laboratoire de Rennes. La chimiothèque Huneck contient 1520 substances numérotées et cataloguées par Siegfried Huneck et ses collaborateurs (dont Guido Benno Feige), ainsi que des extraits supplémentaires de ses recherches. Une liste complète est disponible auprès de B, compilée par Heidi Kümmerling, Stefanie Schöne et Harrie Sipman. La chimiothèque de Rennes contient à ce jour 144 substances de lichen cataloguées par le personnel du laboratoire. Un total de 250 échantillons représentatifs ont été prélevés de ces chimiothèques (voir **Annexe**, **Tableau S-1**). Chacun a été solubilisé dans du méthanol de qualité HPLC à 1 mg/ml et placé dans des flacons HPLC de 1.5 ml avant l'analyse. Les solvants ont été achetés auprès de Sigma-Aldrich.

2.2 Acquisition des données.

Les échantillons ont été analysés au moyen d'un spectromètre de masse Agilent 6530 Accurate-Mass Q-TOF couplé à un système Agilent 1260 Infinity LC. La colonne utilisée était une Waters SunFire C18 (50 x 4.6 mm, d.i. 3.5 µm) avec un débit de 0.5 ml/min. Les solvants d'élution utilisés étaient l'eau Milli-Q + 0.1 % AF (A) et l'acétonitrile de qualité HPLC + 0.1 % AF (B) et le gradient d'élution était le suivant : 0 min à 5% B, 7 min à 100% B, 8 min à 100% B, 9 min à 5% B. La plupart des analytes ont été ionisés par électrospray en polarité négative, les xanthones et les quinones ont été ionisées en plus en mode positif, et les terpènes ont été ionisés exclusivement avec une source APCI. Les conditions de l'ESI ont été fixées avec la température du capillaire à 320 °C, la tension de la source à 3.5 kV, et un débit de gaz de 10 L/min. En ce qui concerne les analyses APCI, le courant corona a été réglé à 4 µA, la pression du nébuliseur était de 35 psig et un débit d'azote de 10 L/min chauffé à 350°C a été utilisé pour la désolvatation. Les tensions de capillaire, du fragmenteur et du *skimmer* ont été réglées respectivement à 3500 V, 175 V et 65 V. Il y a eu quatre événements de scan : scan MS négatif ou positif avec une fenêtre de *m/z* 100-1200, suivi de trois scans MS/MS données dépendants du premier, deuxième et troisième ion les plus intenses du premier scan. Les paramètres MS/MS étaient les suivants : trois énergies de collision pour le mode négatif (10, 25, 40 eV, et trois énergies supplémentaires pour les depsides : 2.5, 5 et 7.5 eV), trois pour les modes ESI^+ et APCI (5, 20 et 35 eV), charge par défaut de 1, une largeur d'isolement à *m/z* 2. La purine ($C_5H_4N_4$, *m/z* 121.050873 (positif)), l'acide trifluoroacétique (CF_3CO_2H, *m/z* 112.98559, négatif) et HP-0921 (hexakis(*1 H, 1 H, 3 H*-tétrafluoropropoxy)-phosphazène $C_{18}H_{18}F_{24}N_3O_6P_3$, *m/z* 922.009798 (positif), 1033.988109 (négatif, adduit trifluoroacétate) ont été utilisés comme *lock mass* internes. Les scans complets ont été acquis avec une résolution de 10

000 (*m/z* 922) et 4000 (*m/z* 121) (polarité positive) et 10 000 (*m/z* 1033) et 4 800 (*m/z* 112).

2.3 Constitution de la base de données.

309 fichiers au format *d* d'Agilent ont ainsi été générés et convertis dans un format *mzXML* en utilisant le module MSConvert (Adusumilli and Mallick 2017) de Proteowizard (M. C. Chambers et al. 2012; Kessner et al. 2008). Les données brutes converties ont ensuite été traitées à l'aide d'un script personnalisé sur R 3.6.0 (R Development Core Team 2008) avec le package MSnBase (Gatto and Lilley 2012) pour isoler les spectres MS/MS à chaque énergie de collision pour chaque métabolite et enregistrer chaque spectre dans un fichier *mzXML* individuel. Pour chaque molécule, un spectre MS/MS consensus a été généré à partir des spectres à différentes énergies de collision pour être exporté au format *mzXML*. Les spectres ont été examinés manuellement pour évaluer la qualité de la fragmentation et confirmer l'identité de la molécule. Les spectres consensus ont ensuite été réunis dans un fichier MGF unique à l'aide d'un algorithme privé du GNPS et envoyés sur leur plateforme avec un fichier de métadonnées

2.4 Création d'un réseau moléculaire avec les données de la LDB.

La LDB ont été soumise au GNPS pour créer un réseau à partir de ses spectres et évaluer les regroupements qui pourraient être attendus avec ces métabolites lichéniques. Pour permettre cette évaluation, chaque spectre a été associé à sa classe Hun&Yosh96 dans un fichier d'attributs qui sera chargé avec les nœuds. Les paramètres utilisés pour la création de ce réseau sont présentés dans le **Tableau 3**.

Tableau 3 – Paramètres utilisés pour générer le réseau moléculaire à partir des spectres de la LDB.

Paramètre	*Valeurs*
Minimum pairs cosine	0.6
Parent mass ion tolerance	0.02
Fragment ion mass tolerance	0.02
Minimum matched fragment ions	6
Top K	10
Minimum cluster size	1
Maximum connected component size	0
Run MSCluster	No
Library search score threshold	0.6
Library search minimum matched peaks	6

2.5 Validation technique avec trois extraits de lichens.

La capacité de la LDB à dérépliquer des composés lichéniques a été évaluée en lui soumettant les données LC-MS de trois lichens prélevés dans l'herbier de Rennes. Ces trois lichens sont *Ophioparma ventosa* (JB/14/211), *Evernia prunastri* (JB/13/156) et *Hypogymnia physodes* (JB18/234) (**Tableau 4**).

Tableau 4 – Lichens utilisés pour la validation technique dans l'herbier de Rennes.

Lichen	*Référence*	*Détails*	*Collecteur*
Evernia prunastri (L.) Ach.	JB/13/156	Sur chêne, Jayac, Périgord (France), 170 m (01/2013) N 45.031'486'' E 1.36'156	Joël Boustie, Rennes
Ophioparma ventosa (L.) Norman	JB/14/211	Sur roche siliceuse, lac Großer Winterleitensee, Styria (Autriche), 2000 m N 47.005'15'' E 14.33'45''	Walter Obermayer, Graz
Hypogymnia physodes (L.) Nyl.	JB/18/234	Sur l'écorce de résineux, forêt de Liffré, près de Rennes (France), 110 m N 48.12'32.5'' W1.33'24.3''	Damien Olivier, Rennes

Ces trois lichens ont été extraits avec de l'acétone, puis analysés par LC-MS en polarité négative avec les mêmes paramètres que la LDB. Les fichiers produits ont été convertis au format *mzXML*, puis traités sur MZmine (**Tableau 5**) et via le GNPS où la LDB sera utilisée (**Tableau 6**). Cette déréplication a été effectuée par trois méthodes :

- En n'utilisant que la LDB.
- En utilisant toutes les bases de données spectrales du GNPS, y compris la LDB.
- En utilisant toutes les bases de données spectrales du GNPS, excepté la LDB.

Tableau 5 – Paramètres MZmine utilisés pour traiter les fichiers LC-MS/MS des trois lichens.

Module	*Paramètres & valeurs*
Mass detection	MS1: 1.0E3 MS2: 5.0E1
ADAP Chromatogram builder (Myers et al. 2017)	Min. group size: 1 Group intensity threshold: 1.0E3 Min highest intensity: 1.0E3 *m/z* tolerance: 20 ppm
Chromatogram deconvolution (ADAP Wavelets) (Myers et al. 2017)	*m/z* range for MS2 scan pairing (DA): 0.3 RT range for MS2 scan pairing (min): 1 S/N threshold: 2 S/N estimator: Intensity window SN min feature height: 1000 coefficient/area threshold: 5 Peak duration range: 0.01-7.00 RT wavelet range: 0.01-0.50
Isotopic peaks grouper	*m/z* tolerance: 20 ppm Retention time tolerance: 0.2 min Monotonic shape: unchecked Maximum charge: 2 Representative isotope: Most intense
Join aligner	*m/z* tolerance: 20 ppm Weight for *m/z*: 1 Retention time tolerance: 0.7 Weight for RT: 1 Require same charge state: unchecked Require same ID: unchecked
Same RT and m/z range gap-filler	*m/z* tolerance: 20 ppm
Peak list rows filter	Only checked values: Keep only peaks with MS2 scan (GNPS) Reset the peak number ID

Tableau 6 – Paramètres utilisés pour générer le réseau moléculaire à partir des trois lichens.

Paramètre	*Valeurs*
Minimum pairs cosine	0.6
Parent mass ion tolerance	0.02
Fragment ion mass tolerance	0.02
Minimum matched fragment ions	6
Top K	10
Minimum cluster size	1
Maximum connected component size	0
Run MSCluster	No
Library search score threshold	0.5
Library search minimum matched peaks	4

3 Résultats

3.1 Constitution de la base de données.

La LDB contient 309 spectres MS/MS consensus de 250 métabolites lichéniques ionisés par ESI en mode négatif (226 spectres), positif (68 spectres) et en APCI (15 spectres). De plus, les 1011 spectres MS/MS à énergies de collision individuelles sont disponibles sur MetaboLights. Les composés analysés couvrent une grande majorité des structures décrites dans Hun&Yosh96 (**Figure 32-A et B**) et devraient fournir des informations précieuses pour le profilage chimique des lichens. La présence de fonctions acides dans une grande majorité des métabolites de lichen a incité à les analyser en mode négatif, conformément aux études précédentes (Le Pogam et al. 2015).

Si les principaux paramètres d'analyse (les énergies de collision 10/25/40 eV et l'ionisation ESI^-) ont permis d'obtenir des spectres de masse en tandem utilisables pour la plupart des structures étudiées ici, trois cas spécifiques ont nécessité l'application de paramètres d'acquisition alternatifs :

Les depsides : de par leur structure, ces composés ont subi une fragmentation trop importante dans les conditions susmentionnées. Pour y remédier, des analyses supplémentaires utilisant des énergies de collision plus faibles ont été effectuées sur cette classe spécifique (2.5/5/7.5 eV). Les résultats étaient globalement meilleurs et des spectres consensus ont été générés pour ces molécules en prenant en compte les six énergies de collision.

Les quinones, xanthones et chromones : l'ionisation par électrospray facilite principalement la formation de molécules déprotonées et rarement celle d'ions radicalaires (Mann 1990). Par conséquent, dans des environnements moléculaires spécifiques où les groupes phénoliques peuvent initier des liaisons hydrogène intramoléculaires, leur déprotonation est rendue impossible et la molécule ne se prête pas à la détection avec ESI^- sans procédures analytiques dédiées (Rafaëlly et al. 2008). Cette configuration est surtout rencontrée dans les métabolites contenant des γ-pyrones, c'est-à-dire les quinones, les xanthones et les chromones dans le cas des lichens. Leur analyse en ESI^+ a fourni des spectres MS^2 satisfaisants.

Terpènes et stéroïdes : ces composés ont été analysés à l'aide d'une source APCI qui donne de meilleurs résultats que l'ESI pour les composés de faible à moyenne polarité (Holcapek, Volna, and Vanerkova 2007).

La base de données est disponible sur le GNPS pour être utilisée dans les réseaux moléculaires en mode positif (https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=LDB_POSITIVE) et négatif (https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=LDB_NEGATIVE). Les spectres correspondant aux énergies de collision individuelles sont consultables sur

MetaboLights sous l'identifiant MTBLS999 (Olivier-Jimenez, Chollet-Krugler, Rondeau, Beniddir, Ferron, Delhaye, Sipman, et al. 2019) (https://www.ebi.ac.uk/metabolights/MTBLS999).

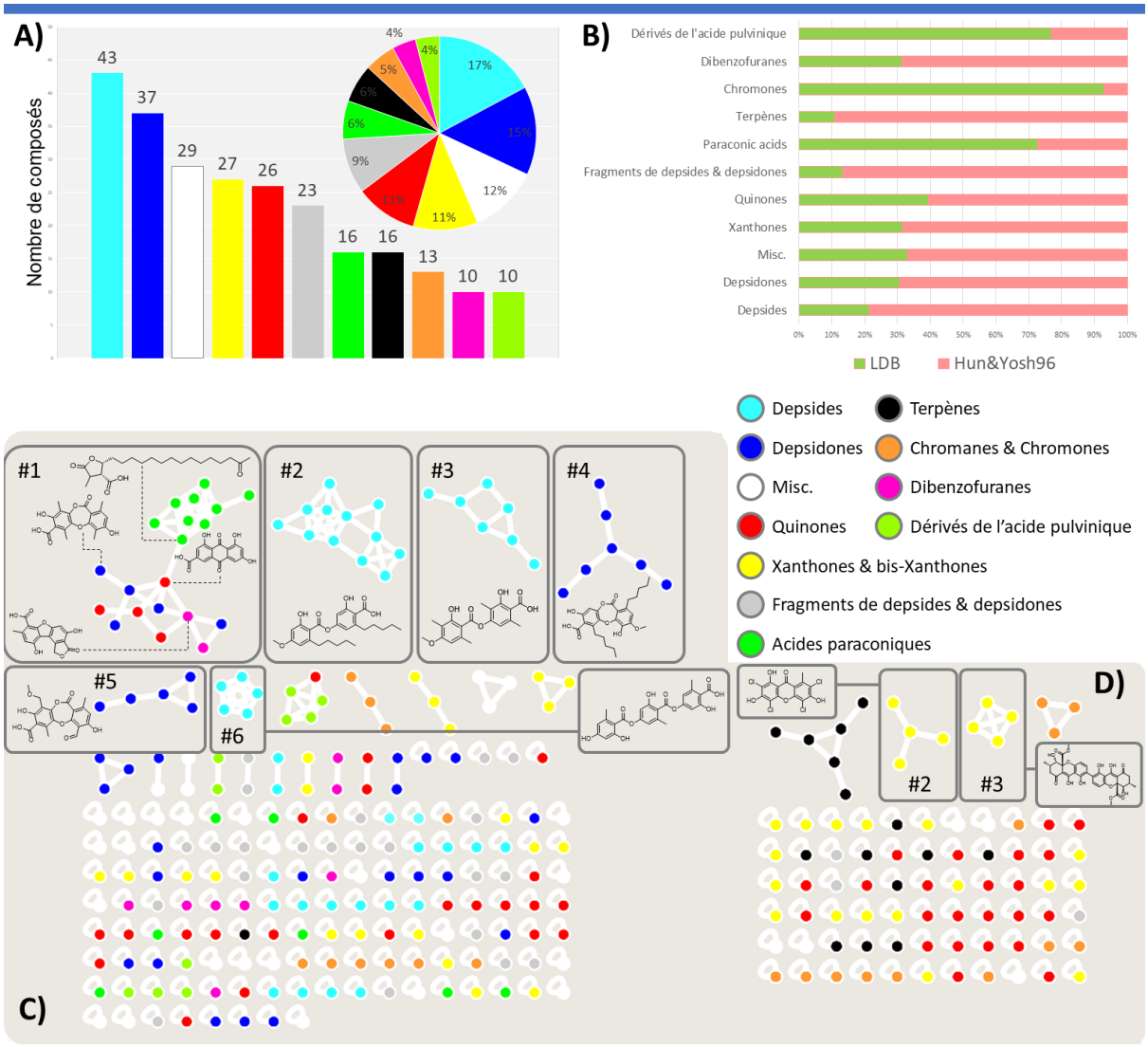


Figure 32 – Métriques et caractéristiques de la LDB. Les métabolites sont classés selon Huneck et Yoshimura(Huneck and Yoshimura 1996) *et la classe Misc. représente les acides, les composés aliphatiques et cycloaliphatiques, les diphényléthers, les mycosporines, les naphtopyranes, les composés contenant de l'azote, les polyols, les monosaccharides et les glucides. (a) : Nombre de représentants pour chaque classe de métabolites et leurs proportions relatives. (b) : Proportion de métabolites couverts pour chaque classe. (c) et (d) : réseaux moléculaires utilisant respectivement les spectres en mode négatif et positif des métabolites de lichen du LDB comme entrée avec un seuil de similarité du cosinus de 0.6. Les blocs principaux sont numérotés et une structure représentative de chacun est affichée.*

3.2 Création d'un réseau moléculaire avec les données de la LDB.

Les **Figures 32-C** et **32-D** montrent deux réseaux moléculaires générés à partir des spectres de la LDB et chaque nœud (ou ion) a été coloré en fonction de sa classe structurale selon Hun&Yosh96 (Huneck and Yoshimura 1996). Le réseau moléculaire généré en mode négatif est plus important que celui en mode positif en raison du nombre

plus élevé de composés de la LDB analysés dans ce mode. Dans l'ensemble, les réseaux obtenus ont eu tendance à regrouper les ions en fonction des classes structurelles. Cependant, chaque classe n'a pas produit un bloc unique et homogène et, à l'inverse, des métabolites apparemment sans rapport se sont parfois regroupés en raison de la similarité de leurs groupes fonctionnels, étant ainsi susceptibles de subir des pertes neutres similaires. Les caractéristiques structurelles qui expliquent la topologie des réseaux moléculaires ne méritent généralement pas d'être étudiées de manière trop approfondie car elle dépend exclusivement de la définition du seuil cosinus et des pics des spectres MS/MS, indépendamment des classifications structurales. Néanmoins, certains regroupements peuvent être rationalisés de manière simple. La **Figure 32** présente des exemples de produits chimiques représentatifs des blocs discutés. À ce titre, un exemple intéressant est celui des depsides (**Figure 32-C**, en cyan) qui ont été divisés en trois grands blocs. Le bloc n°2 contient des depsides avec des chaînes latérales alkyles de trois à sept membres, tandis que le bloc n°3 ne rassemble que des depsides dépourvus de ces longues chaînes latérales. Enfin, le bloc n°6 comprend des tridepsides et un seul didepside - l'acide lécanorique - ce qui s'explique aisément par le fait que son noyau diaromatique se retrouve dans tous les tridepsides représentés dans la LDB. Une discrimination sur la base de la longueur des chaînes latérales peut être constatée pour les depsidones (respectivement les blocs n°4 et n°5). Un dernier exemple de regroupement dépendant de la structure est celui des xanthones qui sont regroupées différemment selon leur statut monomérique ou dimérique (**Figure 32-D**). La partie inférieure du bloc n°1 est un exemple de composés structuralement hétérogènes, qui présentent des processus de fragmentation suffisamment proches pour appartenir à un même bloc (**Figure 32-C**). Ces structures « erratiques » sont représentées par des quinones (**Figure 32-C** et **32-D**, en rouge) qui produisent peu de fragments en MS/MS et se trouvent principalement sous la forme de nœuds *self-looped* ou liés à des composés ayant des ornementations similaires. Chaque réseau moléculaire présenté est disponible sur la plateforme du GNPS : https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=cc4925fa2ccd43b790b708576f47e7b5 pour le mode négatif (**Figure 32-C**) et https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c79d748b515b4357a515bcec1435e6a1 pour le mode positif (**Figure 32-D**). Chaque bloc peut y être retrouvé avec la même numérotation que sur la figure, ainsi que chaque nœud avec ses spectres et ses identifications.

3.3 Déréplications sur le réseau moléculaire des trois lichens.

Les trois lichens (*Ophioparma ventosa*, *Evernia prunastri* et *Hypogymnia physodes*) ont été extraits, analysés et traités pour produire un réseau moléculaire (**Figure 33**). Ce réseau peut être consulté sur la plateforme GNPS à l'adresse https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ee1285c8de3a45cda13d719271570dc7. L'étendue des avantages offerts par la LDB a été évaluée en comparant les résultats obtenus par les trois méthodes de déréplication et en colorant les bords des nœuds en conséquence :

- Vert : nœuds dérépliqués exclusivement par la LDB.
- Jaune : nœuds dérépliqués par la LDB et une autre base du GNPS.
- Rouge : nœuds non dérépliqués.

Aucun nœud n'a été dérépliqué exclusivement par une base de données autre que la LDB. Au total, 15 molécules uniques ont ainsi été dérépliquées, 11 d'entre elles étant exclusivement identifiées avec la LDB, et quatre étant partagées avec d'autres bibliothèques du GNPS (**Tableau 7**). Comme les annotations générées par *gap-filling* sur MZmine (Pluskal et al. 2010) (mêmes *m/z* et TR) sont associées à un degré de confiance moindre, elles sont classées à 5 selon les niveaux de confiance largement acceptés en métabolomique définis par Schymanski et al (Schymanski et al. 2014) (**Figure 34**). Une annotation provisoire de certains nœuds non-annotés a été effectuée à l'aide du logiciel SIRIUS (Böcker et al. 2009).

Dans le contexte de la lichénologie, les réseaux moléculaires sont particulièrement utiles pour mettre en évidence des ensembles caractéristiques de produits, également appelés chémosyndromes. Ces métabolites sont structurellement similaires en raison de leurs interconnexions biosynthétiques et on peut donc s'attendre à ce qu'ils se regroupent pour former des blocs dans un réseau moléculaire. Ces chémosyndromes sont souvent composés d'un métabolite majeur accompagné de plusieurs composés satellites mineurs (C. F. Culberson and Culberson 1976).

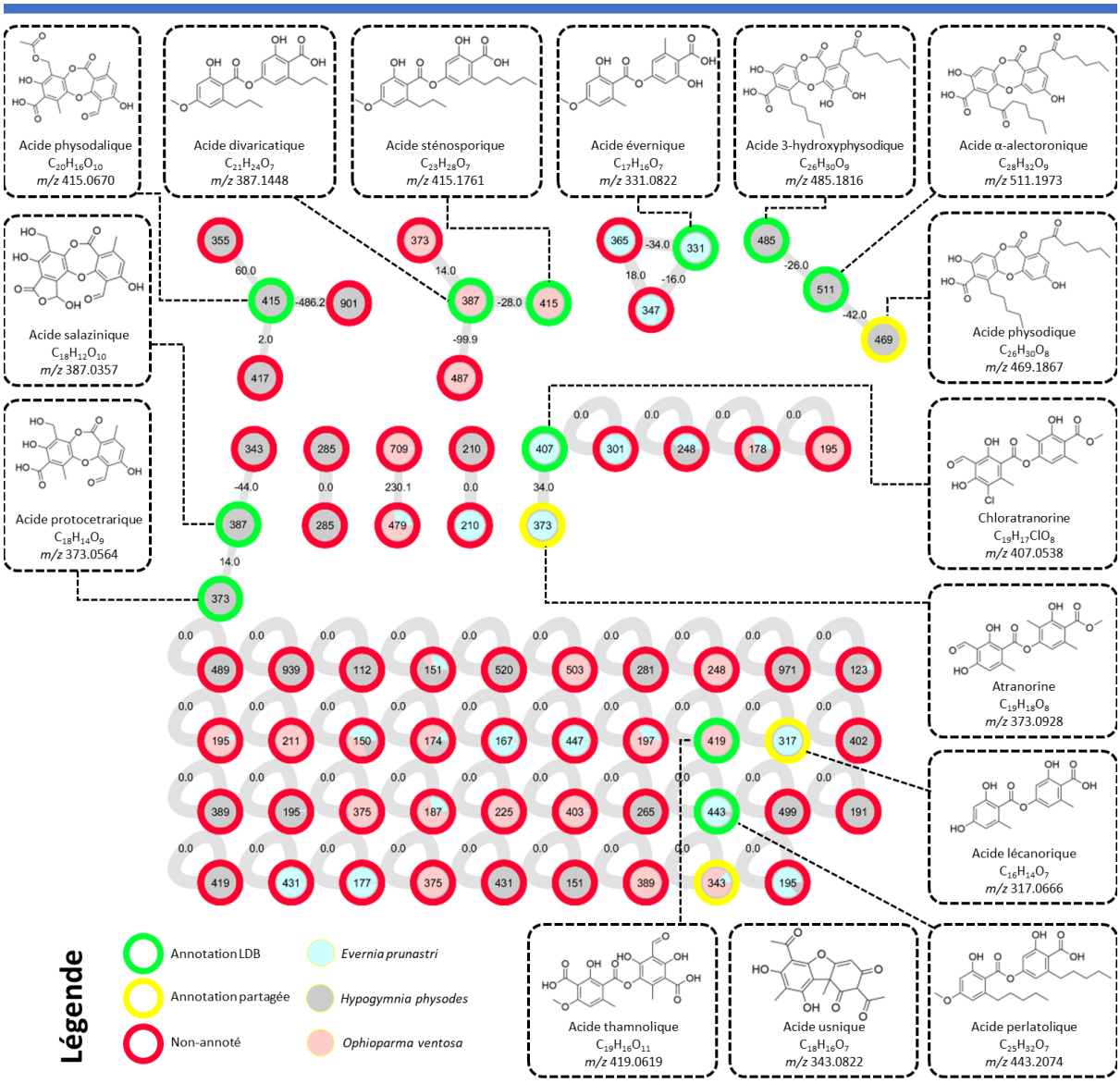


Figure 33 – Réseau moléculaire généré à partir des extraits acétone d'Ophioparma ventosa, Evernia prunastri et Hypogymnia physodes par FBMN. Les nœuds avec des rebords verts représentent les ions dérépliqués uniquement avec la LDB, ceux avec des rebords jaunes des ions dérépliqués par la LDB et d'autres bases de données du GNPS. Les cercles à bords rouges représentent les nœuds qui n'ont pas pu être dérépliqués automatiquement. Les diagrammes circulaires à l'intérieur des nœuds représentent la proportion dans laquelle les ions ont été observés dans chacun des trois lichens en utilisant l'aire de pic. Les ions identifiés sont représentés avec leur structure, leurs formules chimiques et leurs rapports m/z théoriques. Les arêtes entre les nœuds portent la différence de masse entre les deux nœuds connectés.

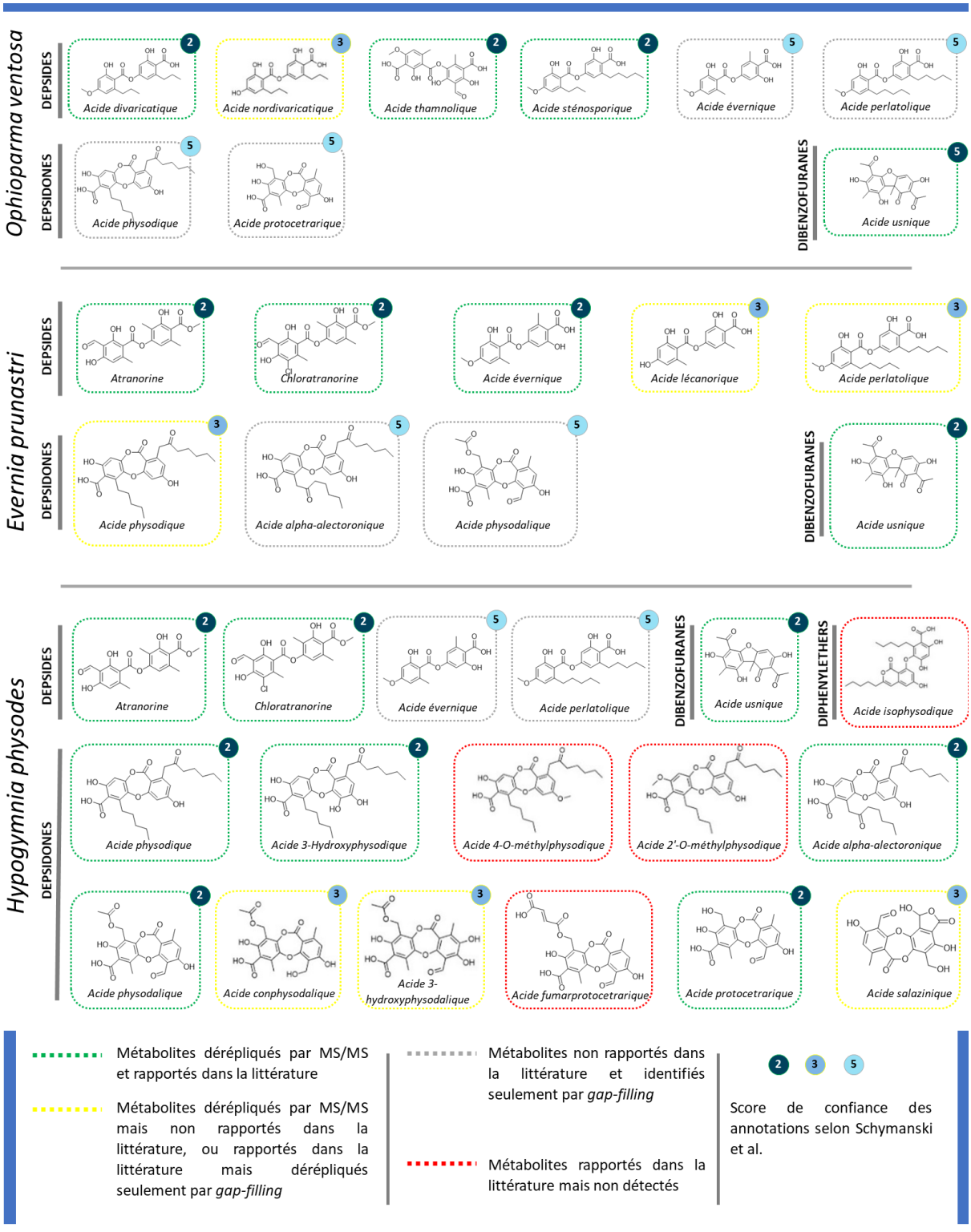


Figure 34 – Résultat de la déréplication des extraits d'Ophioparma ventosa, Evernia prunastri et Hypogymnia physodes. Les métabolites dérépliqués automatiquement par similarité cosinus et signalés dans la littérature pour le lichen étudié sont entourés d'un rectangle vert. Ceux dérépliqués mais non signalés dans la littérature, soit signalés dans la littérature mais uniquement identifiés par gap-filling sont entourés de rectangles jaunes. Les métabolites qui ne sont pas signalés dans la littérature et qui ne sont identifiés que par gap-filling sont entourés par des rectangles gris. Les métabolites signalés dans la littérature mais non identifiés dans l'analyse sont entourés de rectangles rouges. Toutes les annotations

sont accompagnées d'un score de confiance selon Schymanski et al. (Schymanski et al. 2014), de 5 (même masse exacte) à 1 (même spectre MS/MS et RT que la molécule de référence).

3.4 Déréplication de l'extrait d'*Ophioparma ventosa*.

Tableau 7 – Métabolites annotés sur le réseau moléculaire. Les références bibliographiques sont les suivantes : ***(1)*** – (Białońska and Dayan 2005), ***(2)*** – (Latkowska et al. 2015), ***(3)*** – (Ranković et al. 2014), ***(4)*** – (Solhaug et al. 2009), ***(5)*** – (Cansaran-Duman et al. 2010), ***(6)*** – (Avalos and Vicente 1987), ***(7)*** – (C. F. Culberson 1963), ***(8)*** – (Díaz-Guerra and Manrique 1984), ***(9)*** – (Herrero-Yudego et al. 1989), ***(10)*** – (Kosanić et al. 2013), ***(11)*** – (Vicente and Pérez-Urria 1988), ***(12)*** – (Legaz et al. 1986), ***(13)*** – (Bjerke, Lerfall, and Elvebakk 2002), ***(14)*** – (Le Pogam, Le Lamer, Legouin, et al. 2016), ***(15)*** – (Skult 1997).

Source	m/z	TR (min)	Annotation	CScore	Erreur (ppm)	Pics partagés	Ref.
Hypogymnia physodes	485.1810	5.95	Acide 3-Hydroxyphysodique	0.65	10	13	Oui[1,2]
	511.1925	7.33	Acide alpha-alectoronique	0.82	4	19	Oui[2]
	373.0924	8.80	Atranorine	0.85	4	7	Oui[1-4]
	407.0558	8.89	Chloroatranorine	0.87	1	7	Oui[2-4]
	331.0828	8.71	Acide évernique	0.56	2	5	Non
	443.2071	8.46	Acide perlatolique	0.62	17	5	Non
	415.0709	6.59	Acide physodalique	0.93	34	13	Oui[1-4]
	469.1878	8.16	Acide physodique	0.73	15	8	Oui[1,2,4]
	373.0551	4.93	Acide protocetrarique	0.85	18	10	Oui[2,4]
	387.0366	4.29	Acide salazinique	0.88	22	12	Non
	343.0823	8.71	Acide usnique	0.80	18	4	Oui[3,5]
Evernia prunastri	511.1925	7.33	Acide alpha-alectoronique	0.82	4	19	Non
	373.0924	8.80	Atranorine	0.85	4	7	Oui[6-11]
	407.0558	8.89	Chloroatranorine	0.87	1	7	Oui[6-12]
	331.0828	8.71	Acide évernique	0.56	2	5	Oui[6-11]
	317.0660	7.42	Acide lécanorique	0.75	14	4	Non
	443.2071	8.46	Acide perlatolique	0.62	17	5	Non
	415.0709	6.59	Acide physodalique	0.93	34	13	Non
	469.1878	8.16	Acide physodique	0.73	15	8	Oui[10]
	343.0823	8.71	Acide usnique	0.80	18	4	Oui[6-11]
Ophioparma ventosa	387.1474	9.07	Acide divaricatique	0.84	18	12	Oui[13-15]
	331.0828	8.71	Acide évernique	0.56	2	5	Non
	443.2071	8.46	Acide perlatolique	0.62	17	5	Non
	469.1878	8.16	Acide physodique	0.73	15	8	Non
	373.0551	4.93	Acide protocetrarique	0.85	18	10	Non
	415.1786	9.63	Acide sténosporique	0.87	19	7	Oui[15]
	419.0460	10.4	Acide thamnolique	0.86	42	8	Oui[13-15]
	343.0823	8.71	Acide usnique	0.80	18	4	Oui[13-15]

La chimie du lichen crustacé *Ophioparma ventosa* a été étudiée par plusieurs groupes de recherche au cours des dernières décennies et son métabolome spécialisé peut être considéré comme plutôt complexe. Ceci est lié à la présence de certains métabolites constants (acide thamnolique, acide décarboxythamnolique, acide usnique, acide divaricatique, hémoventosine et une multitude d'autres pigments pyranonaphtoquinones) (Holzmann and Leuckert 1990; Le Pogam, Le Lamer, Siva, et al. 2016) ainsi que de certains autres composés, comme l'acide sténosporique, qui peuvent être présents ou non et dont on suppose - au moins en partie - qu'ils proviennent d'espèces de lichens limitrophes (Skult 1997; Le Pogam, Le Lamer, Siva, et al. 2016; May 1997). Comme l'échantillon d'*Ophioparma* utilisé était plutôt pauvre en apothécies, l'hémoventosine et les pyronaphthoquinones qui lui sont associées n'ont pas été détectées. Certaines études récentes traitant de cette espèce de lichen, y compris des

études phytochimiques approfondies menées par notre groupe (Le Pogam, Le Lamer, Siva, et al. 2016; Le Pogam, Le Lamer, Legouin, et al. 2016), permettent d'avoir des données assez fiables sur les métabolites qui y sont présents.

La déréplication automatique à l'aide de la LDB a permis l'annotation de quatre des cinq métabolites attendus en mode négatif : les acides thamnolique, usnique, stenosporique et divaricatique (**Tableau 7**). Quatre autres composés ont été détectés par *gap-filling* : les acides évernique, perlatolique, physodique et protocétrarique (**Figure 34**). Aucun de ceux-ci n'avait été signalé auparavant dans la littérature dans le genre *Ophioparma*.

Ces métabolites identifiés sont regroupés avec certains métabolites non identifiés. Les acides divaricatique et sténosporique sont regroupés avec deux nœuds à *m/z* 373.1315 et 487.0743. Alors que le second ne peut être considéré que comme un depside inconnu, l'acide nordivaricatique (acide 4-*O*-déméthyldivaricatique, **Figure 35-A**), dont la présence a été précédemment signalée en compagnie des acides divaricatique et sténosporique dans plusieurs lichens, dont *O. ventosa*, pourrait être un candidat probable pour expliquer le premier (Skult 1997). Un autre bloc notable pour ce lichen contient l'ion à *m/z* 479.0596, apparemment un métabolite apparenté à l'acide usnique (dibenzofurane) (**Figure 35-B**). La présence conjointe de l'acide divaricatique et de l'acide sténosporique est une illustration d'un chémosyndrome, les deux depsides portant des chaînes latérales de longueur modérée, le dernier étant un composé satellite mineur du premier (Skult 1997).

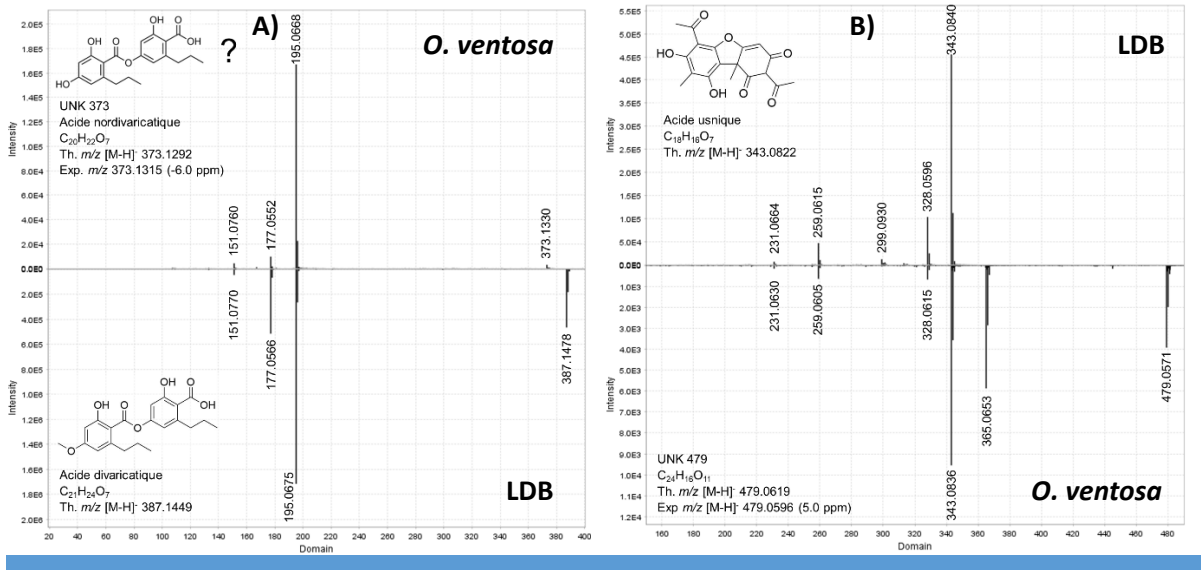


Figure 35 – Représentation en miroir de différents spectres MS/MS d'Ophioparma ventosa. (A) spectre manuellement associé à l'acide nordivaricatique représenté contre l'acide divaricatique de la LDB. (B) représentation du spectre d'O. ventosa à m/z 479 contre l'acide usnique de la LDB.

3.5 Déréplication de l'extrait d'*Evernia prunastri*.

Le lichen fruticuleux *Evernia prunastri* a fait l'objet d'études approfondies concernant son contenu chimique car il est largement utilisé dans l'industrie des parfums. Également appelé mousse de chêne, son odeur typique est liée à l'hydrolyse des depsides inodores

pour produire une série de composés monoaromatiques très odorants (Joulain and Tabacchi 2009a). Les composés attendus, tels que décrits historiquement par C. Culberson (C. F. Culberson 1963), sont l'atranorine, la chloratranorine, l'acide evernique et usnique. En outre, Joulain et Tabacchi ont publié une revue des métabolites signalés pour *Evernia prunastri* qui a atteint plus de 170 structures (Joulain and Tabacchi 2009a). Comme l'ont mentionné les auteurs, certains composés doivent être considérés avec la plus grande prudence car les sources sont parfois un mélange de plantes et de lichens en plus des polluants environnementaux, ce qui a entrainé la détection de plusieurs produits pétroliers par GC.

La déréplication automatique a permis l'identification simple des quatre composés décrits classiquement (atranorine, chloratranorine, acides evernique et usnique) en plus de l'acide lécanorique et de l'acide perlatolique, ces deux derniers métabolites étant nouvellement signalés dans ce lichen déjà très étudié (**Tableau 7**). L'acide physodique, précédemment signalé dans *Evernia prunastri*, a été détecté par *gap-filling* (Kosanić et al. 2013). Les autres métabolites détectés par *gap-filling* étaient l'acide alpha-alectoronique et l'acide physodalique (**Figure 34**). Les nœuds non-annotés, regroupés avec l'acide évernique, comprennent deux ions à *m/z* 365.0443 et 347.0764, qui semblent partager un squelette depside (**Figures 36-A et B**). *Evernia prunastri* semble contenir le même dérivé d'acide usnique qu'*Ophioparma ventosa* à *m/z* 479.0596.

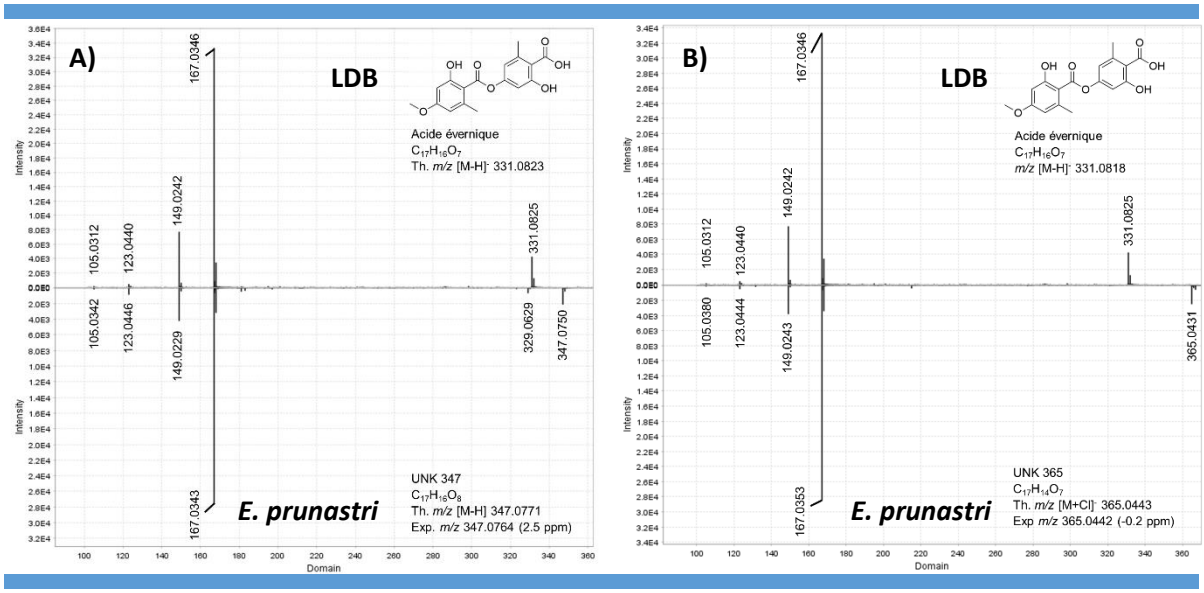


Figure 36 – Représentation en miroir de spectres MS/MS d'Evernia prunastri vis-à-vis de composés proches dans la LDB. (A) représentation du spectre à m/z 347 comparé à l'acide évernique de la LDB, suggérant la détection d'un dérivé avec une fonction hydroxyle additionnelle. (B) représentation du spectre à m/z 365 indiquant à encore une molécule proche de l'acide évernique.

3.6 Déréplication de l'extrait d'*Hypogymnia physodes*.

Cette espèce foliacée est connue pour produire un ensemble de métabolites de structure diverse comprenant divers depsides et des depsidones qui diffèrent par la longueur et l'hydroxylation de leurs chaînes latérales (Molnár and Farkas 2011). Il convient de noter que des études phytochimiques récentes basées sur la LC-MS/MS ont été réalisées sur *H.*

physodes (Latkowska et al. 2015), on peut donc en déduire que la chimie de ce lichen est assez largement couverte pour évaluer le degré d'information véhiculé par la déréplication via la LDB. Les métabolites attendus sont l'acide usnique, l'atranorine et la chloroatranorine, ainsi qu'une vaste gamme de depsidones (à savoir l'acide physodique, l'acide 3-hydroxyphysodique, l'acide 4-*O*-méthylphysodique, l'acide 2'-*O*-méthylphysodique, l'acide isophysodique, l'acide physodique, l'acide 3-hydroxyphysodique, l'acide conphysodique, l'acide protocetrarique et l'acide fumarprotocetrarique), ainsi que l'acide α-alectoronique, un métabolite mineur dont la présence dans ce lichen a récemment été mise en évidence (Latkowska et al. 2015).

Tous ces métabolites ont été dérépliqués à l'exception de l'acide fumarprotocetrarique et des métabolites absents de la LDB, c'est-à-dire l'acide conphysodique, l'acide 4-*O*-méthylphysodique, l'acide 2'-*O*-méthylphysodique, l'acide 3-hydroxyphysodique et l'acide isophysodique (**Tableau 7**). En outre, l'acide salazinique, qui n'avait pas été signalé auparavant dans *H. physodes*, a été dérépliqué. L'acide conphysodique pourrait être présent sous la forme d'un nœud à *m/z* 417.0833 dans le bloc de l'acide physodique (**Figure 33** et **37**), ainsi que l'acide 3-hydroxyphysodique en tant que nœud *self-looped* à *m/z* 431.0617. Les autres annotations de niveau 5 comprennent l'acide évernique et l'acide perlatolique. L'acide fumarprotocetrarique, l'acide 4-*O*-méthylphysodique, l'acide 2'-*O*-méthylphysodique et l'acide isophysodique n'ont pas pu être détectés, même à l'état de traces (**Figure 34**).

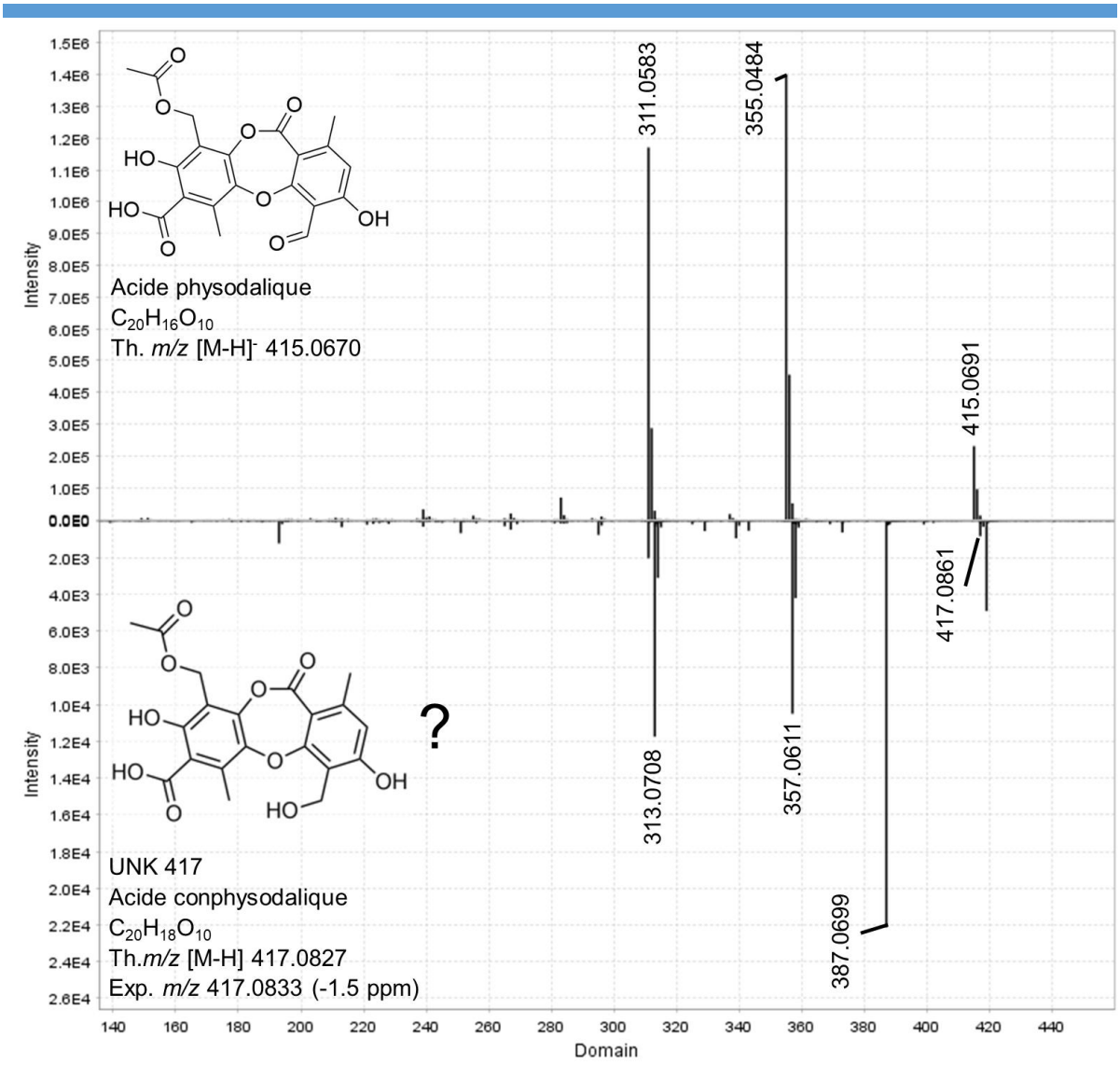


Figure 37 – Réprésentation en miroir de l'acide physodalique de la LDB contre un ion à m/z 417 d'Hypogymnia physodes, soupçonné être l'acide conphysodalique.

4 Conclusion

4.1 Des constituants de la LDB.

La LDB est actuellement constituée des spectres 250 métabolites représentatifs des classes structurales décrites dans les lichens. Ils ont été principalement analysés en ionisation ESI^- et éventuellement en ESI+ et APCI dans des cas particuliers. Ceci se traduit par un total de 309 spectres, 226 en ESI^-, 68 en ESI^+ et 15 en APCI. Ce sont, pour la grande majorité, des spectres de molécules (dé)protonées (96%), les adduits tels que $[M-H+CH_3OH]^-$, $[M-H+HCOOH]^-$ et $[2M-H]^-$ n'ayant été recherchés que lorsque les premiers n'étaient pas détectées. Bien qu'il soit attendu que la plupart des molécules s'ionisent en [M-H]- ou [M+H]+, ne considérer que ces spectres ne permettra jamais d'expliquer l'intégralité d'un réseau par la déréplication seule. L'intégration d'autres types de spectres semble indispensable pour dérépliquer au mieux les données LC-MS.

4.2 Du regroupement par similarité cosinus.

Bien que les blocs formés par similarité cosinus soient plutôt homogènes, soit 100% des nœuds regroupés en mode positif et 71% des nœuds regroupés en mode négatif, la plupart des nœuds restent sous la forme de *self-loops*. Ces nœuds regroupés concernent toutes les classes structurales, avec des pourcentages qui restent élevés (80% en mode positif, 58% en négatif) malgré le seuil relativement bas utilisé pour l'appariement (0.6). Il s'agit d'un phénomène fréquent dans les réseaux moléculaires : ces nœuds nécessiteront soit une déréplication soit l'emploi d'autres méthodes de traitement pour éviter leur examen manuel.

4.3 De l'utilité de la LDB intégrée au FBMN.

L'intégration de la LDB dans le FBMN pour la déréplication des trois lichens a néanmoins permis l'identification de plusieurs métabolites attendus. Quelques molécules supplémentaires ont également été signalées pour la première fois à partir de ces organismes déjà bien connus chimiquement. Ces résultats au premier abord surprenants peuvent être liés à des variations chémosyndromiques non signalées auparavant (C. F. Culberson and Culberson 1976) et/ou à la plus grande sensibilité de cette stratégie analytique par rapport aux techniques utilisées dans les études précédentes. La plupart des annotations étaient uniques à la LDB, ce qui souligne le fait que la plupart de ces molécules n'avaient pas été signalées auparavant dans les bases de données MS/MS du GNPS. Les métabolites absents de la LDB mais détectés dans les lichens étudiés ont pu être annotés putativement, tandis que des métabolites complètement inconnus ont pu être détectés et reliés à des classes structurelles connues dans les lichens.

La LDB, couplée à la LDB-Lit, facilite considérablement l'analyse des données LC-MS de lichens. Un réseau moléculaire simple laisse cependant une quantité considérable d'inconnues, notamment dans les nœuds *self-looped*. Bien qu'il soit toujours possible

d’examiner ces spectres manuellement, il serait préférable d’utiliser des outils complémentaires permettant de les interpréter.

Chapitre III

– Extension de la LDB –

Et étude de l'impact des instruments d'analyse & des adduits sur la déréplication des données LC-MS/MS

Intervenants extérieurs : Simon Ollivier[(b)], Mehdi A. Beniddir[(a)], Solenn Ferron[(b)], Thomas Delhaye[(c)], Pierre-Marie Allard[(d)], Jean-Luc Wolfender[(d)], Harrie J. M. Sipman[(e)], Robert Lücking[(e)], Pierre Le Pogam[(a)].

(a) : CNRS, BioCIS (Biomolécules : Conception Isolement et Synthèse)-UMR 8076, Univ Paris-Sud, Université Paris-Saclay F-92290 Châtenay-Malabry, France
(b) : CNRS, ISCR (Institut des Sciences Chimiques de Rennes)-UMR 6226, Univ Rennes, F-35000 Rennes, France
(c) : CNRS, IETR (Institut d'Électronique et Télécommunications de Rennes)-UMR 6164, Univ Rennes, F-35000 Rennes, France
(d) : EPGL, Université de Genève, Université de Lausanne, CMU, 1 Rue Michel Servet, 1211 Genève 4, Suisse
(e) : Botanischer Garten und Botanisches Museum, Freie Universität Berlin, Königin-Luise-Strasse 6–8, D-14195 Berlin, Allemagne

Contributions externes : HJMS, RL – *Mise à disposition des standards*. SF – *Préparation des standards*. MAB, PLP – *Analyses LC-MS à BioCis (Paris-Sud)*. TD – *Analyses LC-MS à l'IETR (Rennes)*. PMA, JLW – *Analyses LC-MS à Genève*. SO – *Discussion des résultats*.

Résumé

Suite à la création de la LDB dans le chapitre précédent, les mêmes métabolites ont été analysés sur deux instruments LC-MS supplémentaires : le Xevo G2-XS qToF de Waters et l'Orbitrap Q-Exactive Focus de Thermo. En plus des molécules (dé)protonées habituellement recherchées, les scripts ont été modifiés pour rechercher d'autres adduits de façon automatique. Avec ces données, la LDB a été étendue et passe de 309 à 1870 spectres. En utilisant ses données, l'impact des adduits et des instruments d'analyse sur la déréplication a été évalué. Ceci a permis d'établir que les adduits devraient être davantage pris en compte lors des analyses de métabolomique, notamment quand l'instrument utilisé en produit beaucoup. Par ailleurs, en comparant par similarité cosinus les spectres produits sur l'Agilent 6530, le Xevo G2-XS et la Q-Exative, un quart à un tiers des spectres ne pouvaient pas être reconnus.

Sommaire

1 – Introduction 93

2 – Méthodes 94

2.1 Préparation des échantillons 94

2.2 Paramètres pour le Xevo G2-XS Q-ToF de Waters 94

2.3 Paramètres pour la Q-Exactive Focus de Thermo Fisher 94

2.4 Extraction des spectres MS/MS sur R 95

2.5 Mise à jour de la LDB sur le GNPS 97

2.6 Impact de l'instrument d'analyse sur la similarité spectrale 98

2.7 Diversité des adduits dans des bases de données 98

2.8 Impact des adduits sur la déréplication 98

3 – Résultats 99

3.1 Mise à jour de la LDB sur le GNPS 99

3.2 Impact de l'instrument d'analyse sur la similarité spectrale 101

3.3 Diversité des adduits dans des bases de données 105

3.4 Impact des adduits sur la déréplication 107

4 – Conclusion 109

4.1 De l'extension de la LDB 109

4.2 Des spectres acquis sur différents instruments 109

4.3 De l'importance des adduits 109

1 Introduction

La LDB développée précédemment était constituée de 309 spectres MS/MS produits sur un Agilent 6530 Q-ToF (226 en mode d'ionisation négatif, 83 en positif) (Olivier-Jimenez, Chollet-Krugler, Rondeau, Beniddir, Ferron, Delhaye, Allard, et al. 2019). Elle était essentiellement composée de spectres de molécules (dé)protonées, d'autres adduits n'étant recherchés que quand les premiers n'étaient pas trouvés. La prédominance des molécules (dé)protonées dans les bases de données spectrales est fréquente mais pose des soucis en métabolomique non ciblée. Les analyses LC-MS données-dépendantes acquièrent des spectres MS/MS pour tout rapport *m/z* détecté, qu'il s'agisse d'un fragment, d'un adduit ou d'un pic du massif isotopique. Pour ce qui est des adduits, chaque instrument les produit en quantités variables suivant les conditions expérimentales. Comme les librairies spectrales sont souvent limitées aux molécules (dé)protonées, elles sont inadéquates pour la déréplication de spectres expérimentaux, par défaut considérés comme des ions $[M+H]^+$ ou $[M-H]^-$. L'absence d'une forme d'ionisation pour une molécule dans les bases de données entrainera au mieux l'absence d'identification, au pire un faux-positif. En plus de cela, les spectres ESI ne sont pas identiques d'une machine à l'autre et les résultats d'une déréplication automatique sont moins bons.

Dans ce contexte, la LDB a été étendue par l'analyse des mêmes molécules du *Chapitre II* avec deux autres instruments : le Xevo G2-XS qToF Waters de l'IETR, BioEM Rennes et la Q-Exactive Focus Thermo Fisher du laboratoire de phytochimie de Genève. Ces instruments seront désignés par le nom de leur fabriquant : Agilent, Waters et Thermo. Cette fois, les spectres de différents adduits ont été récupérés pour chaque molécule à l'aide de scripts sur R et Python. Après vérification manuelle de chaque spectre, ils ont été soumis au GNPS pour enrichir la LDB. Avec cette base de données améliorée, l'impact des adduits dans la déréplication a pu être évalué, ainsi que l'impact du changement d'appareil d'analyse.

2 Méthodes

2.1 Préparation des échantillons.

Tous les 250 standards (**Annexe**, **Tableau S-1**) ont été dissous dans du méthanol de qualité UPLC à une concentration de 1 mg/mL.

2.2 Paramètres pour le Xevo G2-XS Q-ToF de Waters.

La séparation chromatographique et l'analyse en masse ont été réalisées sur un système Xevo G2-XS (Waters, Milford, MA, USA) composé d'une UPLC et d'un spectromètre de masse Q-ToF. La colonne utilisée était une Waters BEH C18 100x2.1 mm, 1.7 µm. La phase mobile était composée d'un mélange d'eau (A) eau avec 0,1 % d'acide formique et d'acétonitrile (ACN) (B) avec 0.1 % d'acide formique. Le gradient utilisé était le suivant : 5% ACN (0 min), 100% ACN (3.5 min), 100% ACN (5 min), 5% ACN (5.01 min), 5% ACN (5.5 min). Le débit de solvant était de 600 µL/min et volume d'injection de 2 µL. Chaque standard a été analysé en modes positif et négatif avec la méthode Fast-DDA (Data-Dependent Analysis). Le temps de scan a été fixé à 0.067 s/scan. Pour chaque spectre MS^1, les trois ions les plus intenses ont été fragmentés à quatre énergies de collision (10, 20, 30 et 40 eV). La fenêtre *m/z* a été fixée en utilisant une valeur de LM à 16 et de HM à 18. Une liste d'exclusion pour chaque ion fragmenté a été établie, les excluant des fragmentations pendant 2 secondes. La lock-mass interne utilisée était la Leucine-Enképhaline ($C_{28}H_{37}N_5O_7$) et l'appareil a été calibré à l'aide d'une solution de formiate de sodium (HCOONa).

2.3 Paramètres pour la Q-Exactive Focus de Thermo Fisher.

La séparation chromatographique a été réalisée sur un système Acquity UHPLC (Waters, Milford, MA, USA) interfacé à un spectromètre de masse Q-Exactive Focus (Thermo Scientific, Brême, Allemagne), utilisant une source d'ionisation par électrospray chauffée (HESI-II). Les conditions de la LC étaient les suivantes : colonne : Waters BEH C18 100x2.1 mm, 1.7 µm ; phase mobile : (A) eau avec 0,1 % d'acide formique ; (B) acétonitrile avec 0.1 % d'acide formique ; débit : 600 µL/min ; volume d'injection : 2 µL ; gradient. Les paramètres optimisés de la source HESI-II étaient les suivants : tension de source : 3,5 kV, débit de gaz (N_2) : 48 unités ; débit de gaz auxiliaire : 11 unités ; débit de gaz de réserve : 2.0 ; température capillaire : 256.2 °C (pos), niveau RF de S-Lens : 45. Le spectromètre de masse a été calibré à l'aide d'un mélange de caféine, de méthionine-arginine-phénylalanine-alanine-acétate (MRFA), de sulfate de dodécyle de sodium, de taurocholate de sodium et d'Ultramark 1621 dans une solution acétonitrile/méthanol/eau contenant 1% d'acide formique par infusion directe. Les acquisitions MS/MS données-dépendantes ont été réalisés sur les 3 ions les plus intenses détectés dans la MS à balayage complet (expérience Top3). La largeur de la fenêtre d'isolement MS/MS était de 1 Da, et l'énergie de collision normalisée (NCE) a été fixée à

15, 30 et 45 unités. Dans les expériences MS/MS dépendantes des données, des balayages complets ont été acquis à une résolution de 35 000 FWHM (à *m/z* 200) et des balayages MS/MS à 17 500 FWHM avec un temps d'injection maximum automatique. Après avoir été acquis dans les scans MS/MS, les ions parents ont été placés dans une liste d'exclusion dynamique pendant 2.0 secondes.

2.4 Extraction des spectres MS/MS sur R.

Les fichiers LC-MS produits par tous les instruments ont été convertis au format mzXML à l'aide du module msConvert (Adusumilli and Mallick 2017) de ProteoWizard (Kessner et al. 2008; M. C. Chambers et al. 2012). Dans le cas des fichiers Waters, le canal de la *lockmass* (derniers fichiers IDX et STS, correspondant à la Leucine-Enképhaline) a été supprimé au préalable. Pour extraire les spectres individuels des fichiers mzXML, une liste de toutes les valeurs de *m/z* correspondant aux différents adduits attendus pour chaque fichier a été produite (**Tableau 8**). A partir de ces listes, les masses attendues pour chaque molécule ont été calculées : $mz = x \times m + \Delta m$ (mz étant le rapport *m/z* calculé pour l'adduit donné, x étant le nombre de molécules M, Δm la masse apportée par l'adduit et m la masse exacte de la molécule neutre).

Les charges n'ont pas été prises en compte : aucun adduit multichargé n'a été recherché. Ces fenêtres étaient de 15 ppm pour les fichiers issus des instruments Thermo et Agilent, 3000 ppm pour ceux de l'instrument Waters.

Tableau 8 – Adduits recherchés avec les paramètres nécessaires au calcul de leur rapport m/z.

Mode positif			Mode négatif		
Adduit	*Δm*	*x*	*Adduit*	*Δm*	*x*
$[M+H]^+$	+1.0078	1	$[M-H-H_2O]^-$	-19.0189	1
$[M+H+HCOOH]^+$	+47.0133	1	$[M-H]^-$	-1.0078	1
$[M+NH_4]^+$	+18.0343	1	$[M-2H+Na]^-$	+20.9741	1
$[M+Na]^+$	+22.9897	1	$[M-H+CH_3OH]^-$	+31.0183	1
$[M+H+CH_3OH]^+$	+33.0340	1	$[M+Cl]^-$	+34.9688	1
$[M+K]^+$	+38.9637	1	$[M-2H+K]^-$	+36.9480	1
$[M+H+CH_3CN]^+$	+42.0343	1	$[M-H+HCOOH]^-$	+44.9976	1
$[M+2Na-H]^+$	+44.9717	1	$[2M-H]^-$	-1.0078	2
$[M+Na+CH_3CN]^+$	+64.0163	1	$[2M-2H+Na]^-$	+20.9741	2
$[M+H+C_2H_6OS]^+$	+79.0217	1	$[2M-2H+K]^-$	+36.9480	2
$[2M+H]^+$	+1.0078	2	$[2M-H+HCOOH]^-$	+44.9976	2
$[2M+H+CH_3OH]^+$	+33.0340	2			
$[2M+H+HCOOH]^+$	+47.0133	2			
$[2M+NH_4]^+$	+18.0343	2			
$[2M+Na]^+$	+22.9897	2			
$[2M+Na+CH_3OH]^+$	+55.0165	2			
$[2M+K]^+$	+38.9637	2			
$[2M+H+CH_3CN]^+$	+42.0343	2			
$[2M+Na+CH_3CN]^+$	+64.0163	2			

Les fichiers convertis en mzXML ont ensuite été chargés et traités à l'aide de la librairie MSnBase (Gatto and Lilley 2012) sur R 3.6.1 (R Development Core Team 2008). L'objectif de ce traitement est de transformer des fichiers LC-MS (enchaînement de scans MS^1 et MS^2 sur plusieurs minutes d'analyse) en spectres $MS2^2$ individuels au format MGF (**Figure 38**).

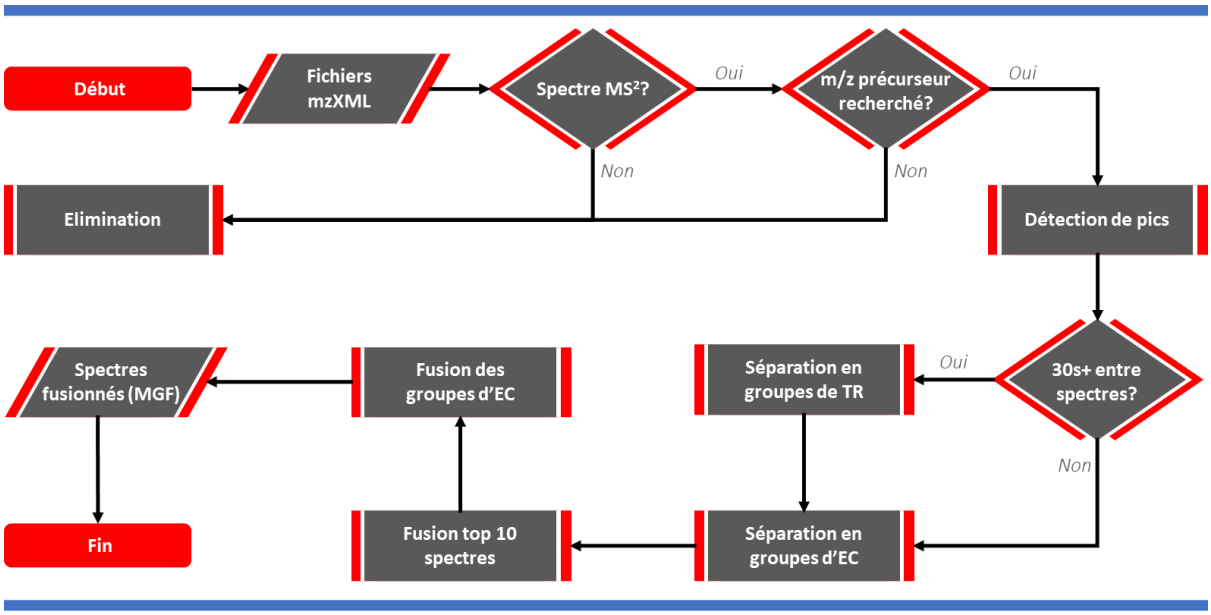


Figure 38 – Traitement des fichiers bruts mzXML sur R. TR : Temps de Rétention, EC : Energie de Collision.

Après avoir importé sur R les fichiers mzXML, les scans MS^1 ont été éliminés. Parmi les scans restants (MS^2), seuls ceux dont le *m/z* de l'ion parent était présent dans liste des masses recherchées pour la molécule donnée ont été conservés. Les fenêtres de *m/z* utilisées pour faire correspondre ces masses étaient de 15 ppm pour les instruments Agilent et Thermo, 3000 ppm pour le Waters. Les rapports *m/z* affichés dans les fichiers *mzXML* pour les ions précurseurs sur Waters étaient systématiquement très éloignés de ce qui était mesuré et il a fallu adapter la fenêtre d'erreur. Une détection de pics rudimentaire a été effectuée pour séparer les ions isobares qui pourraient être détectés dans la même analyse que les molécules recherchées. Cette séparation en pics était effectuée quand deux scans MS^2 étaient espacés de 30 secondes ou plus. Quand les ions ont été fragmentés à différentes énergies de collision (Agilent, Waters), les dix spectres les plus intenses (par TIC) pour chaque énergie de collision ont été fusionnés en un spectre consensus pour réduire les variations ponctuelles. Un spectre consensus global a ensuite été créé en combinant ceux correspondant à chaque énergie de collision et obtenir un seul document représentatif de toutes ces énergies. Les spectres avant et après fusion ont été exportés sous forme d'image pour vérifier la qualité de l'opération. Un exemple de cette vérification manuelle est présenté en **Figure 39**, avec le cas de l'acide glomellique. Dix spectres de chaque énergie de collision ont été sélectionnés pour être fusionnés. Les pics bleus de la partie inférieure des spectres représentent ces dix spectres superposés et les pics rouges dans la partie supérieure, le résultat de la combinaison. Les quatre spectres produits ont été combinés pour produire un spectre consensus global, avec à nouveau dans la partie inférieure les spectres à combiner superposés et dans la

partie supérieure le résultat. Dans le cas des données Thermo, les spectres sont automatiquement combinés durant l'analyse. Il n'était donc pas nécessaire de séparer en groupes d'énergie de collision et les dix spectres les plus intenses pour chaque pic ont été directement fusionnés. Après combinaison, les spectres consensus ont été exportés au format MGF. Une vérification semi-automatique a été effectuée en comparant les spectres d'une même molécule acquis sur les trois appareils par le biais d'une représentation en miroir, de leur similarité cosinus ainsi que de leur temps de rétention et TIC (**Figure 40**). De façon générale, ont été retenus les spectres qui étaient semblables pour les trois instruments avec des similarités cosinus élevées, un TIC élevé, des temps de rétention proches et s'ils ressemblaient à ce qui avait déjà été validé pour l'instrument Agilent auparavant. Ceci a été réalisé avec Python 3.7 (Van Rossum and Drake 2009) et la librairie Libmetgem (Olivon et al. 2018).

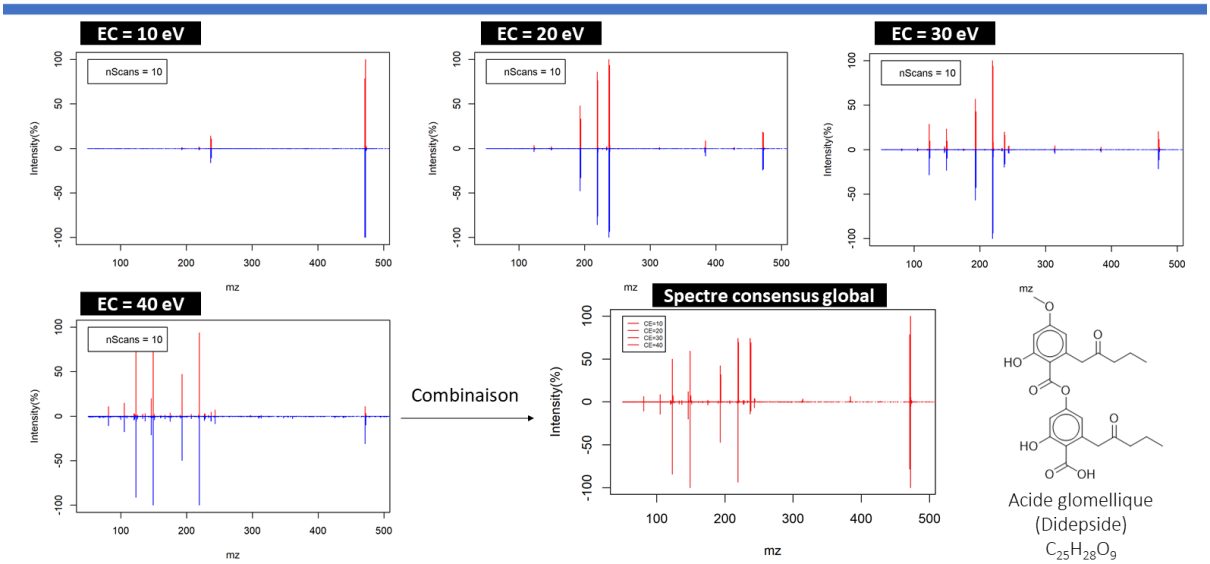


Figure 39 – Création du spectre consensus de l'acide glomellique sous la forme $[M-H]^-$ (Xevo G2-XS). Les quatre premières représentations en miroir correspondent à la combinaison (rouge) des spectres à chaque énergie de collision (bleu). La dernière, correspond au spectre consensus produit à partir des quatre spectres combinés précédents.

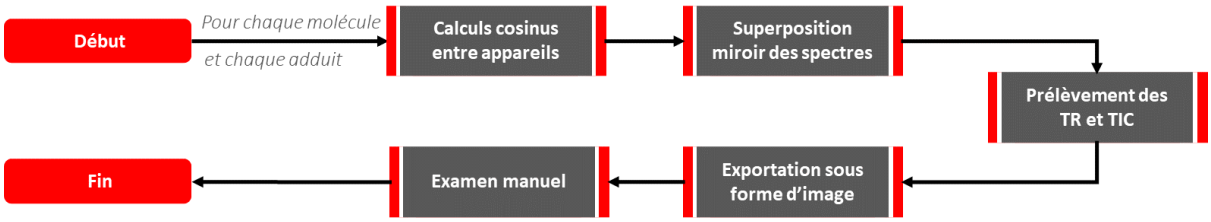


Figure 40 – Vérification manuelle des spectres sur la base de leur similarité d'un appareil à l'autre.

2.5 Mise à jour de la LDB sur le GNPS.

Une fois toutes les étapes de vérification terminées, les spectres ont tous été convertis à nouveau au format mzXML et soumis au GNPS pour enrichir la LDB pré-existante en

suivant les recommandations de la documentation fournie par le site (https://ccms-ucsd.github.io/GNPSDocumentation/batchupload/).

2.6 Impact de l'instrument d'analyse sur la similarité spectrale.

L'intégralité des données de la LDB ont été utilisées pour estimer la proportion des spectres pouvant être reconnus d'un instrument à l'autre. Pour ce faire, des scores de cosinus ont été calculés entre toutes les paires de spectres correspondant à la même molécule et au même adduit tant qu'ils étaient produits par des machines différentes. Le seuil de score utilisé était de 0.7 (seuil habituellement utilisé). Les données des modes positif et négatif ont été traitées séparément pour mettre en évidence les différences éventuelles. Les mêmes calculs ont été réalisés en faisant varier le seuil utilisé de 0.1 à 1.0 par pas de 0.025 pour s'affranchir des limitations d'un seuil unique. Ces calculs ont été effectués sur Python 3.7 à l'aide de la librairie Libmetgem.

2.7 Diversité des adduits dans des bases de données.

Une comparaison de la diversité des adduits entre la LDB et le reste des librairies du GNPS a été réalisée. Les bases de données mises à disposition par le GNPS ont été téléchargées sur la page des librairies (https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp). Les différents types d'adduits ont été relevés ainsi que le nombre de représentants pour chacun d'entre eux.

2.8 Impact des adduits sur la déréplication.

L'efficacité d'une déréplication « aveugle » (sans filtre d'adduits) a ensuite été évaluée en soumettant les spectres de la LDB à elle-même. Pour chaque machine, les spectres ont été comparés à ceux des autres machines par similarité cosinus et l'identification au meilleur score a été retenue. La même opération a été réalisée une seconde fois avec un filtre pour ne permettre de comparer que les spectres qui présentent la même forme d'ionisation. Dans les deux cas, un seuil de score variable a été utilisé, allant de 0.05 à 1.00 par pas de 0.05. La proportion de spectres correctement identifiés, non-identifiés et mal identifiés en fonction du score seuil utilisé a été mesurée.

3 Résultats

3.1 Mise à jour de la LDB sur le GNPS.

Après l'analyse des 250 métabolites lichéniques, un total de 1561 spectres (816 en mode négatif, 745 en mode positif) ont été ajoutés aux 309 déjà présents dans la LDB. Elle répertorie à présent 1870 spectres MS/MS de molécules lichéniques répartis sur trois spectromètres de masse et bien plus d'adduits qu'auparavant. Parmi ces spectres, 342 proviennent d'Agilent (97 pos, 245 neg), 517 de Waters (281 pos, 236 neg) et 702 de Thermo (367 pos, 335 neg). La diversité des adduits est exposée dans la **Figure 41** et le **Tableau 9**.

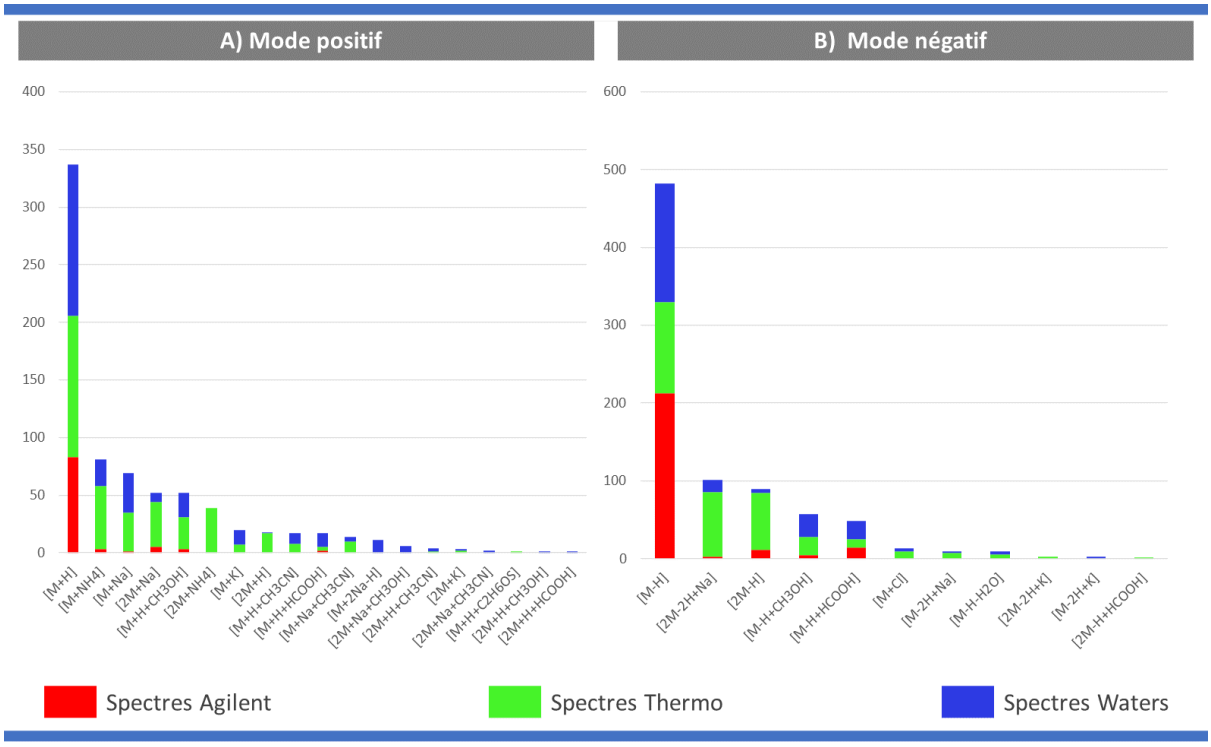


Figure 41 – Nombre de spectres détectés pour chaque catégorie d'adduit pour chaque instrument analytique utilisé en mode positif (A) et négatif (B).

Les molécules (dé)protonées restent majoritaires bien que d'autres adduits soient fréquemment détectés. Une chose se dégage déjà de la **Figure 41** : cette prépondérance des molécules (dé)protonées est surtout vérifiée sur Agilent où presque aucune autre forme n'a été détectée en abondance. Les autres formes sont mieux détectées sur l'instrument Waters bien que les molécules (dé)protonées soient encore majoritaires. Sur l'instrument Thermo, ces ions deviennent presque aussi courants que les autres.

Ceci est vérifié en considérant les chiffres fournis par le **Tableau 9**. Sur Agilent, 85% et 87% des spectres sont des molécules protonées et déprotonées respectivement. Ces pourcentages sont considérablement modifiés sur Waters : 47% de $[M+H]^+$ et 64% de $[M-H]^-$. D'autres formes doivent être considérées comme $[M+Na]^+$ (12%), $[M+NH_4]^+$ (8%) et

$[M+H+CH_3OH]^+$ (7%) en mode positif, $[M-H+CH_3OH]^-$ (12%), $[M-H+HCOOH]^-$ (9%) et $[2M-2H+Na]^-$ (6%) en mode négatif. Sur l'Orbitrap de Thermo les molécules (dé)protonées ne forment plus que 33 et 35% des ions détectés respectivement en mode positif et négatif. Plusieurs autres formes sont détectées en quantités importantes, tels que les adduits $[M+NH_4]^+$ (15%), des dimères $[2M+Na]^+$ $[2M+NH_4]^+$ (10% chacun), $[M+Na]^+$ (9%) et $[M+H+CH_3OH]^+$ (7%) en mode positif, $[2M-2+Na]^-$ (25%), $[2M-H]^-$ (22%) et $[M-H+CH_3OH]^-$ (7%) en mode négatif.

Tableau 9 – Adduits recherchés suivant le mode d'ionisation et le nombre de spectres trouvés pour chacun. Le nombre de spectres par catégorie est représenté en gras et le pourcentage occupé par cette catégorie parmi les autres adduits est représenté en italique entre parenthèses.

Mode Positif				Mode Négatif			
Adduit	*Agilent*	*Thermo*	*Waters*	*Adduit*	*Agilent*	*Thermo*	*Waters*
$[M+H]^+$	**83** *(85.57)*	**123** *(33.51)*	**131** *(46.62)*	$[M-H]^-$	**213** *(86.94)*	**117** *(34.93)*	**152** *(64.41)*
$[M+NH_4]^+$	**3** *(3.09)*	**55** *(14.99)*	**23** *(8.19)*	$[2M-2H+Na]^-$	**3** *(1.22)*	**83** *(24.78)*	**15** *(6.36)*
$[M+Na]^+$	**1** *(1.03)*	**34** *(9.26)*	**34** *(12.10)*	$[2M-H]^-$	**11** *(4.49)*	**74** *(22.09)*	**4** *(1.69)*
$[2M+Na]^+$	**5** *(5.15)*	**39** *(10.63)*	**8** *(2.85)*	$[M-H+CH_3OH]^-$	**4** *(1.63)*	**24** *(7.16)*	**29** *(12.29)*
$[M+H+CH_3OH]^+$	**3** *(3.09)*	**28** *(7.63)*	**21** *(7.47)*	$[M-H+HCOOH]^-$	**14** *(5.71)*	**11** *(3.28)*	**23** *(9.75)*
$[2M+NH_4]^+$	**0** *(0.00)*	**39** *(10.63)*	**0** *(0.00)*	$[M+Cl]^-$	**0** *(0.00)*	**9** *(2.69)*	**4** *(1.69)*
$[M+K]^+$	**0** *(0.00)*	**7** *(1.91)*	**13** *(4.63)*	$[M-2H+Na]^-$	**0** *(0.00)*	**7** *(2.09)*	**2** *(0.85)*
$[2M+H]^+$	**0** *(0.00)*	**17** *(4.63)*	**1** *(0.36)*	$[M-H-H_2O]^-$	**0** *(0.00)*	**5** *(1.49)*	**4** *(1.69)*
$[M+H+CH_3CN]^+$	**0** *(0.00)*	**8** *(2.18)*	**9** *(3.20)*	$[2M-2H+K]^-$	**0** *(0.00)*	**3** *(0.90)*	**0** *(0.00)*
$[M+H+HCOOH]^+$	**2** *(2.06)*	**3** *(0.82)*	**12** *(4.27)*	$[M-2H+K]^-$	**0** *(0.00)*	**0** *(0.00)*	**3** *(1.27)*
$[M+Na+CH_3CN]^+$	**0** *(0.00)*	**10** *(2.72)*	**4** *(1.42)*	$[2M-H+HCOOH]^-$	**0** *(0.00)*	**2** *(0.60)*	**0** *(0.00)*
$[M+2Na-H]^+$	**0** *(0.00)*	**0** *(0.00)*	**11** *(3.91)*				
$[2M+Na+CH_3OH]^+$	**0** *(0.00)*	**0** *(0.00)*	**6** *(2.14)*				
$[2M+H+CH_3CN]^+$	**0** *(0.00)*	**1** *(0.27)*	**3** *(1.07)*				
$[2M+K]^+$	**0** *(0.00)*	**2** *(0.54)*	**1** *(0.36)*				
$[2M+Na+CH_3CN]^+$	**0** *(0.00)*	**0** *(0.00)*	**2** *(0.71)*				
$[M+H+C_2H_6OS]^+$	**0** *(0.00)*	**1** *(0.27)*	**0** *(0.00)*				
$[2M+H+CH_3OH]^+$	**0** *(0.00)*	**0** *(0.00)*	**1** *(0.36)*				
$[2M+H+HCOOH]^+$	**0** *(0.00)*	**0** *(0.00)*	**1** *(0.36)*				

Rapportés au nombre de molécules analysées, l'instrument Agilent a produit en moyenne 1.17 et 1.09 ions par molécule analysée respectivement en mode positif et négatif. Ces taux montent à 1.89 et 1.49 pour Waters et atteignent 2.80 et 2.72 pour Thermo.

Ces différences significatives mettent en évidence les quantités variables d'adduits générés d'un instrument à l'autre. Une analyse telle que conduite ici avec le Q-ToF 6530 d'Agilent ne génère que très peu d'adduits en dehors des molécules (dé)protonées et ceci ne poserait pas de problème lors d'utilisation de librairies spectrales avec peu d'adduits. Pour une analyse effectuée dans les mêmes conditions qu'avec la Q-Exactive Thermo, le décalage serait grand entre une base de données pauvre en adduits et les données générées avec presque trois ions différents par molécule. Ces rapports sont cependant issus de l'analyse de standards concentrés et bien qu'il soit possible de dire que l'appareil Agilent ne produit que peu d'adduits, l'analyse avec la Q-Exactive sur un échantillon complexe pourrait produire en moyenne moins d'adduits. Cependant, la liste des adduits recherchés ici n'est pas exhaustive et ces taux pourraient être revus à la hausse en utilisant une liste complète.

L'origine de ces adduits pourrait être débattue : conséquence de la nature ou de la géométrie des appareils ? De la source d'ions utilisée ? De la matrice des échantillons ? Ceci requiert d'autre part des études plus spécifiques sur l'appareillage.

Comme auparavant, la LDB est consultable sur le site du GNPS en mode positif (https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=LDB_POSITIVE) et en mode négatif (https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=LDB_NEGATIVE).

De par la nature de ses spectres, la LDB offre des opportunités pour étudier le comportement des identifications par similarité de cosinus en fonction de l'instrument d'analyse, ainsi que l'impact des adduits lors des déréplications.

3.2 Impact de l'instrument d'analyse sur la similarité spectrale.

Les spectres acquis sur un instrument ont été tour à tour comparés à leurs équivalents sur les deux autres machines (même molécule, même adduit). La proportion d'ions reconnus d'une machine à l'autre avec un score seuil fixé à 0.7 est présentée en **Figure 42**.

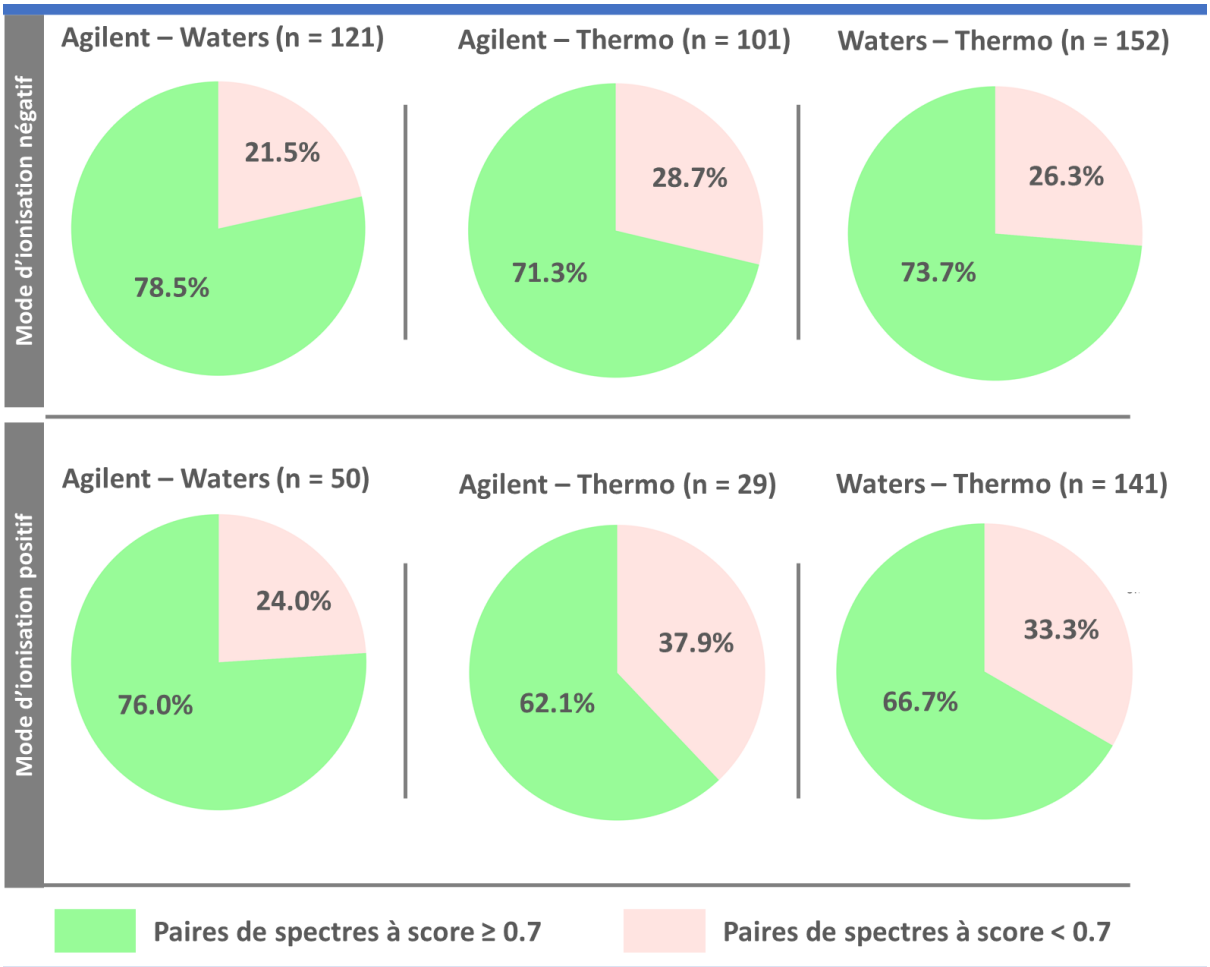


Figure 42 – Pourcentage de spectres identifiés en les comparant d'un instrument à l'autre avec un seuil de similarité cosinus à 0.7. Le nombre de paires utilisées pour chaque comparaison est représenté par « n ».

Qu'il s'agisse du mode positif ou négatif, le pourcentage d'ions identifiés est plus élevé pour la comparaison Agilent – Waters. Ensuite, ce sont les données produites sur Waters qui sont les plus proches de celles de Thermo et les données Agilent – Thermo ont les pourcentages les plus faibles. Ceci peut s'expliquer par la nature des instruments et par les paramètres d'analyse. En premier lieu, les instruments Waters et Agilent utilisent tous deux des spectromètres de masse Q-ToF, alors que le Thermo est une orbitrap. Ceci pourrait être à l'origine de la ressemblance des spectres Waters et Agilent, tous deux plus distants de ceux d'une orbitrap. Du point de vue de l'analyse, la fenêtre utilisée sur l'appareil Waters a permis de produire des spectres MS/MS sans massif isotopique, comme ce qui est observé dans la Q-Exactive suite à la transformée de Fourier. Ceci n'avait pas été fait pour les données Agilent. Cette absence de massif isotopique peut être à l'origine du rapprochement entre les données Waters et Thermo par rapport à celles d'Agilent.

Concernant le mode d'ionisation, les pourcentages d'identification sont supérieurs en mode négatif. Sans certitude, ceci pourrait être attribué aux nombres moins significatifs de spectres en mode positif, ou bien s'expliquer par la nature des molécules étudiées. Habituellement, les molécules lichéniques, des acides pour la plupart, sont analysées en mode négatif ce qui permet de mieux les ioniser. Les analyser en mode positif pourrait produire des spectres de moins bonne qualité pour lesquels il est plus difficile d'obtenir de bons scores de similarité cosinus.

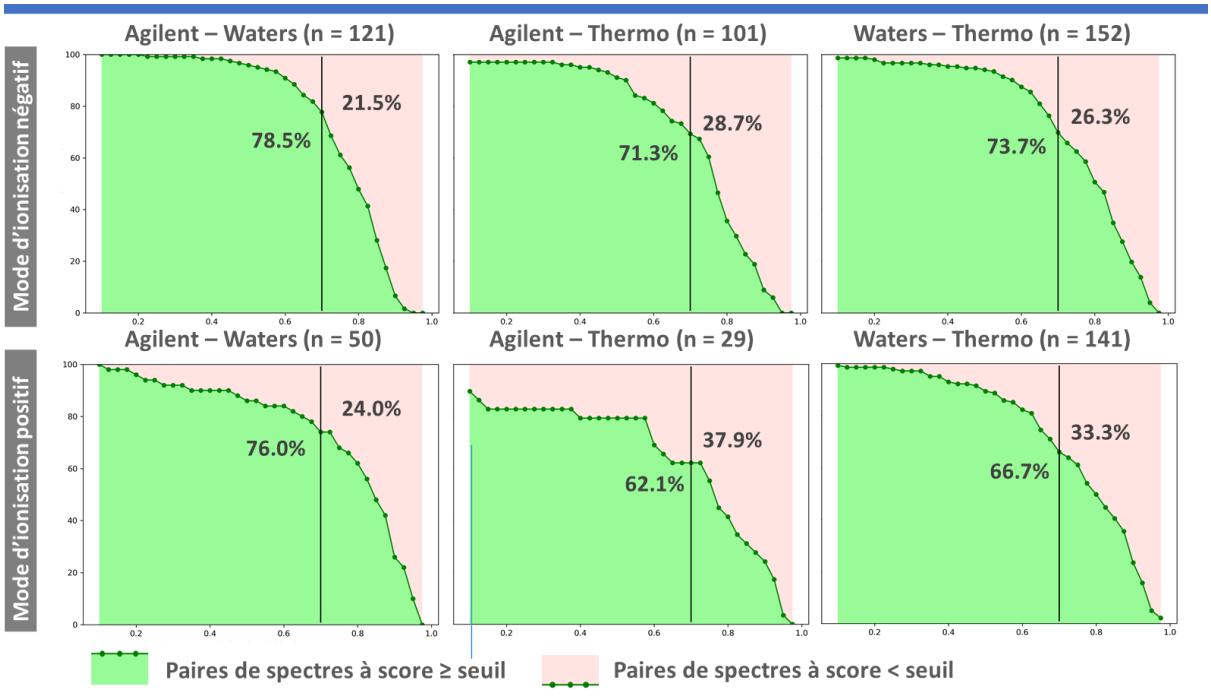


Figure 43 – Efficacité d'identification des spectres de la LDB par similarité cosinus en fonction du seuil fixé. Une ligne noire verticale a été tracée à x = 0.7 pour repérer le seuil habituellement utilisé, et les pourcentages observés à ce seuil ont été précisés.

Ces pourcentages ne sont valides que pour un seuil de score fixé à 0.7. Pour ne pas se limiter à un seul score, la proportion d'ions reconnus d'une machine à l'autre a également été mesurée en faisant varier la valeur seuil (**Figure 43**)

Les proportions de spectres identifiés en fonction du seuil utilisé restent similaires en mode négatif. Ces proportions sont légèrement inférieures en mode positif, même pour les données Waters – Thermo où le nombre de spectres se rapproche de ceux disponibles en mode négatif. Une augmentation du seuil entraine de façon prévisible une diminution du pourcentage d'identification, alors que l'abaissement du seuil augmente ce pourcentage. Cependant, abaisser ce seuil entrainerait également une augmentation du nombre de faux-positifs, non visualisables ici puisque les spectres ont été comparés à leurs équivalents sur les autres machines.

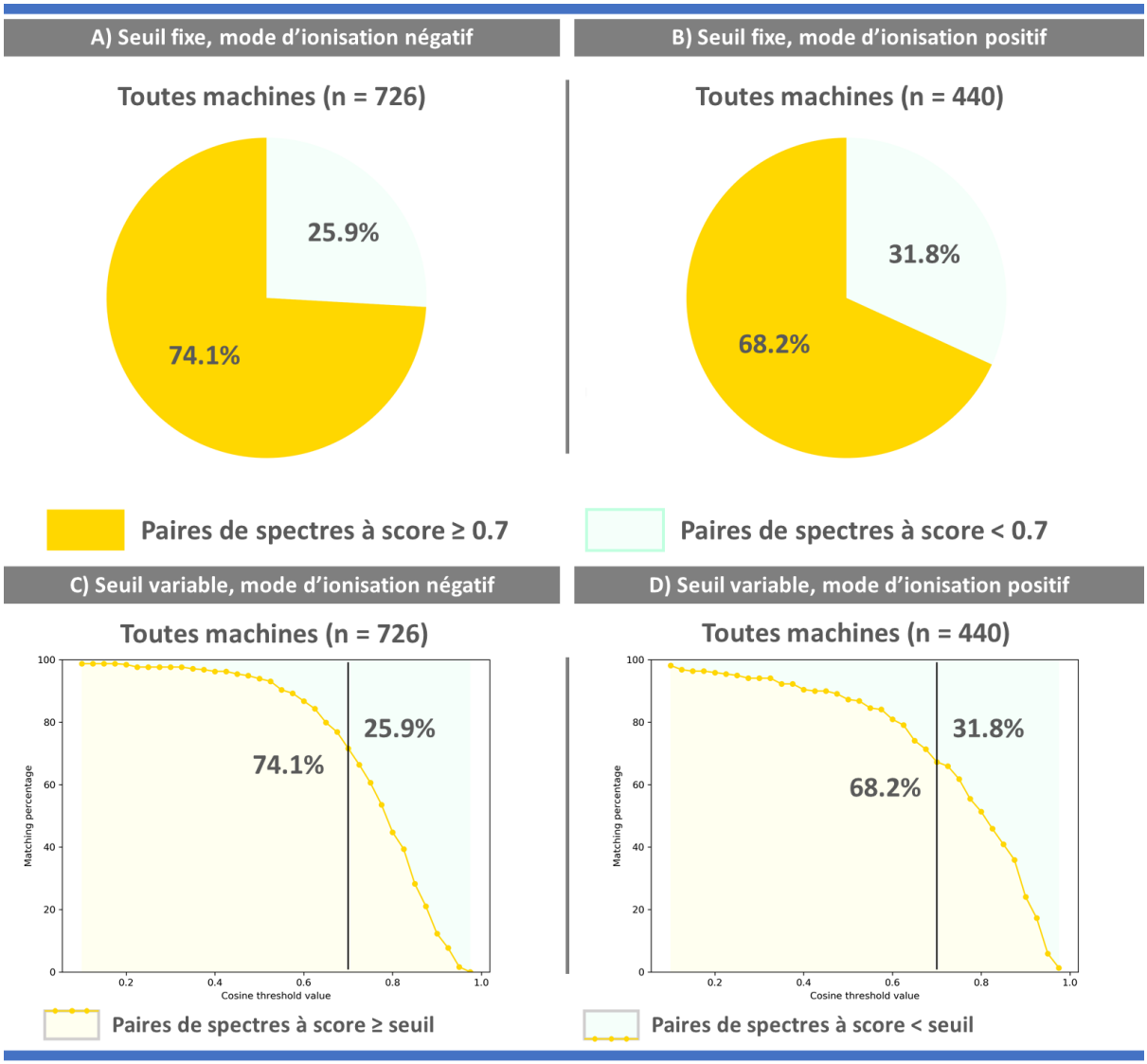


Figure 44 – Pourcentage de spectres reconnus ou non au sein de la LDB. Ces pourcentages sont représentés avec une valeur seuil à 0.7 pour le mode positif en A), pour le mode négatif en B). Ces proportions ont été mesurées en faisant varier le seuil d'identification de 0.1 à 1.0 par pas de 0.025 en mode négatif en C) et en mode positif en D).

En moyenne, 74% des ions peuvent être reconnus en mode négatif contre 68% en mode positif en utilisant un seuil de 0.7 (**Figure 44**). Un quart à un tiers des spectres ne peuvent pas être dérépliquées.

3.3 Diversité des adduits dans des bases de données.

La diversité en adduits des autres librairies spectrales du GNPS a été comparée à celle de la LDB. 9251 spectres du GNPS ont été téléchargés (8999 en mode positif, 252 en mode négatif) contre 1460 spectres de la LDB (717 en mode positif et 743 en mode négatif). 34 adduits ont été repérés en mode positif et 13 en mode négatif (**Figure 45, Tableau 10**).

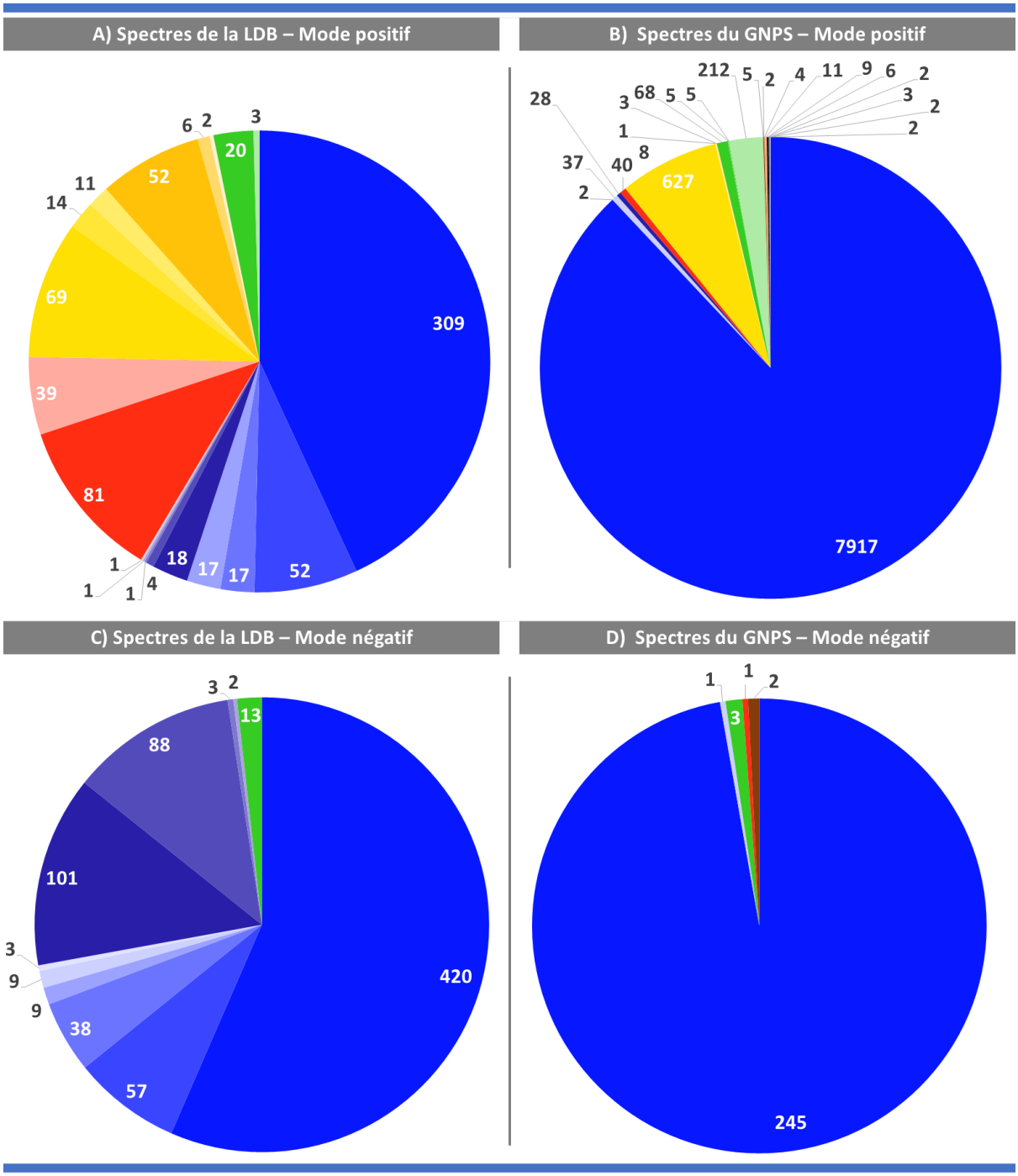


Figure 45 – Représentation de la diversité en adduits de la LDB et des librairies du GNPS suivant le mode d'ionisation. Les codes couleur utilisés correspondent à ceux des adduits du Tableau 10.

Tableau 10 – Adduits considérés pour la LDB et les librairies du GNPS. Les nuances de couleur font référence aux mêmes groupes d'adduits : pour le mode positif, les nuances de bleu correspondent aux variantes des molécules protonées, les rouges aux adduits ammonium, les jaunes aux adduits sodium et les verts aux adduits potassium (etc…). Le nombre de spectres détectés pour chaque adduit est représenté en gras, le pourcentage que représentent ces nombres au sein des bases de données est en italique et gris.

Mode positif				Mode négatif		
	Adduit	*LDB*	*GNPS*	*Adduit*	*LDB*	*GNPS*
	$[M+H]^+$	**309** *(43.10%)*	**7917** *(87.98%)*	$[M-H]^-$	**420** *(56.53%)*	**245** *(97.22%)*
	$[M+H+CH_3OH]^+$	**52** *(7.25%)*	**0** *(0.00%)*	$[M-H+CH_3OH]^-$	**57** *(7.67%)*	**0** *(0.00%)*
	$[M+H+CH_3CN]^+$	**17** *(2.37%)*	**2** *(0.02%)*	$[M-H+HCOOH]^-$	**38** *(5.11%)*	**0** *(0.00%)*
	$[M+H+HCOOH]^+$	**17** *(2.37%)*	**0** *(0.00%)*	$[M-H-H_2O]^-$	**9** *(1.21%)*	**0** *(0.00%)*
	$[M+H-H_2O]^+$	**0** *(0.00%)*	**37** *(0.41%)*	$[M-2H+Na]^-$	**9** *(1.21%)*	**1** *(0.40%)*
	$[2M+H]^+$	**18** *(2.51%)*	**28** *(0.31%)*	$[M-2H+K]^-$	**3** *(0.40%)*	**0** *(0.00%)*
	$[2M+H+CH_3CN]^+$	**4** *(0.56%)*	**0** *(0.00%)*	$[2M-2H+Na]^-$	**101** *(13.59%)*	**0** *(0.00%)*
	$[2M+H+CH_3OH]^+$	**1** *(0.14%)*	**0** *(0.00%)*	$[2M-H]^-$	**88** *(11.84%)*	**0** *(0.00%)*
	$[2M+H+HCOOH]^+$	**1** *(0.14%)*	**0** *(0.00%)*	$[2M-2H+K]^-$	**3** *(0.40%)*	**0** *(0.00%)*
	$[2M+H+C_2H_6OS]^+$	**1** *(0.14%)*	**0** *(0.00%)*	$[2M-H+HCOOH]^-$	**2** *(0.27%)*	**0** *(0.00%)*
	$[M+NH_4]^+$	**81** *(11.3%)*	**40** *(0.44%)*	$[M+Cl]^-$	**13** *(1.75%)*	**3** *(1.19%)*
	$[2M+NH_4]^+$	**39** *(5.44%)*	**0** *(0.00%)*	$[M+OH]^-$	**0** *(0.00%)*	**1** *(0.40%)*
	$[M+Na]^+$	**69** *(9.62%)*	**627** *(6.97%)*	$[M-2H]^{2-}$	**0** *(0.00%)*	**2** *(0.79%)*
	$[M+Na+CH_3CN]^+$	**14** *(1.95%)*	**0** *(0.00%)*			
	$[M+2Na-H]^+$	**11** *(1.53%)*	**1** *(0.01%)*			
	$[M+Na+H_2O]^+$	**0** *(0.00%)*	**8** *(0.09%)*			
	$[2M+Na]^+$	**52** *(7.25%)*	**3** *(0.03%)*			
	$[2M+Na+CH_3OH]^+$	**6** *(0.84%)*	**0** *(0.00%)*			
	$[2M+Na+CH_3CN]^+$	**2** *(0.28%)*	**0** *(0.00%)*			
	$[M+K]^+$	**20** *(2.79%)*	**68** *(0.76%)*			
	$[M+2K-H]^+$	**0** *(0.00%)*	**5** *(0.06%)*			
	$[M+K+H_2O]^+$	**0** *(0.00%)*	**5** *(0.06%)*			
	$[2M+K]^+$	**3** *(0.42%)*	**0** *(0.00%)*			
	$[M+2H]^{2+}$	**0** *(0.00%)*	**212** *(2.36%)*			
	$[M+2Na]^{2+}$	**0** *(0.00%)*	**5** *(0.06%)*			
	$[M+2K]^{2+}$	**0** *(0.00%)*	**2** *(0.02%)*			
	$[M+3H]^{3+}$	**0** *(0.00%)*	**4** *(0.04%)*			
	$[M+H+Na]^{2+}$	**0** *(0.00%)*	**11** *(0.12%)*			
	$[M+H+K]^{2+}$	**0** *(0.00%)*	**9** *(0.10%)*			
	$[M+Na+K-H]^+$	**0** *(0.00%)*	**6** *(0.07%)*			
	$[M]^+$	**0** *(0.00%)*	**3** *(0.03%)*			
	$[M+2Na+K-2H]^+$	**0** *(0.00%)*	**2** *(0.02%)*			
	$[M+Na+2K-2H]^+$	**0** *(0.00%)*	**2** *(0.02%)*			
	$[M+Na+K]^{2+}$	**0** *(0.00%)*	**2** *(0.02%)*			

En mode positif, les différences entre la LDB et les librairies du GNPS (**Figure 45-A et B**) sont bien visibles. Là où seulement 12% des spectres du GNPS étaient formés d'autre chose que de molécules protonées, 56.9% de la LDB est composée d'autres ions. Les autres variantes de molécules protonées passent de 0.74% à 15.48%, principalement par l'ajout des complexes avec les solvants et l'acidifiant utilisés. Il en va de même pour les

adduits sodium qui passent de 7.1% à 21.47%. Les adduits ammonium passent de 0.44% à 16.74% simplement en considérant les ions $[M+NH_4]^+$ et $[2M+NH_4]^+$. Les adduits potassium représentant 3.21% des spectres de la LDB contre 0.88% dans le GNPS. Les multichargés n'ont pas été recherchés en produisant la LDB et ils représentent 2.7% des spectres du GNPS, essentiellement sous la forme $[M+2H]^+$. De même pour d'autres adduits plus complexes cumulant trois à cinq ions tels que les $[M+2Na+K-2H]^-$ qui restent minoritaires dans le GNPS (0.16%).

Le mode négatif (**Figure 45-C et D**) ne compte que peu d'ions dans le GNPS : 252 contre les 8999 du mode positif. 97% de ces spectres sont des molécules déprotonées, réduits à 56% dans la LDB. A la différence du mode positif, les ions recherchés ici sont essentiellement des variantes de molécules déprotonées, étant donné qu'une perte de proton est presque toujours nécessaire pour s'ioniser. Ces variantes sont presque absentes du GNPS mais constituent l'autre moitié des données de la LDB (42%), le reste (2%) étant composé d'ions $[M+Cl]^-$.

3.4 Impact des adduits sur la déréplication.

Au vu des différences d'abondance d'adduits qui peuvent être rencontrées entre des données expérimentales et les librairies d'adduits, des faux-positifs peuvent survenir lors des identifications. Dans le cas d'un instrument produisant peu d'adduits, le manque de diversité des librairies spectrales ne posera que peu de soucis. Dans des situations où les données expérimentales sont composées en grande partie d'autre chose que de molécules (dé)protonées, l'inadéquation risque d'entrainer l'annotation incorrecte de spectres.

Pour estimer l'impact de ces inadéquations, les spectres produits par chaque instrument de la LDB ont été dérépliqués vis-à-vis de ceux des autres appareils. La comparaison à des spectres d'autres instrument est ici nécessaire, car autrement le meilleur *hit* pour un spectre acquis sur un instrument serait lui-même. La déréplication s'est faite en calculant les scores de similarité cosinus avec les autres spectres indépendamment de l'adduit dans un premier temps, puis en ne comparant que les spectres présentant les mêmes formes d'ionisation. Un seuil de score variable a été utilisé, allant de 0.05 à 1.0 par pas de 0.05. La proportion d'ions correctement identifiés, mal identifiés et non identifiés en fonction de la valeur seuil utilisée est présentée en **Figure 46**.

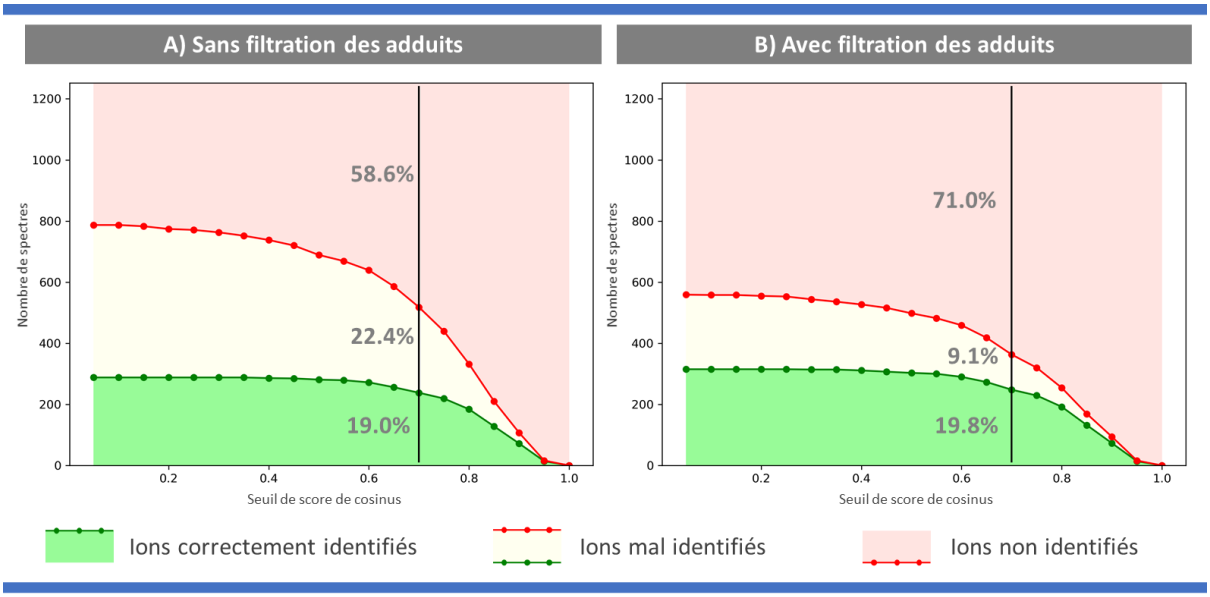


Figure 46 – Impact de l'inadéquation de diversité d'adduits lors de la déréplication, exemple de la LDB. Une ligne verticale à x = 0.7 a été placée pour représenter le seuil de cosinus habituellement utilisé. Les pourcentages correspondant à chaque catégorie d'identification à ce seuil ont été représentés.

En se basant sur un seuil à 0.7, sans appliquer de filtres sur les adduits (**Figure 46-A**), la proportion d'ions correctement identifiés (19%) est d'un ordre de grandeur similaire à celle des mal identifiés (22.4%). Bien que le jeu de données reste relativement petit (limité aux spectres de la LDB), 58% des spectres restent non identifiés. Ceci s'explique par les 25 à 30% de spectres non reconnus d'un appareil à l'autre en plus des spectres qui sont uniques à certains appareils et qui ne peuvent donc pas être dérépliqués. Comme établi précédemment, abaisser ce seuil n'augmente que légèrement le nombre d'ions correctement identifiés. La proportion de faux-positifs augmente plus vite en puisant dans les ions non identifiés.

En appliquant un filtre sur les adduits (**Figure 46-B**), le nombre d'ions correctement identifiés reste sensiblement le même (augmentation de 0.8 points) alors que les faux positifs sont 2.5 fois moins nombreux et passent de 22.4 à 9.1%. L'application du filtre d'adduits permet notamment de préserver de nombreux ions non-identifiés d'une mauvaise identification, leurs nombres passant de 58.6 à 71.0%. En abaissant le seuil, le nombre d'ions mal identifiés atteint plus rapidement un plateau qu'avant.

Quand il y a de nombreux types d'adduits dans les données, l'usage de filtres d'adduits devient intéressant pour réduire l'effort de vérification manuelle après la déréplication. Autrement, une partie non négligeable des données est mal annotée.

4 Conclusion

4.1 De l'extension de la LDB.

L'extension de la LDB par l'analyse sur différents instruments et l'intégration des nombreux adduits a permis de sextupler la taille de cette dernière en passant de 309 à 1870 spectres. A l'aide de ces données, il a été possible d'étudier l'impact du changement d'instrument de mesure sur les spectres MS/MS et l'intérêt de la diversité des adduits dans les librairies spectrales. Une meilleure concordance entre les instruments Agilent et Waters est observée par rapport à celui de Thermo. Ceci est sans doute lié au fait que les deux premiers utilisent des détecteurs ToF et que le dernier soit une Orbitrap.

4.2 Des spectres acquis sur différents instruments.

En utilisant un seuil de score à 0.7, 26% (mode négatif) à 32% (mode positif) des spectres MS/MS ne sont pas reconnus d'un instrument à l'autre. Les différences entre les deux modes est certainement due à la nature des molécules analysées, plus difficiles à ioniser en mode positif. Une meilleure concordance est observée entre les instruments Agilent et Waters, probablement car ce sont des ToF, alors que l'instrument Thermo est une Orbitrap.

4.3 De l'importance des adduits.

La proportion d'adduits produits a varié d'un instrument à l'autre, l'appareil Agilent n'en produisant que très peu et l'Orbitrap Thermo en produisant beaucoup. Ceci est certainement causé par les différences dans les paramètres utilisés, notamment l'usage d'une liste d'exclusion dynamique

. Grâce à celle-ci, d'autres espèces ioniques peuvent être observées pour chaque molécule, en l'absence de celle-ci, les fragmentations resteront superficielles et peu d'adduits seront fragments. Sans même aborder les fragments de source, ceci remet en cause l'importance accordée aux molécules (dé)protonées dans les bases de données pour expliquer les ions des données LC-MS. Une déréplication « aveugle », ne prenant pas en compte cette diversité d'adduits, génère un taux élevé de faux-positifs. La détermination de la forme d'ionisation devrait être une priorité lors de l'interprétation de données LC-MS en métabolomique non-ciblée. En plus d'un filtre d'adduits, d'autres filtres pourraient être envisagés pour réduire cette proportion de faux positifs, comme l'annotation des fragments de source.

Chapitre IV

– LDB-MotifDB & *Classnotator* –

Ou la création d'une base de données de motifs pour la LDB sur MS2LDA et d'un outil de prédiction des classes structurales

Intervenants extérieurs : Justin J. J. van der Hooft[(a)], Kyo Bin Kang[(b)], Simon Rogers[(c)], Joe Wandy[(d)].

(a) : Groupe de Bioinformatique, Université de Wageningen, Pays-Bas
(b) : College of Pharmacy, Sookmyung Women's University, Seoul, République de Corée
(c) : School of Computing Science, Université de Glasgow, Glasgow, Royaume-Uni
(d) : Glasgow Polyomics, Université de Glasgow, Glasgow, Royaume-Uni

Contributions externes : KBK – *Mise en relation avec le personnel de MS2LDA et annotation de certains motifs*. JJJvdH, SR, JW – *Conseils pour l'élaboration de la LDB-MotifDB, génération des annotations MAGMa*. JJJvdH – *Proposition d'améliorations pour Classnotator*.

Résumé

La LDB étendue est utilisée dans ce chapitre pour créer la LDB-MotifDB sur MS2LDA : une base de données de motifs (mass2motifs) récurrents dans les spectres MS^2 de la LDB. Les bases de données de motifs sont utilisées pour faciliter l'interprétation des réseaux moléculaires, en reliant des nœuds partageant ces motifs récurrents. Sur cette base, un outil permettant d'annoter automatiquement les nœuds avec une classe structurale est créé : *Classnotator*. Ici, la classification selon Hun&Yosh96 est utilisée et ceci est démontré sur les spectres de la LDB.

Sommaire

1 – Introduction 112

2 – Méthodes 116

2.1 Création de la LDB-motifDB sur ms2lda.org 116

2.2 Recherche des motifs représentatifs des classes chimiques 116

2.3 Impact des adduits et des instruments d'analyse sur les motifs 118

3 – Résultats 120

3.1 Création de la LDB-motifDB sur ms2lda.org 120

3.2 Recherche des motifs représentatifs des classes chimiques 120

3.3 Impact des adduits et des instruments d'analyse sur les motifs 123

4 – Conclusion 126

4.1 De la création de *Classnotator* 126

4.2 Impact des adduits et des instruments d'analyse sur les motifs 126

1 Introduction

Les méthodes de déréplication actuelles sont basées sur la comparaison de spectres MS/MS un à un à des bases de données, ce qui conduit pour chaque spectre soit à l'identification correcte de la molécule, une identification fausse, ou une absence d'identification. Comme il a été démontré dans le *Chapitre III* (et évoqué par d'autres auteurs (van der Hooft et al. 2017; da Silva, Dorrestein, and Quinn 2015; Cantrell et al. 2019)), l'absence d'identification est le cas de figure le plus fréquent et il peut arriver qu'il y ait autant de faux-positifs que d'identifications correctes même en utilisant une base de données spécialisée. Les faux-positifs, à défaut de donner une identification correcte, informent généralement sur la classe structurale de la molécule. L'absence d'identification en revanche ne permet aucune interprétation directe par l'utilisateur, et comme la plupart des ions seront dans ce cas de figure, l'utilisateur se retrouve alors avec une grande majorité de signaux sur lesquels il ne sait rien. Le fait que la majorité des ions ne soient pas annotés peut s'expliquer deux facteurs : 1) la majorité des molécules demeurent à ce jour inconnues, et 2) seulement une petite partie des molécules connues présentent un spectre MS/MS disponible dans des bases de données en libre accès. Ceci donne lieu à deux cas de figure principaux pour ces ions sans annotation : ce sont soit des molécules connues mais sans spectre MS/MS dans les bases de données utilisées (*known unknowns*), soit ce sont de vraies molécules inconnues jamais décrites (*unknown unknowns*). Des annotations structurales pour ces molécules seraient d'une grande aide à l'utilisateur pour déterminer s'il a affaire à des *known unknowns* ou à des *unknown unknowns*. Dans ce contexte, les réseaux moléculaires permettent d'informer sur la portion non-identifiée des ions en les regroupant par similarité spectrale avec des ions connus. Bien que ceci permette de guider l'utilisateur dans l'interprétation des données, aucune information structurale ne lui est fournie sur les nœuds inconnus si ce n'est qu'ils peuvent avoir une structure proche avec un ion dérépliqué du même bloc. Il aura tout au plus le score de similarité cosinus entre chaque paire et pour chaque déréplication. Un autre outil de regroupement a été créé pour résoudre ce problème : MS2LDA. Il permet d'annoter de façon non-supervisée les nœuds d'un réseau avec des données structurales. Il se base sur un algorithme de *machine learning* utilisé dans le *text mining* (Latent Dirichlet Allocation), et permet ici d'extraire les motifs récurrents (mass2motifs) dans les spectres MS/MS, à l'instar des thèmes détectés dans les textes par *text mining* (**Figure 47**).

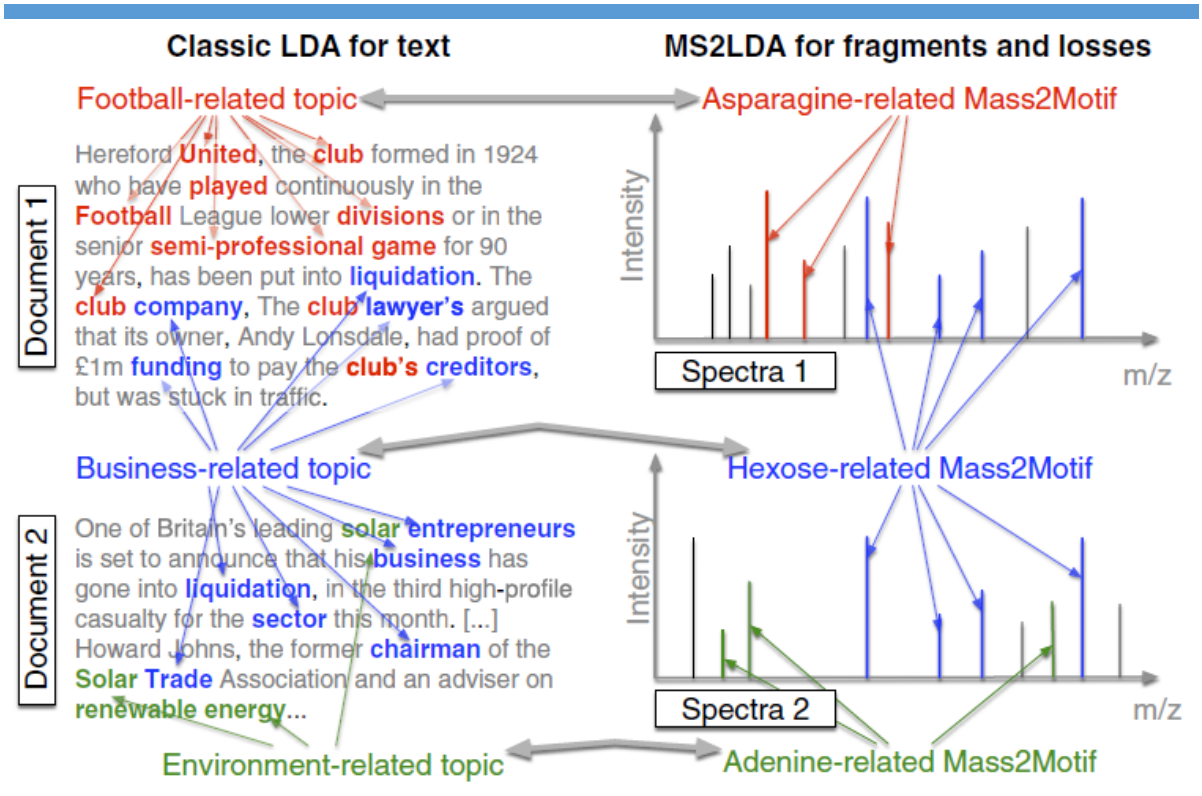


Figure 47 – Analogie entre LDA pour le texte et MS2LDA (extrait de J. J. J. van der Hooft et al., 2016).

MS2LDA est disponible sur http://ms2lda.org/ et permet à n'importe quel utilisateur de traiter ses données. Il est également implémenté sur le site du GNPS. Il va rechercher dans les spectres des motifs, appelés Mass2Motifs, qui représentent un ensemble de fragments et/ou pertes de neutres, assimilables à des marqueurs biochimiques ou structuraux. Dans les faits, ces motifs représentent les signaux de fragmentation d'une ou plusieurs sous-structures partagées par des molécules. Ceci permettra de regrouper les molécules déjà dérépliquées ou non à partir de sous-structures communes, parfois représentatives d'une voie de biosynthèse commune. A chaque fois qu'un motif est détecté dans un spectre MS/MS, deux scores y sont associés : le *probability score* qui représente la proportion du spectre MS/MS expliquée par le motif, et l'*overlap score* qui représente la proportion du motif présente dans le spectre. Dans un réseau moléculaire, il est devenu conventionnel de figurer les motifs comme des arêtes, représentant le motif commun reliant deux nœuds (**Figure 48**). Dans cet exemple, plusieurs blocs ont été formés par similarité cosinus et l'annotation MS2LDA permet visuellement d'observer l'homogénéité de ces blocs. Certains blocs seront totalement homogènes avec une seule couleur, indiquant qu'ils représentent des molécules très similaires, alors que d'autres blocs peuvent être divisés en sous-blocs en fonction des différents motifs qui relient ses ions. Ces motifs peuvent même relier indirectement des nœuds de deux blocs différents. L'un des plus grands avantages de cette méthode est l'annotation des nœuds inconnus : même les blocs avec des nœuds grisés (aucune molécule dérépliquée) présentent des motifs qui peuvent les relier à des molécules dérépliquées sans qu'elles n'aient à être dans le même bloc.

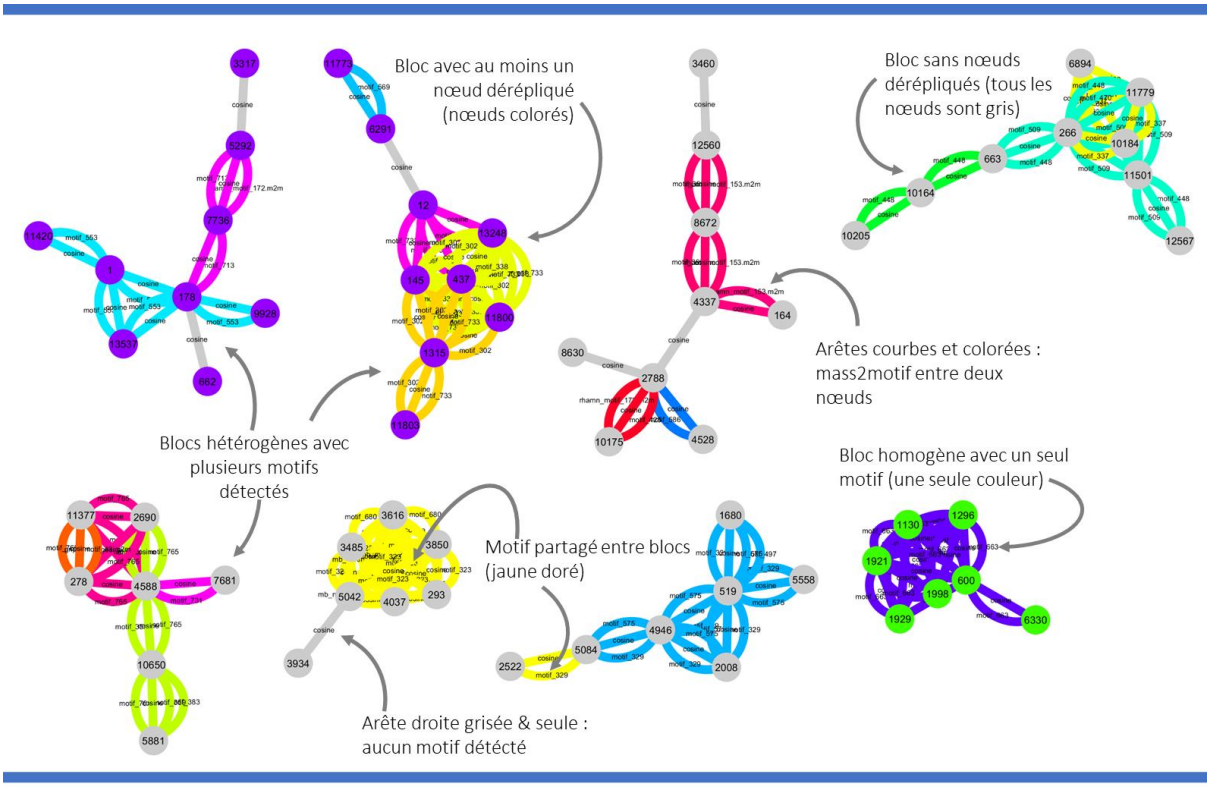


Figure 48 – Exemple d'annotation MS2LDA sur un réseau moléculaire.

Une approche proposée par les créateurs de MS2LDA est de créer des bases de données de motifs à partir des données des utilisateurs. Ainsi il est possible de se baser sur les motifs d'autres utilisateurs, les ayant déjà annotés avec quelques informations structurales. Il est par ailleurs possible de fournir à MS2LDA directement une base de données MS/MS plutôt qu'une analyse LC-MS, ce qui permet d'associer directement certains motifs à des structures connues et leurs sous-structures. Pour faciliter l'annotation de ces motifs avec des sous-structures, l'utilisation de MAGMa et de ClassyFire est également proposée par MS2LDA (Wandy et al. 2017; S. Rogers et al. 2019).

L'annotation de données LC-MS par MS2LDA est régulièrement utilisée depuis 2019 pour l'exploration de la diversité chimique d'organismes dans le domaine de la chimie des produits naturels. Ainsi, de nouveaux sménamides ont pu être identifiés dans l'éponge *Smenospongia aurea* par association de motifs entre des ions inconnus et des sménamides connus (Cantrell et al. 2019), des dérivés d'intérêt dans le genre *Euphorbia* ont été détectés en combinant MS2LDA à d'autres outils de métabolomiques (MolNetEnhancer) (Ernst, Kang, et al. 2019; Ernst, Nothias, et al. 2019; Nothias-Esposito et al. 2019). La diversité chimique d'espèces du genre *Alnus* a été étudiée (Kang et al. 2020) ainsi que celle de la famille des Rhamnacées, ce qui a abouti à la publication d'une base de données de motifs spécifique à cette famille (Kang et al. 2019). L'influence des marées noires sur le métabolisme des microorganismes sous-marins a également été étudiée à l'aide de MS2LDA pour détecter les molécules dérivées du pétrole (Moreno-Ulloa et al. 2019). MS2LDA a par ailleurs été utilisé dans le domaine de la santé pour identifier les biomarqueurs de maladies à l'aide des différences d'intensité observées pour les motifs entre des échantillons (Ernst et al. 2020; van der Hooft et al. 2017; McAvoy

et al. 2020; McLuskey et al. 2020). Les derniers développement indiquent un couplage avec les analyses de génomique et métagénomique pour étudier au mieux les organismes et leurs phénotypes (Mohanty et al. 2020; van der Hooft et al. 2020).

Dans le cadre de cette thèse, la LDB offre une opportunité unique pour la création d'une base de données de motifs à l'aide de ses 1870 spectres représentant de façon assez exhaustive l'ionisation de 250 molécules sous différentes formes et par différents instruments LC-MS.

Ces 250 molécules ne représentent que 15% des molécules recensées dans la LDB-Lit et bien moins en considérant l'étendue de l'inconnu dans les organismes lichéniques. L'annotation d'un réseau de molécules lichénique par déréplication sur la LDB reste donc limitée au mieux à ces 250 molécules. La LDB est en revanche représentative de la diversité structurelle connue dans les lichens, et devrait donc permettre de créer une base de données de motifs représentative de la diversité des molécules lichéniques. Par ailleurs, une nouvelle approche est proposée ici pour l'annotation structurale des nœuds par MS2LDA. L'approche classique consiste à extrapoler des données structurales entre un nœud déréplicué vers un nœud non-déréplicué sur la base d'un ou plusieurs motifs partagés. Cette approche peut être automatisée pour dépasser les limitations humaines, et ainsi annoter automatiquement la classe chimique des nœuds inconnus de façon supervisée grâce à des combinaisons de motifs.

La LDB étendue produite dans le *Chapitre III* précédente est utilisée pour créer la LDB motifDB : une liste de motifs MS/MS générées à l'aide de MS2LDA permettant de faire le lien entre des molécules présentant de similarités spectrales autrement que par similarité cosinus. Ceci permet de regrouper les molécules présentant des sous-structures similaires et dans certains cas, des molécules issues des mêmes voies de biosynthèse. Un outil est développé pour faciliter l'interprétation des données MS2LDA : *Classnotator*. Il se base sur la correspondance entre des classes structurales et des motifs ou combinaisons de motifs, permettant de produire des « *motifs purs* » spécifiques à ces classes. *Classnotator* va alors automatiser les extrapolations et annoter les nœuds avec les classes structurales les plus probables. Dans cet exemple, la classification des molécules lichéniques selon Hun&Yosh96 est utilisée. Puisque la LDB en donne la possibilité, le partage des motifs entre les spectres de différents adduits de la même molécule est étudié, ainsi que le partage des motifs entre spectres MS/MS acquis sur différents instruments.

2 Méthodes

2.1 Création de la LDB-motifDB sur ms2lda.org.

Les deux fichiers MGF contenant la LDB en mode négatif et en mode positif ont été modifiés pour rajouter quelques attributs à chaque spectre : le nom de la molécule associée à chaque spectre, l'instrument d'acquisition, la forme d'ionisation (molécule (dé)protonée, ou autres formes d'adduits) et la structure sous forme de SMILES, d'InChI et d'InChIKey. Les deux fichiers MGF modifiés ont ensuite été soumis à ms2lda.org. Les paramètres utilisés pour produire les motifs à partir de ces spectres sont les suivants : intensité minimale MS^2 : 100, *bin size* : 0.01, nombre de motifs produits (*free motifs*) : 100, nombre d'itérations : 1000. Les autres paramètres ne sont pas pris en compte lorsqu'un fichier MGF de spectres déjà identifiés est soumis et ont par conséquent été conservés à leurs valeurs par défaut.

2.2 Recherche des motifs représentatifs des classes chimiques.

Les motifs ou combinaisons de motifs permettant d'identifier des classes structurales ont été recherchés. Ils sont désignés par le nom de *motifs purs* et sont caractérisés par leur appartenance exclusive à une seule classe chimique. Un algorithme a été conçu sur Python 3.7 à cet effet (**Figure 49**). Dans un premier temps, les motifs de la LDB sont importés ainsi qu'un tableau d'attributs de la LDB. Le tableau de motifs contient l'intégralité des motifs détectés pour chaque molécule, alors que le tableau d'attributs contient les classes structurales de chacune selon Hun&Yosh96. Les deux sont associés pour créer un tableau *molmotif* qui contient pour chaque molécule les différents motifs détectés, la classe chimique et le nombre de motifs détectés *n_motifs* (**Tableau 11**)

Tableau 11 – Exemple d'un tableau molmotif.

Molécule	*Motifs*	*Classes*	*n_motifs*
Thuringione	motif_149|motif_61	Xanthones et bis-Xanthones	2
Acide planaïque	motif_124|motif_75|motif_89|motif_97	Depsides (Didepsides)	4
Acide nornotatique	motif_105|motif_108|motif_11|motif_121|motif_124|motif_131|motif_29|motif_48|motif_65|motif_96	Depsidones	10
Acide néphromopsique	motif_108|motif_124|motif_136|motif_20|motif_5|motif_55|motif_72|motif_78	Acides paraconiques	8
Acide glomelliférique	motif_123|motif_132|motif_31	Depsides (Didepsides)	3
Acide acaranoïque	motif_105|motif_106|motif_108|motif_124|motif_136|motif_21|motif_5|motif_89	Acides	8
3-déchloro-4-*O*-méthyldiploïcine	motif_111|motif_12|motif_148|motif_94	Depsidones	4
Roccanine	motif_83|motif_89	Composés azotés	2

Ce tableau *molmotif* servira de base pour la recherche de motifs : les *motifs purs* en seront progressivement extraits jusqu'à ce qu'il soit vide. Une variable *n_combi* est initialisée avec une valeur de 1 : elle représente le nombre de motifs consécutifs à combiner. Comme

sa première valeur est 1, les motifs seront testés individuellement : la liste de ces motifs uniques sera utilisée pour créer un tableau *Motifs x Classes* présentant les motifs uniques (lignes) et les classes selon Hun&Yosh96 (colonnes) (**Tableau 12**).

Tableau 12 – Exemple de tableau Motifs x Classes répertoriant le nombre de fois que chaque motif a été détecté dans chaque classe.

motif	Xanthones et bis-Xanthones	Depsides (Didepsides)	Depsidones	Acides paraconiques	Acides	Composés azotés	Dibenzofuranes	Diphényléthers	Depsides (Tridepsides)	Chromanes et Chromones	Fragments de Depsides et Depsidones	Dérivés de l' acide pulvinique	Quinones	Polyols, Monosaccharides et Carbohydrates	Depsones	Benzyldepsides	Mycosporines	Composés Aliphatiques et Cycloaliphatiques	Naphthopyranes	Terpenoides : Triterpenes	Nombre de molécules	Max n	Pourcentage maximal
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	100
2	0	1	16	9	2	0	2	0	0	0	3	0	2	0	0	0	0	0	0	0	35	16	45
3	0	5	1	0	0	0	0	0	0	3	2	0	0	1	0	0	0	0	0	0	12	5	41
4	0	7	0	0	0	0	1	0	4	0	2	0	0	0	0	0	0	0	0	0	14	7	50
5	0	6	0	0	0	0	0	0	3	0	4	0	0	0	0	0	0	0	0	0	13	6	46
6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	100
7	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	100
8	2	1	18	16	4	0	5	2	0	2	13	3	12	0	1	0	0	0	1	0	80	18	22
9	0	0	0	0	0	0	1	0	0	0	0	8	1	0	0	0	0	0	0	0	10	8	80

A l'aide du tableau *molmotif* qui contient tous les motifs de chaque molécule et leur classe chimique, le nombre d'occurrences motif-classe sont comptés et rapportés dans le tableau *Motifs x Classes*. A partir de ces données, les *motifs purs* sont sélectionnés en ne conservant que les motifs observés exclusivement dans une classe chimique. Ces *motifs purs* sont conservés dans un tableau de *motifs purs* et ils sont éliminés du tableau *molmotif* pour la prochaine itération, leur présence étant déjà synonyme de l'appartenance à une classe chimique. La variable *n_combi* est incrémentée de 1 et les molécules présentant moins de motifs que la valeur actuelle de *n_combi* sont éliminées du tableau *molmotif*. Dans cet exemple, *n_combi* est à présent égal à 2, et la combinaison de motifs uniques va cette fois créer toutes les combinaisons de 2 motifs uniques observables dans chaque molécule (ex : *motif_10|motif_18*). Le tableau *Motifs x Classes* contiendra cette fois pour chaque ligne une combinaison de deux motifs, et les occurrences de ces combinaisons seront comptées dans le tableau *molmotif*. La boucle continue ainsi jusqu'à ce que toutes les molécules aient été éliminées du tableau *molmotif*. Le tableau de *motifs purs* ainsi complété est exporté au format CSV et pourra servir à l'identification des classes structurales d'un jeu de données LC-MS annoté avec des motifs de la LDB.

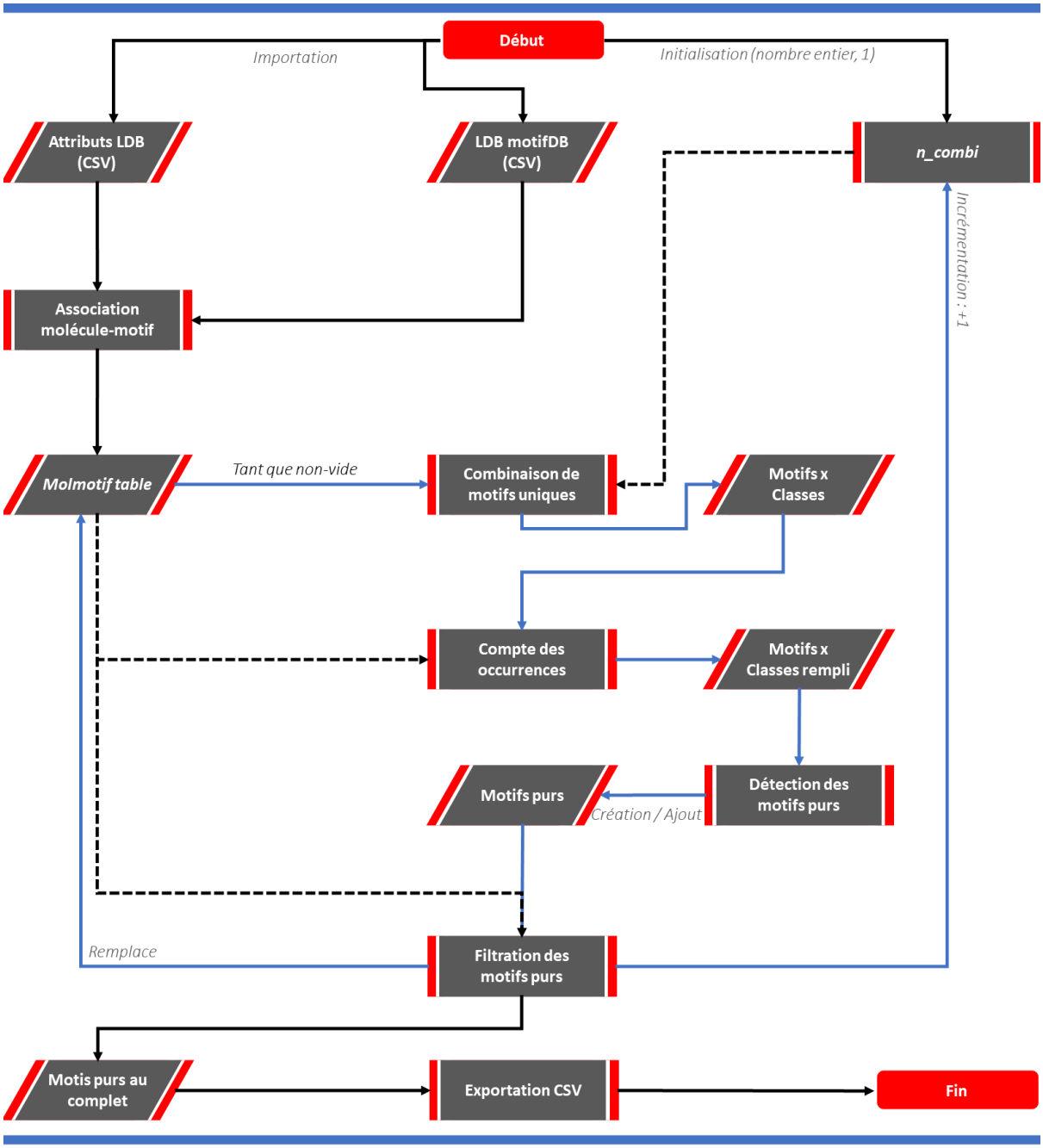


Figure 49 – Obtention des motifs purs à partir des motifs de la LDB.

2.3 Impact des adduits et des instruments d'analyse sur les motifs.

Etant donné que les spectres soumis à ms2lda.org comprenaient différents adduits pour une même molécule ainsi que des spectres produits par différents instruments LC-MS, l'impact de ces facteurs sur la production des motifs a été évalué. Les différents adduits issus d'un même instrument pour une même molécule ont été pris en compte dans un premier temps : il a été vérifié si les motifs présents dans un adduit se retrouvaient dans les autres adduits de la même molécule (**Figure 50-A**). L'impact des instruments a ensuite été étudié en considérant les spectres de mêmes adduits acquis sur différents appareils LC-MS et en observant si les motifs détectés pour un spectre sur un instrument peuvent être retrouvés sur un autre pour le même adduit (**Figure 50-B**). Par la même

occasion, le nombre de paires d'adduits de la même molécule partageant au moins un motif a été compté.

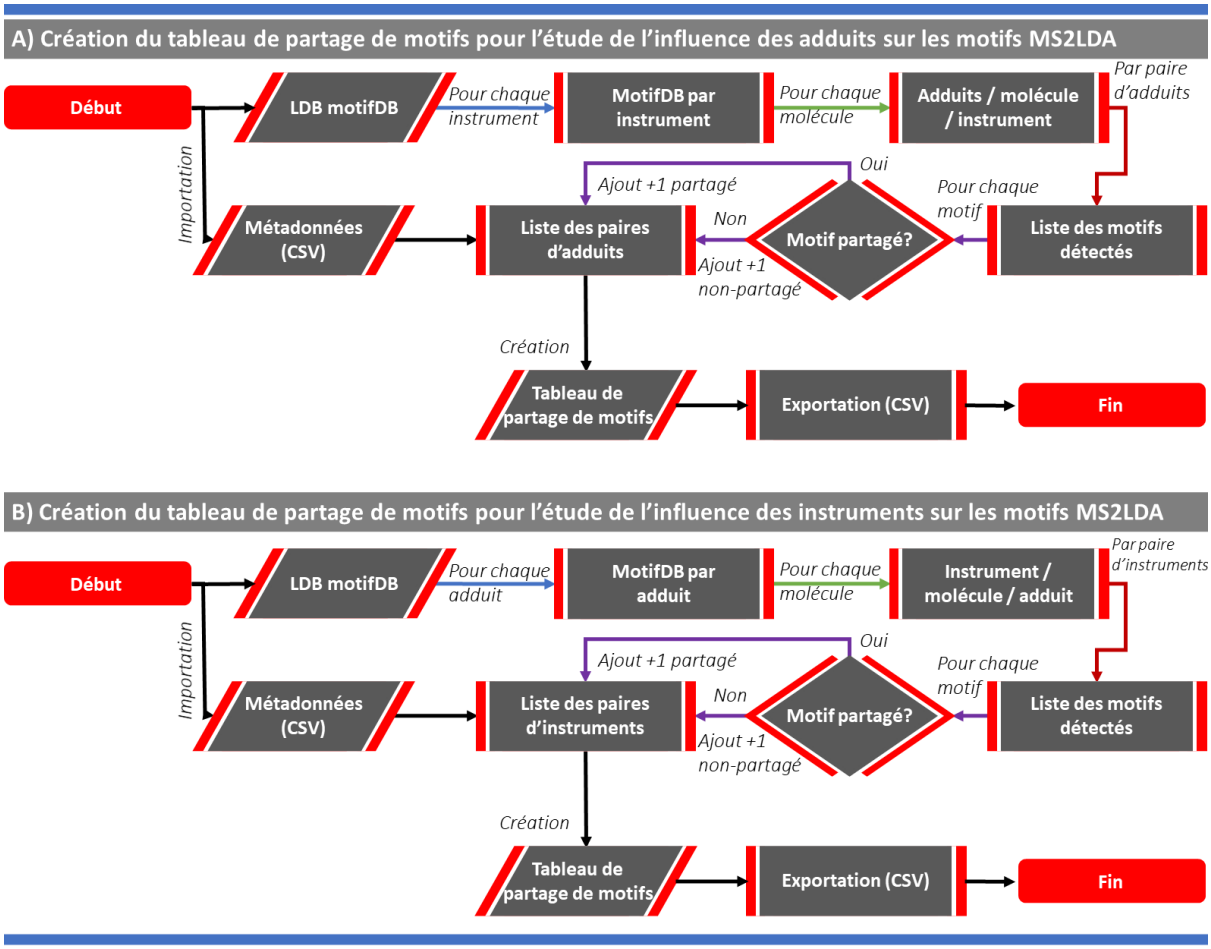


Figure 50 – Algorithmes utilisés pour évaluer le partage des motifs entre adduits (A) et entre instruments (B).

3 Résultats

3.1 Création de la LDB-motifDB sur ms2lda.org.

Les bases de données de motifs pour la LDB ont été créées et rendues publiques sur ms2lda.org pour le mode négatif (http://ms2lda.org/motifdb/motif_set/33/) et le mode positif (http://ms2lda.org/motifdb/motif_set/37/). L'intégralité des paramètres utilisés peut être consultée sur http://ms2lda.org/basicviz/summary/1282/ et http://ms2lda.org/basicviz/summary/1281/ pour le mode négatif et positif respectivement. Le nom des molécules de la LDB y est également consultable avec pour chaque molécule ses motifs, sa structure ainsi que les annotations MAGMa fournies par S. Rogers, J. Wandy et J. van der Hooft. Les annotations MAGMa permettent d'interpréter les fragments et les pertes de neutres grâce aux structures fournies par la LDB. Ces annotations permettent d'annoter efficacement les pertes simples (-H_2O, -CO_2, -CO…) mais l'identification des fragmentations complexes est plus difficile.

3.2 Recherche des motifs représentatifs des classes chimiques.

Classnotator a produit 842 *motifs purs* en mode positif et 704 en mode négatif. Parmi les 21 classes, le mode positif couvre 20 classes, ne trouvant aucun *motif pur* pour les benzyldepsides. Le mode négatif couvre 16 des 21 classes, aucun *motif pur* n'ayant été trouvé pour les benzyldepsides, les composés aliphatiques, les naphthopyranes, les diterpènes et les triterpènes (**Tableau 13**).

Tableau 13 – Motifs purs retrouvés pour chaque classe chimique en fonction du mode d'ionisation. Le nombre de molécules représentées par la LDB pour chaque classe en fonction du mode d'ionisation ainsi que le pourcentage pouvant être identifié avec des motifs purs ont également été rapportés.

	Mode négatif			Mode positif		
Classes Hun&Yosh96	*Molécules*	*Pourcentage*	*Motifs purs*	*Molécules*	*Pourcentage*	*Motifs purs*
Acides	7	78	52	7	86	20
Acides paraconiques	16	100	92	15	100	86
Benzyldepsides	2	0	0	0	0	0
Chromanes et Chromones	11	46	14	14	86	51
Composés Aliphatiques et Cycloaliphatiques	4	0	0	3	100	12
Composés azotés	5	80	23	4	100	18
Depsides (Didepsides)	37	57	36	14	86	39
Depsides (Tridepsides)	4	25	1	1	100	1
Depsidones	36	92	188	28	96	201
Depsones	1	100	10	1	100	1
Dérivés de l'acide pulvinique	11	82	36	11	100	63
Dibenzofuranes	10	80	32	5	80	8
Diphényléthers	3	67	7	2	100	23
Fragments de Depsides et Depsidones	22	91	110	19	74	17
Mycosporines	3	67	11	3	67	5
Naphthopyranes	1	0	0	1	100	5
Polyols, Monosaccharides et Carbohydrates	4	75	6	1	100	1
Quinones	28	71	67	25	84	158
Terpenoides : Diterpenes	0	0	0	2	100	3
Terpenoides : Triterpenes	1	0	0	20	85	47
Xanthones et bis-Xanthones	21	71	19	30	80	83

Chaque *motif pur* est associé à un nombre de molécules *n_mol* dans lesquelles il a été retrouvé et à un nombre *n_combi* de motifs consécutifs combinés pour l'obtenir (**Tableau 14, Tableau 15**).

Tableau 14 – Répartition des motifs purs du mode négatif en fonction du nombre de combinaisons successives et du nombre de molécules dans lesquelles elles ont été décrites.

		Nombre de molécules (*n_mol*)						
		1	**2**	**3**	**4**	**5**	**6**	**7**
n_combi	**1**	3	1	3	1	0	0	0
	2	579	72	13	6	2	1	1
	3	20	3	0	0	0	0	0

Tableau 15 – Répartition des motifs purs du mode positif en fonction du nombre de combinaisons successives et du nombre de molécules dans lesquelles elles ont été décrites.

		Nombre de moléculés (*n_mol*)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n_combi	**1**	1	3	2	2	1	1	0	0	1	0
	2	669	62	25	22	2	3	1	0	0	1
	3	44	2	0	0	0	0	0	0	0	0

Un *motif pur* détecté dans de nombreuses molécules est plus représentatif de sa classe structurale qu'un *motif pur* repéré qu'une seule fois (*n_mol* = 1). La grande majorité des *motifs purs* se trouvent dans la catégorie *n_combi* = 2, *n_mol* = 1 (82% des cas en mode négatif, 79% en mode positif). Bien qu'ils soient d'un apport moindre, leur spécificité n'est pas négligeable puis qu'ils n'ont été détectés que dans une seule molécule dans les 1870 spectres de la LDB. Les *motifs purs* peuvent être utilisés pour identifier la classe chimique de composés lichéniques inconnus. Si plusieurs motifs purs sont détectés pour une même molécule, une pondération est effectuée en cumulant les valeurs de *n_mol* pour les différentes classes. Les *motifs purs* sont associés à leurs classes dans le **Tableau 16**. Au vu de la quantité d'information, ceux avec un *n_combi* de 2 et un *n_mol* de 1 ont été omis.

Tableau 16 *– Classes structurales et leurs motifs pur. Ceux pour lesquels n_combi = 2 et n_mol = 1 ont été omis. Les noms de motifs ont été simplifiés en éliminant leur préfixe, « LDB_NEG_motif_ » ou « LDB_POS_motif_ » selon le mode d'ionisation. Dans le cas des combinaisons de motifs, les motifs constituants sont séparés par un « | ». Est représenté entre parenthèses après chaque motif, son n_mol.*

Classes	*Mode négatif*	*Mode positif*
Acides	93 *(3)*, 13|15 *(2)*, 39|78 *(2)*, 61|65|78 *(1)*, 61|65|86 *(1)*, 61|78|86 *(1)*, 65|78|86 *(1)*	43 *(3)*, 49|79 *(2)*, 49|53|67 *(1)*, 49|53|83 *(1)*, 49|53|85 *(1)*, 49|67|83 *(1)*, 49|67|85 *(1)*, 49|83|85 *(1)*, 53|67|83 *(1)*, 53|67|85 *(1)*, 53|83|85 *(1)*
Acides paraconiques	10|18 *(2)*, 10|60 *(2)*, 10|78 *(2)*, 11|14 *(2)*, 11|6 *(2)*, 11|94 *(2)*, 12|33 *(2)*, 12|45 *(2)*, 12|94 *(2)*, 14|45 *(2)*, 14|49 *(2)*, 14|6 *(2)*, 14|60 *(2)*, 15|53 *(2)*, 18|32 *(2)*, 18|45 *(2)*, 18|53 *(2)*, 32|45 *(2)*, 32|53 *(2)*, 32|60 *(2)*, 32|90 *(2)*, 33|60 *(2)*, 45|78 *(2)*, 49|6 *(2)*, 49|94 *(2)*, 53|60 *(2)*, 53|94 *(2)*, 78|90 *(2)*, 0|15 *(3)*, 11|49 *(3)*, 14|53 *(3)*, 32|78 *(3)*, 45|60 *(3)*, 49|86 *(3)*, 53|78 *(3)*, 53|86 *(3)*, 0|10 *(4)*, 0|14 *(4)*, 0|86 *(4)*, 33|45 *(4)*, 45|94 *(6)*, 14|15 *(7)*, 60|78|86 *(1)*, 60|78|94 *(1)*, 60|86|90 *(1)*, 60|90|94 *(1)*, 65|86|94 *(1)*, 60|86|94 *(2)*, 78|86|94 *(2)*, 86|90|94 *(2)*	0|31 *(2)*, 0|55 *(2)*, 0|85 *(2)*, 15|55 *(2)*, 55|83 *(2)*, 55|98 *(2)*, 67|93 *(2)*, 83|93 *(2)*, 85|98 *(2)*, 88|98 *(2)*, 15|85 *(3)*, 31|98 *(3)*, 49|98 *(3)*, 55|67 *(3)*, 55|88 *(3)*, 88|90 *(3)*, 88|92 *(3)*, 49|90 *(4)*, 49|92 *(4)*, 49|93 *(4)*, 53|98 *(4)*, 55|90 *(4)*, 55|92 *(4)*, 55|93 *(4)*, 85|90 *(4)*, 85|92 *(4)*, 85|93 *(4)*, 90|92 *(4)*, 49|88 *(5)*, 31|49 *(6)*, 31|85 *(6)*, 49|55 *(10)*
Chromanes et Chromones	30|90 *(2)*, 38|59 *(2)*, 38|65 *(2)*, 59|61 *(2)*	23|9 *(2)*, 65|96 *(2)*
Composés azotés	0|61 *(2)*	
Depsides (Didepsides)	23|84 *(4)*, 37|84 *(2)*, 45|52 *(2)*, 45|55 *(2)*, 45|79 *(2)*, 52|55|79 *(1)*, 55|7|79 *(1)*	46|6 *(2)*, 11|7 *(3)*
Depsidones	96 *(3)*, 1 *(1)*, 95 *(4)*, 15|17 *(2)*, 15|25 *(2)*, 15|33 *(3)*, 15|37 *(2)*, 15|56 *(2)*, 15|80 *(2)*, 15|83 *(4)*, 17|94 *(2)*, 25|86 *(3)*, 33|37 *(2)*, 33|83 *(2)*, 37|83 *(2)*, 37|94 *(2)*, 46|94 *(2)*, 56|83 *(2)*, 56|94 *(2)*, 66|86 *(2)*, 70|86 *(2)*, 83|94 *(5)*, 33|86|94 *(1)*	66 *(1)*, 20 *(2)*, 41 *(4)*, 32 *(5)*, 64 *(6)*, 16|17 *(2)*, 1|33 *(2)*, 21|65 *(2)*, 21|75 *(2)*, 21|86 *(2)*, 31|33 *(2)*, 31|70 *(2)*, 33|48 *(2)*, 33|65 *(2)*, 44|65 *(2)*, 44|75 *(2)*, 44|86 *(2)*, 70|84 *(2)*, 75|79 *(2)*, 1|16 *(3)*, 1|21 *(3)*, 1|44 *(3)*, 1|75 *(3)*, 21|55 *(3)*, 33|44 *(3)*, 33|70 *(3)*, 40|79 *(3)*, 44|55 *(3)*, 48|79 *(3)*, 21|31 *(4)*, 21|6 *(4)*, 31|44 *(4)*, 33|53 *(4)*, 33|6 *(4)*, 33|79 *(4)*, 44|6 *(4)*, 6|70 *(4)*, 6|79 *(4)*, 79|84 *(4)*, 16|21|95 *(1)*, 17|21|44 *(1)*, 17|21|48 *(1)*, 17|44|48 *(1)*, 1|53|65 *(1)*, 21|44|48 *(1)*, 44|48|53 *(1)*, 53|59|65 *(1)*, 53|59|79 *(1)*, 53|59|95 *(1)*, 53|65|95 *(1)*, 53|79|95 *(1)*, 59|65|95 *(1)*, 59|79|95 *(1)*, 53|65|79 *(2)*
Dérivés de l'acide pulvinique	32|51 *(2)*, 36|51 *(2)*, 51|65 *(2)*, 51|54 *(5)*	91 *(2)*, 74 *(3)*, 31|35 *(2)*, 31|47 *(2)*, 35|50 *(2)*, 47|50 *(2)*, 50|65 *(2)*, 58|79 *(2)*, 31|50 *(6)*, 56|70|79 *(1)*
Dibenzofuranes	85 *(1)*, 46|52 *(2)*, 4|46 *(2)*, 4|52 *(2)*, 58|80|86 *(1)*, 58|80|94 *(1)*, 58|86|90 *(1)*, 58|86|94 *(1)*	81 *(2)*, 39|53 *(2)*, 39|79 *(2)*, 53|55|65 *(1)*, 53|55|82 *(1)*
Diphényléthers		53|55|84 *(1)*
Fragments de Depsides et Depsidones	64|90 *(2)*, 79|90 *(2)*, 86|99 *(3)*, 87|99 *(2)*, 90|99 *(3)*, 19|86|90 *(1)*, 34|86|90 *(1)*, 69|86|90 *(1)*, 7|86|90 *(1)*	4|53 *(5)*, 37|59|65 *(1)*, 53|6|82 *(1)*
Naphthopyranes		21|37|58 *(1)*
Polyols, Monosaccharides et Carbohydrates	40|61 *(2)*, 40|74 *(2)*, 40|88 *(2)*, 61|88 *(2)*, 74|88 *(2)*	
Quinones	92 *(1)*, 55|58 *(2)*, 55|72 *(2)*, 55|97 *(2)*, 55|80 *(3)*	13|16 *(2)*, 13|70 *(2)*, 16|37 *(2)*, 30|5 *(2)*, 37|70 *(2)*, 37|8 *(2)*, 37|95 *(2)*, 38|53 *(2)*, 5|65 *(2)*, 65|89 *(2)*, 13|5 *(3)*, 13|95 *(3)*, 5|95 *(3)*, 13|30|53 *(1)*, 13|30|65 *(1)*, 13|53|65 *(1)*, 30|53|73 *(1)*, 55|59|65 *(1)*, 55|59|79 *(1)*, 55|59|82 *(1)*, 55|65|79 *(1)*, 55|79|82 *(1)*, 59|65|82 *(1)*, 59|79|82 *(1)*, 65|79|82 *(1)*, 30|53|65 *(2)*
Terpénoïdes : Triterpenes		68 *(9)*, 17|3 *(2)*, 21|88 *(2)*, 30|87 *(2)*, 48|88 *(2)*, 6|88 *(2)*, 3|88 *(3)*, 44|88 *(3)*, 53|87 *(3)*, 87|88 *(3)*, 17|88 *(4)*
Xanthones et bis-Xanthones	8 *(2)*, 41|47 *(2)*	19 *(4)*, 13|25 *(2)*, 13|80 *(2)*, 13|97 *(2)*, 25|29 *(2)*, 25|51 *(2)*, 25|65 *(2)*, 29|51 *(2)*, 29|65 *(2)*, 2|25 *(2)*, 2|80 *(2)*, 2|97 *(2)*, 13|2 *(7)*, 44|53|90 *(1)*, 53|65|94 *(1)*

Le nombre de *motifs purs* détectés dépend de la quantité de spectres & molécules disponibles pour chaque classe et de l'abondance des fragments dans les spectres. Ainsi, certaines classes avec peu de représentants n'ont pas de *motifs purs* associés. D'autres,

bien que peu représentées, possèdent des spectres particuliers et produisent plusieurs *motifs purs*. La couverture de la LDB par les *motifs purs* a été représentée graphiquement en **Figure 51**, en prenant en compte non-seulement les molécules détectées pour chaque mode mais également l'ensemble des molécules de chaque classe présentes dans la LDB.

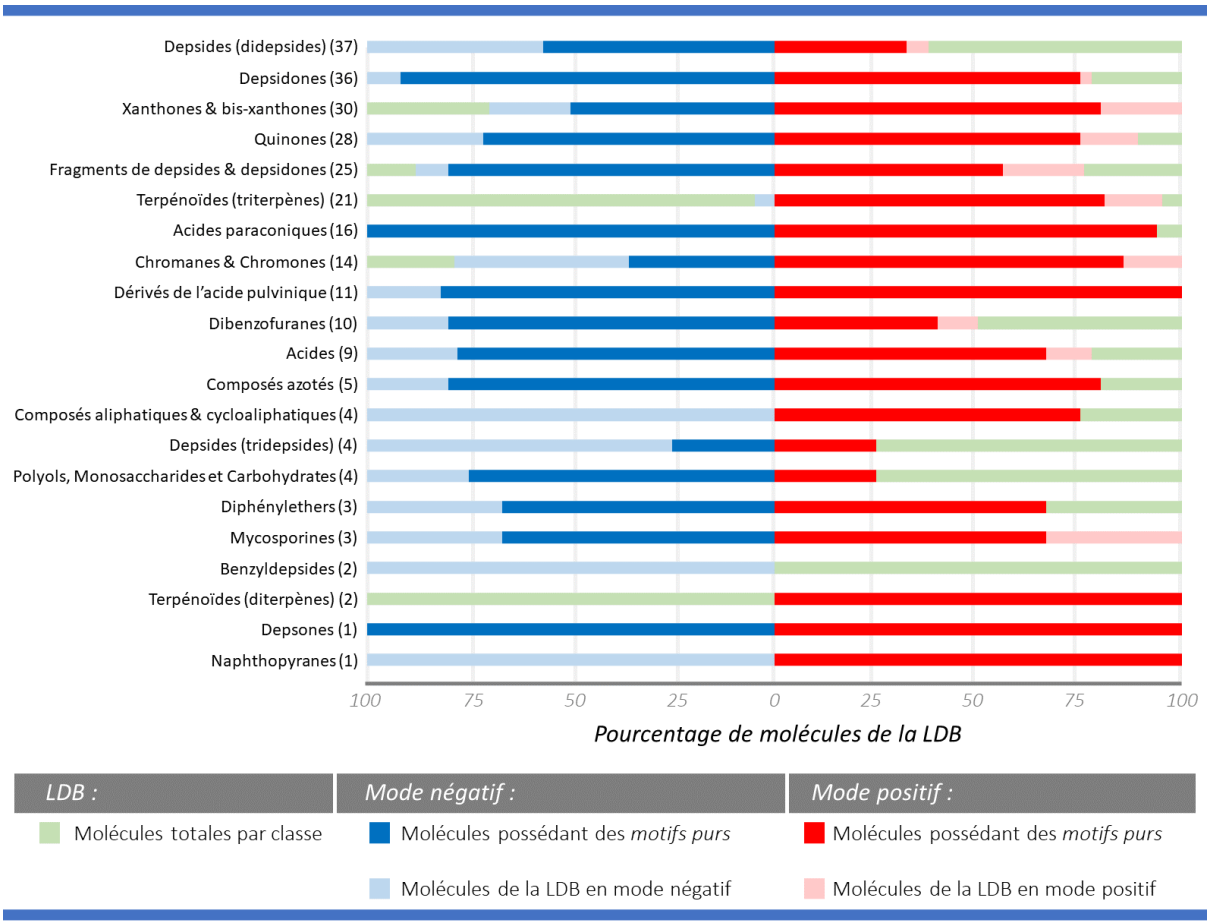


Figure 51 – Couverture de la LDB par les motifs purs dans les deux modes d'ionisation.

Globalement, les classes avec assez de représentants sont facilement identifiables par leurs *motifs purs* dans les deux modes, à l'exception des terpènes et des chromones. Les terpènes n'ont pratiquement pas été détectés en mode négatif et les chromones qui, bien que détectées, ne produisent que peu de *motifs purs*, ce qui favorise le mode positif pour ces deux classes

Ce genre d'approche se prête bien à l'accumulation des données pour une même molécule : si tous ses adduits peuvent être reliés, le nombre de *motifs purs* augmentera ainsi que la solidité de l'identification. Une identification fragile dans un mode d'ionisation pourra être renforcée par l'identification dans l'autre mode d'ionisation si un couplage positif-négatif est réalisé. Un exemple d'application de cette méthode est démontré dans le *Chapitre V*.

3.3 Impact des adduits et des instruments d'analyse sur les motifs.

La présence de chaque motif détecté pour une molécule a été vérifiée parmi tous ses spectres d'adduits acquis sur un même instrument en mode négatif (**Tableau 17**) et positif (**Tableau 18**). Un score a été calculé, représentant le nombre de motifs retrouvés

divisé par le nombre de motifs total considérés pour chaque paire d'adduits (« Total » dans le tableau). Les combinaisons d'adduits pour lesquelles aucune occurrence n'a été trouvée n'ont pas été représentés (Total = 0).

Tableau 17 – Motifs retrouvés entre adduits en mode négatif.

Adduit 1	Adduit 2	Score	Total	Adduit 1	Adduit 2	Score	Total
$[M+Cl]^-$	$[2M-H]^-$	1	7	$[M-H-H_2O]^-$	$[2M-2H+Na]^-$	0.5	4
$[M-H+HCOOH]^-$	$[2M-H]^-$	1	4	$[M-H-H_2O]^-$	$[M-2H+Na]^-$	0.5	2
$[M+Cl]^-$	$[2M-2H+Na]^-$	1	2	$[M-H-H_2O]^-$	$[M+Cl]^-$	0.5	2
$[2M-2H+K]^-$	$[2M-H]^-$	1	2	$[2M-2H+K]^-$	$[M-2H+Na]^-$	0.5	2
$[M+Cl]^-$	$[M-2H+Na]^-$	1	1	$[2M-H+HCOOH]^-$	$[M-H]^-$	0.5	2
$[M-2H+K]^-$	$[M+Cl]^-$	1	1	$[2M-H]^-$	$[M-H]^-$	0.48	153
$[2M-H+HCOOH]^-$	$[M+Cl]^-$	1	1	$[M-H+CH_3OH]^-$	$[2M-H]^-$	0.45	29
$[2M-2H+Na]^-$	$[2M-H]^-$	0.97	66	$[M+Cl]^-$	$[M-H]^-$	0.44	18
$[2M-2H+K]^-$	$[2M-2H+Na]^-$	0.83	6	$[2M-2H+Na]^-$	$[M-H+CH_3OH]^-$	0.39	33
$[2M-2H+Na]^-$	$[M-H+HCOOH]^-$	0.75	4	$[M-H-H_2O]^-$	$[M-H]^-$	0.38	21
$[M-2H+Na]^-$	$[2M-2H+Na]^-$	0.67	9	$[M-H+HCOOH]^-$	$[M-H]^-$	0.35	52
$[2M-H+HCOOH]^-$	$[M-H+HCOOH]^-$	0.67	3	$[M+Cl]^-$	$[M-H+CH_3OH]^-$	0.33	3
$[M-2H+Na]^-$	$[2M-H]^-$	0.6	5	$[M-H-H_2O]^-$	$[M-H+HCOOH]^-$	0.25	4
$[M-2H+Na]^-$	$[M-H]^-$	0.56	18	$[M-H+CH_3OH]^-$	$[M-H]^-$	0.23	95
$[2M-2H+Na]^-$	$[M-H]^-$	0.54	163	$[M-H+HCOOH]^-$	$[M-H+CH_3OH]^-$	0.13	15
$[M+Cl]^-$	$[M-H+HCOOH]^-$	0.5	8	$[M-2H+K]^-$	$[M-H]^-$	0	1
$[2M-2H+K]^-$	$[M-H]^-$	0.5	6	$[M-2H+K]^-$	$[M-H+CH_3OH]^-$	0	1
$[M-H-H_2O]^-$	$[2M-H]^-$	0.5	4				

Tableau 18 – Motifs retrouvés entre adduits en mode positif.

Adduit 1	Adduit 2	Score	Total	Adduit 1	Adduit 2	Score	Total
$[2M+K]^+$	$[M+Na]^+$	1	1	$[M+2Na-H]^+$	$[M+Na]^+$	0.2	5
$[2M+K]^+$	$[M+H]^+$	1	1	$[M+K]^+$	$[M+NH_4]^+$	0.17	12
$[2M+K]^+$	$[M+K]^+$	1	1	$[M+K]^+$	$[M+H]^+$	0.15	20
$[2M+K]^+$	$[2M+H+CH_3CN]^+$	1	1	$[M+H+CH_3OH]^+$	$[M+Na]^+$	0.14	21
$[2M+H+CH_3OH]^+$	$[2M+Na+CH_3CN]^+$	1	1	$[2M+H]^+$	$[M+Na]^+$	0.14	14
$[2M+H+HCOOH]^+$	$[2M+Na+CH_3CN]^+$	1	1	$[M+Na+CH_3CN]^+$	$[2M+NH_4]^+$	0.14	14
$[2M+H+HCOOH]^+$	$[2M+H+CH_3OH]^+$	1	1	$[M+H+HCOOH]^+$	$[M+Na]^+$	0.14	7
$[M+H+C_2H_6OS]^+$	$[2M+H]^+$	1	1	$[M+Na+CH_3CN]^+$	$[M+H+CH_3OH]^+$	0.14	7
$[M+H+C_2H_6OS]^+$	$[M+NH_4]^+$	1	1	$[M+NH_4]^+$	$[M+Na]^+$	0.11	56
$[M+H+C_2H_6OS]^+$	$[M+H]^+$	1	1	$[2M+NH_4]^+$	$[2M+Na]^+$	0.09	33
$[M+H+C_2H_6OS]^+$	$[2M+NH_4]^+$	1	1	$[M+Na+CH_3CN]^+$	$[M+NH_4]^+$	0.08	12
$[2M+NH_4]^+$	$[2M+H]^+$	0.88	16	$[M+Na+CH_3CN]^+$	$[M+H]^+$	0.06	17
$[2M+NH_4]^+$	$[M+NH_4]^+$	0.75	52	$[2M+Na]^+$	$[2M+H]^+$	0.06	16
$[2M+NH_4]^+$	$[M+H+CH_3OH]^+$	0.71	14	$[2M+Na]^+$	$[M+H]^+$	0.05	75
$[2M+NH_4]^+$	$[M+H]^+$	0.7	63	$[2M+Na]^+$	$[M+NH_4]^+$	0.05	37
$[M+NH_4]^+$	$[2M+H]^+$	0.67	21	$[M+H+CH_3OH]^+$	$[2M+Na]^+$	0	20
$[M+H]^+$	$[2M+H]^+$	0.61	31	$[M+H+HCOOH]^+$	$[2M+Na]^+$	0	6
$[M+Na+CH_3CN]^+$	$[2M+Na]^+$	0.56	18	$[M+H+CH_3CN]^+$	$[2M+Na]^+$	0	6
$[M+K]^+$	$[2M+H]^+$	0.5	2	$[2M+Na+CH_3OH]^+$	$[M+H]^+$	0	6
$[2M+K]^+$	$[M+NH_4]^+$	0.5	2	$[M+K]^+$	$[M+H+HCOOH]^+$	0	4
$[2M+K]^+$	$[2M+Na]^+$	0.5	2	$[2M+Na+CH_3CN]^+$	$[M+H]^+$	0	4
$[2M+Na]^+$	$[M+Na]^+$	0.49	47	$[M+H+CH_3CN]^+$	$[M+Na]^+$	0	3
$[M+H]^+$	$[M+NH_4]^+$	0.45	165	$[M+2Na-H]^+$	$[M+H+HCOOH]^+$	0	3
$[M+H+CH_3CN]^+$	$[M+NH_4]^+$	0.41	22	$[M+2Na-H]^+$	$[M+H+CH_3CN]^+$	0	3
$[M+H+CH_3OH]^+$	$[2M+H]^+$	0.4	10	$[M+2Na-H]^+$	$[M+K]^+$	0	3
$[M+H+CH_3CN]^+$	$[M+H+CH_3OH]^+$	0.38	8	$[2M+H+CH_3OH]^+$	$[M+H]^+$	0	3
$[M+H+CH_3OH]^+$	$[M+H]^+$	0.36	113	$[M+Na+CH_3CN]^+$	$[2M+H]^+$	0	2
$[M+H+CH_3OH]^+$	$[M+NH_4]^+$	0.35	54	$[M+K]^+$	$[M+H+CH_3OH]^+$	0	2
$[M+K]^+$	$[2M+NH_4]^+$	0.33	6	$[2M+H+CH_3CN]^+$	$[M+H+HCOOH]^+$	0	2
$[M+2Na-H]^+$	$[M+NH_4]^+$	0.33	3	$[2M+Na+CH_3OH]^+$	$[M+K]^+$	0	2
$[2M+K]^+$	$[2M+NH_4]^+$	0.33	3	$[M+2Na-H]^+$	$[M+H+CH_3OH]^+$	0	2
$[M+H+CH_3CN]^+$	$[M+H]^+$	0.32	38	$[M+2Na-H]^+$	$[2M+Na+CH_3OH]^+$	0	2
$[M+Na+CH_3CN]^+$	$[M+Na]^+$	0.32	22	$[M+H+CH_3CN]^+$	$[M+H+HCOOH]^+$	0	1
$[M+K]^+$	$[M+Na]^+$	0.27	22	$[M+K]^+$	$[M+Na+CH_3CN]^+$	0	1
$[M+H+HCOOH]^+$	$[M+H+CH_3OH]^+$	0.25	16	$[M+K]^+$	$[M+H+CH_3CN]^+$	0	1
$[M+K]^+$	$[2M+Na]^+$	0.25	8	$[2M+H+CH_3CN]^+$	$[M+H]^+$	0	1
$[M+H+HCOOH]^+$	$[M+H]^+$	0.24	25	$[2M+Na+CH_3OH]^+$	$[M+NH_4]^+$	0	1
$[M+2Na-H]^+$	$[M+H]^+$	0.23	13	$[2M+Na+CH_3OH]^+$	$[M+H+HCOOH]^+$	0	1
$[M+H+HCOOH]^+$	$[M+NH_4]^+$	0.22	9	$[2M+Na+CH_3CN]^+$	$[2M+Na+CH_3OH]^+$	0	1
$[2M+NH_4]^+$	$[M+Na]^+$	0.21	33	$[M+2Na-H]^+$	$[2M+Na]^+$	0	1
$[M+H]^+$	$[M+Na]^+$	0.2	92	$[2M+H+HCOOH]^+$	$[M+H]^+$	0	1
$[M+H+HCOOH]^+$	$[2M+H]^+$	0.2	5	$[M+H+C_2H_6OS]^+$	$[M+Na]^+$	0	1

En dehors de quelques exceptions, les spectres MS/MS d'une même molécule sont assez différents pour ne pas partager systématiquement tous leurs motifs. Ils ne peuvent pas se résumer à une différence de rapport *m/z* pour l'ion précurseur dans le spectre MS/MS. Quelques paires restent en revanche très similaires : en mode négatif, $[2M-2H+Na]^-$ & $[2M-H]^-$ partagent les mêmes motifs dans 97% des cas (66 spectres). En mode positif d'autres paires sont remarquables, notamment $[2M+NH_4]^+$ combiné avec $[2M-H]^+$ (88%), avec $[M+NH_4]^+$ (75%), avec $[M+H+CH_3OH]^+$ (71%) et avec $[M+H]^+$ (70%), respectivement pour 16, 52, 14 et 63 spectres comparés.

Le lien entre plusieurs adduits d'une même molécule par au moins un motif a également été vérifié. En mode négatif, sur un total de 912 paires analysées, 74% (678) des adduits d'une même molécule pouvaient être reliés par au moins un adduit. Ce pourcentage diminue drastiquement en mode positif, avec seulement 48% des adduits d'une même molécule qui partagent au moins un motif sur 1590 paires analysées.

Le même genre de comparaison a été effectué avec les spectres issus d'une même molécule partageant le même adduit mais acquis sur des instruments LC-MS différents (**Tableau 19**).

Tableau 19 – Motifs retrouvés entre instruments LC-MS.

Mode négatif				Mode positif			
Machine 1	*Machine 2*	*Score*	*Total*	*Machine 1*	*Machine 2*	*Score*	*Total*
Thermo Q-Exactive Focus	Waters Q-ToF Xevo G2-XS	0.68	291	Thermo Q-Exactive Focus	Waters Q-ToF Xevo G2-XS	0.72	307
Agilent Q-ToF 6530	Waters Q-ToF Xevo G2-XS	0.53	301	Agilent Q-ToF 6530	Waters Q-ToF Xevo G2-XS	0.5	107
Agilent Q-ToF 6530	Thermo Q-Exactive Focus	0.59	263	Agilent Q-ToF 6530	Thermo Q-Exactive Focus	0.49	77

De façon inattendue, que ce soit en mode positif ou négatif, les instruments Thermo et Waters partagent le plus de motifs (72% et 68% respectivement). Les autres combinaisons d'instruments restent autour de 50% de motifs partagés. En moyenne, 60% et 63% des motifs sont reconnus d'un instrument à l'autre respectivement en mode négatif et positif.

Au vu des résultats de comparaison entre adduits il est raisonnable de considérer que chaque adduit de chaque molécule va générer un mélange de motifs communs ainsi que d'autres motifs qui leur sont propres. Par ailleurs, la majorité des motifs sont retrouvés d'un instrument LC-MS à l'autre bien qu'une proportion non négligeable ne le soit pas.

4 Conclusion

4.1 De la création de *Classnotator.*

De nombreux motifs caractéristiques de certaines classes, ou *motifs purs*, ont été trouvés grâce à un algorithme développé spécialement à cet effet. Une application de l'annotation par *motifs purs* est démontrée dans le *Chapitre V*. Ce genre d'approche peut servir à trianguler la classe structurale d'une molécule grâce à d'autres méthodes. Il reste encore plusieurs possibilités à explorer pour exploiter cette approche à son potentiel maximal. Ici, les combinaisons de motifs par exclusion n'ont pas été étudiées (par exemple : motif_1 et pas de motif_2). Le postulat au moment de cette expérience est que les motifs sans exclusion suffiront à atteindre les objectifs de *Classnotator*, sans risquer l'explosion combinatoire que pourrait provoquer la prise en compte des exclusions. Les résultats de cet algorithme sont par ailleurs hautement dépendants de la base de données de motifs utilisée. D'autres paramètres avaient été testés pour la production de la motifDB (*free motifs* : 50, 75, 100, 125, 150, 300, *bin size* : 0.1, 0.01, 0.005 Da), mais en l'absence de résultats nettement concluants sur les *motifs purs*, des paramètres intermédiaires ont été choisis (100 *free motifs* et *bin size* de 0.01 Da). L'agrandissement de la LDB de 250 à quelques milliers de moléculaires serait désirable pour améliorer l'efficacité de *Classnotator*. Ici la classification selon Hun&Yosh96 a été utilisée car spécifique aux composés lichéniques. *Classnotator* est déjà capable de reconnaitre la plupart des molécules de la LDB en utilisant cette classification. Il devrait cependant être envisagé d'effectuer la même opération sur des classifications plus universelles comme celle proposée par ClassyFire. Une validation plus large de la méthode pourrait également se faire en prenant des bases de données MS/MS plus grande et plus diversifiées que la LDB et faire fonctionner l'outil avec ClassyFire. L'utilisation de ces motifs spécifiques à des classes structurales bénéficie grandement des possibilités de triangulation. Par exemple, la prise en compte des adduits dans les réseaux moléculaires permettrait de regrouper plusieurs spectres MS/MS de la même molécule, chacun contribuant de façon complémentaire avec ses propres *motifs purs* à déterminer la classe de la molécule inconnue. De même, un couplage des données LC-MS en mode positif et négatif renforcerait l'identification par *motifs purs*.

4.2 Impact des adduits et des instruments d'analyse sur les motifs.

L'étude sur les adduits démontre que, bien qu'ils ne partagent pas toujours l'intégralité de leurs motifs, deux adduits de la même molécule seront souvent reliés par au moins un motif commun. Ceci peut permettre de confirmer des relations entre les ions de la même molécule. Concernant les différences dues aux instruments d'analyse, la majorité des motifs sont partagés sur deux instruments différents. Le fait que certains motifs ne le soient pas incite à produire des bases de données de motifs sur différents instruments pour recueillir de façon exhaustive toutes les variantes de motifs pouvant servir aux

différents utilisateurs. Une analyse détaillée des motifs de la LDB n'a pas été effectuée ici par manque de temps. Les annotations MAGMa permettent d'interpréter automatiquement les fragments et pertes de neutres simple, et d'orienter l'utilisateur vers les sous-structures communes à certains évènements de fragmentation. Une annotation détaillée de ces évènements sera à envisager à l'avenir pour permettre d'attribuer des sous-structures de façon rationnelle aux motifs.

Chapitre V

– Molnotator –

Ou la création d'un outil pour la prédiction des molécules en LC-MS et leur visualisation dans leur contexte d'ionisation

Intervenants extérieurs : Zakaria Bouchoureb[(a)], Simon Ollivier[(b)]

(a) : CNRS, CEISAM (Chimie et Interdisciplinarité : Synthèse, Analyse, Modélisation)-UMR 6230, Univ Nantes, F-44322 Nantes, France
(b) : CNRS, ISCR (Institut des Sciences Chimiques de Rennes)-UMR 6226, Univ Rennes, F-35000 Rennes, France

Contributions externes : ZB – *Conseils techniques & optimisation de certains passages du code*. SO – *Conseils théoriques & discussion des résultats*.

Résumé

Pour aller plus loin que la déréplication des signaux connus en LC-MS, un outil est développé dans ce chapitre : *Molnotator*. Il est composé de plusieurs modules permettant d'interpréter au mieux les données LC-MS : *Fragnotator* pour la détection des fragments de source, *Adnotator* pour l'annotation des adduits et la prédiction des molécules qui les ont générés, *Classnotator* pour la prédiction des classes structurales (ici avec la LDB-MotifDB et les classes Hun&Yosh96), une déréplication guidée et *Mixmoder* pour combiner les informations des modes d'ionisation positif et négatif. *Molnotator* permet ainsi de regrouper autour d'une molécule prédite tous ses ions, adduits ou fragments de source, et d'améliorer la qualité de la déréplication en guidant celle-ci sur la base des adduits, du temps de rétention, et en proposant une classe structurale pour chaque molécule. Cet outil validé en utilisant les données de la LDB-Orbitrap et a permis de détecter 273 spectres d'adduits supplémentaires ainsi que 808 spectres de fragments de source.

Sommaire

1 – Introduction 130

2 – Méthodes 133

2.1 Fonctionnement général 133

2.2 Traitement par MZmine 134

2.3 Division du fichier MGF 135

2.4 *Fragnotator* : annotation & regroupement des fragments 135

2.5 Création de la liste d'adduits à recherche pour *Adnotator* 137

2.6 *Adnotator* : Prédiction des adduits & des molécules 139

2.7 Identification des motifs par MS2LDA 142

2.8 Prédiction des classes structurales par *Classnotator* 142

2.9 Déréplication orientée 143

2.10 Couplage des modes d'ionisation par *Mixmoder* 144

3 – Résultats 145

3.1 Traitement par MZmine 145

3.2 Adduits pour le deuxième traitement par *Adnotator* 145

3.3 *Fragnotator* : annotation & regroupement des fragments 147

3.4 *Adnotator* : Prédiction des adduits & des molécules 148

3.5 *Classnotator*, déréplication orientée & *Mixmoder* 149

3.6 Interprétation des résultats & pistes d'amélioration 151

4 – Conclusion 156

1 Introduction

Dans les chapitres précédents, il a été évoqué que seul un faible pourcentage des nœuds d'un réseau moléculaire pouvait être annoté par des bases de données. Ceci peut être attribué à juste titre à des manques dans les bases de données MS/MS qui restent à ce jour de taille limitée. Un autre point, pourtant aussi important, est régulièrement ignoré : la prise en compte des ions autres que les formes (dé)protonées que peut prendre chaque molécule. Bien qu'il existe des algorithmes permettant l'annotation d'adduits, de fragments de source et de complexes, force est de constater leur absence dans les bases de données MS/MS, souvent dominées par des spectres de molécules protonées, et donc en mode positif. L'augmentation du degré d'annotation des réseaux peut certes être amélioré par la découverte de nouvelles molécules et l'ajout de leur spectres MS/MS aux bases de données, mais ceci suppose que les nœuds non-annotés soient essentiellement des *unknown unknowns*. La convolution des données en LC-MS/MS est pourtant bien reconnue, impliquant qu'une même molécule peut générer plusieurs dizaines de signaux différents. Ainsi, si seul son spectre (dé)protoné est répertorié, tous ses autres signaux non dérépliqués correspondront à des *known unknowns* et non à des *unknown unknowns*. La différenciation entre ces deux types d'inconnues est cruciale pour éviter à l'utilisateur de se fourvoyer lors d'un repérage de nouvelles molécules qui ne sont en réalité que d'autres signaux non-répertoriés de molécules connues. Le postulat formulé dans ce chapitre est qu'une déconvolution robuste et fiable de ces *known unknowns* permettra une simplification considérable des réseaux moléculaires et rendra l'utilisateur plus à même de repérer les *unknown unknowns*.

Pour permettre cette déconvolution, des algorithmes ont été développés ici pour créer des blocs regroupant les différents adduits appartenant à une même molécule, appelés blocs moléculaires. Ces algorithmes se basent sur des calculs combinatoires permettant de générer des molécules neutres hypothétiques et d'y associer les signaux des modes positif et négatif. Il ne s'agit pas simplement de trouver des ions partageant des rapports *m/z* pouvant correspondre à une molécule, mais de trouver les combinaisons d'hypothèses les plus probables et les plus à mêmes d'expliquer les ions d'un réseau. Ensuite, si l'un seul des ions d'un bloc est correctement dérépliqué, l'ensemble des ions composant ce bloc correspondent à d'autres signaux non-répertoriés de cette même molécule. Si aucun ion du bloc ne peut être dérépliqué, il peut s'agir d'un *unknown unknown* ou encore d'un *known unknown* pour lequel aucun spectre n'a été reconnu. Dans les deux cas, l'analyse sera simplifiée car au lieu d'avoir un seul spectre à partir duquel établir des conclusions, l'utilisateur aura tous les spectres du bloc à disposition. L'annotation sera augmentée par l'utilisation des *motifs purs* du *Chapitre IV*, permettant d'orienter l'utilisateur vers une classe structurale (**Figure 52**).

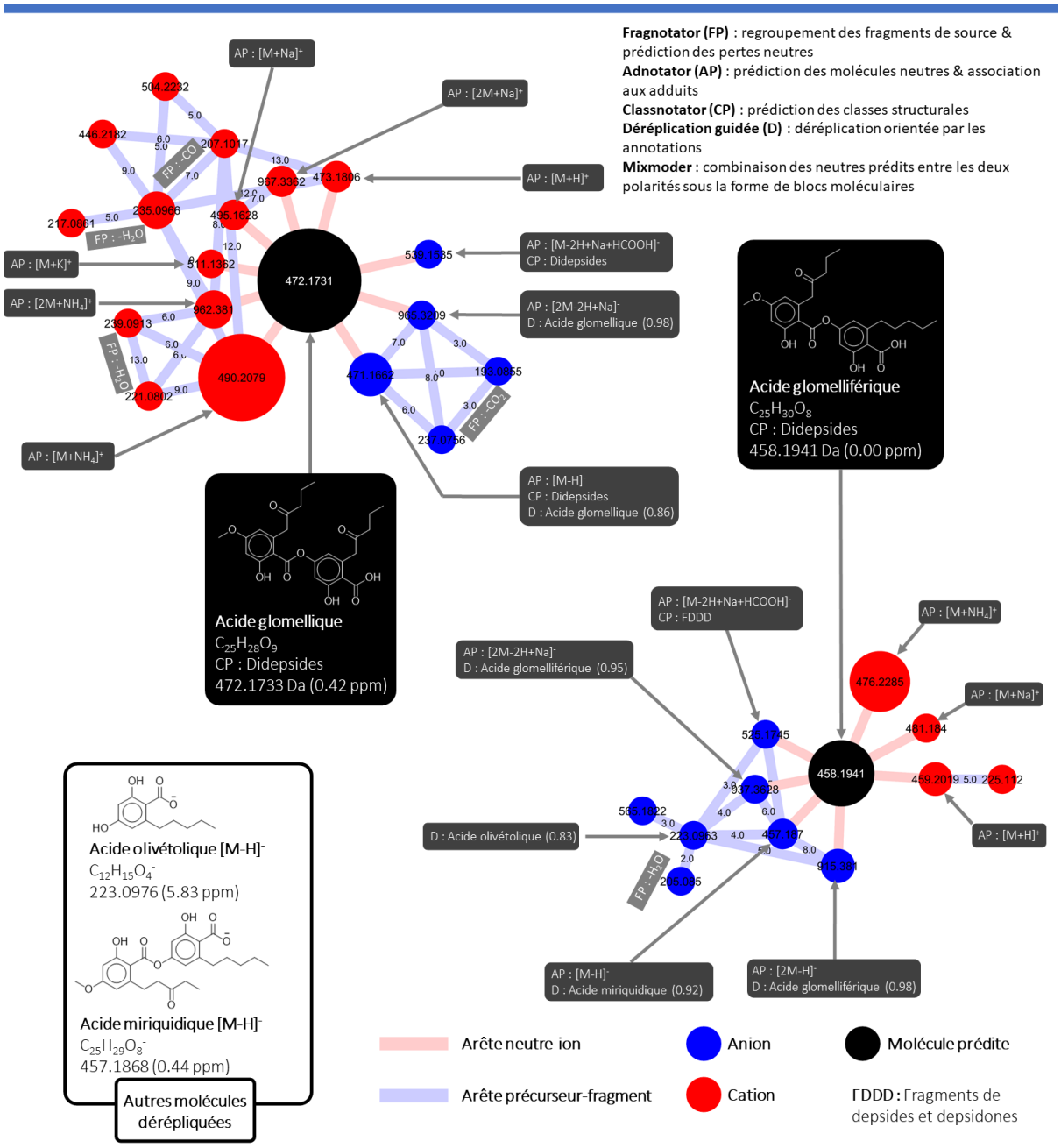


Figure 52 – Exemples de blocs moléculaires générés par Molnotator. Les blocs ont été générés en combinant les spectres positifs et négatifs acquis lors d'analyses d'un standard d'acide glomellique. Il a été possible de détecter un contaminant, dérivé du standard : l'acide glomelliférique. L'identité du nœud neutre peut être déterminée par la déréplication de ses ions : ici l'acide miriquidique a été dérépliqué une fois, très proche structurellement de l'acide glomelliférique qui lui a été dérépliqué deux fois avec des meilleurs score cosinus. Les fragments de source sont également dérépliqués dans chaque bloc, l'un d'entre eux ayant pu être identifié comme étant acide olivétolique, dont la structure est cohérente en tant que fragment de l'acide glomelliférique. Les fragments sont également annotés pour leurs pertes de neutres par Fragnotator et il devient particulièrement facile d'annoter manuellement chaque nœud avec des structures.

Les autres spectres de molécules connues peuvent être rajoutés aux bases de données en ligne pour réduire la proportion de *known unknowns* dans les réseaux moléculaires. Ce

genre d'outil voit également une utilité lors de la production de bases de données MS/MS à partir de standards : une fois le réseau créé, l'utilisateur n'aura qu'à rechercher le nœud neutre présentant la même masse que la molécule analysée, éventuellement la même classe et une déréplication appropriée le cas échéant. Les différents ions que cette molécule aura produits seront reliés à son nœud neutre et il suffira de vérifier la qualité des spectres.

Cette approche est particulièrement appropriée pour des instruments LC-MS qui génèrent un degré élevé d'adduits pour une même molécule, comme l'Orbitrap Q-Exactive Focus. Un grand nombre d'ions permet de déterminer facilement une molécule neutre hypothétique par triangulation. Si, à l'inverse, la diversité d'adduits n'est pas prise en compte lors de la déréplication, celle-ci sera bien moins efficace et génèrera un taux élevé de faux positifs, comme établi dans le *Chapitre III*.

Dans ce chapitre, la méthode est démontrée à l'aide des fichiers LC-MS bruts ayant servi à produire la LDB-Orbitrap (sur l'Orbitrap Q-Exactive Focus). *Molnotator* permettra de produire des réseaux moléculaires représentant, à proprement parler, des molécules associées aux ions qu'elles ont produits et ce, en mode positif et négatif. Cet outil regroupe cinq modules Python développés ici : *Fragnotator* pour l'annotation et le regroupement de fragments de source, *Adnotator* pour l'annotation des adduits et la prédiction des molécules qui les relient, *Classnotator* pour la prédiction supervisée des classes structurales par le biais des *motifs purs*, une déréplication guidée qui tient compte des annotations, et *Mixmoder* pour combiner les modes positif et négatif d'un même échantillon sans avoir eu recours à un mode d'acquisition à polarité mixte.

2 Méthodes

2.1 Fonctionnement général.

Le fonctionnement général de *Molnotator* est présenté en **Figure 53**. Le début du traitement reste similaire à un FBMN classique depuis l'acquisition des données jusqu'au traitement des données sur MZmine et leur exportation. *Molnotator* intervient après MZmine : les données traitées au format MGF sont soumises à *Fragnotator* qui permettra de relier les fragments de source à leurs ions précurseurs, accomplissant deux objectifs : déconvoluer davantage les données traitées & réduire les temps de calcul pour la détection des adduits en ignorant les fragments de source. *Adnotator* permettra ensuite par calculs combinatoires et triangulations de produire les molécules neutres hypothétiques les plus probables et d'y associer les différents ions et fragments de source qu'elles ont générés. En parallèle, les fichiers MGF ont été soumis à MS2LDA pour être annotés avec la LDB motifDB développée dans le *Chapitre IV*. Les résultats d'*Adnotator* et de MS2LDA sont utilisés par *Classnotator* pour prédire la classe structurale de chaque ion ainsi que celle des nœuds neutres. Une déréplication guidée est réalisée sur l'ensemble des ions avec la LDB en tenant compte des adduits prédits. Les résultats de la déréplication sont propagés sur les molécules hypothétiques. Les données des modes positif et négatif, jusque-là traitées en parallèle, sont alors combinées par *Mixmoder* pour créer un réseau de polarité mixte sur la base des molécules neutres de chaque mode et de leurs temps de rétention. Le réseau peut alors être généré et exploré sur un logiciel approprié (ici, Cytoscape).

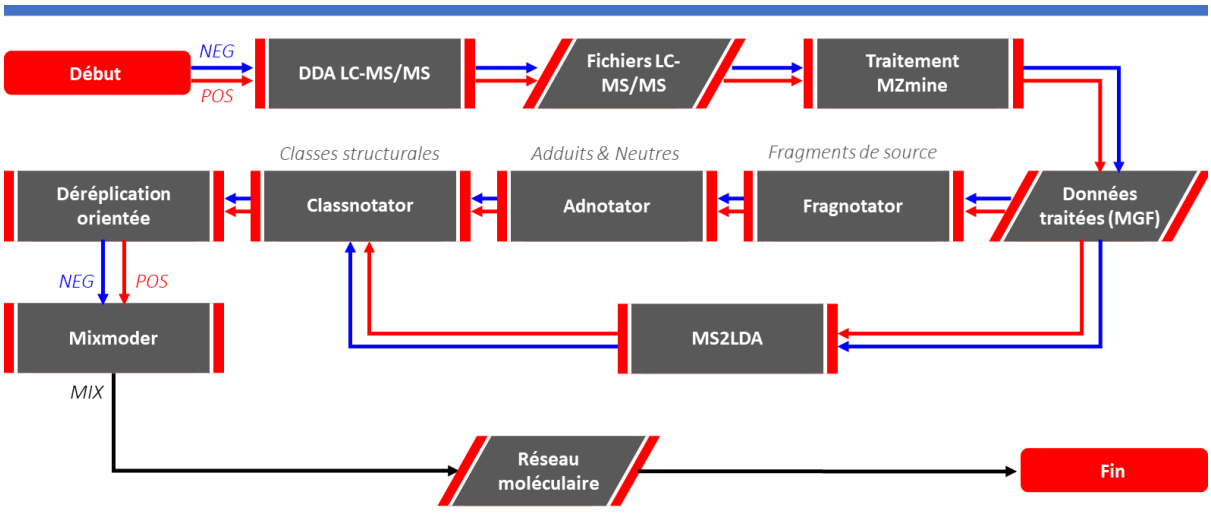


Figure 53 – Fonctionnement général de Molnotator. Molnotator est divisé en cinq modules : Fragnotator pour le regroupement et l'annotation des fragments de source, Adnotator pour la détection des adduits d'une même molécule et le calcul de la masse hypothétique de cette molécule, Classnotator pour la prédiction de la classe structurale des ions et des neutres (ici selon Hun&Yosh96), la déréplication orientée à partir des annotations et de la LDB, et Mixmoder qui permet de combiner les données des deux polarités pour créer un réseau moléculaire.

2.2 Traitement par MZmine.

Les 191 fichiers mzXML en mode positif et négatif ont été traités séparément sur MZmine. Ils ont été envoyés sur le serveur de GenOuest où ils ont été traités en *batch mode* sur MZmine 3.8 pour Linux (**Tableau 20**). Les paramètres ont été choisis de façon à retrouver dans les listes de pics les ions de la LDB-Orbitrap.

Tableau 20 – Paramètres MZmine. Si des paramètres différents ont été utilisés pour chaque mode d'ionisation, ils ont été marqués en rouge pour le mode positif et bleu pour le négatif.

Module	*Paramètres*
Raw data import	Importation de tous les fichiers mzXML
Raw data methods > Mass detection	Scans: MS level: 1 Mass detector: centroid Noise level: 2.0E4 Mass list name: masses
Raw data methods > Mass detection	Scans: MS level: 2 Mass detector: centroid Noise level: 0 Mass list name: masses
Peak list methods > Peak detection > ADAP Chromatogram builder (Myers et al. 2017)	Scans: MS level: 1 Mass list: masses Min group size in # of scans: 2 Group intensity threshold: 2.0E4 Min highest intensity: 2.0E4 m/z tolerance: 10 ppm
Peak list methods > Peak detection > Chromatogram deconvolution	Algorithm: Wavelets (ADAP): S/N threshold: 2 S/N estimator: Wavelet Coeff. SN: Peak width mult.: 3 abs(wavelet coeffs.): checked min feature height: 2.0E4 coefficient/area threshold: 170, 75 Peak duration range: 0.01-1.00 RT wavelet range: 0.008-0.005, 0.01-0.05 m/z center calculation: MEDIAN m/z range for MS2 scan pairing (Da): 0.005 RT range for MS2 scan pairing (min): 0.1
Peak list methods > Isotopes > Isotopic peaks grouper	m/z tolerance: 5.0ppm Retention time tolerance: 0.1 absolute (min) Monotonic shape: unchecked Maximum charge: 2 Representative isotope: Lower m/z
Peak list methods > Alignment > RANSAC aligner	m/z tolerance: 10.0ppm RT tolerance : 0.3 absolute (min) RT tolerance after correction: 0.05 absolute (min) RANSAC iterations: 0 Minimum number of points: 10.0 % Threshold value: 0.3 Linear model: unchecked Require same charge state: unchecked
Peak list methods > Filtering > Peak list rows filter	Keep only peaks with MS2 scan (GNPS): checked Reset the peak number ID: checked (rest is unchecked / unused)

Tableau 20 – *Suite.*

Module	*Paramètres*
Peak list methods > Export/Import > Export for/Submit to GNPS	Mass list: masses Merge MS/MS (experimental): checked Select spectra to merge: across samples m/z merge mode: weighted average (remove outliers) intensity merge mode: sum intensities Expected mass deviation: 5.0ppm Cosine threshold (%): 70.0 Peak count threshold (%): 40.0 Isolation window offset (m/z): 0.0 Isolation windows width (m/z): 3.0 Filter rows: ALL Submit to GNPS: unchecked Open folder: Unchecked
Peak list methods > Export/Import > Export to CSV file	Field separator: , Export common elements: Export row ID: checked Export row m/z: checked Export row retention time: checked Export data file elements: Peak area: checked Export quantitation results and other information: unchecked Identification separator: ";" Filter rows: ALL
Project > Save project	Enregistrement du projet

A l'issue du traitement, les fichiers MGF (spectres MS/MS) et CSV (attributs des ions) sont téléchargés de GenOuest.

2.3 Division du fichier MGF.

Les fichiers MGF obtenus en sortie de MZmine sont divisés de façon à obtenir un fichier MGF propre à chaque échantillon. Ceci est fait pour deux raisons : 1) la réduction du temps de calcul en réduisant le nombre d'ions qu'il est possible de combiner par fichier MGF, 2) réduire le risque de faux-positifs que génèrerait la recherche d'adduits et de fragments à travers plusieurs analyses LC-MS. La librairie *Libmetgem* (Olivon et al. 2018) (Python) est utilisée pour l'importation du fichier MGF, mais n'étant pas dotée de fonction d'exportation, le fichier importé est converti au format utilisé par la librairie *Pyteomics* (Goloborodko et al. 2013; Levitsky et al. 2019) (dotée d'une fonction d'exportation mais ne pouvant pas lire les fichiers MGF produits par MZmine). Grâce au fichier CSV produit par MZmine qui contient les aires de pics de chaque ion pour chaque échantillon, un fichier MGF est produit pour chacun en ne conservant que les ions dont les aires de pic sont non-nulles pour l'échantillon donné. Chaque fichier MGF pourra ensuite être traité par *Fragnotator*.

2.4 *Fragnotator* : annotation & regroupement des fragments.

Le fichier MGF de chaque échantillon est importé sur *Fragnotator* à l'aide de *Libmetgem* et est traité comme indiqué en **Figure 54**. Deux variables sont fixées par l'utilisateur au début du traitement : le $\Delta m/z$ et le ΔTR, ici à 4 ppm et 5 secondes respectivement. A partir du fichier MGF, un *node table* est d'abord créé en extrayant pour chaque ion son rapport *m/z*, son TR, son TIC et sa charge. Un *edge table* est initialisé (vide) pour contenir toutes

les paires reliant un ion précurseur (*Ion 1*) à son fragment de source (*Ion 2*). Pour chaque *Ion 1* du *node table*, un *Tableau de coélution* est produit, rassemblant tous les ions coélués avec l'*Ion 1* à ΔTR près. Ces ions (*Ions 2*) seront analysés l'un après l'autre et formeront une paire *Ion 1 – Ion 2* (ion précurseur – fragment de source), si, pour chacun de ces *Ions 2* :

- Leur rapport *m/z* est inférieur à celui de l'*Ion 1.*
- Leur rapport *m/z* est détecté dans le spectre MS/MS de l'*Ion 1* à Δ*m/z* près.
- S'ils présentent au moins deux fragments en commun avec l'*Ion 1* à Δ*m/z* près.
- Si l'ensemble des fragments communs représente au moins 10% des pics de l'*Ion 2* (*matching score* ≥ 0.1).

Lorsqu'une telle paire est formée, elle est rajoutée à l'*edge table*. Après avoir traité tous les ions du *node table*, un statut y est ajouté pour chacun : « *Fragment* » s'il a été détecté au moins une fois comme fragment de source, « *Parent* » s'il a été annoté exclusivement comme précurseur d'un fragment de source. Les ions ne se trouvant dans aucune des catégories sont annotés « *Unpaired* » et sont rajoutés à l'*edge table* sous la forme d'une paire *self-looped* (paire formée d'un ion relié à lui-même). Le *node table* et l'*edge table* sont tous deux exportés au format CSV pour être utilisés par *Adnotator*. Ils peuvent par ailleurs déjà être utilisés à ce stade pour créer un réseau et vérifier la qualité du traitement.

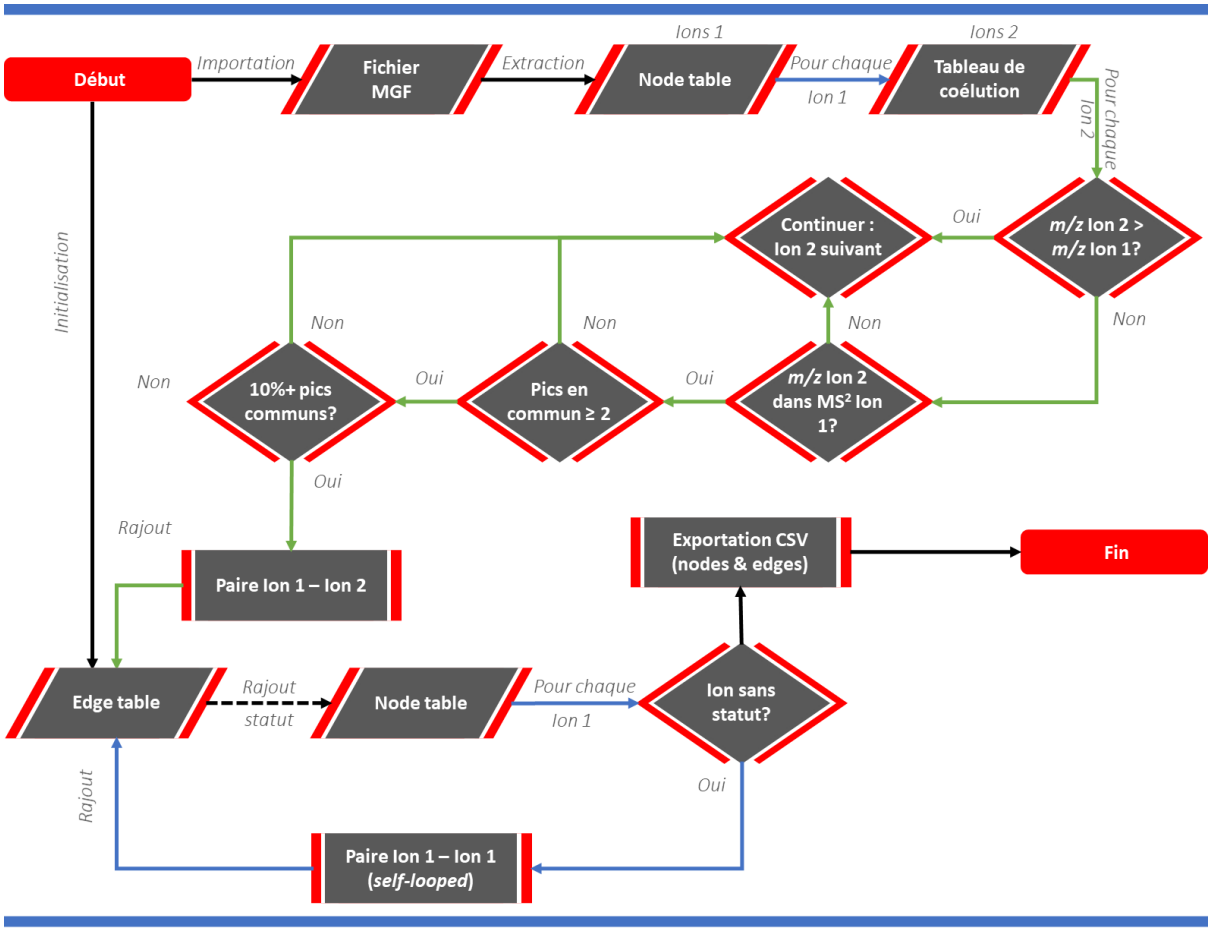


Figure 54 – Fonctionnement de Fragnotator.

2.5 Création de la liste d'adduits à rechercher pour *Adnotator*.

Adnotator se base sur une liste d'adduits pour annoter les ions d'une analyse LC-MS. Comme constaté dans le *Chapitre III*, l'abondance et la nature des adduits produits changent d'un instrument à l'autre. Si l'utilisateur connait les adduits qu'il est susceptible de trouver, il peut fournir sa propre liste. Autrement, un algorithme combinatoire a été créé pour générer toutes les possibilités d'adduits à partir de paramètres fournis par l'utilisateur, de façon à n'en manquer aucun. Cependant, utiliser une combinatoire d'adduits peut générer plusieurs dizaines voire centaines d'adduits, ce qui rallonge considérablement les temps de calcul et augmente le risque de faux-positifs. A l'inverse, une liste d'adduits incomplète génèrera des faux-négatifs, ce qui influencera la qualité du traitement. C'est pourquoi, dans cette démonstration, *Adnotator* sera utilisé deux fois : une première fois avec la combinatoire d'adduits, une deuxième uniquement avec les adduits retenus. Après un premier traitement avec la combinatoire, seuls les adduits détectés pour les standards de la LDB seront retenus. Le deuxième traitement avec ces adduits retenus est ensuite réalisé, avec des temps de calcul plus courts et moins de faux-positifs. Le fonctionnement de l'algorithme combinatoire pour la génération des adduits est exposé ci-dessous.

La formule d'un adduit peut se résumer à son nombre x de molécules M, associé à une combinaison de y espèces chargées I et une combinaison de z espèces neutres N, ainsi que de la charge q résultant de l'ensemble des combinaisons :

$$Adduit = \left[\binom{M}{x} + \binom{I}{y} + \binom{N}{z}\right]^{q}$$

Les espèces chargées ont été sélectionnées parmi les ions communément trouvés dans les bases de données MS/MS. Les espèces neutres choisies sont l'acétonitrile (CH_3CN) et l'acide formique (HCOOH), respectivement un solvant de la phase mobile et l'acidifiant. L'eau (H_2O), autre composant de la phase mobile, n'a pas été sélectionnée car elle entrerait en conflit avec les pertes de neutre ($-H_2O$) repérées par *Fragnotator*. Une valeur *qmax* est fixée à 1, représentant la valeur absolue maximale que peut atteindre la charge q. Bien qu'il soit possible de traiter les ions multichargés, ils ont été ignorés pour le moment étant donné qu'ils restent largement minoritaires. Chaque forme d'ion est associée à un *score de complexité C* ($C = x + y + z$). Un seuil *Cmax* est utilisé pour éliminer les adduits avec un score C supérieur à ce seuil. L'ensemble des adduits générés sera filtré en fonction de leurs charges et du mode d'ionisation utilisé. Les paramètres utilisés sont présentés dans le **Tableau 21**.

Tableau 21 – Paramètres utilisés pour l'algorithme combinatoire.

Paramètre	*Valeurs*
x	[1, 2, 3]
y	[1, 2, 3]
I	[$H^{+/-}$, NH_4^+, Na^+, Cl^-, K^+]
z	[1]
N	[HCOOH, CH_3CN]
Qmax	1
Cmax	5

Avec ces paramètres, l'algorithme combinatoire produit 76 adduits en mode positif (**Tableau 22**) et 54 en mode négatif (**Tableau 23**). Ils sont accompagnés d'une masse *Δm* (somme des masses de ses espèces chargées et neutres), d'un nombre de molécules *x* et d'un score du score de complexité *C*. Ces tableaux sont ensuite exportés au format CSV pour être utilisés par *Adnotator*.

Tableau 22 – Adduits produits en mode positif par l'algorithme combinatoire.

Adduit	*Δm*	*x*	*C*	*Adduit*	*Δm*	*x*	*C*
$[M+H+HCOOH]^+$	47.0133	1	3	$[M+Na+Cl+K]^+$	96.9223	1	4
$[M+H+CH_3CN]^+$	42.0343	1	3	$[M+Cl+2K+HCOOH]^+$	158.9017	1	5
$[M+H]^+$	1.0078	1	1	$[M+Cl+2K+CH_3CN]^+$	153.9228	1	5
$[M+NH_4+HCOOH]^+$	64.0398	1	3	$[M+Cl+2K]^+$	112.8962	1	4
$[M+NH_4+CH_3CN]^+$	59.0609	1	3	$[2M+H+HCOOH]^+$	47.0133	2	4
$[M+NH_4]^+$	18.0343	1	2	$[2M+H+CH_3CN]^+$	42.0343	2	4
$[M+Na+HCOOH]^+$	68.9952	1	3	$[2M+H]^+$	1.0078	2	3
$[M+Na+CH_3CN]^+$	64.0163	1	3	$[2M+NH_4+HCOOH]^+$	64.0398	2	4
$[M+Na]^+$	22.9897	1	1	$[2M+NH_4+CH_3CN]^+$	59.0609	2	4
$[M+K+HCOOH]^+$	84.9691	1	3	$[2M+NH_4]^+$	18.0343	2	3
$[M+K+CH_3CN]^+$	79.9902	1	3	$[2M+Na+HCOOH]^+$	68.9952	2	4
$[M+K]^+$	38.9637	1	2	$[2M+Na+CH_3CN]^+$	64.0163	2	4
$[M+2H+Cl+HCOOH]^+$	82.9899	1	5	$[2M+Na]^+$	22.9897	2	3
$[M+2H+Cl+CH_3CN]^+$	78.0110	1	5	$[2M+K+HCOOH]^+$	84.9691	2	4
$[M+2H+Cl]^+$	36.9845	1	4	$[2M+K+CH_3CN]^+$	79.9902	2	4
$[M+H+NH_4+Cl+HCOOH]^+$	100.0165	1	5	$[2M+K]^+$	38.9637	2	3
$[M+H+NH_4+Cl+CH_3CN]^+$	95.0376	1	5	$[2M+2H+Cl]^+$	36.9845	2	5
$[M+H+NH_4+Cl]^+$	54.0110	1	4	$[2M+H+NH_4+Cl]^+$	54.0110	2	5
$[M+H+Na+Cl+HCOOH]^+$	104.9719	1	5	$[2M+H+Na+Cl]^+$	58.9664	2	5
$[M+H+Na+Cl+CH_3CN]^+$	99.9929	1	5	$[2M+H+Cl+K]^+$	74.9403	2	5
$[M+H+Na+Cl]^+$	58.9664	1	4	$[2M+2NH_4+Cl]^+$	71.0376	2	5
$[M+H+Cl+K+HCOOH]^+$	120.9458	1	5	$[2M+NH_4+Na+Cl]^+$	75.9929	2	5
$[M+H+Cl+K+CH_3CN]^+$	115.9669	1	5	$[2M+NH_4+Cl+K]^+$	91.9669	2	5
$[M+H+Cl+K]^+$	74.9403	1	4	$[2M+2Na+Cl]^+$	80.9483	2	5
$[M+2NH_4+Cl+HCOOH]^+$	117.0430	1	5	$[2M+Na+Cl+K]^+$	96.9223	2	5
$[M+2NH_4+Cl+CH_3CN]^+$	112.0641	1	5	$[2M+Cl+2K]^+$	112.8962	2	5
$[M+2NH_4+Cl]^+$	71.0376	1	4	$[3M+H+HCOOH]^+$	47.0133	3	5
$[M+NH_4+Na+Cl+HCOOH]^+$	121.9984	1	5	$[3M+H+CH_3CN]^+$	42.0343	3	5
$[M+NH_4+Na+Cl+CH_3CN]^+$	117.0195	1	5	$[3M+H]^+$	1.0078	3	4
$[M+NH_4+Na+Cl]^+$	75.9929	1	4	$[3M+NH_4+HCOOH]^+$	64.0398	3	5
$[M+NH_4+Cl+K+HCOOH]^+$	137.9724	1	5	$[3M+NH_4+CH_3CN]^+$	59.0609	3	5
$[M+NH_4+Cl+K+CH_3CN]^+$	132.9934	1	5	$[3M+NH_4]^+$	18.0343	3	4
$[M+NH_4+Cl+K]^+$	91.9669	1	4	$[3M+Na+HCOOH]^+$	68.9952	3	5
$[M+2Na+Cl+HCOOH]^+$	126.9538	1	5	$[3M+Na+CH_3CN]^+$	64.0163	3	5
$[M+2Na+Cl+CH_3CN]^+$	121.9749	1	5	$[3M+Na]^+$	22.9897	3	4
$[M+2Na+Cl]^+$	80.9483	1	4	$[3M+K+HCOOH]^+$	84.9691	3	5
$[M+Na+Cl+K+HCOOH]^+$	142.9278	1	5	$[3M+K+CH_3CN]^+$	79.9902	3	5
$[M+Na+Cl+K+CH_3CN]^+$	137.9488	1	5	$[3M+K]^+$	38.9637	3	4

Tableau 23 – Adduits produits en mode négatif par l'algorithme combinatoire.

Adduit	Δm	x	C	Adduit	Δm	x	C
$[M-H+HCOOH]^-$	44.9976	1	3	$[M+Na+2Cl+HCOOH]^-$	138.9329	1	5
$[M-H+CH_3CN]^-$	40.0187	1	3	$[M+Na+2Cl+CH_3CN]^-$	133.9540	1	5
$[M-H]^-$	-1.0078	1	1	$[M+Na+2Cl]^-$	92.9274	1	4
$[M+Cl+HCOOH]^-$	80.9743	1	3	$[M+2Cl+K+HCOOH]^-$	154.9068	1	5
$[M+Cl+CH_3CN]^-$	75.9954	1	3	$[M+2Cl+K+CH_3CN]^-$	149.9279	1	5
$[M+Cl]^-$	34.9688	1	2	$[M+2Cl+K]^-$	108.9014	1	4
$[M-2H+NH_4+HCOOH]^-$	62.0242	1	5	$[2M-H+HCOOH]^-$	44.9976	2	4
$[M-2H+NH_4+CH_3CN]^-$	57.0452	1	5	$[2M-H+CH_3CN]^-$	40.0187	2	4
$[M-2H+NH_4]^-$	16.0187	1	4	$[2M-H]^-$	-1.0078	2	3
$[M-2H+Na+HCOOH]^-$	66.9796	1	5	$[2M+Cl+HCOOH]^-$	80.9743	2	4
$[M-2H+Na+CH_3CN]^-$	62.0006	1	5	$[2M+Cl+CH_3CN]^-$	75.9954	2	4
$[M-2H+Na]^-$	20.9741	1	4	$[2M+Cl]^-$	34.9688	2	3
$[M-2H+K+HCOOH]^-$	82.9535	1	5	$[2M-2H+NH_4]^-$	16.0187	2	5
$[M-2H+K+CH_3CN]^-$	77.9746	1	5	$[2M-2H+Na]^-$	20.9741	2	5
$[M-2H+K]^-$	36.9480	1	4	$[2M-2H+K]^-$	36.9480	2	5
$[M-H+NH_4+Cl+HCOOH]^-$	98.0008	1	5	$[2M-H+NH_4+Cl]^-$	51.9954	2	5
$[M-H+NH_4+Cl+CH_3CN]^-$	93.0219	1	5	$[2M-H+Na+Cl]^-$	56.9507	2	5
$[M-H+NH_4+Cl]^-$	51.9954	1	4	$[2M-H+Cl+K]^-$	72.9247	2	5
$[M-H+Na+Cl+HCOOH]^-$	102.9562	1	5	$[2M+NH_4+2Cl]^-$	87.9720	2	5
$[M-H+Na+Cl+CH_3CN]^-$	97.9773	1	5	$[2M+Na+2Cl]^-$	92.9274	2	5
$[M-H+Na+Cl]^-$	56.9507	1	4	$[2M+2Cl+K]^-$	108.9014	2	5
$[M-H+Cl+K+HCOOH]^-$	118.9302	1	5	$[3M-H+HCOOH]^-$	44.9976	3	5
$[M-H+Cl+K+CH_3CN]^-$	113.9512	1	5	$[3M-H+CH_3CN]^-$	40.0187	3	5
$[M-H+Cl+K]^-$	72.9247	1	4	$[3M-H]^-$	-1.0078	3	4
$[M+NH_4+2Cl+HCOOH]^-$	133.9775	1	5	$[3M+Cl+HCOOH]^-$	80.9743	3	5
$[M+NH_4+2Cl+CH_3CN]^-$	128.9986	1	5	$[3M+Cl+CH_3CN]^-$	75.9954	3	5
$[M+NH_4+2Cl]^-$	87.9720	1	4	$[3M+Cl]^-$	34.9688	3	4

2.6 *Adnotator* : Prédiction des adduits & des molécules.

Après la détection des fragments de source par *Fragnotator*, le *nodes table* et l'*edge table* sont traités par *Adnotator*. Le *node table* est complété avec la forme d'ionisation prédite pour chaque ion & les molécules neutres hypothétiques leur correspondant sont rajoutées. Les liens entre les nœuds de molécules neutres et les nœuds des différents ions que ces molécules génèrent sont rajoutés à l'*edge table*. Les variables $\Delta m/z$ et ΔTR sont ici fixées à 4 ppm et 7 secondes respectivement. Le fonctionnement général d'*Adnotator* est présenté en **Figure 55**. Chaque sous-partie sera détaillée par la suite.

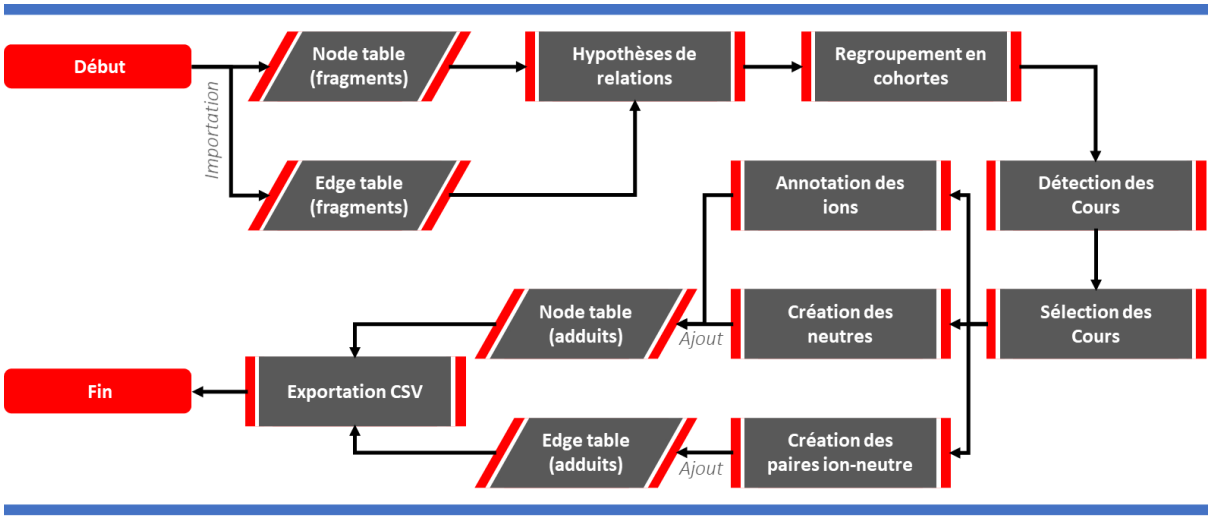


Figure 55 – Fonctionnement général d'Adnotator.

Des *Hypothèses de relations* sont générées à partir des tableaux produits par *Fragnotator*. Chaque ion (*Ion 1*) du *node table*, s'il n'est pas annoté comme fragment de source, sera

associé à 0, 1 ou plusieurs de ces hypothèses. Ces *Hypothèses de relations* se composent d'une première *Hypothèse d'adduit* sur l'*Ion 1*, d'une hypothèse sur le neutre associé à cette *Hypothèse d'adduit 1*, et d'une *Hypothèse d'adduit 2* sur un ion coélué qui présente le bon rapport *m/z* correspondant à un autre adduit possible pour la molécule neutre hypothétique (**Figure 56**). Un *Ion 2* peut être un ion fragment, à la différence des *Ions 1*, ce qui permet de traiter les fragments indirectement et de réduire les temps de calcul en évitant les données redondantes. Dans les faits, si une molécule ne génère qu'un adduit, aucune *Hypothèse de relations* ne pourra être associée à cet ion. Une molécule générant deux adduits ou plus permettra de relier ces ions par une *Hypothèse de relations*.

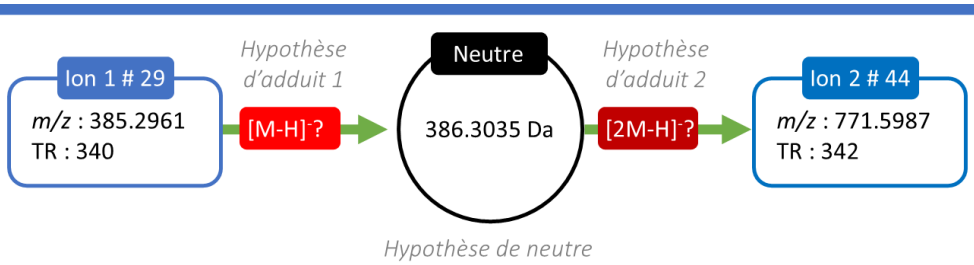


Figure 56 – Exemple d'une Hypothèse de relation. Elle peut être décomposée de la façon suivante : Ion 1 (29) [M-H]⁻ - Neutre 386.3035 – Ion 2 (44) [2M-H]⁻. Les numéros associés à chaque ion représentent le Feature ID de MZmine.

Chaque *Hypothèse de relation* est associée à un score *HRS* reflétant sa simplicité, les *Hypothèses* les plus simples étant considérées plus vraisemblables. Il est calculé de la façon suivante :

$$HRS = \frac{1 + F + N_c}{C}$$

Les éléments permettant de calculer HRS sont *F* (Fragmentation), étant égal à 1 si l'*Ion 1* et *2* sont déjà reliés par *Fragnotator* (souvent le cas pour deux adduits de la même molécule) ou à 0 dans le cas contraire, N_c (confirmation de la fraction neutre) étant égal 0 par défaut ou à 1 si l'*Ion 2* possède une partie neutre et qu'un autre ion sans la partie neutre est détecté parmi les ions coélués, et *C* (complexité) étant le score de complexité de l'*Ion 2* dans le *Tableau d'adduits*. Dans l'exemple en **Figure 56**, [M-H]⁻ – [2M-H]⁻ est plus probable que [2M-H]⁻ – [4M-H]⁻.

Les *Hypothèses dépendantes* sont ensuite regroupées en *Cohortes*. Dans l'exemple de la **Figure 57**, quatre *Hypothèses de relation* sont représentées : HR 1, 2, 3 et 4. HR 1, 2 et 3 sont *dépendantes* car elles partagent toutes au moins un ion, ici l'ion #29. Elles sont toutes indépendantes de HR 4 car elles ne partagent aucun ion avec cette dernière. Une première *Cohorte* pourra être créée avec HR 1, 2 et 3 et une deuxième avec HR 4. L'étape suivante consistera à étudier les relations entre les différentes *Hypothèses de relation* composant une *Cohorte*. Ici par exemple, HR 3 est *incompatible* avec HR 1 et 2 car elle présente une annotation conflictuelle pour l'ion #29 : [2M-H]⁻ et non [M-H]⁻ comme dans les deux autres. HR 1 et 2 sont en revanche *compatibles* car elles ne présentent aucune annotation conflictuelle pour leurs ions, et portent même une annotation consensuelle pour l'ion

#29. Le but du regroupement en *Cohortes* est de rassembler les hypothèses dont les annotations ont un impact sur d'autres ions, ce qui est démontré avec les hypothèses *incompatibles* : si l'ion #29 est un $[M-H]^-$ (HR 1, HR 2), il ne peut pas être un $[2-M]^-$ et l'ion #84 n'est donc probablement pas un $[3M-H]^-$ (HR 3).

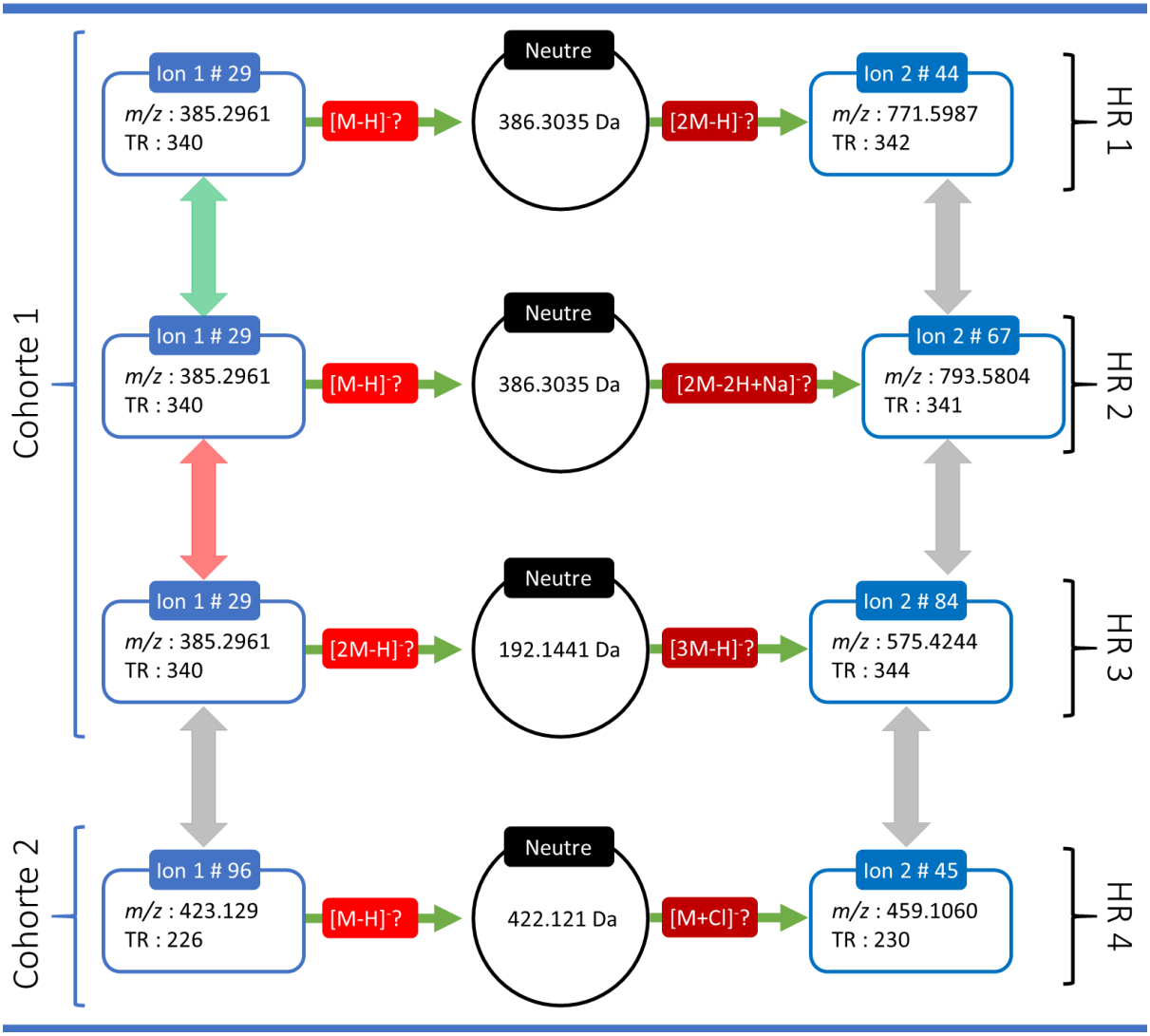


Figure 57 – Regroupement des Hypothèses de relation dépendantes en Cohortes. Les flèches entre les ions représentent des annotations consensuelles (vert), conflictuelles (rouge) ou indépendantes (gris). HR : Hypothèse de relation.

Pour chaque *Cohorte*, toutes les *Hypothèses de relation compatibles* sont regroupées en *Cours*. Chaque *Cour* sera associée à un score, somme des scores de toutes les *Hypothèses de relation* qui la composent. Ainsi, la *Cour* avec le meilleur score permettra d'expliquer le plus grand nombre d'ions dans sa Cohorte avec les annotations les plus simples.

Dans l'exemple de la **Figure 57**, la *Cohorte 1* admettrait deux *Cours* : la première constituée de HR 1 et 2, la deuxième constituée de HR 3. La *Cour 1* permet d'expliquer 3 ions (#29, 44 et 67) avec des hypothèses relativement simples : $[M-H]^-$, $[2M-H]^-$ et $[2M-2H+Na]^-$. La *Cour 2* ne permet d'expliquer que deux ions (#29 et 84) avec des hypothèses plus complexes : $[2M-H]^-$ et $[3M-H]^-$, l'ion #67 demeurant sans annotation. La *Cour 1* sera

donc la plus probable, étant la plus simple et la plus à même d'expliquer les ions, alors que la *Cour 2* n'explique que 2 ions sur 3 de façon moins convaincante.

Une fois que la meilleure *Cour* dans chaque *Cohorte* est sélectionnée, le chemin inverse est parcouru : pour chaque *Cour* sélectionnée, ses *Hypothèses de relation* sont utilisées pour annoter tous les ions qui les composent et créer un nœud correspondant à leur molécule neutre hypothétique. Les adduits prédits sont rajoutés au *node table* ainsi que les molécules neutres. Les liens entre les molécules neutres et les différents ions associés sont rapportés dans l'*edge table*. Les deux tableaux sont alors exportés au format CSV pour être utilisés par les autres modules de la méthode. Comme auparavant, il est déjà possible de créer un réseau à partir des données générées pour évaluer la qualité de la détection d'adduits.

2.7 Identification des motifs par MS2LDA.

Avant de procéder à la prédiction des classes structurales par *Classnotator*, il est nécessaire d'annoter les ions en sortie de MZmine avec MS2LDA et la LDB motifDB. Les fichiers MGF (modes positif et négatif) obtenus après traitement par MZmine sont soumis à ms2lda.org. Une détection de motifs est réalisée pour chacun des modes avec les paramètres suivants : *minimum intensity of MS2 peaks* : 5000, *bin width :* 0.01 Da, *number of Mass2Motifs :* 100, *number of iterations :* 1000. Les *motifsets* utilisés pour initier la création de motifs sont la LDB motifDB en mode négatif (http://ms2lda.org/motifdb/motif_set/33/) et positif (http://ms2lda.org/motifdb/motif_set/37/) suivant la polarité des ions du fichier MGF.

2.8 Prédiction des classes structurales par *Classnotator*.

Les résultats de l'annotation par MS2LDA avec la LDB motifDB ainsi que la liste des *motifs purs* générés par la LDB sont importés par *Classnotator*. Un tableau *molmotif* est créé par l'association à chaque molécule de tous les motifs qui y ont été détectés. Une détection des *motifs purs* est ensuite réalisée et chaque molécule sera annotée avec la classe structurale correspondant à son/ses *motif(s) pur(s),* ainsi qu'avec la somme des valeurs *n_mol* de ces motifs. Ces valeurs cumulées de *n_mol* serviront de score dans le cas où plusieurs classes structurales sont détectées pour un même ion. Les nœuds neutres sont annotés avec l'ensemble des classes prédites pour chacun des nœuds qui leur sont directement reliés. Ils accumuleront également les scores *n_mol* pour chaque classe structurale, permettant de sélectionner la classe la plus probable. Les annotations des ions sont ensuite exportées au format CSV (**Figure 58**).

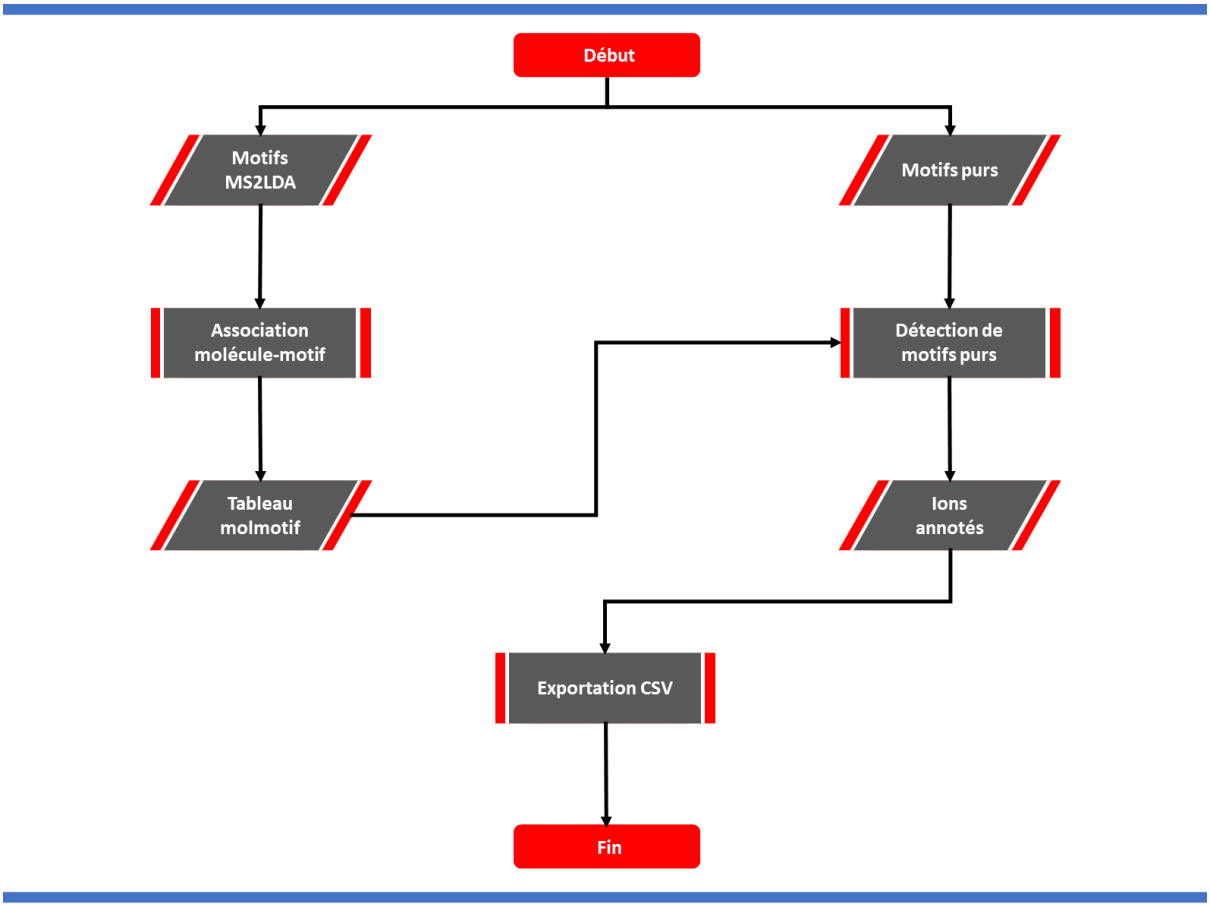


Figure 58 – Annotation des ions/nœuds par Classnotator.

2.9 Déréplication orientée.

Les ions sont dérépliqués de façon guidée grâce à la LDB étendue du *Chapitre III* (**Figure 59**). Ainsi, les nœuds pour lesquels les adduits ont été prédits n'ont été dérépliqués que contre les éléments de la LDB ionisés de la même façon. Une option permet d'utiliser les temps de rétention des standards de la LDB pour améliorer la déréplication. Elle a été utilisée ici avec une fenêtre de temps de rétention ΔTR de 8 secondes. Les ions pour lesquels aucun adduit n'a été prédit (nœuds *self-looped* et certains fragments) ne sont pas soumis à ces filtres. Un filtre supplémentaire réduit les spectres candidats de la LDB à ceux présentant le même rapport *m/z* à Δ*m/z* près, ici 8 ppm. Le seuil de cosinus est fixé à 0.7 et le nombre *top_h* des meilleurs candidats retenus dans LDB est fixé à 3. Les résultats de la déréplication sont rajoutés au *node table* et sont propagés aux nœuds neutres de façon à conclure sur l'identité du bloc moléculaire. Ceci se fait en cumulant les scores de cosinus pour chaque identification sur les nœuds neutres directement reliés et en ne conservant que les *top_h* meilleurs résultats.

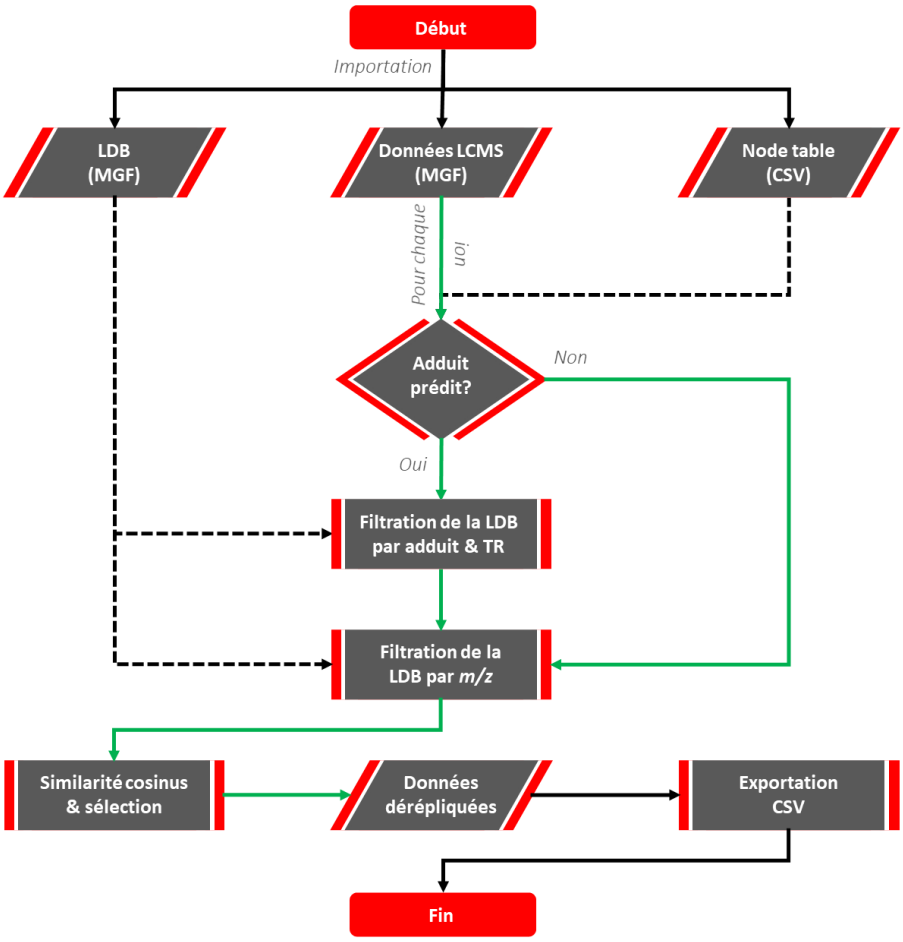


Figure 59 – Déréplication orientée par la LDB de données LC-MS/MS annotées.

2.10 Couplage des modes d'ionisation par *Mixmoder*.

Les données des deux modes d'ionisation sont combinées sur la base des nœuds neutres. Deux blocs moléculaires dont les nœuds neutres coïncident à $\Delta m/z$ et ΔTR près sont considérés être la même molécule. Ces seuils sont fixés respectivement à 10 ppm et 8 secondes. Si pour un bloc d'une polarité contenant un neutre, aucune correspondance n'a pu être retrouvée dans l'autre mode, une recherche simplifiée via une *Hypothèse d'adduit 2* sur son neutre sera effectuée sur les nœuds des blocs non-moléculaires. Ceci permet de récupérer les ions uniques générés par cette molécule dans la polarité opposée et qui n'ont donc pas pu être intégrés dans une *Hypothèse de relations*. Le nœud neutre résultant présentera une masse et un temps de rétention calculé avec la moyenne de ceux des deux modes. L'annotation *Classnotator* et la déréplication orientée pour chacun des modes y seront également reportées en cumulant les scores de chacun des modes. Les *node table* et *edge table* résultants sont exportés au format CSV.

3 Résultats

3.1 Traitement par MZmine.

Les paramètres utilisés sur MZmine sont peu sélectifs pour permettre la détection d'un grand nombre de pics et de retrouver la plupart des spectres de la LDB. Ceci engendre néanmoins des temps de calculs prohibitifs sur une machine locale et force l'utilisation d'un cluster de calculs pour le traitement, ici GenOuest. Après traitement, 19 698 ions ont été détectés en mode négatif et 26 917 en mode positif. Ces ions sont traités successivement par *Fragnotator*, *Adnotator*, *Classnotator*, la déréplication orientée et *Mixmoder*. Au vu de la quantité d'information générée, seuls les résultats finaux en sortie de *Mixmoder* seront présentés dans le détail (**Annexe, Figures S-2 à S-136**). Les résultats pour les autres molécules seront ici illustrés en suivant le traitement d'une molécule représentative : l'acide glomellique.

3.2 Adduits pour le deuxième traitement par *Adnotator*.

Après avoir utilisé *Adnotator* avec la liste d'adduits générée par l'algorithme combinatoire, seuls les adduits retrouvés dans les blocs des molécules de la LDB sont conservés. 89 molécules ont été examinées en mode négatif et 62 en mode positif, permettant respectivement la sélection de 12 et 14 formes d'ionisation dans chaque polarité (**Tableau 24**).

Tableau 24 – Adduits retenus en mode négatif et positif.

Mode négatif		*Mode positif*	
Adduit	Occurrences	Adduit	Occurrences
$[M-H]^-$	89	$[M+H]^+$	53
$[2M-2H+Na]^-$	63	$[M+Na]^+$	44
$[2M-H]^-$	55	$[M+NH_4]^+$	34
$[M-2H+Na+HCOOH]^-$	39	$[2M+Na]^+$	31
$[M+Cl]^-$	11	$[M+K]^+$	26
$[M-H+HCOOH]^-$	10	$[M+Na+CH_3CN]^+$	26
$[M-2H+Na]^-$	5	$[2M+NH_4]^+$	21
$[2M-H+HCOOH]^-$	4	$[2M+H]^+$	15
$[3M-H]^-$	2	$[M+NH_4+CH_3CN]^+$	11
$[2M+Cl]^-$	1	$[2M+K]^+$	10
$[2M-2H+K]^-$	1	$[M+H+CH_3CN]^+$	2
$[2M-2H+Na+HCOOH]^-$	0	$[3M+Na]^+$	2
		$[3M+NH_4]^+$	1
		$[M+H+CH_3OH]^+$	0

Ceci a permis de réduire significativement la quantité d'ions à rechercher par *Adnotator*, faisant passer le mode négatif de 54 à 12 formes et le mode positif de 76 à 14.

Dans le mode négatif, les changements notables par rapport à la production de la LDB-étendue dans le *Chapitre III* sont l'ajout des adduits $[M-2H+Na+HCOOH]^-$, [2M-

2H+Na+HCOOH]$^-$, [3M-H]$^-$, [2M+Cl]$^-$ et [2M-2H+K]$^-$, et la suppression de l'adduit [M-H+CH_3OH]$^-$. L'adduit [M-2H+Na+HCOOH]$^-$ s'est avéré être l'une des formes d'ionisation principales du mode négatif et en conséquent, sa forme dimérisée [2M-2H+Na+HCOOH]$^-$ a également été prise en compte. Les autres formes, bien que mineures, ont été conservées. La forme [M-H+CH_3OH]$^-$ en revanche ne correspondait pas à des adduits mais à des dérivés méthanoliques, avec un ΔTR plus élevé que les autres adduits, résultant en leur séparation du bloc moléculaire et à la formation de leurs propres blocs pour lesquels d'autres adduits étaient détectables. Concernant les complexes ion-neutre, aucun n'a été détecté avec de l'acétonitrile (CH_3CN), à la différence de l'acide formique (HCOOH) qui contribue certainement à l'ionisation en se déprotonant.

En mode positif, les changements les plus notables correspondent à l'ajout des trimères [3M+H]$^+$ et [3M+NH_4]$^+$ ainsi qu'à la suppression de toutes les formes comprenant HCOOH. Ces dernières se sont avérées être des fragments de source et leur prise en compte a généré des blocs faux-positifs.

Les adduits ainsi sélectionnés et réutilisés par *Adnotator* sont présentées dans les **Tableaux 25** et **26**. Cette liste réduite permet d'augmenter considérablement la vitesse de calcul et de réduire le nombre de faux-positifs. Il reste encore des adduits à déterminer, souvent détectés par *Fragnotator* comme des ions précurseurs d'adduits connus, ce qui permet de les repérer même sans les avoir dans le tableau d'adduits.

Tableau 25 – Liste d'adduits réutilisés par Adnotator en mode négatif.

Adduit	*Charge*	*Δm*	*x*	*C*
[M-H+HCOOH]$^-$	-1	44.9976	1	3
[M-H]$^-$	-1	-1.0078	1	1
[M+Cl]$^-$	-1	34.9688	1	2
[M-2H+Na+HCOOH]$^-$	-1	66.9796	1	5
[M-2H+Na]$^-$	-1	20.9741	1	4
[2M-H+HCOOH]$^-$	-1	44.9976	2	4
[2M-H]$^-$	-1	-1.0078	2	3
[2M+Cl]$^-$	-1	34.9688	2	3
[2M-2H+Na+HCOOH]$^-$	-1	66.9796	2	6
[2M-2H+Na]$^-$	-1	20.9741	2	5
[2M-2H+K]$^-$	-1	36.9480	2	5
[3M-H]$^-$	-1	-1.0078	3	4

Tableau 26 – Liste d'adduits réutilisés par Adnotator en mode positif.

Adduit	*Charge*	*Δm*	*x*	*C*
[M+H+CH_3CN]$^+$	1	42.0343	1	3
[M+H+CH_3OH]$^+$	1	33.0340	1	3
[M+H]$^+$	1	1.0078	1	1
[M+NH_4+CH_3CN]$^+$	1	59.0609	1	3
[M+NH_4]$^+$	1	18.0343	1	2
[M+Na+CH_3CN]$^+$	1	64.0163	1	3
[M+Na]$^+$	1	22.9897	1	1
[M+K]$^+$	1	38.9637	1	2
[2M+H]$^+$	1	1.0078	2	3
[2M+NH_4]$^+$	1	18.0343	2	3
[2M+Na]$^+$	1	22.9897	2	3
[2M+K]$^+$	1	38.9637	2	3
[3M+NH_4]$^+$	1	18.0343	3	4
[3M+Na]$^+$	1	22.9897	3	4

3.3 *Fragnotator* : annotation & regroupement des fragments.

Les fragments de source ont été détectés pour chaque fichier MGF. Ceci permet également de regrouper, dans plusieurs cas, les différents adduits d'une même molécule, du fait de leurs similarités spectrales. Ceci peut être observé dans l'exemple de l'acide glomellique. En mode négatif (**Figure 60-A**) et en amont de la détection d'adduits qui sera faite par la suite, l'ion à *m/z* 471 correspond en réalité à la molécule déprotonée $[M-H]^-$, détectée comme fragment de source de $[2M-H]^-$ *m/z* 943, de $[2M-2H+Na]^-$ *m/z* 965 et d'ion ion à *m/z* 583. Ce même ion $[M-H]^-$ produit un fragment de source à *m/z* 193 partagé avec un autre ion à *m/z* 427. Le bloc produit en mode positif (**Figure 60-B**) est plus complexe : les ions à *m/z* 473, 490 et 962 correspondent respectivement à $[M+H]^+$, $[M+NH_4]^+$ et à $[2M+NH_4]^+$. Les autres nœuds correspondent à des fragments de source de ces ions ou à d'autres adduits non répertoriés. Ces annotations, ici faites manuellement, seront confirmées par *Adnotator* dans la suite du traitement. *Fragnotator* permet ainsi de regrouper les différents fragments de source d'une molécule dans un même bloc, ainsi que ses différents adduits sur la base d'un nombre de pics partagés et d'un *matching score*.

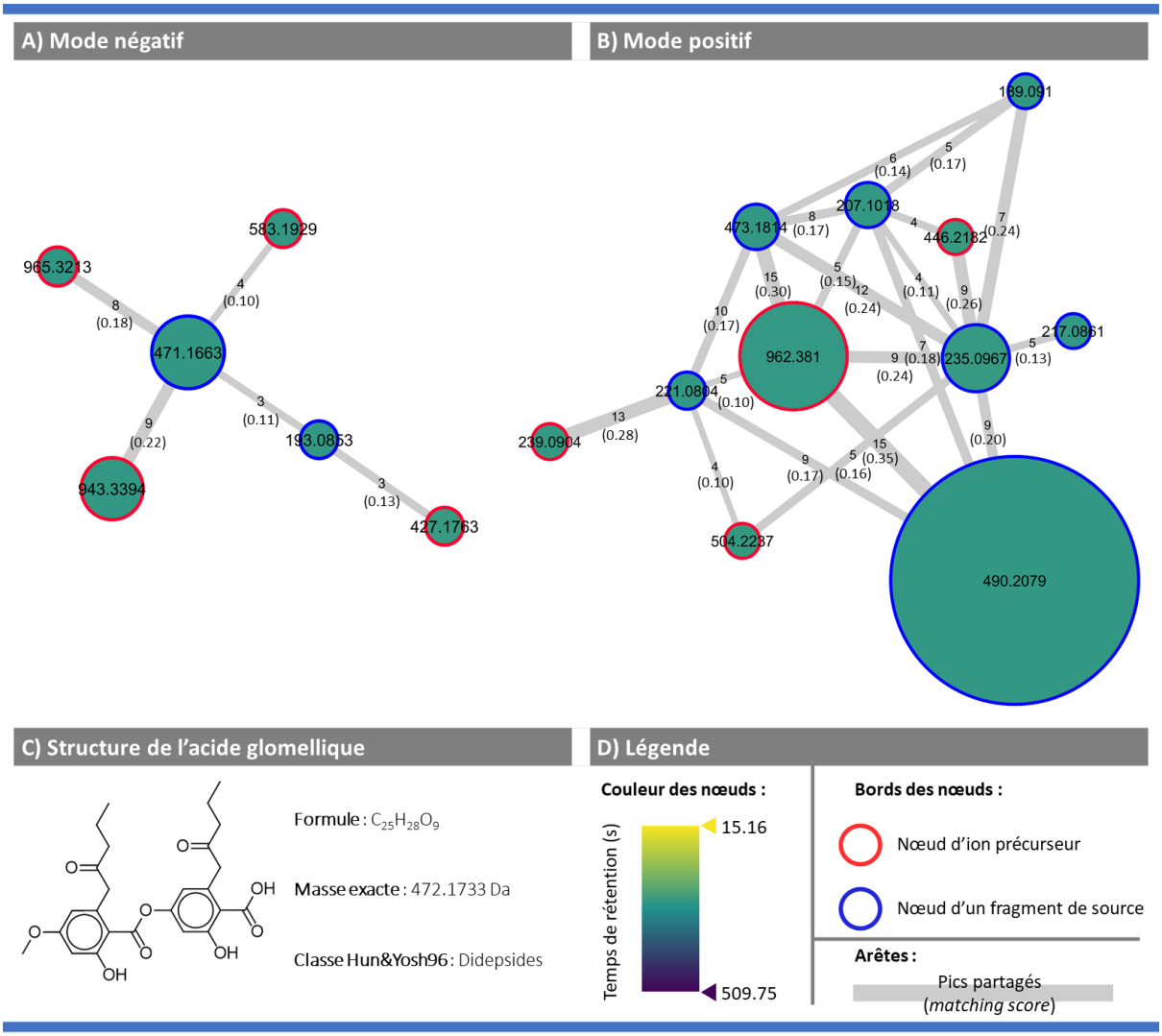


Figure 60 – Blocs de fragmentation correspondant aux ions de l'acide glomellique. Ces blocs sont visualisables dans les réseaux de fragmentation générés par Fragnotator. Les taille des nœuds est

porportionelle au TIC de chaque spectre MS/MS. A) représente le bloc du mode négatif, B) celui du mode positif et C) la structure de l'acide glomellique. L'épaisseur des arêtes est proportionnelle au matching score.

3.4 *Adnotator* : Prédiction des adduits & des molécules.

Après traitement par *Fragnotator*, les ions non annotés en tant que fragments de source sont traités par *Adnotator* pour prédire leur forme d'ionisation à partir du *Tableau d'adduits* réduit et générer les molécules hypothétiques les plus probables. Le cas de l'acide glomellique est illustré en **Figure 61**. Quatre types de nœuds sont observables dans ces réseaux : les nœuds bleus et rouges correspondant respectivement aux fragments de source et aux précurseurs en sortie de *Fragnotator*, les nœuds gris correspondant aux *self-looped* de *Fragnotator*, et les nœuds noirs correspondant aux molécules neutres hypothétiques. Par rapport aux blocs de la **Figure 60**, le nombre de nœuds est ici augmenté, de 6 à 10 pour le mode négatif (**Figure 61-A**) et de 11 à 19 pour le mode positif (**Figure 61-B**). Ceci est dû au rajout des nœuds neutres (noirs) mais aussi au regroupement avec des nœuds précédemment *self-looped* (gris) ainsi qu'avec d'autres blocs de fragmentation (partie inférieure du bloc positif avec les ions à *m/z* 495 et 967). Les ions issus d'une même molécule ne présentent pas tous une fragmentation similaire, du moins avec les paramètres choisis pour *Fragnotator*. En amont de la déréplication orientée, la précision de la prédiction des masses de neutres a été rapportée pour l'acide glomellique : 1.2 et -0.2 ppm pour les modes négatif et positif respectivement. Par ailleurs, d'autres nœuds neutres ont été générés : ceci peut arriver avec les nœuds rouges qui sont également soumis à la recherche de neutres s'ils ne sont pas déjà directement connectés à un nœud noir. Dans ce cas, l'adduit $[M+H+HCOOH]^-$ a été trouvé en complément du $[M-H]^-$ en mode négatif, et l'adduit $[M+H]^+$ combiné au $[M+NH_4]^+$ pour le mode positif, permettant dans les deux cas de mettre en évidence un fragment de source de l'acide glomellique, en l'occurrence suite à une perte de son acide carboxylique. Les erreurs de mesure ont également été calculées et les neutres hypothétiques ont une masse proche de la structure proposée.

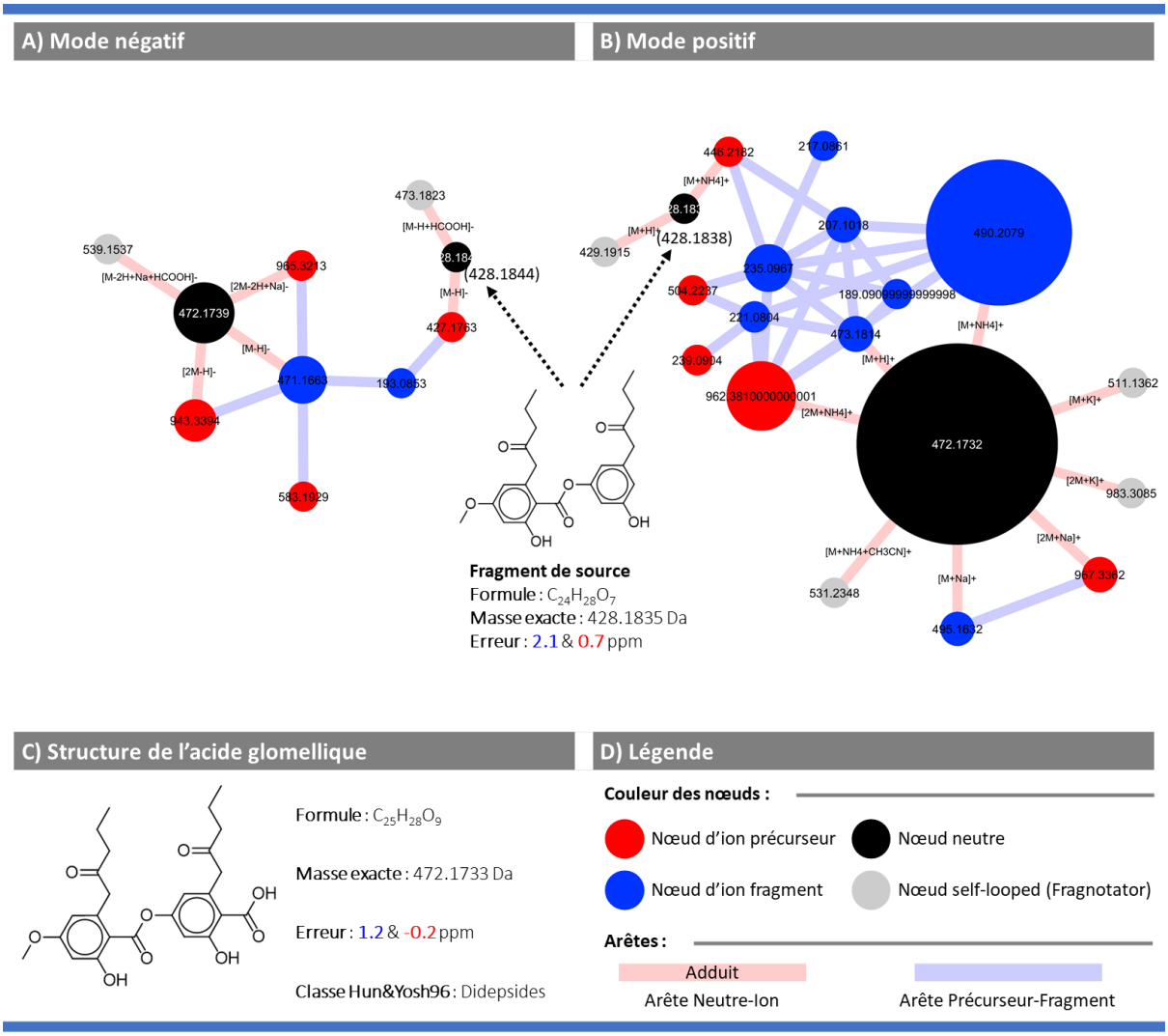


Figure 61 – Blocs moléculaires de l'acide glomellique après traitement par Adnotator. A) mode positif, B) mode négaitf. Un fragment de source a également été détecté sous la forme de nœud neutre. En amont de la déréplication orientée, les erreurs de mesure sont rapportées pour chaque structure en bleu (mode négatif) et en rouge (mode positif).

3.5 *Classnotator*, déréplication orientée & *Mixmoder*.

Les fichiers MGF ont été soumis à MS2LDA pour détecter les motifs de la LDB motifDB qui y étaient présents. Grâce aux *motifs purs*, les classes structurales Hun&Yosh96 ont été prédites pour les ions et les neutres. Les nœuds ont été davantage annotés avec la LDB en orientant la déréplication par les adduits prédits et le temps de rétention. Les données des deux modes d'ionisation ont ensuite été combinées pour ne former qu'un bloc par molécule, de nature mixte. L'exemple de l'acide glommellique est présenté en **Figure 62**.

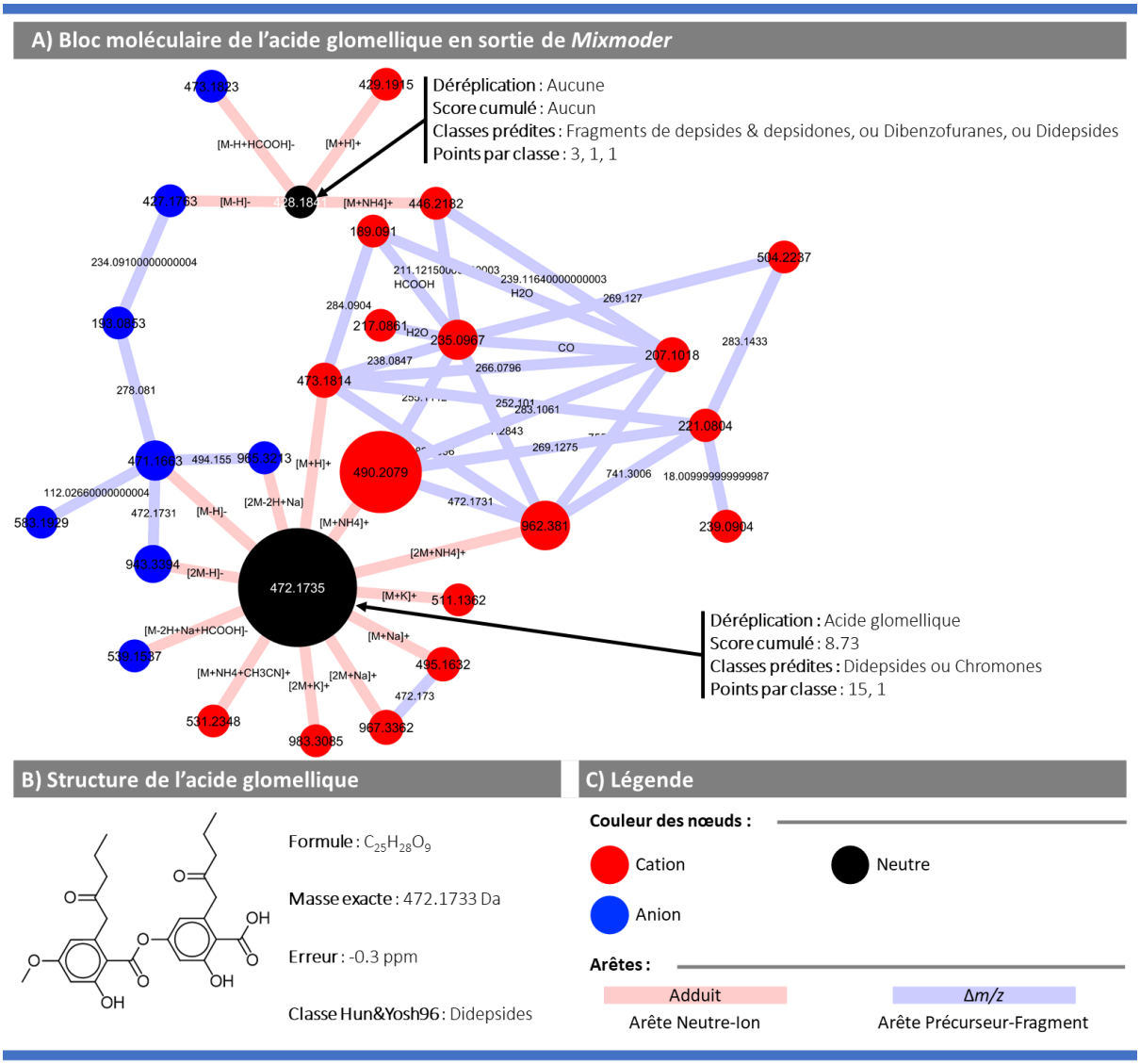


Figure 62 – Bloc moléculaire de l'acide glomellique après traitement par Mixmoder. Les déréplications et les classes prédites n'ont été représentées que pour les nœud neutres, représentatifs de leurs blocs.

Concernant les formes détectées, *Molnotator* a permis de mettre en évidence un adduit supplémentaire pour l'acide glomellique par rapport à la LDB : $[M-2H+Na+HCOOH]^-$ *m/z* 539.1537. Pour le mode positif, ce sont deux adduits supplémentaires qui peuvent être rajoutés : $[2M+K]^+$ m/z 983.3085 et $[M+NH_4+CH_3CN]^+$ *m/z* 531.2348.

Plusieurs nœuds ont été dérépliqués en tant qu'acide glomellique et ceci peut être observé sur le nœud neutre qui porte l'ensemble des identifications et la somme des scores de cosinus pour chacune. Ici, seulement l'acide glomellique a été trouvé avec un score cumulé de 8.73. La classe Hun&Yosh96 prédite est celle des depsides (didepsides) avec 15 points contre celle des chromones qui n'ont qu'un point. Son fragment de source n'a pas pu être dérépliqué, chose attendue au vu de l'absence de la plupart des fragments de source des bases de données. *Classnotator* a en revanche pu le classifier en tant que Fragment de depsides et depsidones, ce qui semble cohérent dans ce contexte, bien qu'il puisse également être considéré comme un didepside au vu de sa structure. Les autres fragments de sources mériteraient également d'être étudiés dans le détail pour les associer à une structure.

A titre de comparaison avec ces réseaux de polarité mixte, des réseaux par similarité cosinus ont été produits pour l'acide glomellique (**Figure 63**). Il peut y être observé que les nœuds du bloc de l'acide glomellique seraient non seulement répartis sur deux polarités différentes, mais aussi dispersés dans différents blocs, mêlés à d'autres nœuds qui ne sont pas attribuables à cette molécule, ou encore sous la forme de nœuds *self-looped*.

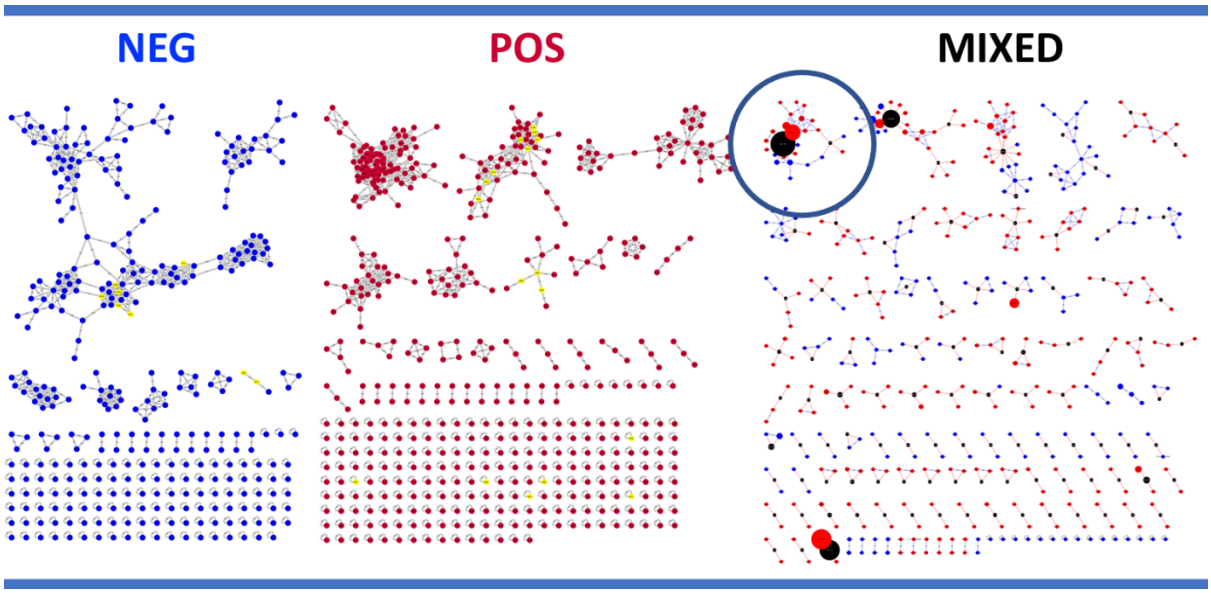


Figure 63 – Comparaison des réseaux générés par similarité cosinus au réseau de polarité mixte généré par Molnotator. Le bloc moléculaire de l'acide glomellique est encerclé et les nœuds correspondants ont été colorés en jaune dans les deux autres réseaux.

Le reste des molécules de la LDB traitées par *Molnotator* sont consultables en **Annexe** (**Figures S-2** à **S-136**) sous la forme de leurs blocs moléculaires.

3.6 Interprétation des résultats & pistes d'amélioration.

Quelques chiffres sur le traitement de la LDB par *Molnotator* sont présentés dans le **Tableau 27**.

Tableau 27 – Récapitulatif du traitement de la LDB-Orbitrap par Molnotator.

	Mode négatif	Mode positif	Total		Réussites	Echecs	Total
Spectres LDB-Orbitrap	335	367	702	Blocs moléculaires	133 (86%)	21 (14%)	154
Molnotator : adduits	420	555	975	*Classnotator*	90 (68%)	43 (32%)	133
Molnotator : fragments	401	407	808	Déréplication guidée	121 (91%)	12 (9%)	133
Molnotator : spectres totaux	821	1329	2150				

Sur les 167 molécules analysées sur l'Orbitrap, 154 avaient pu être intégrées dans la LDB, représentées par au moins un spectre (*Chapitre III*). Ces 154 molécules ont été traitées par *Molnotator* et 133 ont formé des blocs moléculaires. L'absence de blocs pour les 21 molécules restantes (consultables en annexe : **Tableau S-1**, **Figure S-1**) s'explique par deux facteurs principaux : l'absence d'ions essentiels pour la formation du bloc moléculaire ainsi que des erreurs dans la LDB.

L'absence d'ions essentiels peut être due à des Δ*m/z* et ΔTR trop élevés pour les fenêtres utilisées, ce qui est notamment observé pour les molécules de faible poids moléculaire. 10 de ces 21 molécules présentent une masse en dessous de 250 Da et à ce stade, une fenêtre utilisant une valeur fixe en ppm n'est plus adaptée. Une fenêtre mouvante en fonction du rapport *m/z* des ions comparés pourrait être envisagée. Dans d'autres cas, l'absence d'ions essentiels est due à leur suppression par MZmine, chose inévitable même avec les paramètres utilisés.

En ce qui concerne les erreurs dans la LDB, certaines ont pu être constatées après l'utilisation de *Molnotator* : la prise en compte du contexte dans lequel chaque ion est détecté a permis de remettre en cause l'usage de certains adduits complexés au méthanol (CH_3OH) et à l'acide formique (HCOOH), notamment par des ΔTR trop élevés. Ces quelques entrées erronées sont à éliminer de la LDB.

Parmi les 133 molécules ayant produit un bloc moléculaire, les masses calculées des molécules hypoéthétiques présentaient une valeur proche de celle qui était attendue, avec une erreur moyenne de 1.8 ppm. 43% des prédictions avaient une précision sous 1 ppm en valeur absolue (**Figure 64-A**). La précision avec laquelle les masses des neutres peuvent être prédites dépend directement des ions associés et donc à la précision de mesure de l'instrument. L'utilisation d'une valeur moyenne à partir des rapports *m/z* de tous les ions permet de réduire les erreurs de mesure.

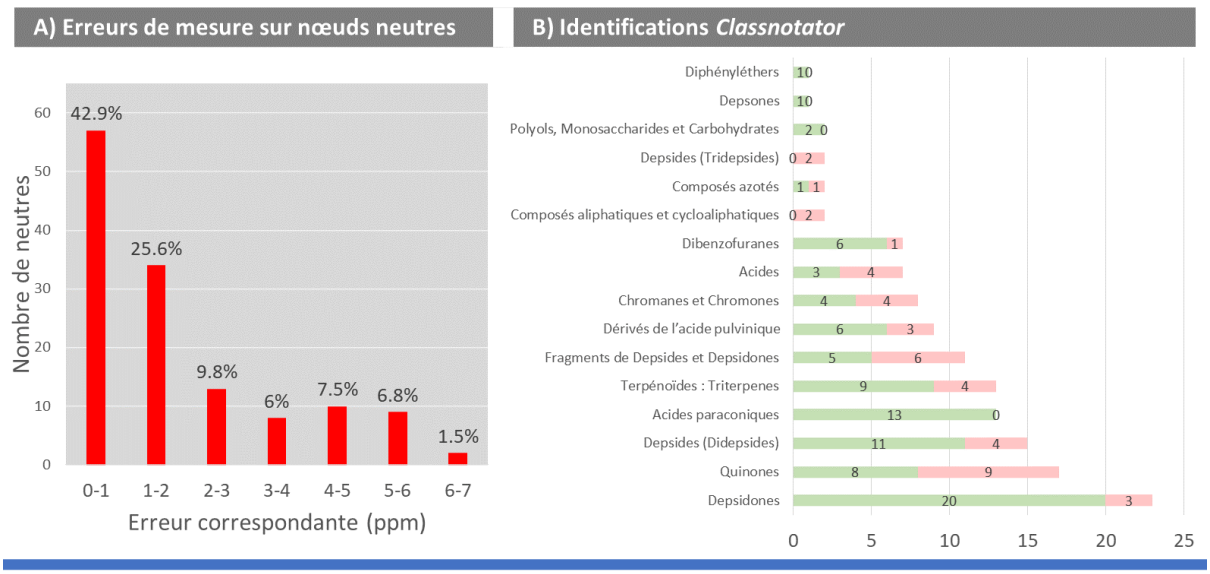


Figure 64 – Résultats de Molnotator sur la prédiction des nœuds neutres et les identifications par Classnotator. A) Histogramme représentant la répartition du degré de précision en ppm (valeurs absolues) avec lequel les masses des nœuds neutres ont été prédites, B) Proportions, pour chaque classe Hun&Yosh96 représentée dans la LDB-Orbitrap, de molécules dont la classe a été correctement prédite (vert) sur le total des molécules (rouge).

Un total de 975 spectres (420 ESI^-, 555 ESI^+) ont pu être détectés par *Molnotator*, contre les 702 déjà présents dans la LDB pour l'Orbitrap (335 ESI^-, 367 ESI^+). Cette augmentation est attribuable à une meilleure définition des adduits à rechercher. Une vérification manuelle des spectres sera cependant nécessaire avant l'intégration à la LDB. Additionnellement, 808 fragments de source ont également pu être détectés (401 ESI^-,

407 ESI^+), bien qu'il puisse s'agir également d'autres adduits non répertoriés ou de molécules proches coéluées. Une étude minutieuse de chaque spectre permettrait de leur attribuer une structure. Leur intégration dans des bases de données doit cependant attendre l'adaptation des algorithmes de déréplication utilisés par la communauté des produits naturels.

En ce qui concerne la prédiction des classes Hun&Yosh96, *Classnotator* a permis leur prédiction pour 90 neutres sur 133, soit 68% (**Figure 64-B**). L'efficacité de la prédiction est hautement dépendante de la base de données MS/MS utilisée par *Classnotator* : des classes distinguables en MS/MS et hautement représentées comme les depsidones et les acides paraconiques sont facilement repérables. A l'inverse, des classes peu représentées, peu différenciables par MS/MS ou hétérogènes comme les « Composés azotés », les « Fragments de depsides et depsidones » et les « Acides » sont plus difficiles à prédire. L'efficacité de *Classnotator* dépend également de la taille du bloc moléculaire généré, des classes comme les quinones génèrent des blocs discrets ce qui réduit la quantité de données disponible pour la triangulation. Les classes Hun&Yosh96 pourraient être davantage revues pour créer des classes plus homogènes, comme ce qui a été fait pour les « Acides paraconiques » et les « Mycosporines ». Un dernier point permettrait d'améliorer l'efficacité de *Classnotator* : l'agrandissement de la LDB avec une meilleure représentation des classes structurales des molécules lichéniques. Le tableau de *motifs purs* serait plus riche et permettrait de mieux trianguler les classes des ions.

Pour finir, sur les 133 molécules, 121 ont pu être correctement identifiées grâce à la déréplication guidée, réduisant le taux de faux positifs comme établi dans la *Chapitre III*. Les mauvaises identifications sont dues à des isomères coélués avec les analytes, rendant la déréplication difficile (eugénitine et isoeugénitine, plusieurs déchloro-4-*O*-méthyldiploïcines ainsi que des isomères de terpènes et d'acides paraconiques).

En l'état actuel, *Molnotator* traite les échantillons individuellement bien qu'ils soient issus du même traitement sur MZmine. Une triangulation entre échantillons serait grandement bénéfique pour la création des blocs moléculaires et permettrait de faire un réseau unique pour l'ensemble des échantillons. Ceci permettrait de résoudre le principal problème derrière le fonctionnement de *Molnotator* : la création des blocs moléculaires dépend du nombre d'ions identifiables générés par la molécule. Ainsi, une molécule hautement concentrée / ionisable devrait générer assez d'ions pour former un bloc, alors qu'une molécule peu abondante ou difficilement ionisable formera peu d'ions et ses chances de créer un bloc sont amoindries. S'il s'agit d'un problème de concentration, la triangulation sur un autre échantillon où elle sera plus concentrée permettra de régler ce problème. Ceci a pu être observé pendant cette expérience, par exemple avec le diacétylpyxinol qui, détecté dans un autre échantillon, est ionisé sous d'autres formes (**Figure 65**). La détection d'autres formes d'ionisation d'un échantillon à l'autre s'explique par le contexte d'ionisation de la molécule, par la suppression d'ions et également parce que les ions sont en concurrence pour être sélectionnés en DDA. Dans un autre échantillon où le nombre d'ions détectés au même moment est moindre, d'autres adduits peuvent être fragmentés.

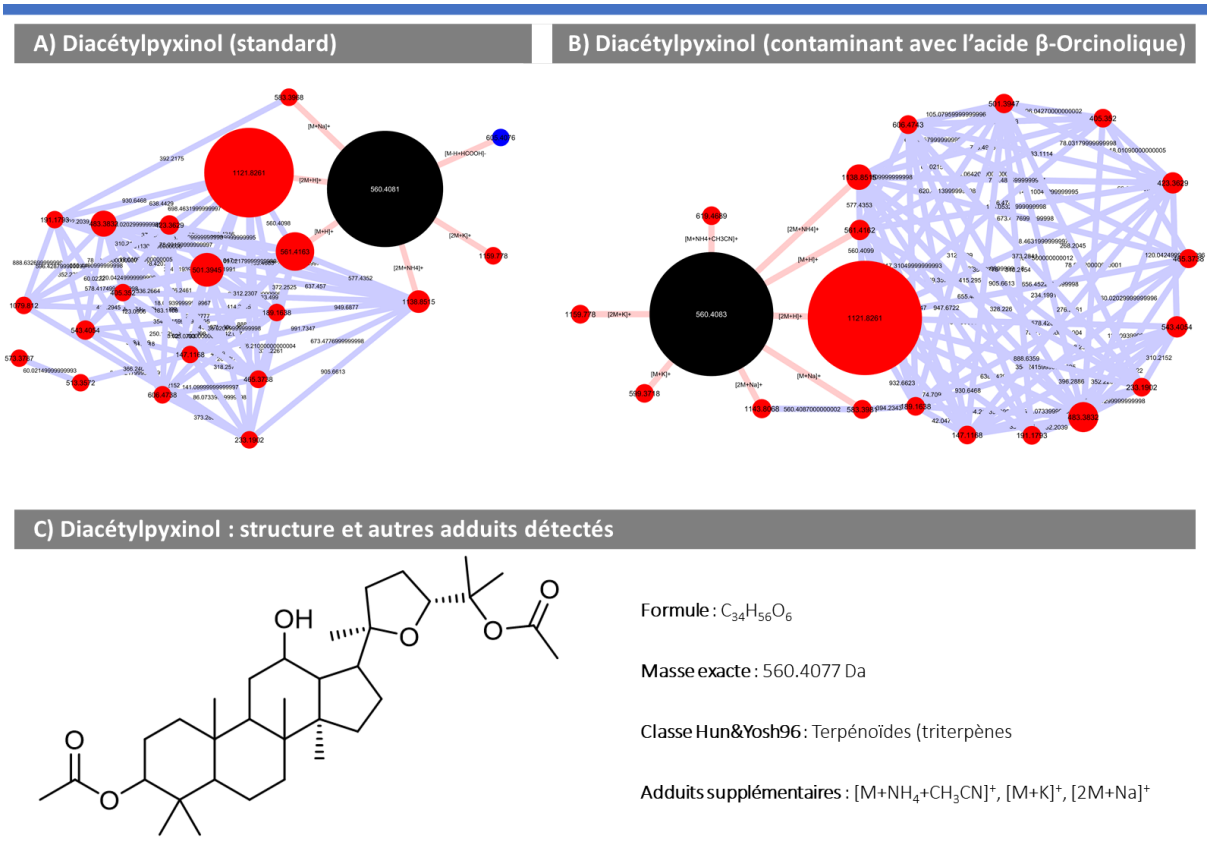


Figure 65 – Blocs moléculaires du diacétylpyxinol à partir de l'analyse de son standard puis du standard d'une autre molécule, mettant en évidence d'autres formes d'ionisation.

L'intégration d'une triangulation entre échantillons augmente cependant de façon considérable le risque d'explosion combinatoire : *Molnotator* bénéficierait de l'usage de la théorie des graphes pour faciliter le traitement d'un ensemble d'échantillons

La mise en place de critères minimaux pour la formation d'un bloc moléculaire pourrait être envisagée pour réduire les chances de créer un faux bloc de façon fortuite. Par exemple, un bloc composé d'un ion $[2M+Cl]^-$ et d'un $[M-2H+K]^-$ semble peu probable et pourrait être éliminé car il ne contiendrait pas des formes ubiquitaires comme $[M-H]^-$ et/ou $[2M-H]^-$.

Un défaut mineur d'*Adnotator* pourrait se corriger facilement : la différenciation entre les monomères et les dimères. La sélection des hypothèses par leur simplicité implique que si, pour une molécule, seuls des dimères ont été ionisés, ils seront considérés être des monomères car il s'agit de l'explication la plus simple des données. Bien que ce soit souvent vrai, dans certains cas il pourrait bel et bien s'agir de dimères au vu du contexte d'ionisation (ex : **Figure S-24**, neutre à 914.3704 ionisé sous forme $[M+NH_4]^+$ & $[M+H]^+$, **Figure S-27**, neutre à 986.4277 ionisé sous forme $[M+Na]_+$ & $[M+NH_4]^+$ etc…). Ceci pourrait être corrigé par une simple inspection automatique des pics de leurs spectres MS/MS.

En plus des spectres d'adduits et des fragments de source, *Molnotator* a permis de détecter quelques dérivés structuraux non répertoriés dans la LDB. Ces dérivés se trouvent parfois dans le même bloc que l'analyte de par des fragments de source partagés.

Une étude plus approfondie du contenu de ces standards pourrait permettre d'isoler ces dérivés.

Les ions multichargés n'ont pas été abordés ici mais des modules appropriés pourraient être rajoutés pour les intégrer au traitement.

4 Conclusion

Molnotator a permis de détecter de façon rapide et automatique la plupart des ions de la LDB-Orbitrap et quelques centaines de spectres d'adduits supplémentaires pourraient être rajoutés à la LDB. La représentation des molécules dans leur contexte d'ionisation permet de mieux comprendre les données qu'avec un simple regroupement par similarité cosinus. Ces regroupements ne sont pas pour autant mutuellement exclusifs et pourraient être combinés lors d'interprétation de données de métabolomique. Ceci est vrai pour nombre d'autres outils utilisés en métabolomique, comme la déréplication avec des spectres *in silico*, des pondérateurs sur la base de la taxonomie (Taxoponderator, Chromanot), MolNetEnhancer etc…

L'usage un tel outil pourrait remettre à l'ordre du jour l'utilisation des acquisitions LC-MS avec des polarités mixtes pour une même analyse, telles que proposées sur certains instruments. Ceci, à condition d'adapter les débits de solvant et la durée d'analyse pour compenser la quantité de données à acquérir.

Il pourrait être soulevé que l'utilisation d'adduits ne permet d'appliquer cet outil qu'aux composés majoritaires / fortement ionisés dans un échantillon de métabolomique. Cependant, la triangulation entre échantillons permettrait de compenser les différences d'intensité pour un ion d'un échantillon à l'autre, comme il sera démontré dans le chapitre suivant.

Des approches similaires commencent à émerger et même si *Molnotator* reste au stade expérimental, certains des principes exposés ici peuvent être transposés à l'avenir sur d'autres logiciels.

Chapitre VI

– Diversité chimique des lichens –

Et l'élucidation de celle-ci par l'analyse de 300 échantillons de lichens par LC-MS et le traitement par *Molnotator*

Intervenants extérieurs : Zakaria Bouchouireb[(a)], Simon Ollivier[(b)], Françoise Lohézic-Le Dévéhat[(b)], Joël Esnault[(b)], Pierre-Marie Allard[(c)], Jean-Luc Wolfender[(c)].

(a) : CNRS, CEISAM (Chimie et Interdisciplinarité : Synthèse, Analyse, Modélisation)-UMR 6230, Univ Nantes, F-44322 Nantes, France
(b) : CNRS, ISCR (Institut des Sciences Chimiques de Rennes)-UMR 6226, Univ Rennes, F-35000 Rennes, France
(c) : EPGL, Université de Genève, Université de Lausanne, CMU, 1 Rue Michel Servet, 1211 Genève 4, Suisse

Contributions externes : ZB – *Conseils techniques & optimisation de certains passages du code*. SO – *Conseils théoriques & discussion des résultats*. FLD – *Mise à disposition de certains échantillons de lichens.* JE – *Identification, préparation et mise à disposition de la plupart des échantillons lichéniques*. PMA, JLW – *Accueil à Genève, partage du protocole d'extraction et de préparation des échantillons, analyses LC-MS*.

Résumé

Après avoir validé *Molnotator* sur la LDB, celui-ci est utilisé pour étudier la composition chimique de 300 échantillons de lichen et ainsi donner une image approximative de ce qui reste à découvrir dans ces champignons. Ces échantillons ont été prélevés dans plusieurs herbiers de l'Université de Rennes 1 pour être analysés par LC-MS sur une Orbitrap Q-Exactive de l'Université de Genève. Les données dans les deux modes d'ionisation ont été traitées sur MZmine comme dans un FBMN tel que décrit précédemment, suite à quoi elles ont été soumises à *Molnotator*. Ce dernier été amélioré pour traiter le volume conséquent de données et a permis de mettre en évidence un nombre important de molécules inconnues. L'interprétation finale nécessitera cependant une vérification manuelle des données, notamment l'intégration des molécules connues mais absentes des bases de données à la LDB. Néanmoins, *Molnotator* permet d'ores et déjà de facilement cibler les molécules inconnues dans un lichen donné dans le but de les isoler.

Sommaire

1 – Introduction 159

2 – Méthodes 160

2.1 Solvants utilisés 160

2.2 Echantillons de lichens utilisés 160

2.3 Extraction des échantillons 160

2.4 Filtration sur SPE 160

2.5 Préparation des échantillons pour l'analyse 160

2.6 Analyse LC-MS 160

2.7 Conversion des données et traitement par MZmine 161

2.8 Paramètres *Molnotator* 162

3 – Résultats 164

3.1 Traitement MZmine & *Molnotator* 165

3.2 Interprétation des résultats pour *Evernia prunastri* 168

3.3 Interprétation des résultats pour *Cladonia gracilis* 177

4 – Conclusion 186

4.1 De la contribution de chaque mode d'ionisation 186

4.2 Des constituants des analyses LC-MS 186

4.3 Des avantages de *Molnotator* & pistes d'amélioration 186

4.4 La diversité chimique des lichens & comment mieux l'étudier 187

1 Introduction

Dans les chapitres précédents, des outils pour faciliter les études des lichens dans le contexte de la métabolomique ont été développés. Dans le *Chapitre I*, la LDB-Lit a été créée, regroupant 1662 molécules lichéniques issues de la bibliographie avec leur source biologique. Dans le *Chapitre II*, la LDB a été créée, une base de données spectrale couvrant 250 composés lichéniques. Le couplage de la LDB-Lit à la LDB permet de grandement faciliter l'interprétation des données LC-MS/MS, mais cela reste insuffisant à ce stade. Dans le *Chapitre III*, la LDB a été étendue avec de nombreux spectres en considérant les différents adduits de chaque molécule. L'étude de la diversité de ces adduits suggère qu'ils devraient être systématiquement intégrés aux bases de données et que l'identification de la forme d'ionisation est indispensable pour pouvoir interpréter l'ensemble des données LC-MS. La comparaison des spectres acquis sur différents instruments souligne l'importance d'avoir plusieurs signaux pour une même molécule, puisque certains de ces spectres ne seront pas reconnus. Dans le *Chapitre IV*, une base de données de motifs à été créée sur MS2LDA grâce à la LDB (la LDB-motifDB) dans le but de faciliter l'annotation des nœuds inconnus. Une méthode pour automatiser cette annotation des nœuds selon leur classe structurale a été développée : *Classnotator*. Dans le *Chapitre V*, une nouvelle méthode d'interprétation des données LC-MS a été développée, tenant compte des fragments de source, des adduits, des classes structurales et permettant de prédire les molécules en les représentant dans leur contexte d'ionisation : *Molnotator*. Dans ce VIe et dernier chapitre, *Molnotator* est utilisé pour annoter les données LC-MS issues de l'analyse de 300 échantillons de lichens sur l'Orbitrap Q-Exactive Focus de Genève. L'outil a été améliorer pour permettre le traitement d'un tel volume de données : *Adnotator* a été utilisé en deux fois pour éviter l'explosion combinatoire et un module a été créé (*File merger*) pour trianguler les résultats d'*Adnotator* entre les échantillons. Cette dernière triangulation permet, pour chaque échantillon, d'apporter des informations sur ses nœuds non-interprétés en les cherchant dans les autres échantillons où ils ont été mieux détectés. La quantité d'information générée sur ces 300 échantillons est trop grande pour être abordée textuellement. Un résumé de ces résultats a néanmoins été produit en **Annexe**, et deux exemples ont été utilisés plus bas pour illustrer l'intérêt de *Molnotator* : *Evernia prunastri* et *Cladonia gracilis*.

2 Méthodes

2.1 Solvants utilisés.

Plusieurs solvants de qualité UPLC ont été utlisés pour les extractions et les analyses : méthanol, dichlorométhane, acétone et acétonitrile. L'eau utilisée était de l'eau MilliQ et le DMSO pour solubiliser les extraits était de qualité biologique.

2.2 Echantillons de lichens utilisés.

Un total de 299 échantillons a été prélevé dans différents herbiers, principalement ceux de Rennes. 123 espèces sont représentées, ainsi que 90 autres échantillons qui n'ont pas été déterminés avec certitude. La liste des espèces ainsi que celle des échantillons est disponible en **Annexe** dans les **Tableaux S-3** et **S-4**. Quelques dizaines de milligrammes de chacun des échantillons ont été prélevés et stockés dans des tubes Eppendorf safe-lock de 2 mL.

2.3 Extraction des échantillons.

Les tubes Eppendorf sont plongés dans l'azote liquide puis l'échantillon à l'intérieur est broyé avec un pilon en téflon. Après le broyage, 1 mL de dichlorométhane a été ajouté dans chaque tube avec des billes en verre. L'extraction est réalisée sous agitation à l'aide d'un broyeur à billes Qiagen (TissueLyser II) à 3000 rpm pendant 3 minutes. L'extrait est transféré dans des plaques deep-well. Une deuxième extraction sur les mêmes échantillons est faite, cette fois avec de l'acétone, et les extraits sont réunis avec ceux au dichlorométhane. L'ensemble est évaporé sous flux d'azote jusqu'à séchage des échantillons avant de passer à l'étape suivante.

2.4 Filtration sur SPE.

Une plaque SPE Discovery DSC-18 575603-U (Supelco) est conditionnée avec 3x1 mL d'eau MilliQ puis 3x1 mL de méthanol. Les extraits séchés sont dissous dans 1 mL de méthanol et déposés sur la plaque SPE pour être récupérés dans des plaques à tubes séparables pré-tarés. L'élution est favorisée à l'aide d'un *Vacuum manifold* permettant d'isoler le système de façon hermétique et de mettre sous vide.

2.5 Préparation des échantillons pour l'analyse.

Chacun des extraits a été dissous dans du DMSO à une concentration de 5 mg/mL avant leur transfert dans des plaques 96 puits UPLC-MS.

2.6 Analyse LC-MS.

La séparation chromatographique a été réalisée sur un système Acquity UHPLC (Waters, Milford, MA, USA) interfacé à un spectromètre de masse Q-Exactive Focus (Thermo Scientific, Brême, Allemagne), utilisant une source d'ionisation par électrospray chauffée (HESI-II). Les conditions de la LC étaient les suivantes : colonne : Waters BEH C18 100x2.1

mm, 1.7 µm ; phase mobile : (A) eau avec 0.1 % d'acide formique ; (B) acétonitrile avec 0.1 % d'acide formique ; débit : 600 µL/min ; volume d'injection : 2 µL ; gradient. Les paramètres optimisés de la source HESI-II étaient les suivants : tension de source : 3,5 kV, débit de gaz (N_2) : 48 unités ; débit de gaz auxiliaire : 11 unités ; débit de gaz de réserve : 2.0 ; température capillaire : 256.2 °C (pos), niveau RF de S-Lens : 45. Le spectromètre de masse a été calibré à l'aide d'un mélange de caféine, de méthionine-arginine-phénylalaninephénylalanine-alanine-acétate (MRFA), de sulfate de dodécyle de sodium, de taurocholate de sodium et d'Ultramark 1621 dans une solution acétonitrile/méthanol/eau contenant 1% d'acide formique par infusion directe. Les acquisitions MS/MS données-dépendantes ont été réalisés sur les 3 ions les plus intenses détectés dans la MS à balayage complet (expérience Top3). La largeur de la fenêtre d'isolation MS/MS était de 1 Da, et l'énergie de collision normalisée (NCE) a été fixée à 15, 30 et 45 unités. Dans les expériences MS/MS dépendantes des données, des balayages complets ont été acquis à une résolution de 35 000 FWHM (à *m/z* 200) et des balayages MS/MS à 17 500 FWHM avec un temps d'injection maximum automatique. Après avoir été acquis dans les scans MS/MS, les ions parents ont été placés dans une liste d'exclusion dynamique pendant 2.0 secondes.

2.7 Conversion des données et traitement par MZmine.

Les fichiers propriétaires raw sont convertis au format mzXML par le module MSConvert de ProteoWizard. Tous les fichiers sont importés sur MZmine 2.39 pour Linux et sont traités à l'aide du cluster BAOBAB de Genève. Les paramètres utilisés sont présentés dans le **Tableau 28**.

Tableau 28 – Paramètres MZmine. Si des paramètres différents ont été utilisés pour chaque mode d'ionisation, ils ont été marqués en rouge pour le mode positif et bleu pour le négatif.

Module	*Paramètres*
Raw data import	Importation de tous les fichiers mzXML
Raw data methods > Mass detection	Scans: MS level: 1 Mass detector: centroid Noise level: 5.0E5, 2.0E5 Mass list name: masses
Raw data methods > Mass detection	Scans: MS level: 2 Mass detector: centroid Noise level: 0 Mass list name: masses
Peak list methods > Peak detection > ADAP Chromatogram builder (Myers et al. 2017)	Scans: MS level: 1 Mass list: masses Min group size in # of scans: 5 Group intensity threshold: 5.0E5, 2.0E5 Min highest intensity: 5.0E5, 2.0E5 m/z tolerance: 15 ppm

Tableau 28 – Suite.

Module	*Paramètres*
Peak list methods > Peak detection > Chromatogram deconvolution	Algorithm: Wavelets (ADAP): S/N threshold: 10 S/N estimator: Wavelet Coeff. SN: Peak width mult.: 3 abs(wavelet coeffs.): checked min feature height: 5.0E5, 2.0E5 coefficient/area threshold: 150 Peak duration range: 0.00-1.00 RT wavelet range: 0.00-0.20 m/z center calculation: MEDIAN m/z range for MS2 scan pairing (Da): 0.03 RT range for MS2 scan pairing (min): 0.1
Peak list methods > Isotopes > Isotopic peaks grouper	m/z tolerance: 8.0ppm Retention time tolerance: 0.08 absolute (min) Monotonic shape: unchecked Maximum charge: 2 Representative isotope: Lower m/z
Peak list methods > Alignment > Join aligner	m/z tolerance: 15.0ppm Weigth for m/z: 1 Retention time tolerance: 0.2 absolute (min) Weight for RT: 1 Require same charge state: unchecked Require same ID: unchecked Compare isotope pattern: unchecked
Peak list methods > Filtering > Peak list rows filter	Keep only peaks with MS2 scan (GNPS): checked Reset the peak number ID: checked (rest is unchecked / unused)
Peak list methods > Export/Import > Export for/Submit to GNPS	Mass list: masses Merge MS/MS (experimental): checked Select spectra to merge: across samples m/z merge mode: weighted average (remove outliers) intensity merge mode: sum intensities Expected mass deviation: 5.0ppm Cosine threshold (%): 70.0 Peak count threshold (%): 40.0 Isolation window offset (m/z): 0.0 Isolation windows width (m/z): 3.0 Filter rows: ALL Submit to GNPS: unchecked Open folder: Unchecked
Peak list methods > Export/Import > Export to CSV file	Field separator: “,” Export common elements: Export row ID: checked Export row m/z: checked Export row retention time: checked Export data file elements: Peak area: checked Export quantitation results and other information: unchecked Identification separator: “;” Filter rows: ALL
Project > Save project	Enregistrement du projet

Les paramètres utilisés ici sont différents de ceux utilisés lors de la démonstration de *Molnotator* dans le *Chapitre V* : compte tenu de la quantité de données, le risque d’explosion combinatoire est bien trop grand avec des paramètres aussi permissifs. Les listes de pics de MZmine est exportée au format MGF et les attributs de chaque pic au format CSV.

2.8 Paramètres Molnotator.

Les fichiers MGF sont divisés en autant d’échantillons avec le module *Sample Slicer*.

Les fragments de source ont été annotés par *Fragnotator* avec les paramètres suivants : min_shared_peaks = 2, matching_score = 0.1, $\Delta m/z$ = 5 ppm et ΔTR = 7 secondes.

Les adduits ont été détectés et les neutres générés par *Adnotator* en deux étapes, d'abord en utilisant les adduits les plus fréquents (**Tableau 29** & **30**, batch 1) et ensuite en n'appliquant que les *Hypothèses d'Ion 2* sur les neutres déjà générés et le reste des adduits (**Tableau 29** & **30**, batch 2). Ce compromis permet d'éviter l'explosion combinatoire en réduisant légèrement la qualité du traitement.

Tableau 29 – Liste d'adduits utilisés par Adnotator en mode négatif. 8 adduits sur 12 utilisés dans le premier batch.

Adduit	*Charge*	*Δm*	*x*	*C*	*Batch*
$[M-H+HCOOH]^-$	-1	44.9976	1	3	1
$[M-H]^-$	-1	-1.0078	1	1	1
$[M+Cl]^-$	-1	34.9688	1	2	1
$[M-2H+Na+HCOOH]^-$	-1	66.9796	1	5	1
$[M-2H+Na]^-$	-1	20.9741	1	4	1
$[2M-H+HCOOH]^-$	-1	44.9976	2	4	2
$[2M-H]^-$	-1	-1.0078	2	3	1
$[2M+Cl]^-$	-1	34.9688	2	3	2
$[2M-2H+Na+HCOOH]^-$	-1	66.9796	2	6	2
$[2M-2H+Na]^-$	-1	20.9741	2	5	1
$[2M-2H+K]^-$	-1	36.9480	2	5	2
$[3M-H]^-$	-1	-1.0078	3	4	1

Tableau 30 – Liste d'adduits utilisés par Adnotator en mode positif. 10 adduits sur 14 sont utilisés dans le premier batch.

Adduit	*Charge*	*Δm*	*x*	*C*	*Batch*
$[M+H+CH_3CN]^+$	1	42.0343	1	3	2
$[M+H+CH_3OH]^+$	1	33.0340	1	3	2
$[M+H]^+$	1	1.0078	1	1	1
$[M+NH_4+CH_3CN]^+$	1	59.0609	1	3	1
$[M+NH_4]^+$	1	18.0343	1	2	1
$[M+Na+CH_3CN]^+$	1	64.0163	1	3	1
$[M+Na]^+$	1	22.9897	1	1	1
$[M+K]^+$	1	38.9637	1	2	1
$[2M+H]^+$	1	1.0078	2	3	1
$[2M+NH_4]^+$	1	18.0343	2	3	1
$[2M+Na]^+$	1	22.9897	2	3	1
$[2M+K]^+$	1	38.9637	2	3	1
$[3M+NH_4]^+$	1	18.0343	3	4	2
$[3M+Na]^+$	1	22.9897	3	4	2

L'ajout du nouveau module *File Merger* (**Figure 66**) permet de compenser les éventuelles pertes dues à ce compromis en combinant des traitements de chaque échantillon pour ne garder que les meilleurs résultats et produire un réseau unique pour chaque polarité. Dans un premier temps, les neutres générés par *Adnotator* sont collectés dans chacun des *node tables* (*Neutral Collector*). Ensuite, la liste de l'ensemble des neutres trouvés est simplifiée de façon ce qu'ils aient le même identifiant à travers tous les échantillons (*Neutral Merger*). Une sélection des neutres est réalisée : un tableau est généré, en comptant pour chaque combinaison Ion-Neutre le nombre de fois qu'elle a été repérée sur l'ensemble des analyses. Pour chaque ion, ne sera conservé que le lien Ion-Neutre

pour le neutre présentant 1) le plus grand nombre d'adduits et 2) le plus grand nombre d'occurrences pour la paire Ion-Neutre (*Neutral Selection*). Après cette étape de reformation des paires Ion-Neutre, les neutres n'étant reliés qu'à 1 ou 0 ions sont supprimés et les paires survivantes sont reportées dans un *edge table* (*Neutral Filtering*). Les paires précurseur-fragment de *Fragnotator* sont importées et conservées telles-quelles (*Fragnotation Merger*). Les nœuds n'étant pas associés dans des paires avec un neutre ou avec un fragment/précurseur sont rapportés dans l'*edge table* en tant que *self-looped nodes* (*Self-looper*). Les statuts de chaque nœud sont mis à jour et les *node table & edge table* représentant l'ensemble des échantillons combinés sont exportés au format CSV.

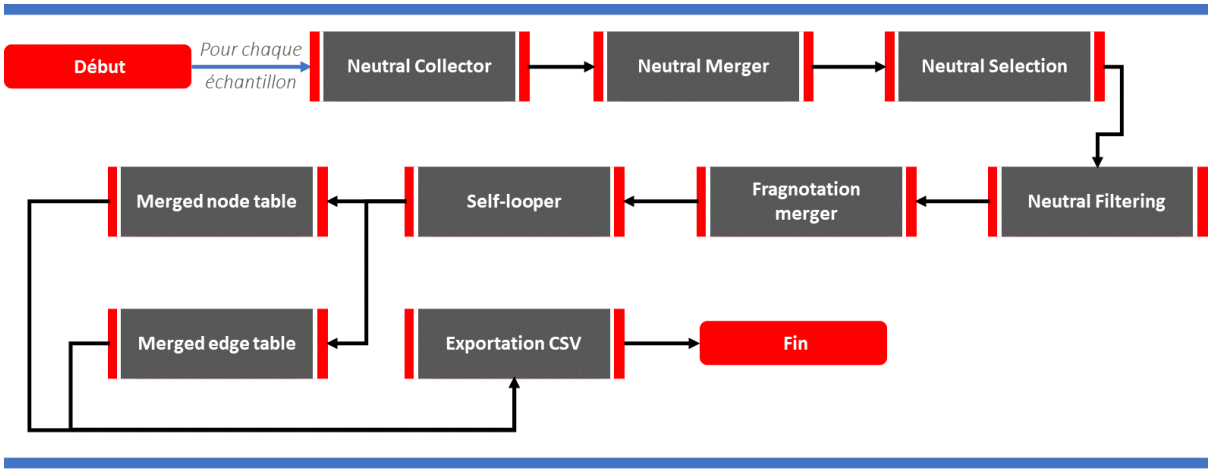


Figure 66 – Fonctionnement général de File Merger.

Pour prédire les classes avec *Classnotator*, les fichiers MGF en sortie de MZmine (POS et NEG) ont été soumis aux serveurs de MS2LDA pour identifier les motifs de la LDB MotifDB (ms2bins = 0.01 Da, Number of Mass2Motifs = 100, Number of iterations for LDA = 1000). Les motifs LDB détectés pour chaque ion sont ensuite importés par *Classnotator* et la prédiction des classes structurales selon Hun&Yosh96 est réalisée en utilisant les mêmes tableaux de *motifs purs* que précédemment, produits dans le *Chapitre IV*.

La déréplication guidée est réalisée avec un seuil de cosinus fixé à 0.7, des fenêtres $\Delta m/z$ = 8 ppm et ΔTR = 8 secondes. Les 3 meilleurs *hits* de déréplication sont conservés (*top_h* = 3).

Les fichiers des deux modes d'ionisation sont combinés par *Mixmoder* ($\Delta m/z$ = 10 ppm et ΔTR = 8 secondes). En plus de la déréplication par la LDB, les neutres sont dérépliqués contre la LDB-Lit en utilisant les masses prédites et un $\Delta m/z$ à 7 ppm.

Le réseau total est ensuite divisé de façon à obtenir un réseau propre à chaque échantillon pour déterminer leur contenu chimique.

3 Résultats

3.1 Traitement MZmine & Molnotator.

En sortie de MZmine, 24 434 ions ont été détectés en mode négatif et 31 747 en positif, soit un total de 56 181 ions.

Après traitement par *Molnotator*, un réseau de 64 134 nœuds est créé. Un réseau d'une telle taille se prête peu à une interprétation visuelle, une forme simplifiée a tout de même été produite à titre indicatif, ne présentant que les neutres et les adduits (**Figure 67**).

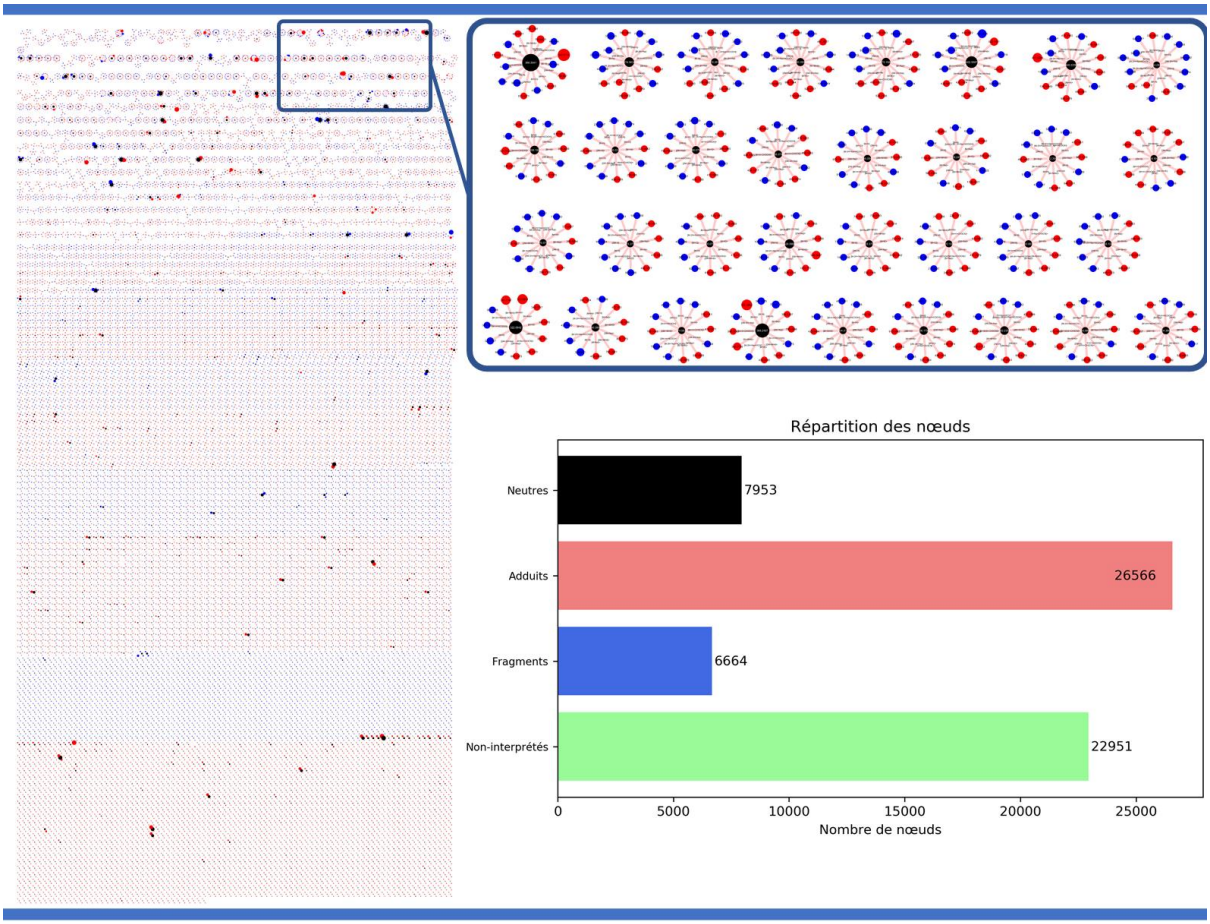


Figure 67 – Réseau final sous une forme simplifiée ne représentant que les neutres et les adduits. Les nœuds bleus représentent des anions, les rouges des cations, et les noirs, des molécules prédites. Un histogramme a été rajouté pour représenter la répartition des nœuds dans les quatre classes d'un réseau Molnotator : neutres, adduits, fragments ou non-interprétés (self-looped).

Le traitement a permis de prédire 7 953 molécules associées à 26 566 adduits et 6664 autres ions. 22 951 ions (35%) n'ont pas pu être interprétés, n'étant regroupés avec aucun autre nœud (*self-looped*). Parmi les adduits, 43% sont des anions et 57% sont des cations. La répartition des différents adduits est présentée en **Figure 68** (mode négatif) et **69** (mode positif)

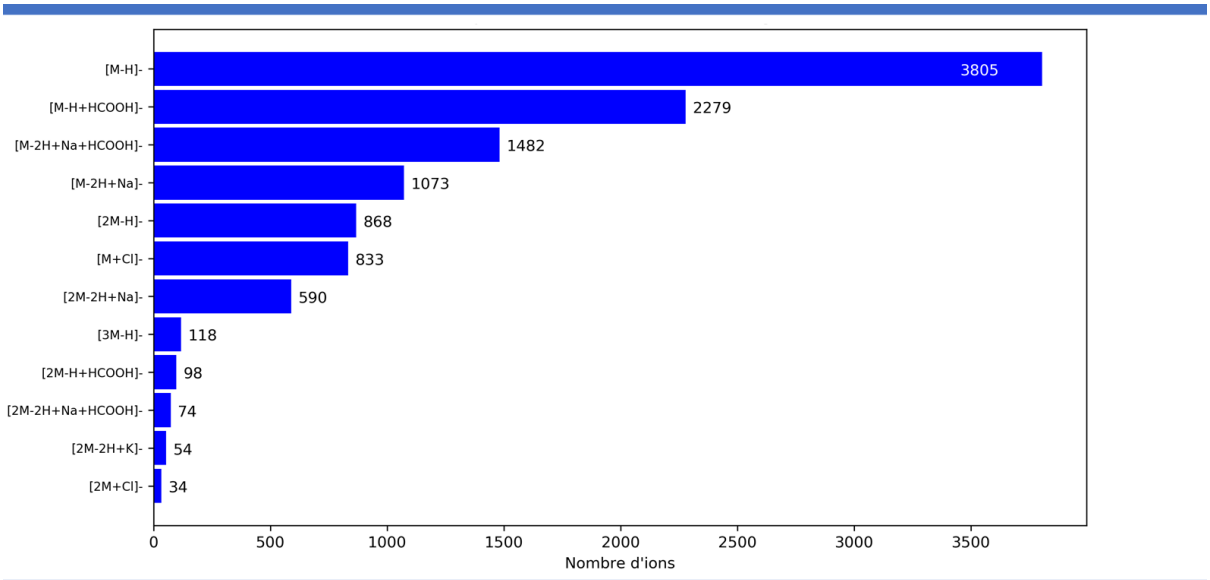


Figure 68 – Distribution des adduits en mode négatif.

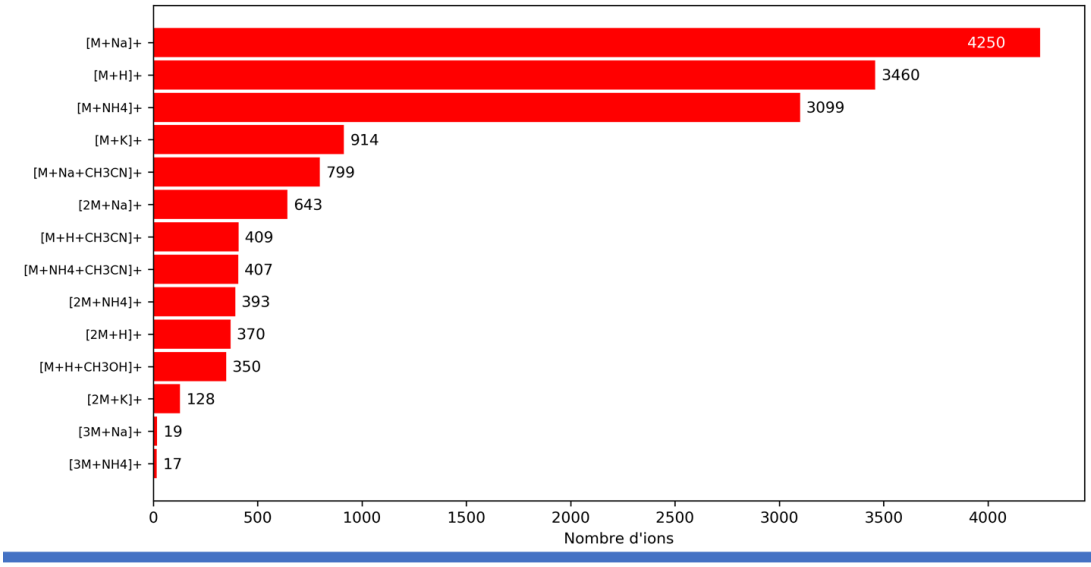


Figure 69 – Distribution des adduits en mode positif.

Comme établi dans le *Chapitre III*, les molécules (dé)protonées sont loin d'être la forme d'ionisation unique dans l'Orbitrap utilisée ici. En mode positif, les adduits sodium sont même plus nombreux que les molécules protonées. Par ailleurs, plus de cations ont pu être détectés que d'anions, malgré qu'il soit communément convenu que les substances lichéniques soient mieux étudiées en mode négatif. La diversité des adduits générés en mode positif étant supérieure à celle générée en mode négatif, il n'est donc pas étonnant d'observer plus de cations que d'anions pour une même molécule.

A ce stade, la fiabilité de la prédiction des neutres dépend du nombre d'adduits qui leur sont associés : une molécule associée à un nombre élevé d'adduits sera plus probable qu'une autre avec un faible nombre d'adduits. Le nombre d'adduits par neutres a été représenté en **Figure 70-A**. De façon assez prévisible, la majorité des neutres sont

associés à un faible nombre d'adduits. Lors de l'interprétation des résultats, ces neutres seront à prendre en compte avec précaution.

Les 7 953 neutres ont été soumis à une déréplication contre la LDB et la LDB-Lit (**Figure 70-B**). La LDB ne contient qu'un faible nombre de molécules et pour compenser ceci, une déréplication moins précise sur la base des masses prédites a été réalisée avec la LDB-Lit. Un nombre relativement faible des molécules a pu être dérépliqué par la LDB : 87, soit 1% des neutres. La déréplication par la LDB-Lit, moins stricte, a permis d'annoter 1307 neutres (16.5%). En utilisant des informations complémentaires telles que l'origine biologique des molécules, le contexte d'ionisation et les classes structurales prédites, il sera possible de confirmer certaines de ces annotations et de rajouter les spectres à la LDB. La majorité des neutres (6559, 82.5%) reste non-annotée.

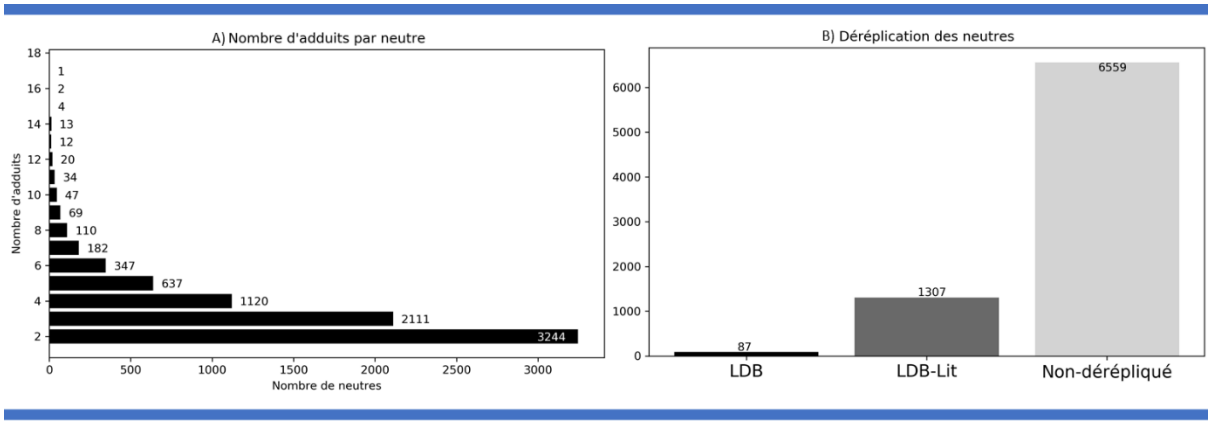


Figure 70 – Données sur le nombre d'adduits et la déréplication des neutres prédits.

La prédiction des classes structurales par *Classnotator* a permis d'annoter 3 955 des molécules avec des classes selon Hun&Yosh96 (**Figure 71**). La classe avec le meilleur score pour chaque neutre a été sélectionnée pour produire cette figure (chaque neutre présente en réalité plusieurs annotations avec différents scores). A ce stade, étant donné que *Classnotator* est basé sur le faible nombre de molécules de la LDB, les prédictions de classe ne sont qu'indicatives. Le faible nombre de depsides dans les neutres (152) est un résultat inattendu au vu de ce qui est connu des lichens, leur nombre se rapprochant de celui des Depsidones (ici 1149). Ceci est en revanche attendu d'un point de vue performance : comme observé dans le *Chapitre IV*, *Classnotator* identifie encore assez mal les depsides. Une partie de ceux-ci a sans doute été assimilée à des depsidones, ce qui explique également la quantité de depsidones prédites. Ces annotations restent néanmoins utiles, couplées aux déréplications par la LDB-Lit et aux origines biologiques des molécules, pour guider les annotations.

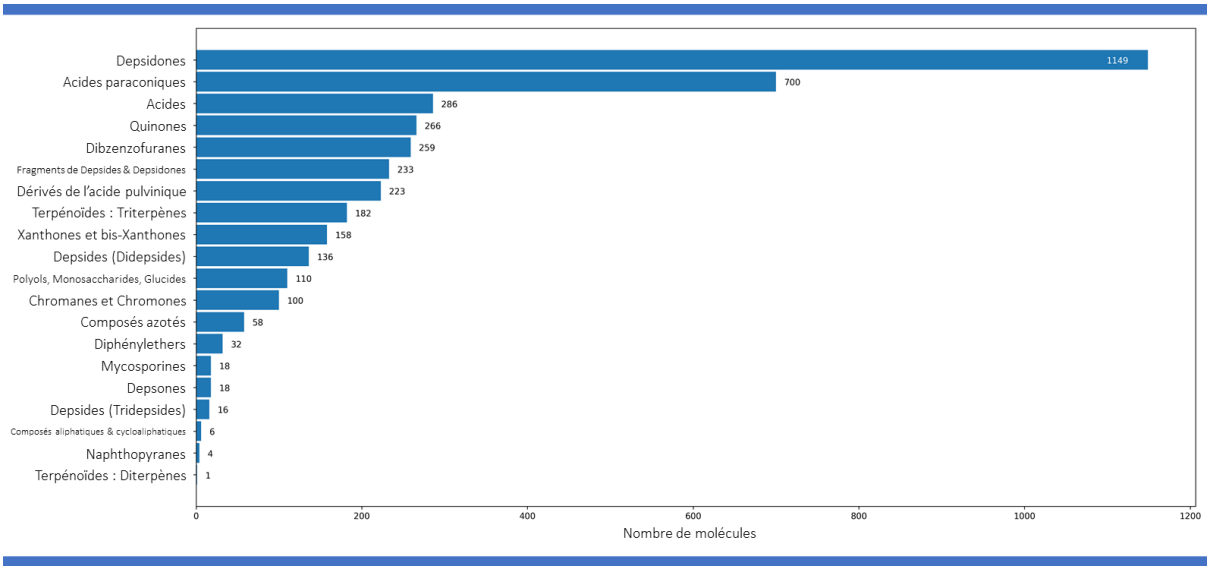


Figure 71 – Distribution des neutres dans les classes prédites par Classnotator.

Ces résultats représentent l'ensemble des données pour les échantillons analysés. Les résultats spécifiques pour chaque espèce / échantillon sont consultables en **Annexe**. Seuls quelques chiffres sont exposés dans chaque cas, étant donné que ces données doivent encore être interprétées manuellement pour évaluer avec plus de précision la qualité du traitement.

A titre représentatif, l'interprétations des réseaux pour *Evernia prunastri* (L.) Ach. et *Cladonia gracilis* (L.) Willd sera détaillée ci-dessous. *Evernia prunastri* est un lichen qui a été très étudié par le passé, comme vu dans le *Chapitre II*. A l'inverse, *Cladonia gracilis* n'a été étudié que rarement étudié et exclusivement par CCM. La liste des molécules auxquelles il sera fait référence est disponible dans les **Figures S-349** et **S-350** dans l'**Annexe**. Les déréplications et autres informations sur les neutres sont issues du réseau global. Il est donc possible de trouver des blocs moléculaires constitués de deux nœuds : un neutre et un ion. Ces blocs sont en réalités plus développés dans le réseau global et présente une déréplication : si un seul des ions de ces blocs est repéré dans le réseau d'un lichen, il sera annoté avec les données du réseau global. Avec l'exemple de ces deux lichens, l'utilité de *Molnotator* lors d'études de métabolomique non-ciblée par LC-MS sera démontrée.

3.2 Interprétation des résultats pour *Evernia prunastri*.

L'échantillon *d'Evernia prunastri* étudié ici porte la référence JB/13/156 et est désigné en **Annexe** par la référence S025. Un récapitulatif de l'analyse est présenté en **Figure 72**.

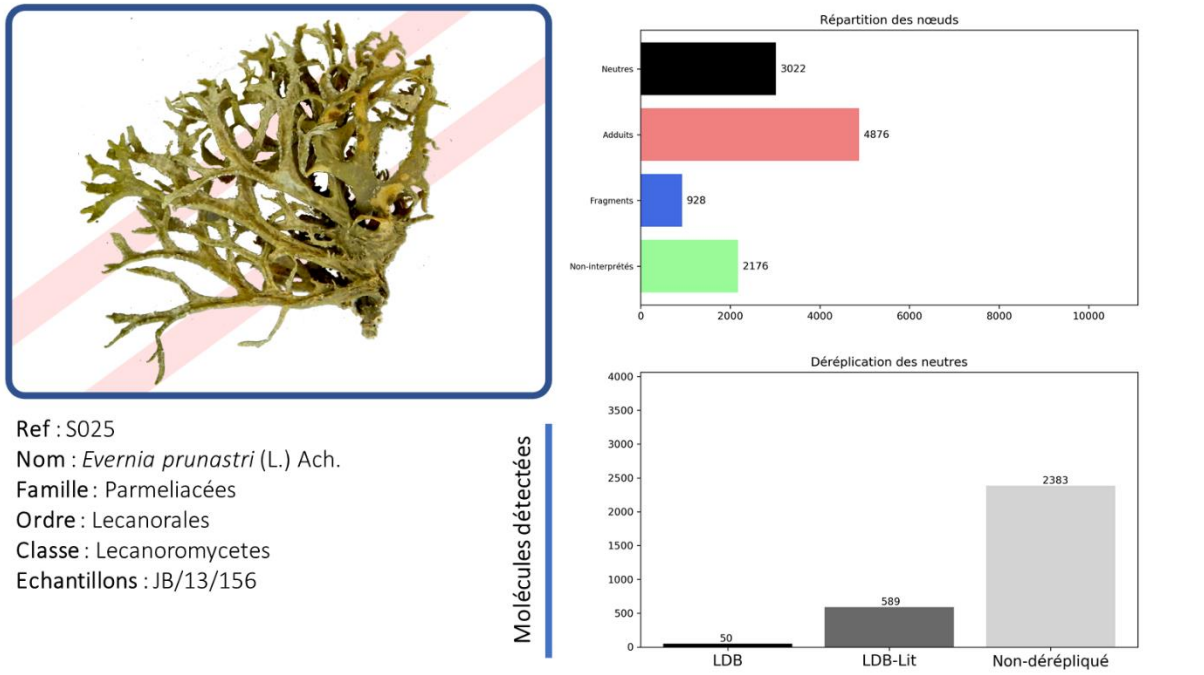


Figure 72 – Résultats bruts de l'analyse d'Evernia prunastri (S025).

3022 molécules ont pu y être prédites, accompagnées de 5804 adduits et fragments de source. 27% des ions n'ont pas pu être interprétés par *Molnotator*. Seulement 50 molécules ont pu être dérépliquées par la LDB, ce qui représente tout de même 1/5e de ses composés. 589 ont pu être annotées à l'aide de la LDB-Lit, ce qui pourra guider la réflexion lors de l'interprétation des résultats. La liste des molécules dérépliquées par la LDB (automatiquement) est présentée dans le **Tableau 31** avec pour chacune le score cumulé représentant la somme des scores de similarité cosinus pour chaque adduit du neutre, la classe Hun&Yosh96, le mode dans lequel le neutre a été prédit (MIX quand il a été détecté dans les deux modes) et le nombre d'adduits qui ont pu être associés au neutre dans le réseau global. Lorsque plusieurs déréplications sont possibles pour le neutre, elles sont séparées par un « | ».

Tableau 31 – Déréplication guidée (automatique) des molécules prédites dans Evernia prunastri. FDD : Fragments de Depsides et Depsidones, PMG : Polyols, Monosaccharides, Glucides, DAP : Dérivés de l'acide pulvinique.

Molécule	*Score*	*Classe*	*Mode*	*Adduits*
Acide 3,5-dichlororsellinique	2.79	FDD	NEG	3
Arabitol	0.84	PMG	MIX	5
Atranol	2.95	FDD	MIX	5
Acide barbatique	1.99	Depsides (Didepsides)	NEG	4
Acide bourgeanique	8.46	Acids	MIX	15
Acide capératique	3.8	Acides	MIX	10
Acide constictique	4.83	Depsidones	MIX	14
Acide crustinique	1.96	Depsides (Tridepsides)	NEG	9
Acide diffractaïque	4.81	Depsides (Didepsides)	MIX	7
Acide divaricatique	1.93	Depsides (Didepsides)	NEG	6
D-mannitol	3.54	PMG	MIX	6
Érythine	5.83	Depsides (Didepsides)	MIX	11

Tableau 31 – Suite.

Molécule	*Score*	*Classe*	*Mode*	*Adduits*
Acide gyrophorique	2.92	Depsides (Tridepsides)	MIX	7
Acide hypoprotocetrarique	0.85	Depsidones	MIX	4
Acide isoéverninique	0.93	FDD	MIX	7
Acide isomurolique|acide dihydromuronique|Acide murolique	3.28|2.21|1.53	Acides paraconiques	MIX	11
Acide lécanorique	2.93	Depsides (Didepsides)	MIX	9
Acide leprarique	0.93	Chromanes et Chromones	MIX	16
Acide lichestérinique	2.55	Acides paraconiques	MIX	11
Acide lobarique	6.57	Depsidones	MIX	12
Acide nephromopsique|Acide roccellarique|Acide néodihydroprotolichestérinique	2.84|2.73|2.54	Acides paraconiques	NEG	11
Acide orsellinique	3.96	FDD	MIX	11
Acide ovoïque	0.72	Depsides (Tridepsides)	POS	4
Acide physodique	5.49	Depsidones	MIX	14
Acide porphyrilique	3.7	Dibenzofuranes	MIX	9
Acide pseudoplacodiolique	1.86	Dibenzofuranes	NEG	4
Acide roccellique	2.94	Acides	NEG	6
Rugulosine	1.94	Quinones	MIX	4
Acide schizopeltique	3.73	Dibenzofuranes	POS	12
Acide squamatique	5.79	Depsides (Didepsides)	MIX	14
Acide stictique	1.84	Depsidones	MIX	17
Strepsiline	2.74	Dibenzofuranes	MIX	11
Acide usnique	3.92	Dibenzofuranes	NEG	6
Acide virensique	0.95	Depsidones	MIX	11
Acide vulpinique	1.98	DAP	MIX	8

Compte tenu de la quantité de neutres prédits, seuls quelques neutres remarquables seront considérés dans cette démonstration. Les métabolites rapportés pour *Evernia prunastri* dans la litérature sont nombreux (255), mais ceux qui lui sont habituellement associés sont l'atranorine, la chloratranorine, l'acide évernique, l'acide usnique ainsi que l'acide vulpinique (d'après le LIAS, consulté le 08/2020). Après la déréplication guidée par la LDB, seuls les acides vulpinique et usnique ont pu être retrouvés. Ceci peut s'expliquer simplement : l'atranorine, la chloratranorine et l'acide évernique ne sont pas présents dans la LDB-Orbitrap, ce qui réduit leurs chances d'être dérépliqués puisque seulement 63 à 75% des spectres sont reconnus d'un instrument à l'autre, comme établi dans le *Chapitre III*. Dans ce cas, elles ne l'ont pas été mais la déréplication par la LDB-Lit permettra de les retrouver. L'acide évernique par exemple est facilement repérable dans le réseau (**Figure 73**).

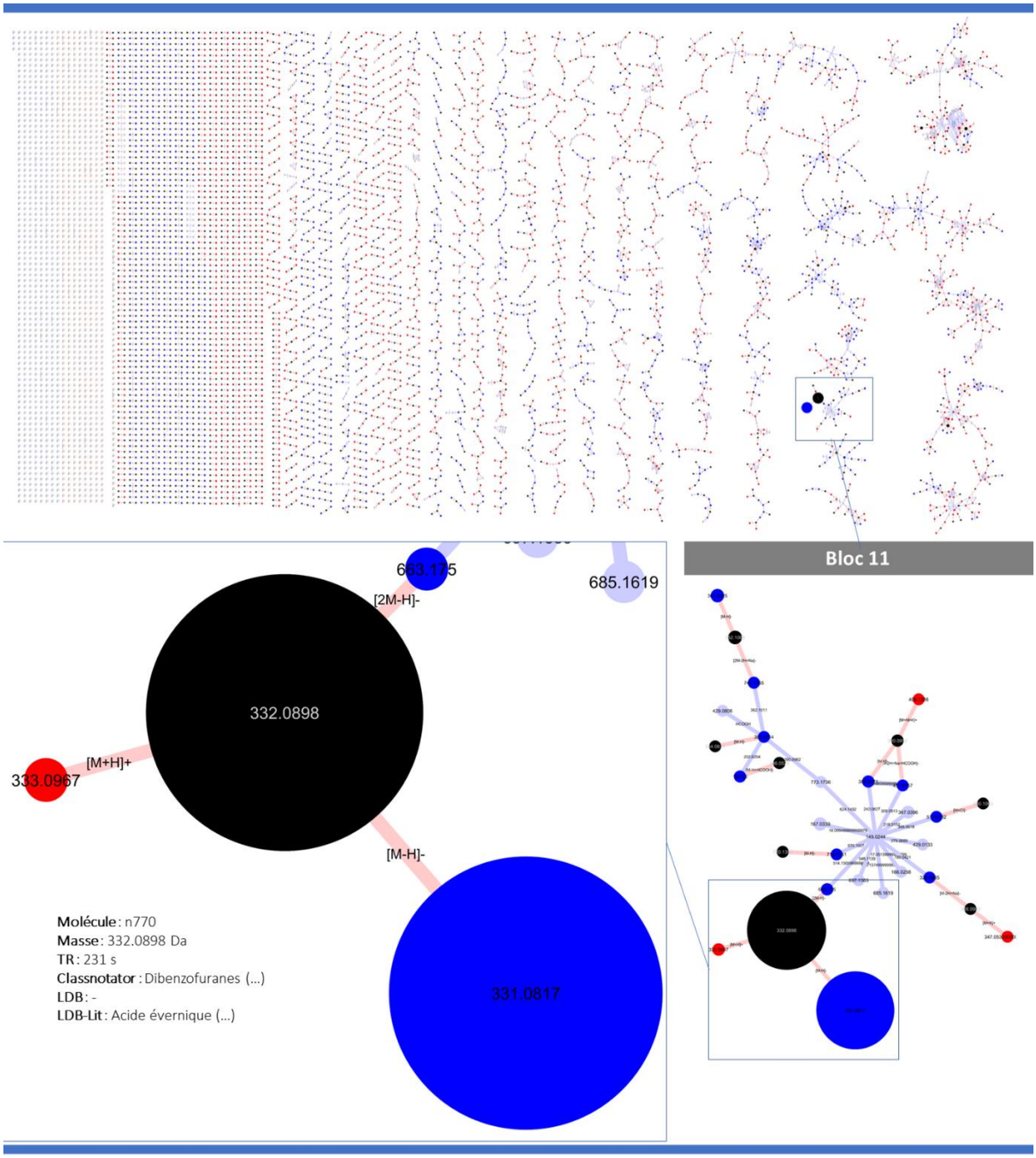


Figure 73 – Neutres remarquables du bloc 11 d'Evernia prunastri.

L'acide évernique est ici le nœud le plus important du réseau d'*Evernia prunastri* : n770. Bien qu'il n'ait pas été dérépliqué par la LDB ni par *Classnotator*, *Molnotator* a permis de le repérer sous la forme de nœud neutre. La LDB-Lit a ensuite permis de l'annoter avec la bonne identification.

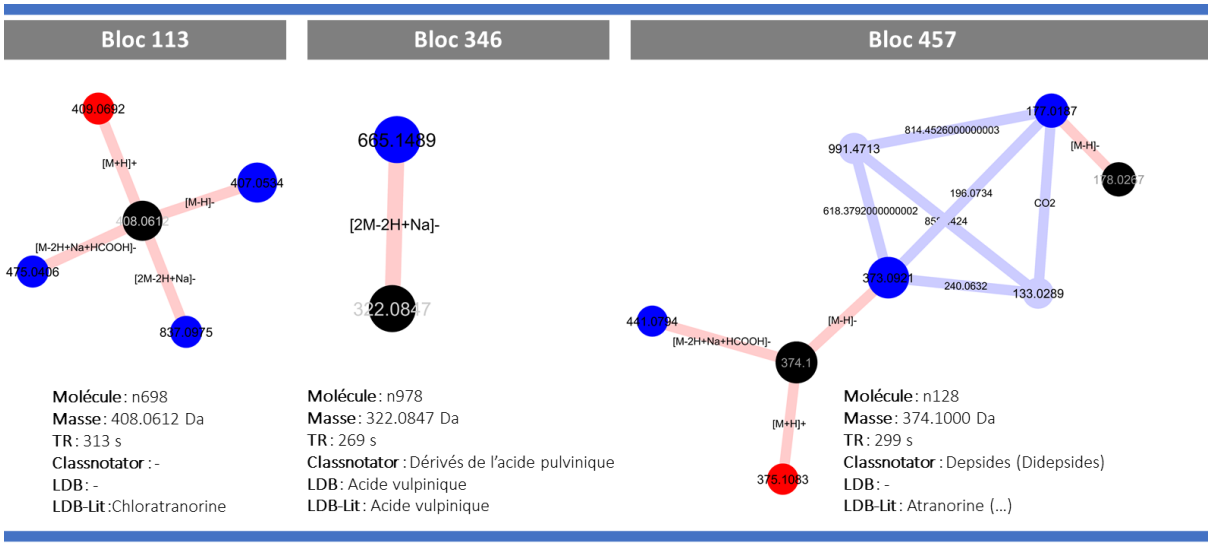


Figure 74 – Autres nœuds remarquables d'Evernia prunastri.

L'atranorine et la chloratranorine sont plus discrètes dans le réseau, mais ont pu être reconnues grâce à la LDB-Lit et dans le cas de l'atranorine, grâce à ses fragments de source caractéristiques qui lui ont été associés : *m/z* 177 et 133. Les nœuds de l'atranorine et de la chloratranorine sont donc n128 et n698. Dans ces exemples, *Molnotator* permet de prédire des molécules qui sont absentes des bases de données spectrales utilisées pour la déréplication et d'interpréter les données sans avoir recours à une identification (**Figure 74**).

L'acide vulpinique a quant à lui été repéré grâce à la LDB sous la forme de l'adduit $[2M-2H+Na]^-$ seul. Le bloc de ce neutre (n978) est bien plus développé dans le réseau global, avec 8 adduits associés (**Tableau 31**). *Molnotator* permet ici, grâce à la triangulation entre échantillons par *File merger*, d'annoter un ion unique qu'il aurait été difficile d'identifier même en observant les spectres MS^2. D'autres blocs contiennent des molécules inconnues ou du moins absentes de la LDB.

Dans le premier bloc (**Figure 75**), quatre neutres ressortent grâce au nombre d'adduits qui leur sont associés : n402 (474.3555 Da), n243 (418.3293 Da), n7 (518.3819 Da) et n366 (458.3605 Da). Il est cependant difficile de tirer d'autres informations sur ces molécules sans étudier leurs spectres MS^2. La molécule n243 a été prédite comme étant une depsidone et la seule molécule dans la LDB-Lit à laquelle sa masse pourrait correspondre serait l'acide 9,10,12,13-tetrahydroxytricosanoïque.

Le bloc 4 (**Figure 76**) porte plusieurs molécules, notamment n2468 avec une masse remarquablement élevée (602.3309 Da). Des molécules avec des masses moléculaires aussi élevées sont rarement rapportées dans les lichens. Son association à un grand nombre d'adduits dont un monomère $[M+NH_4]^+$ et un dimère $[2M+NH_4]^+$ suggère qu'il ne s'agit pas d'un dimère détecté en monomère par *Adnotator*, et donc que sa masse est correctement prédite. Cette molécule a été détectée dans les deux modes d'ionisation mais est mieux détectée en mode positif au vu du nombre de cations, adduits ou fragments qui lui sont associés. Un autre neutre notable est observable dans ce bloc :

n641, relié à n2468 par l'intermédiaire de l'ion fragment pos929 à *m/z* 195.1388. La topologie du réseau évoque un depside à haute masse moléculaire, ou du moins une molécule formée de plusieurs sous-unités détectables sous la forme de fragments de source, n2468 étant la molécule entière, n641 son fragment de source principal et pos929 l'un de ses blocs élémentaires. Les blocs produits par *Molnotator* permettent ainsi de faciliter l'identification des molécules inconnues grâce aux ions proches détectés en même temps. Il est alors commun de voir plusieurs neutres appartenant en réalité à la même molécule.

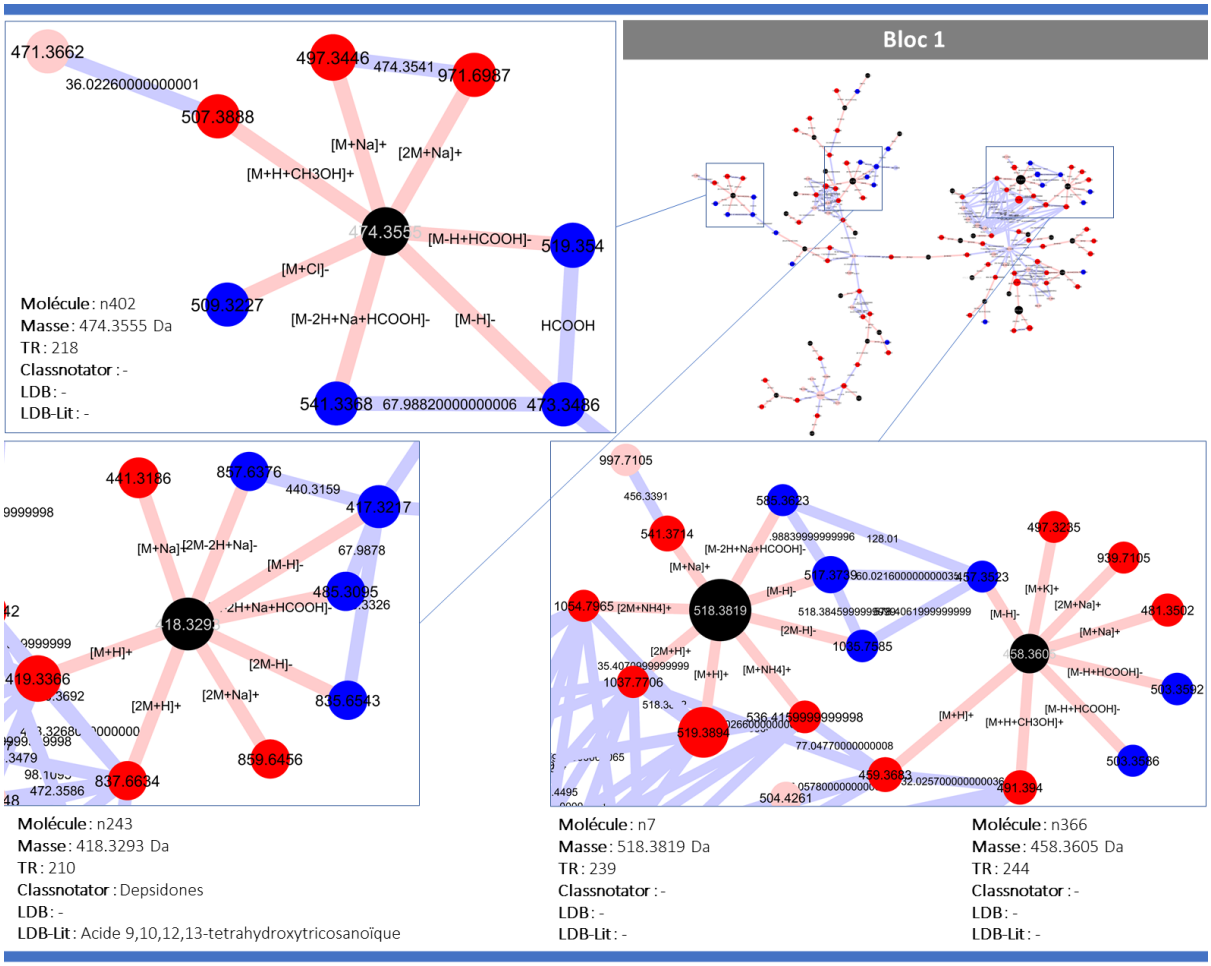


Figure 75 – Neutres remarquables du bloc 1 d'Evernia prunastri.

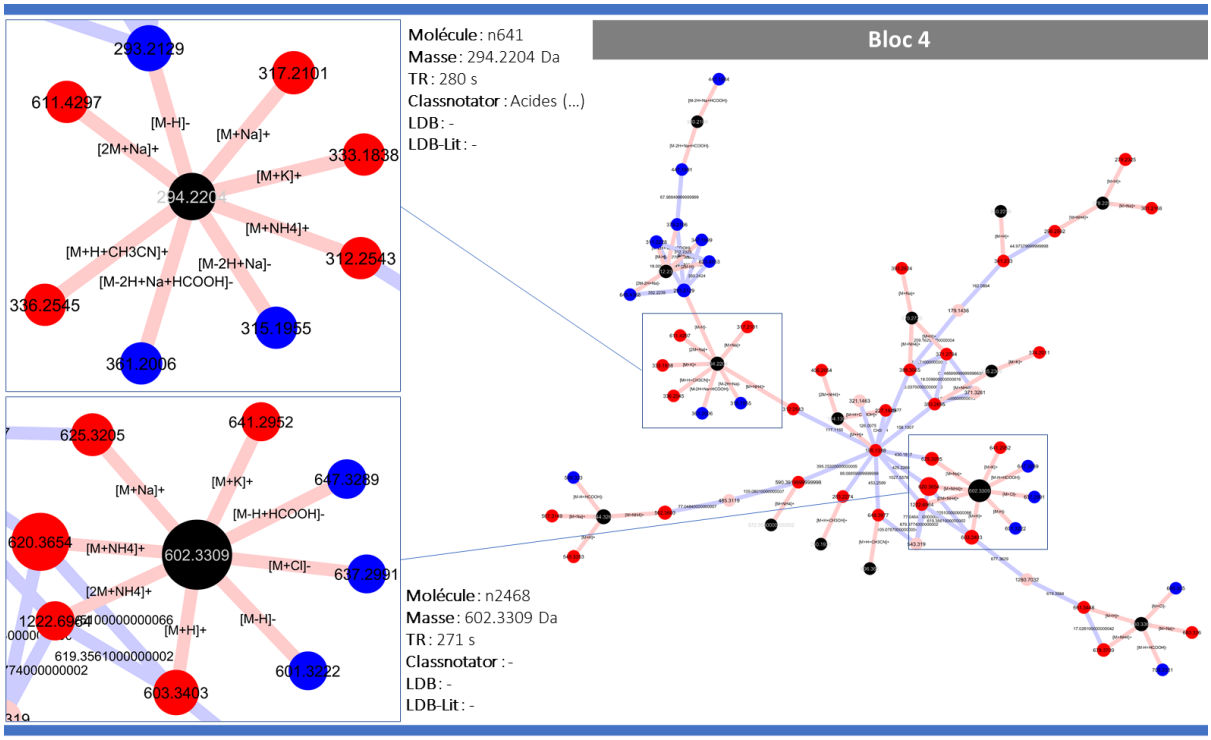


Figure 76 – *Neutres remarquables du bloc 4 d'Evernia prunastri.*

Le bloc 9 (**Figure 77**) est plus complexe. La seule molécule identifiée par la LDB est n1290 : l'acide isoéverninique. D'autres neutres ont été annotés par la LDB-Lit : n570, n1698, n92, n1552, et n2. En tenant compte de ces annotations et de l'organisation des nœuds du réseau, il est possible de trouver les annotations les plus probables. Comme vu précédemment, ces neutres peuvent correspondre à des molécules coéluées, ou à alors à des fragments récurrents d'une même molécule. Le neutre avec la masse la plus élevée est n570 et la LDB-Lit propose deux annotations : les acides confumarprotocetrarique et succinprotocetrarique. Ces deux molécules ne diffèrent que par la position d'une insaturation. Ce neutre admet un fragment de source détecté dans les deux modes d'ionisation, représenté ici par n1698. La perte de neutre suggèrerait que n570 corresponde à l'acide succinprotocetrarique et n1698 aurait alors une structure coïncidant avec l'acide virensique, comme proposé par la LDB-Lit. Si la déréplication n'avait pas été guidée sur *Molnotator*, les ions de n1698 auraient été reconnus comme l'acide virensique et non des fragments d'une autre molécule. En revanche, n92 semble être une depsidone proche coéluée, partageant le fragment n1698 et ne différant de celui-ci que par un acide carboxylique supplémentaire. L'anion de n1698 est relié à un autre ion appartenant au neutre n1552. Ce lien semble fortuit, n1552 s'apparentant davantage à un didepside d'après les annotations de la LDB-Lit, la prédiction *Classnotator* et le fragment de source qui est relié à son anion : n2. La LDB-Lit propose cinq depsides pour n1552 : la cladonioïdésine et les acides squamatique, cryptothamnolique, dissectique et échinocarpique. Deux structures sont également proposées pour n2 : l'acide hématommique ou isohématommique. La seule combinaison de structure qui aurait du sens dans ce contexte serait l'acide dissectique (n1552) avec l'acide hématommique (n2).

Ceci est par ailleurs également compatible avec l'annotation choisie pour n570, l'acide succinprotocetrarique pouvant également se fragmenter en acide hématommique.

Après interprétation superficielle du bloc et en l'absence de meilleures options, n570 serait le nœud de l'acide succinprotocetrarique, n1698 son fragment, n92 une depsidone coéluée, n1552 l'acide dissectique et n2 l'acide hématommique. L'analyse du bloc pourrait être approfondie pour identifier d'autres nœuds et pousser l'identification grâce aux spectres MS^2, mais il est également possible d'interpréter d'autres réseaux dans lesquels le même bloc pourra être retrouvé avec plus d'informations.

Le bloc 12 (**Figure 78**) contient l'acide orsellinique, déjà décrit dans ce lichen. Dans ce même bloc, son fragment peut être observé (-CO_2) : l'orcinol (n1280). L'acide orsellinique est retrouvé ici de façon native et non sous la forme d'un fragment de source, puisqu'il a été dérépliqué par la LDB et donc grâce à son temps de rétention (TR = 81 s).

Dans le bloc 15 (**Figure 79**), un nœud pourrait correspondre à l'acide fumarprotocetrarique d'après la LDB-Lit : n439. Ceci est compatible avec la classe prédite. Cette annotation devra cependant être confirmée par une analyse plus poussée des spectres MS^2 et des autres ions du bloc.

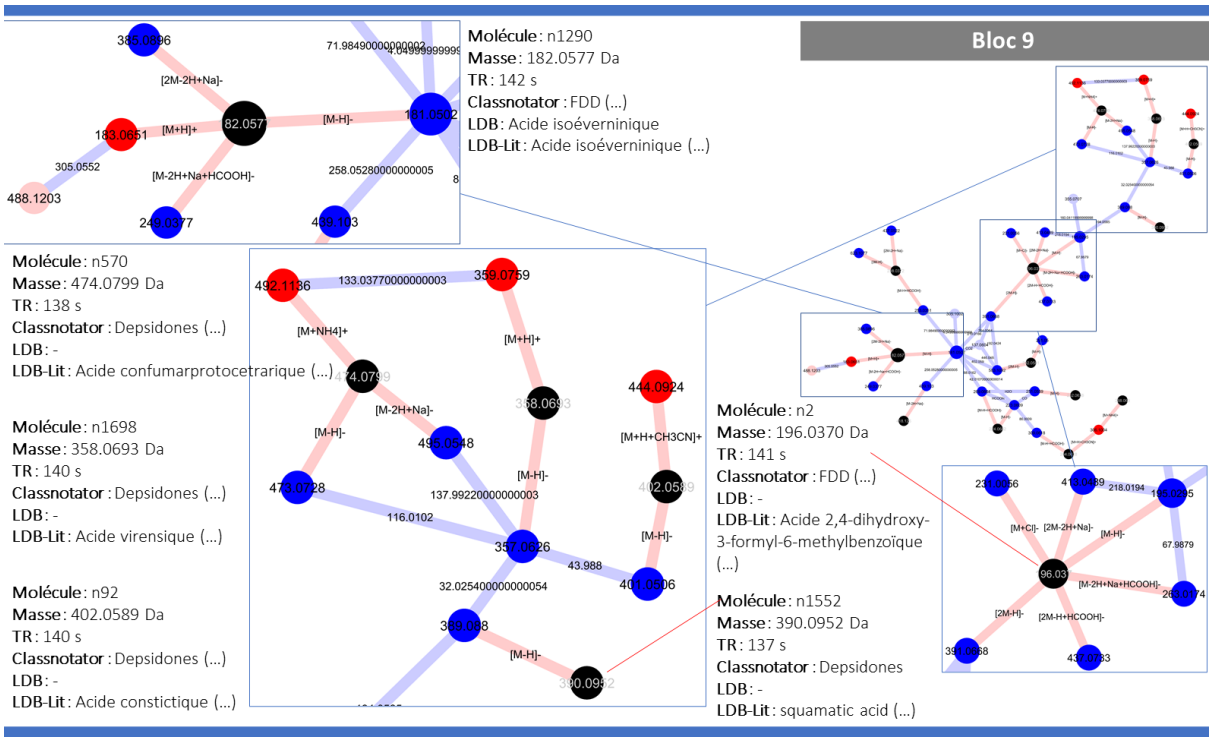


Figure 77 – Neutres remarquables du bloc 9 d'Evernia prunastri.

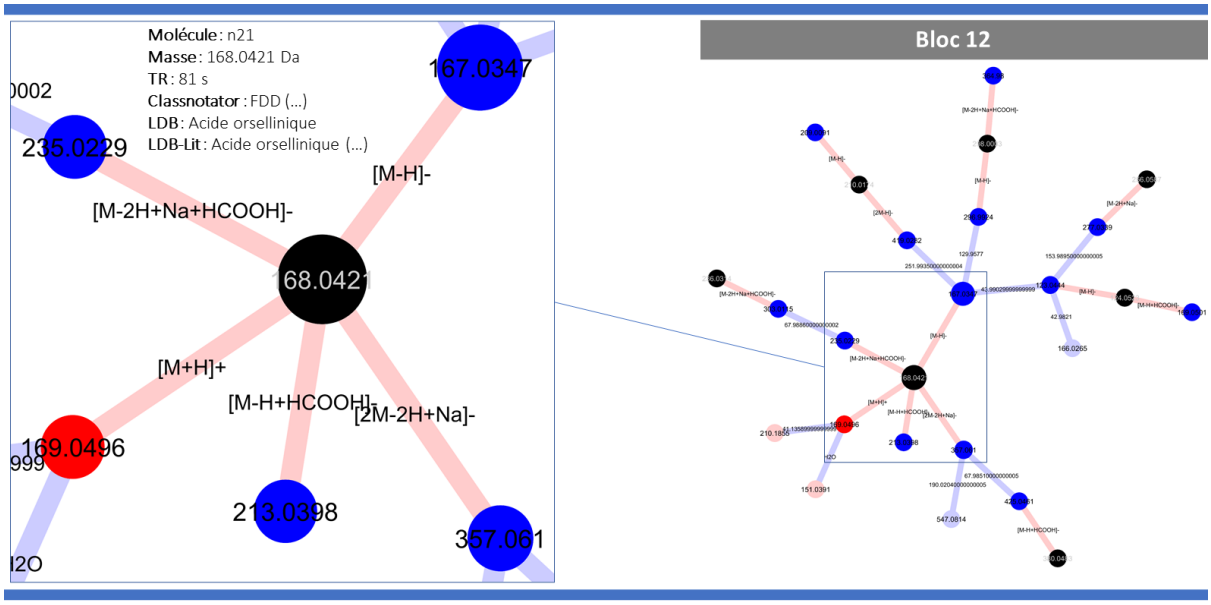


Figure 78 – *Neutres remarquables du bloc 12 d'Evernia prunastri.*

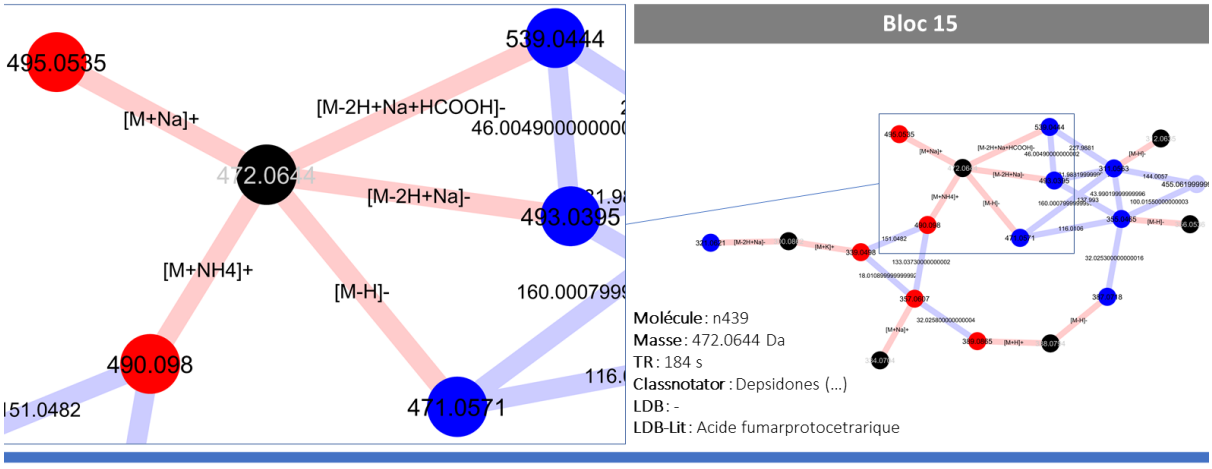


Figure 79 – *Neutres remarquables du bloc 15 d'Evernia prunastri.*

L'acide usnique a été dérépliqué par la LDB dans le bloc 18 (**Figure 80**). C'est l'un des composés habituellement attendus dans *Evernia prunastri*, ici présent sous la forme du neutre 210, dérépliqué par la LDB au bon temps de rétention et la LDB-Lit également. La classe prédite est bien celle des dibenzofuranes. Ce bloc contient plusieurs autres ions qui mériteraient d'être davantage étudiés.

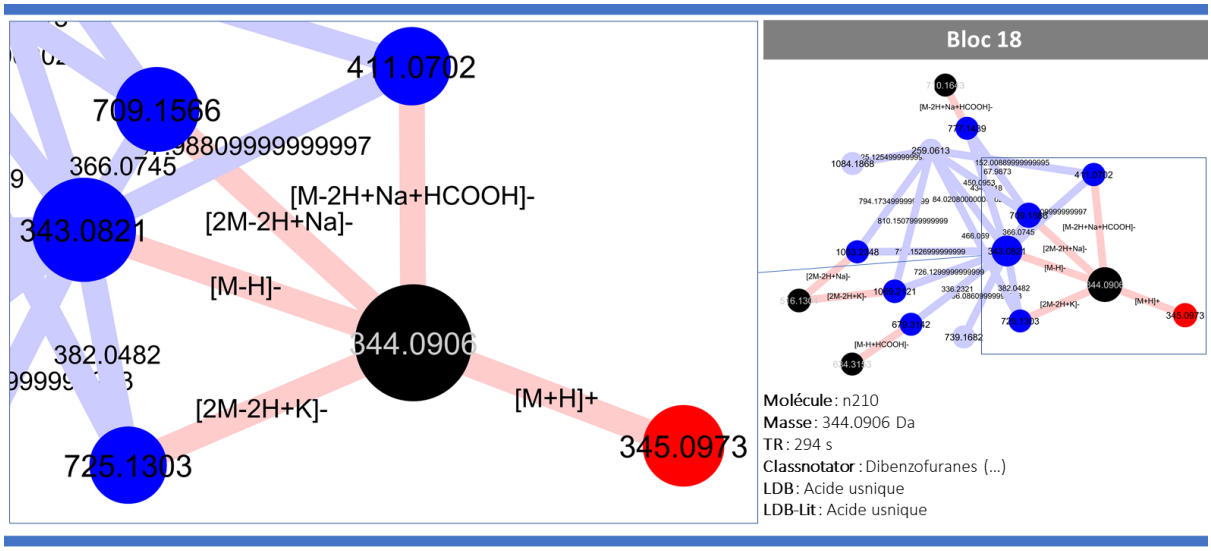


Figure 80 – Neutres remarquables du bloc 18 d'Evernia prunastri.

L'interprétation superficielle du réseau d'*Evernia prunastri* a permis de retrouver les quelques molécules attendues dans le lichen et de mettre en évidence la présence de nombreuses autres molécules encore non-identifiées ou simplement absentes de la LDB. Parmi les exemples abordés, l'acide lécanorique (n449, déjà détecté dans le *Chapitre II*), l'acide isoéverninique (n1290), l'acide orsellinique (n21), l'acide usnique (n210) et l'acide vulpinique (n978) ont pu être dérépliqués automatiquement par MS/MS. D'autres nœuds ont pu être annotés avec les molécules les plus probables au vu de leur contexte d'ionisation : n570 (acide succinprotocetrarique), n1698 (fragment de n570), n1552 (acide dissectique), n2 (acide hématommique), n770 (acide évernique), n1280 (orcinol), n439 (acide fumarprotocetrarique ?), n698 (chloratranorine) et n128 (atranorine). Ces annotations pourront être confirmées dans d'autres réseaux et les spectres seront rajoutés à la LDB le cas échéant. La prise en compte du contexte d'ionisation a permis d'éviter des annotations erronées : n1698 aurait été dérépliqué comme étant l'acide virensique, s'il n'était pas associé à son précurseur n570 et s'il n'y avait pas eu de filtre par temps de rétention. Dans un réseau normal, ce contexte n'aurait pas été apparent et l'acide virensique serait apparu plusieurs fois dans le réseau.

Bien qu'*Evernia prunastri* ait été largement étudié dans le passé, l'analyse LC-MS de ce lichen couplée au traitement de données rigoureux proposé ici a permis la détection de molécules encore inconnues, comme notamment n2468, un potentiel depside à haute masse moléculaire (**Figure 76**).

Ceci pourra être contrasté avec l'interprétation du réseau de *Cladonia gracilis*, lichen ayant été bien moins étudié.

3.3 Interprétation des résultats pour *Cladonia gracilis*.

L'échantillon de *C. gracilis* étudié ici porte la référence JB/03/6 et correspond à l'espèce S015 (**Figure 81**). A l'inverse d'*E. prunastri*, ce lichen n'a été que très peu étudié chimiquement, essentiellement en CCM par Santiago *et al* en 2010. Parmi ses molécules se trouvent l'atranorine (Wegrzyn et al. 2019), le peroxyde d'ergosterol (Dembitsky

2015), les acides usnique, stictique, norstictique, salazinique, diffractaïque, barbatique et galbinique (Santiago et al. 2010). Seules celles déjà présentes dans la LDB ont pu être dérépliquées, le peroxyde d'ergostérol, les acides norstictique, diffractaïque et galbinique en étant absents (**Tableau 32**). Les autres ont pu être repérées grâce à la LDB-Lit et à *Classnotator* : l'acide diffractaïque (n897), le peroxyde d'ergostérol pouvant correspondre à n4134 ou n4331 et l'acide norstictique pour lequel cinq neutres pourraient correspondre : n129, n1317, n596, n889 ou n123.

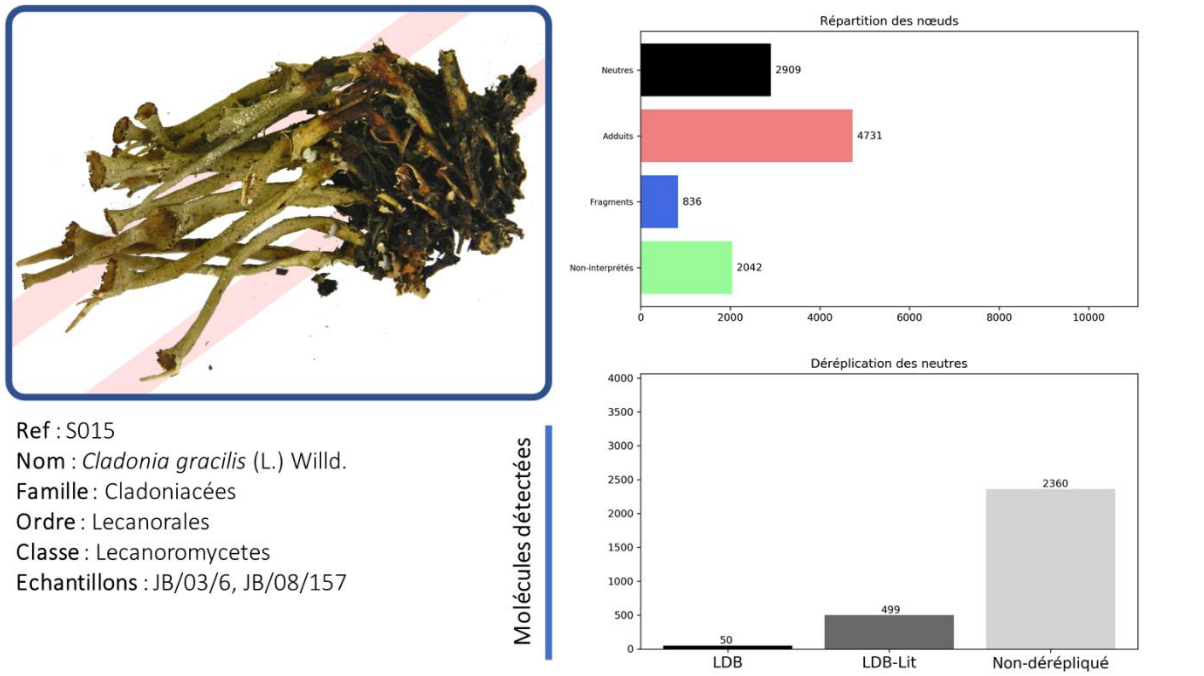


Figure 81 – Résultats bruts de l'analyse de Cladonia gracilis (S015).

Tableau 32 – Déréplication guidée (automatique) des molécules prédites dans Cladonia gracilis. FDD : Fragments de Depsides et Depsidones, PMG : Polyols, Monosaccharides, Glucides, DAP : Dérivés de l'acide pulvinique.

Molécule	Score	Classe	Mode	Adduits
Acide 3-hydroxyphysodique	5.77	Depsidones	MIX	13
Acide α-alectoronique	4.75	Depsidones	MIX	10
Arabitol	0.84	PMG	MIX	5
Atranol	2.95	FDD	MIX	5
Acide barbatique	1.99	Depsides (Didepsides)	NEG	4
Acide β-orcinolique	3.86	FDD	NEG	6
Caloploïcine	1.84	Depsidones	NEG	3
Acide capératique	3.8	Acides	MIX	10
Acide confluentique	2.67	Depsides (Didepsides)	MIX	4
Acide crustinique	1.96	Depsides (Tridepsides)	NEG	9
Acide diffractaïque	4.81	Depsides (Didepsides)	MIX	7
Acide divaricatique	1.93	Depsides (Didepsides)	NEG	6
D-mannitol	3.54	PMG	MIX	6
Emodine	0.97	Quinones	NEG	5
Erythine	5.83	Depsides (Didepsides)	MIX	11
Acide glomelliferique|Acide miriquidique	2.84|1.91	Depsides (Didepsides)	MIX	6
Acide gyrophorique	2.92	Depsides (Tridepsides)	MIX	7
Acide isoeverninique	0.93	FDD	MIX	7
Acide isomurolique|Acide dihydromuronique|Acide murolique	3.28|2.21|1.53	Acides paraconiques	MIX	11
Acide lécanorique	2.93	Depsides (Didepsides)	MIX	9

Tableau 32 – Suite.

Molécule	Score	Classe	Mode	Adduits
Acide léprarique	0.93	Chromanes et Chromones	MIX	16
Acide lichestérinique	2.55	Acides paraconiques	MIX	11
Acide lobarique	6.57	Depsidones	MIX	12
Acide mérochlorophéïque	0.96	Depsides (Didepsides)	POS	6
Acide nephromopsique|Acide roccellarique|Acide néodihydroprotolichestérinique	2.84|2.73|2.54	Acides paraconiques	NEG	11
Olivétonide	2.82	FDD	NEG	5
Acide olivétorique	3.86	Depsides (Didepsides)	MIX	8
Acide orsellinique	3.96	FDD	MIX	11
Acide physodalique	3.79	Depsidones	MIX	12
Acide physodique	5.49	Depsidones	MIX	14
Acide porphyrilique	3.7	Dibenzofuranes	MIX	9
Acide rangiformique|Acide isorangiformique	5.54|5.45	Acides	MIX	10
Acide rhizocarpique	2.88	DAP	NEG	5
Acide rhizonique	1.97	FDD	NEG	5
Acide roccellique	2.94	Acides	NEG	6
Acide salazinique	2.92	Depsidones	MIX	13
Acide stictique	1.84	Depsidones	MIX	17
Strepsiline	2.74	Dibenzofuranes	MIX	11
Acide usnique	3.92	Dibenzofuranes	NEG	6
Acide virensique	0.95	Depsidones	MIX	11
Volémitol	1.88	PMG	NEG	3

En observant le bloc 1 (**Figure 82**) plusieurs neutres ressortent, sans aucune proposition par la LDB et seulement trois de la LDB-Lit. La molécule n643 pourrait bien être un depside ou un fragment de depside en considérant son contexte. Les mêmes remarques peuvent être appliquées pour n366 et n7 qui ont par ailleurs également été observés dans le bloc 1 d'*Evernia prunastri* sans pouvoir les identifier. Le neutre n366 est probablement un fragment de n7. En revanche, n10 ne semble pas être un depside au premier abord, *Classnotator* propose notamment la classe des acides paraconiques mais il serait nécessaire de mieux étudier les spectres et le réseau pour déterminer ces molécules. Une annotation d'un neutre proche de n10 conforte son annotation en tant qu'acide.

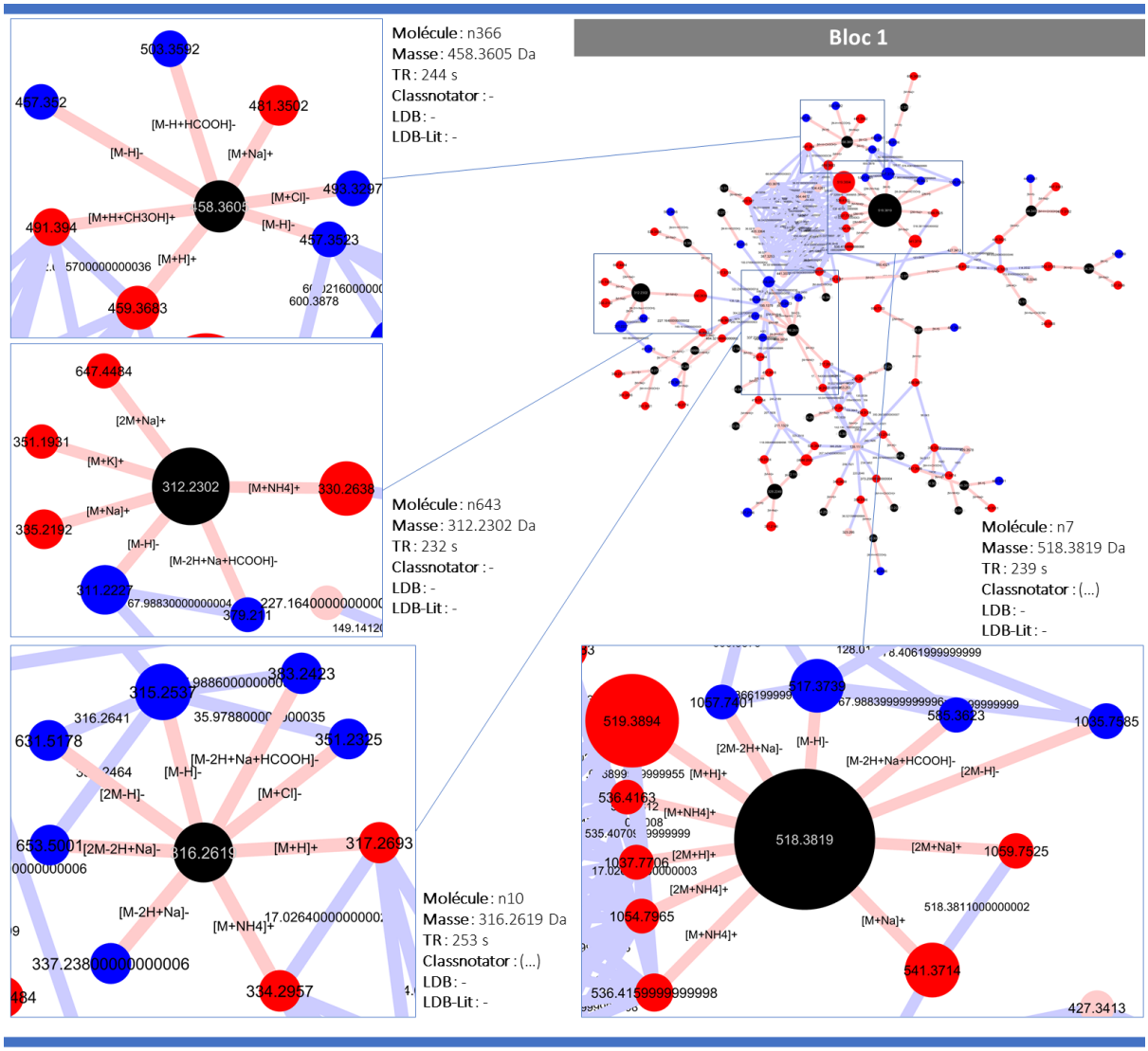


Figure 82 – *Neutres remarquables du bloc 1 de Cladonia gracilis.*

Dans le bloc 3 (**Figure 83**), le neutre n310 déjà observé dans le bloc 10 d'*Evernia prunastri* est retrouvé. Il ne s'agit probablement pas de l'acide norjackinique mais d'un depside inconnu. Les autres neutres sont ici observés pour la première fois mais peu de choses peuvent être conclues à leur sujet sans une étude approfondie du bloc et des spectres : là encore, aucune déréplication par la LDB n'a pu être faite et seulement trois par la LDB-Lit, peu informatives.

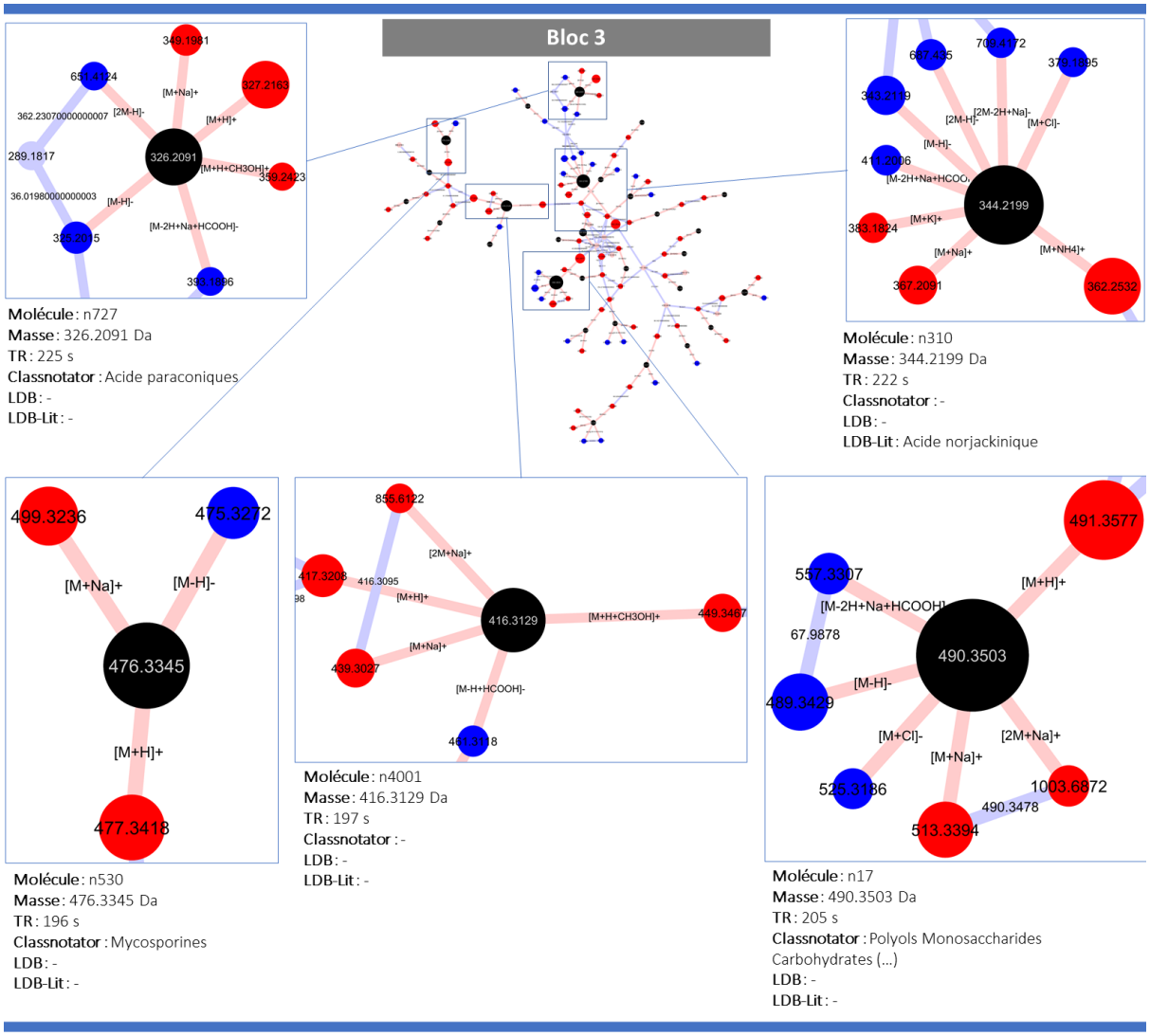


Figure 83 – *Neutres remarquables du bloc 3 de Cladonia gracilis.*

Le bloc 4 (**Figure 84**) se démarque par n2, fortement détecté grâce à son ion $[M-H]^-$ *m/z* 195.0295. Il s'agit du même fragment vu précédemment dans *Evernia prunastri* relié par conséquent au même neutre : n1552 (l'acide dissectique). Sont également retrouvés n570 et n1698, à savoir l'acide succinprotocetrarique et son fragment. Par ailleurs, d'autres neutres sont reliés aux ions de n2, indiquant d'autres potentiels depsides pouvant contribuer à l'intensité de l'ion $[M-H]^-$ *m/z* 195. Le neutre n391 a été dérépliqué comme étant un acide paraconique, ce qui pourrait bien être le cas au vu des fragments observés dans le réseau. Quelques identifications par la LDB-Lit des neutres proches indiqueraient qu'il s'agirait plutôt d'un fragment de despides / depsidones à longue chaîne, ce qui pourrait expliquer qu'il ait été dérépliqué en tant qu'acide paraconique. Il ne semble en revanche pas connecté à son précurseur dans le réseau, indiquant qu'il s'agit plutôt de la molécule libre que d'un fragment de source.

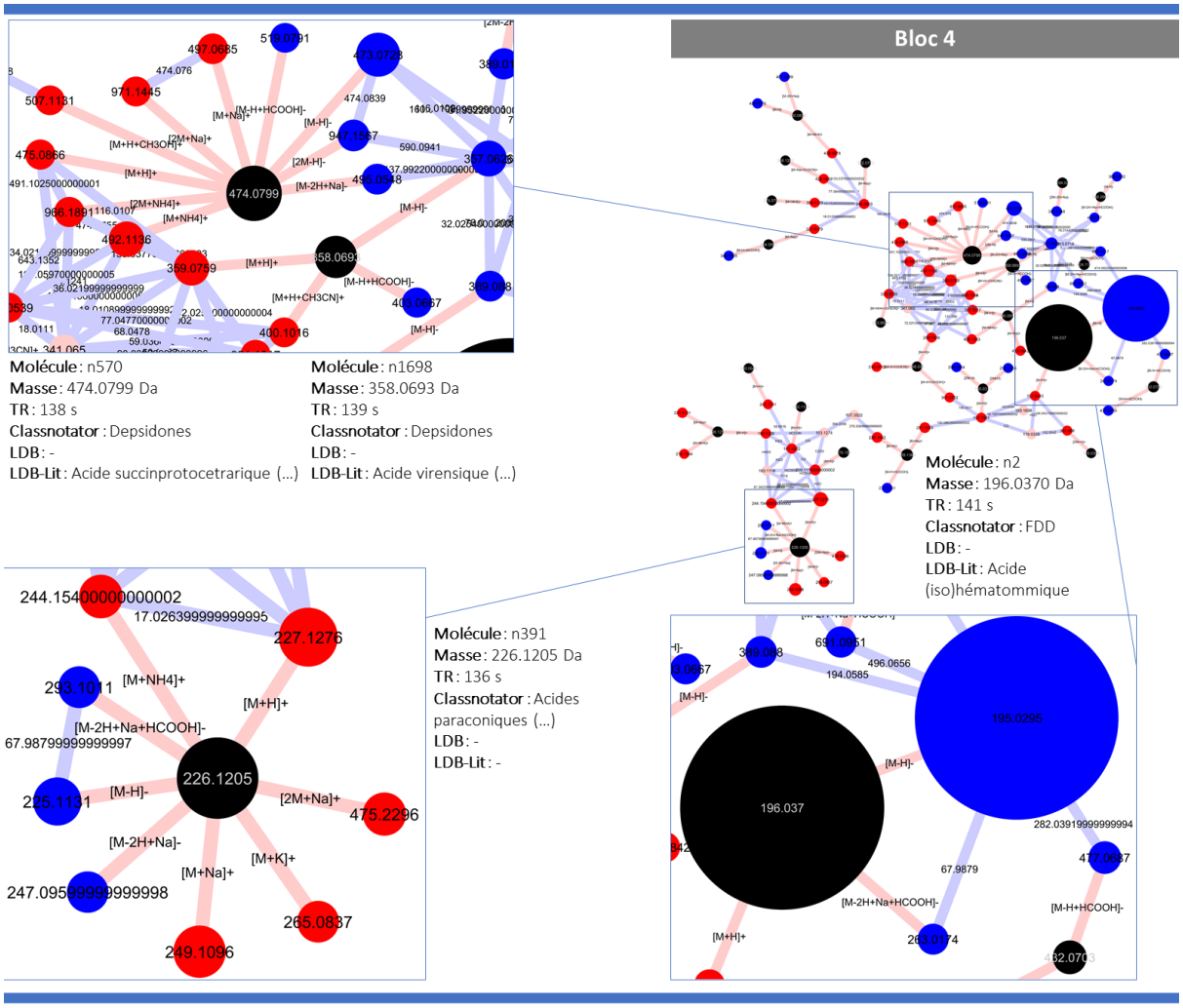


Figure 84 – *Neutres remarquables du bloc 4 de Cladonia gracilis.*

Le bloc 5 (**Figure 85**) porte le neutre n439, supposément l'acide fumarprotocetrarique, cette fois avec des ions plus intenses et plus nombreux. Le deuxième plus grand neutre du bloc est n518, annoté par la LDB-Lit avec 6 possibilités : les acides cryptostictique, hypoconstictique, 8'-méthylménégazziaïque, 2-méthoxypsoromique, 9'-méthylprotocetrarique ou la neuropogonine B. Tous deux partagent par l'intermédiaire de leur molécule déprotonée le fragment à *m/z* 355. L'acide 9'-méthylprotocetrarique est déjà dans la LDB et le fait qu'il n'ait pas été dérépliqué signifie qu'il s'agit ici certainement d'un fragment de n439 (TR = 314 s dans la LDB contre 191 s ici). En tenant compte de ceci, n439 a plus de chances d'être l'acide fumarprotocetrarique et n518 son fragment, partageant la même structure que l'acide 9'-méthylprotocetrarique. Il peut être raisonnable de penser que tout ce bloc correspond à des ions produits par l'acide fumarprotocetrarique.

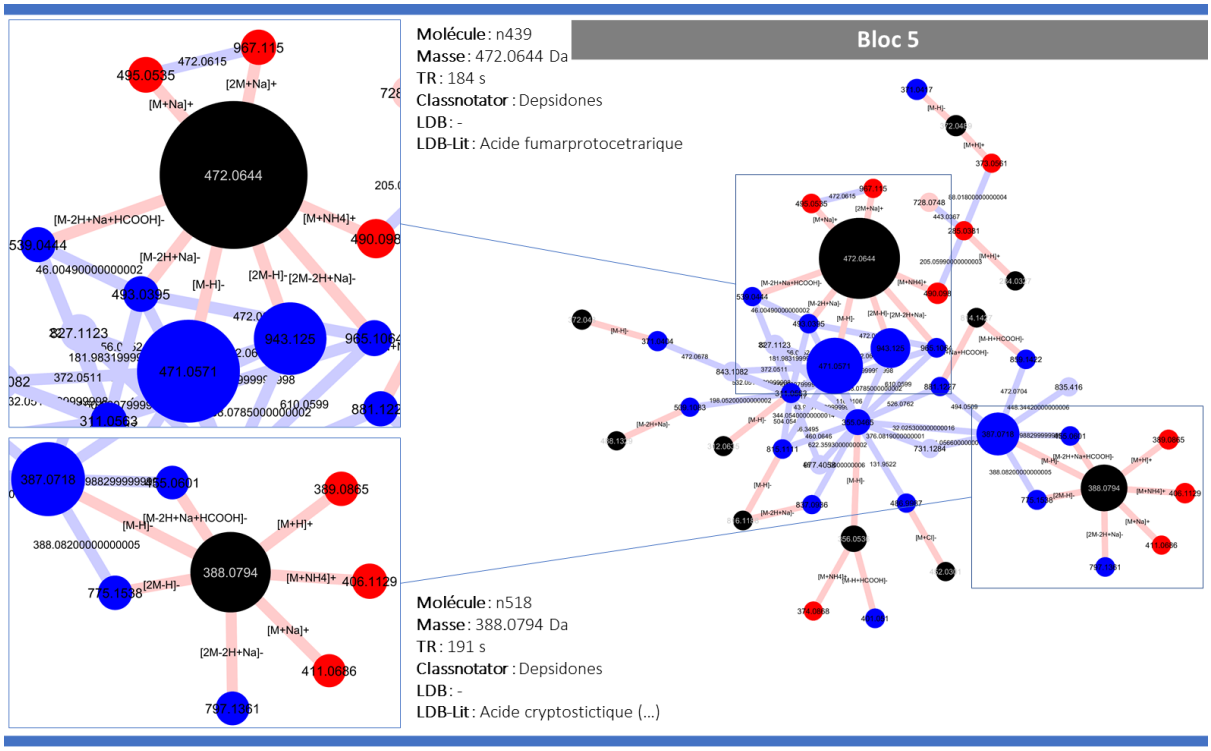


Figure 85 – *Neutres remarquables du bloc 5 de Cladonia gracilis.*

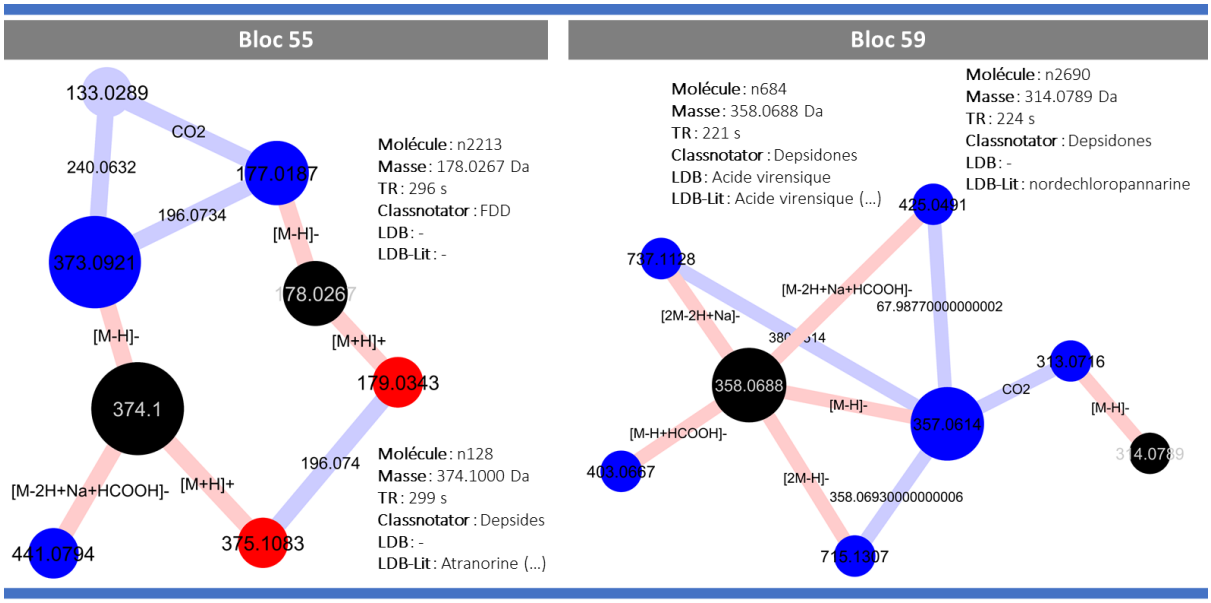


Figure 86 – *Neutres remarquables du bloc 55 et 59 de Cladonia gracilis.*

Le bloc 55 (**Figure 86**) contient le neutre n128 déjà observé précédemment et identifié comme étant l'atranorine. Le bloc 59 correspond à l'acide virensique, dérépliqué par la LDB (n684) et à son fragment de source (n2690).

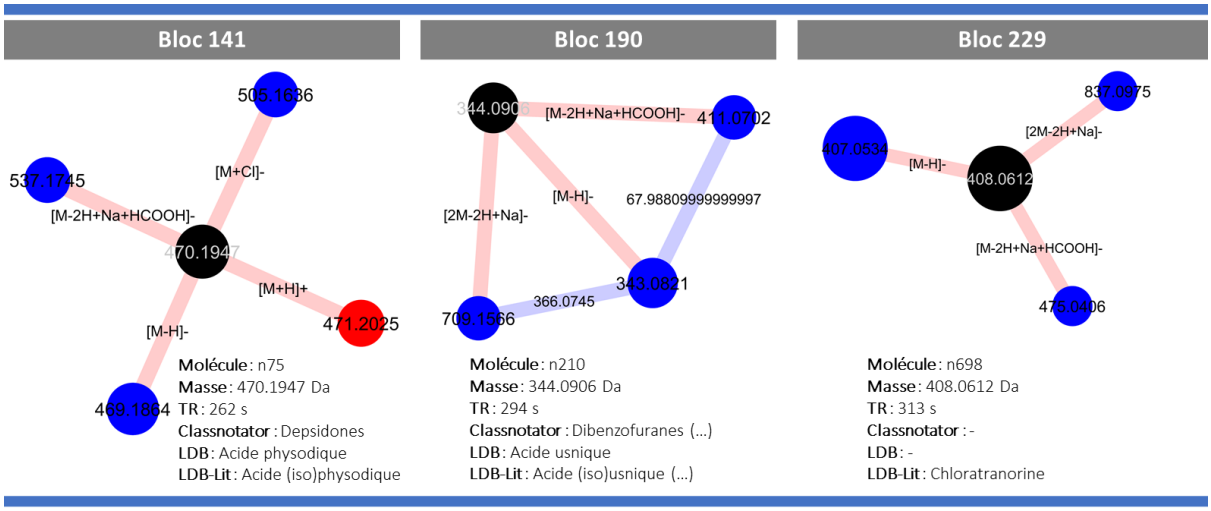


Figure 87 – Neutres remarquables du bloc 141,190 et 229 de Cladonia gracilis.

Les blocs de la **Figure 87** correspondent aux acides physodique et usnique, tous deux dérépliqués par la LDB, ainsi que le neutre n698 précédemment établi comme étant la chloratranorine.

Nombre d'autres neutres notables restent particulièrement difficiles à interpréter du fait du manque de déréplications et donc de l'inadéquation des bases de données face à la réalité du contenu chimique des lichens. Plusieurs de ces molécules peuvent être observées dans les **Figures 88** à **90**.

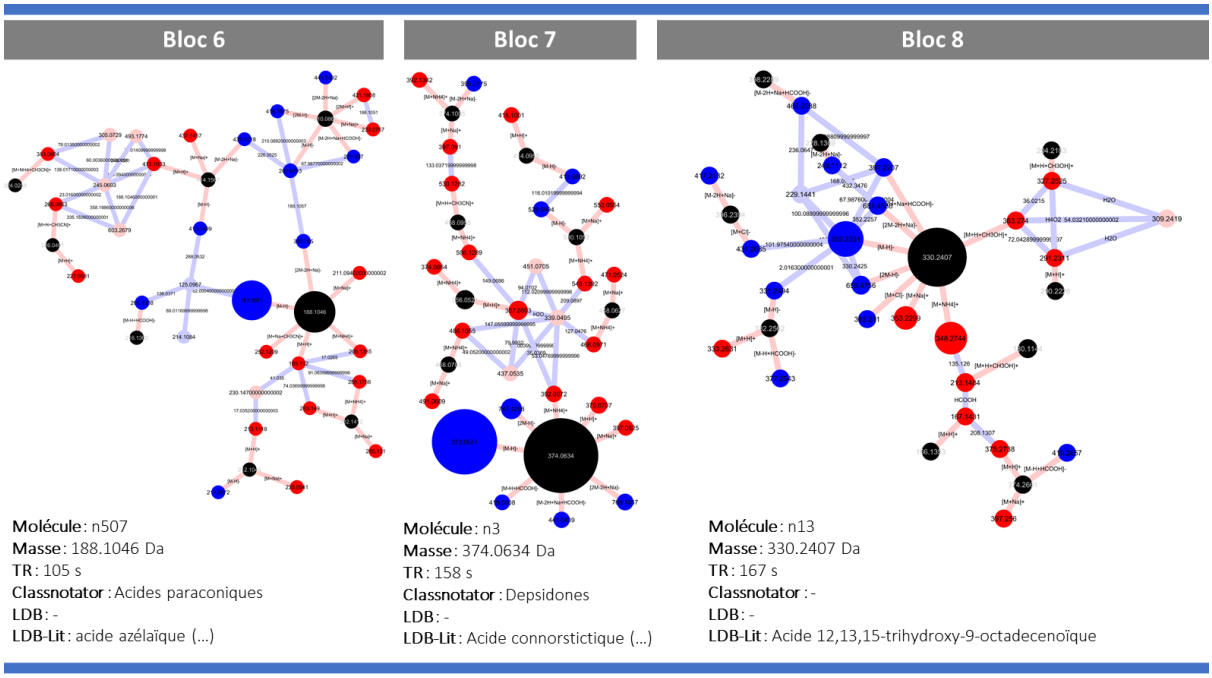


Figure 88 – Autres neutres notables de Cladonia gracilis non interprétés (blocs 6, 7 et 8).

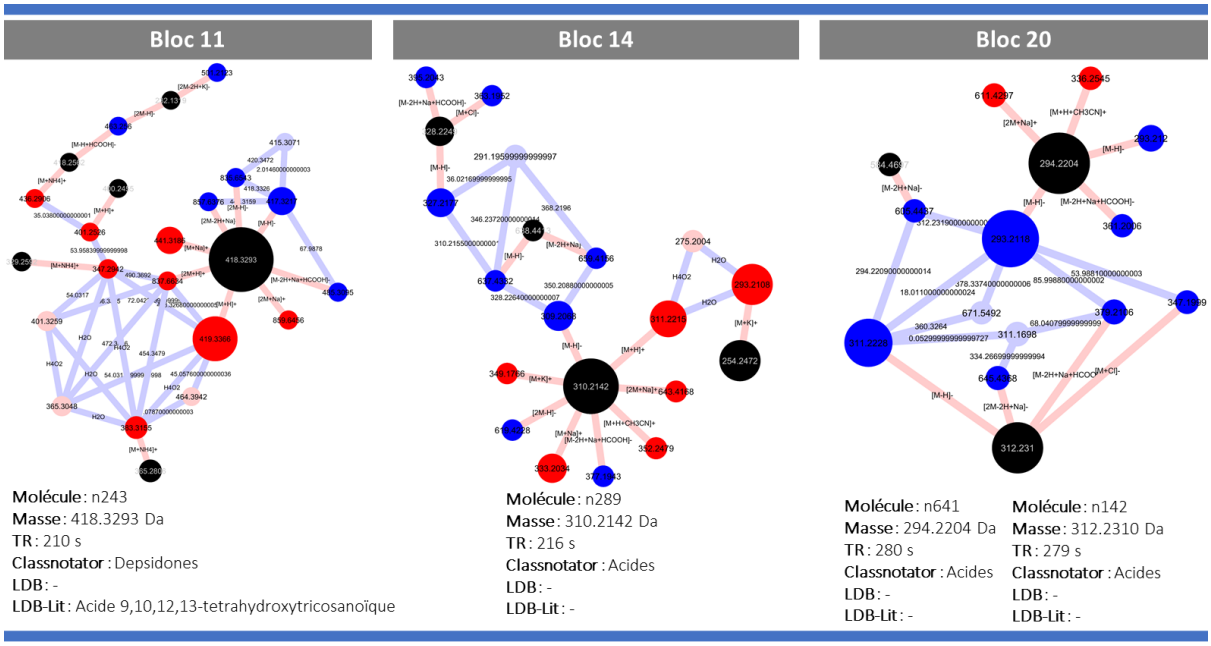


Figure 89 – Autres neutres notables de Cladonia gracilis non interprétés (blocs 11, 14 et 20).

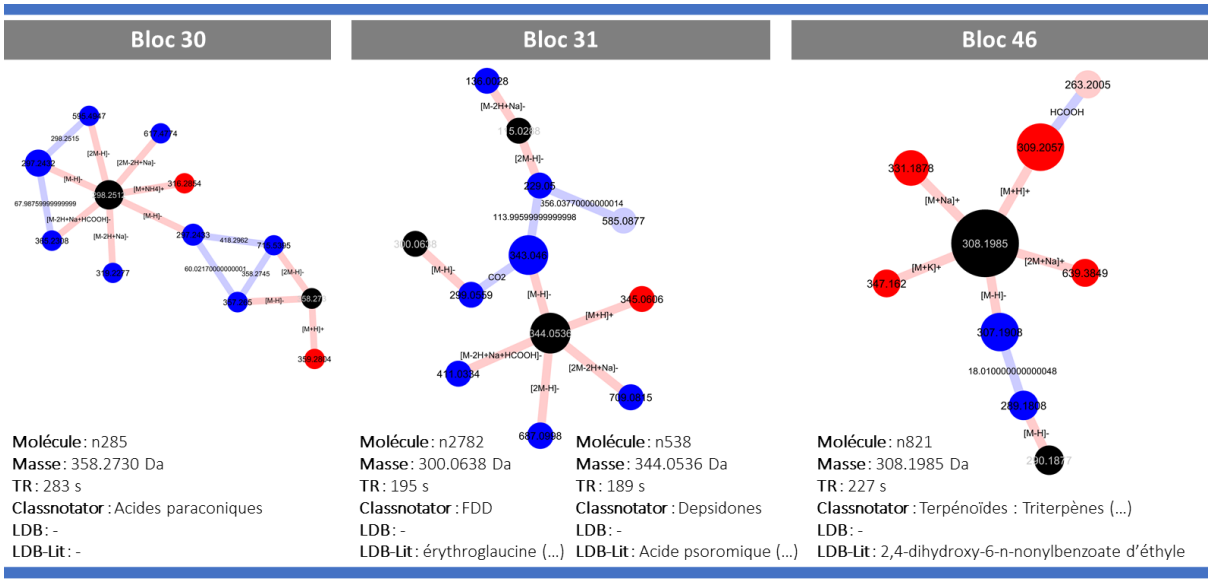


Figure 90 – Autres neutres notables de Cladonia gracilis non interprétés (blocs 30, 31 et 46).

Après l'analyse superficielle du réseau de *C. gracilis*, toutes les molécules qui lui sont attribuées dans la littérature ont pu être retrouvées avec des degrés de fiabilité variables. Nombre d'autres molécules ont pu être dérépliquées automatiquement grâce à la LDB et certaines grâce à la LDB-Lit comme l'atranorine et la chloratranorine déjà repérées dans *Evernia prunastri*. Bien plus de molécules restent cependant totalement inconnues, comme en témoignent les **Figures 88** à **90**. Ces nœuds majoritaires sont en apparence plus nombreux que dans *E. prunastri*, mais ceci ne signifie pas nécessairement que ces molécules seront faciles à isoler. Des molécules majoritaires comme l'atranorine forment des blocs discrets dans les deux réseaux, d'autres comme l'acide dissectique ne ressortent que grâce à leur ion fragment fortement ionisé, dont l'intensité est éventuellement renforcée par la fragmentation d'autres depsides proches qui produisent le même fragment.

4 Conclusion

Dans ce chapitre, le fonctionnement de *Molnotator* a été démontré avec le traitement de données d'une taille supérieure à ce qui est observé dans la plupart des études de métabolomique. L'intégralité des 300 échantillons a pu être traitée en rajoutant deux modules à *Molnotator* : l'utilisation d'*Adnotator* en deux étapes et la combinaison du traitement de l'ensemble des échantillons pour ne former qu'un réseau global avec *File merger*.

4.1 De la contribution de chaque mode d'ionisation.

Le réseau final est constitué de 64 134 nœuds : 7953 neutres, 26 566 adduits, 6664 fragments et 22 951 *self-looped*. Contrairement à ce qui pourrait être attendu, plus d'ions ont été détectés en mode positif (31 747) qu'en mode négatif (24 434), ce qui peut s'expliquer par la plus grande quantité d'adduits générés pour une même molécule dans le premier mode.

En termes de neutres, 5290 molécules ont pu être prédites en mode positif contre les 3476 du mode négatif (813 partagées par deux modes). Le mode d'ionisation négatif est pourtant établi comme celui de prédilection pour l'analyse des lichens, or ceci va à l'encontre des résultats obtenus dans cette étude. Quoiqu'il en soit, une analyse avec les deux modes est à privilégier pour permettre à *Molnotator* d'exploiter au mieux l'ensemble des données.

4.2 Des constituants des analyses LC-MS.

En ce qui concerne les adduits et les fragments, il peut être estimé que chaque molécule produira en moyenne 4 ions dans ces conditions en cumulant les deux polarités. Les molécules protonées ne représentent que 22% des ions du mode positif et les déprotonées 33% de ceux du mode négatif. Limiter les bases de données à ces molécules (dé)protonées revient à ignorer la majorité des ions d'une analyse, sans même tenir compte du taux de faux-positifs qu'une déréplication « aveugle » entrainerait.

4.3 Des avantages de *Molnotator* & pistes d'amélioration.

La dereplication guidée de *Molnotator* permet notamment d'éviter ces faux positifs et de regrouper les différents signaux d'une molécule dans un même bloc. Comme observé lors des interprétations, certains de ces neutres correspondent eux-mêmes à des fragments d'autres molécules du même bloc. Ces fragments facilitent l'interprétation en fournissant directement dans le réseau des données sur les précurseurs, les depsides étant particulièrement faciles à repérer par cette méthode.

Molnotator a permis dans cette démonstration d'interpréter 60% des nœuds du réseau en les reliant à des molécules, ce qui est bien supérieur aux capacités des méthodes déréplicatives classiquement utilisées. 40% des ions restent néanmoins sous la forme de nœuds *self-looped*. Plusieurs solutions pourraient être envisagées pour ceux-ci, comme le regroupement avec d'autres ions par similarité cosinus. Par ailleurs, la déréplication ici

ne s'est opérée que sur les ions pour lesquels des adduits ont été prédits, ou exclusivement sur les neutres dans le cas de la LDB-Lit. Un troisième type de déréplication moins stricte pourrait être fait avec la LDB sur tous les ions sans appliquer de filtre, ce qui renseignerait certainement sur l'identité de certains *self-looped*.

La déréplication par la LDB-Lit, bien que moins précise, permet de repérer certains ions absents de la LDB et éventuellement de rajouter leurs spectres après confirmation de l'identité de la molécule. Il pourrait aussi être envisagé de rajouter les fragments de source, mais ceci nécessiterait un module spécifiquement conçu à cet effet dans les algorithmes de déréplication pour les différencier des molécules entières. Une dernière amélioration pourrait être apportée à la déréplication : l'annotation des pertes de neutres. A ce stade, *Fragnotator* annote les arêtes avec quelques pertes simples, mais des molécules et/ou fragments de la LDB-Lit pourraient y être rajoutés, étant donné qu'il y a tout de même une classe de « Fragments » au sein de la classification de Hun&Yosh96.

4.4 La diversité chimique des lichens & comment mieux l'étudier.

La quantité de molécules détectées suite à ces analyses va à l'encontre de la vision traditionnelle de la chimie des lichens. Il a été démontré ici que les lichens contiennent quelques centaines voire milliers de métabolites secondaires chacun, plutôt que quelques dizaines. Ces molécules restent cependant inconnues compte tenu de nos connaissances actuelles sur la chimie des lichens. La première étape pour mieux étudier ces composés serait d'annoter dans le réseau toutes les molécules déjà connues dans ces lichens et de reporter leurs spectres dans la LDB. Avec quelques milliers de molécules, la LDB pourra être réutilisée pour refaire la LDB-motifDB et améliorer l'efficacité de *Classnotator*. Une fois ceci fait, les réseaux seront bien mieux annotés et les nœuds inconnus seront associés à des classes structurales d'un meilleur degré de fiabilité. Bien qu'il soit ensuite possible de cibler ces nœuds pour isoler les composés, leur nombre devrait favoriser des techniques alternatives puisqu'isoler des milliers de molécules potentiellement à l'état de traces est laborieux et très peu efficace. Avec l'amélioration des méthodes de traitement, il devrait être bientôt possible d'assigner des structures de façon rigoureuse à ces molécules sans avoir à passer par l'isolement, une stratégie bien élégante et efficace. L'isolement devrait être réservé aux nœuds présentant un intérêt de par leur activité supposée ou leur originalité structurale.

Conclusion

Sommaire

1 – De la création des bases de données 189

2 – Des différences entre CCM et LC-MS 192

3 – A propos du concept de chimiotype 193

4 – Des solutions apportées par les approches molécule-centrées 195

5 – Quant à la diversité chimique des lichens 196

6 – Perspectives 197

Le mot de la fin

– 199 –

Ces travaux avaient pour objectif principal de développer l'analyse des métabolites lichéniques dans le contexte de la métabolomique. L'exploration de leur métabolome à la lumière des techniques analytiques modernes a encouragé le développement d'outils spécialisés. Parmi ceux-ci, des bases de données ont été créées à partir de données de la littérature et également à partir d'analyses de standards sur différents instruments LC-MS. Ces bases de données ont servi de point de départ pour l'analyse de quelques centaines d'extraits lichéniques produits à partir d'une dizaine de milligrammes de thalle. Pour faciliter l'interprétation de ces données, un autre outil a été développé pour prédire la masse des molécules présentes dans l'extrait analysé à partir des adduits dans leur contexte d'ionisation. L'exploration détaillée des réseaux reste à faire, mais les résultats obtenus sont d'ores et déjà suffisants pour donner un aperçu du métabolome inexploré des lichens.

1. De la création des bases de données.

La première contribution apportée ici est la création de bases de données adaptées à l'étude des champignons lichénisés.

La LDB-Lit, produite à partir de la bibliographie, fournit des données structurales pour 1662 métabolites lichéniques. Il peut être estimé que de nombreuses autres molécules restent à ajouter, sans compter celles pour lesquelles aucun taxon n'a été associé (notamment dans Hun&Yosh96) et les molécules sans structures décrites. Même sans être exhaustive, elle remet en cause le décompte de 1050 composés communément accepté pour les lichens. L'intégration d'une composante taxonomique permet de mettre en évidence les différents degrés d'étude de ces organismes : certains étant décrits avec de nombreuses molécules (intérêt industriel, études phytochimiques) et d'autres pour lesquels ce nombre reste faible (profilages CCM simples) voire nul. Dans le contexte des analyses LC-MS, la LDB-Lit est utilisable dans des déréplications automatiques en calculant les rapports *m/z* des ions générés par chaque molécule. Par souci de simplicité, ceci est généralement réduit aux molécules (dé)protonées, la forme d'ionisation principale des molécules soumises à une source électrospray. Cependant, l'identification ne se fait que sur la base du rapport *m/z* et éventuellement des temps de rétention. La fiabilité des annotations devait être améliorée avec des données spectrales.

La LDB-Lit n'est pas encore publiée et ses fonctionnalités dépendront du choix d'hébergement. En plus des recherches de molécules par nom, masse, formule chimique, la recherche par sous-structure devra être développée. Puisqu'elle présente des données chimiques, taxonomique et bibliographiques, des requêtes appropriées seraient souhaitables pour faciliter l'extraction des données. Par exemple, trouver les sources bibliographiques rapportant dans un taxon donné la molécule recherchée, trouver toutes les molécules pour un taxon ou tous les organismes produisant une molécule (**Figure 91**).

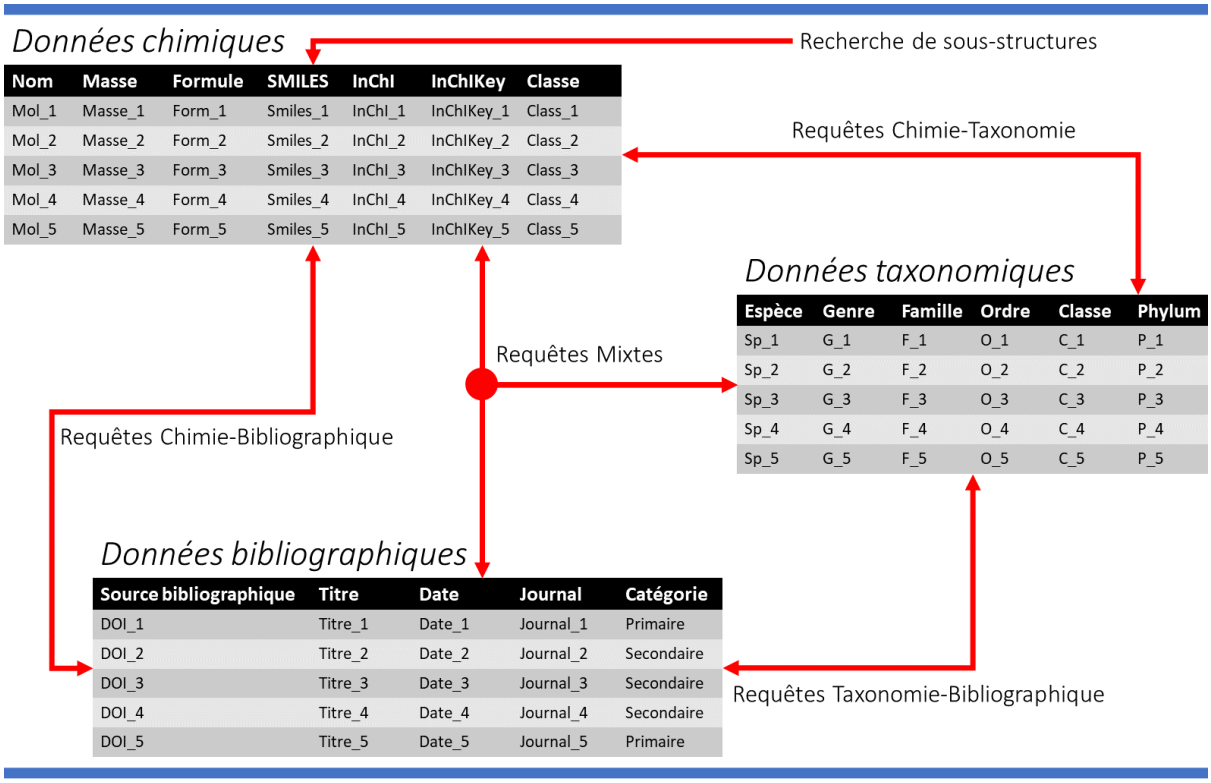


Figure 91 – Requêtes à développer pour la LDB-Lit.

Par la suite, la LDB a été produite, réunissant les spectres MS^2 de 250 substances lichéniques représentatives. Ces spectres ont été acquis sur trois instruments LC-MS différents (Agilent 6530 q-ToF, Thermo Q-Exactive Focus, Waters Xevo G2-XS) et couvrent de nombreuses espèces ioniques en plus des molécules (dé)protonées. Les annotations peuvent désormais être faites sur la base de similarités spectrales, ici grâce à un score de similarité cosinus.

Cependant, comme établi au début des années 2000, l'usage des bases de données doit être combiné à des outils de traitement adaptés (Fiehn 2001). La déréplication seule n'explique qu'une faible portion des données LC-MS et les raisons principales derrière ceci ont été soulevées :

- La quantité de molécules absentes des librairies spectrales, connues ou inconnues.
- La prédominance des spectres $[M-H]^-/[M+H]^+$ dans les bases de données.
- La portion non-négligeable de spectres non reconnus d'un instrument à l'autre.

Le premier point ne peut être résolu qu'en élucidant la structure de composés encore inconnus et le rajout des spectres de molécules connues ou inconnues aux bases de données.

Le deuxième point a été abordé avec l'ajout de différentes espèces ioniques à la LDB. Une prédiction robuste de la nature de chaque ion (adduit, fragment de source, complexe…) pour guider la déréplication restait encore à mettre en place. Ce dernier point est celui qui a été étudié ici, mais avant de conclure à ce sujet, les avantages et défauts d'une

stratégie de déréplication basée sur la LC-MS par rapport à la CCM utilisée historiquement doivent être exposés.

Le troisième point a été partiellement abordé en analysant les molécules sur différents instruments. Cependant, avec en moyenne 25% des spectres qui ne sont pas reconnus d'un instrument à l'autre (*Chapitre III*), ce problème de reproductibilité devient un enjeu majeur de métabolomique. En pratique, ceci a empêché la déréplication automatique par similarité cosinus de l'atranorine et de la chloratranorine dans le *Chapitre VI*. Bien que ceci n'apparaisse pas ici, une stratégie d'harmonisation des énergies de collision des différents instruments a été explorée l'aide d'ions thermomètres. Les différences d'un appareil à l'autre résident, au moins en partie, dans une différence d'énergie de collision réelle produite par chacun. Ceci peut être observé dans la **Figure 92**, où sont présentés les rendements de survie (*survival yields*) de la leucine enképhaline à différentes énergies de collision pour les trois instruments LC-MS utilisés. Ce rendement permet d'évaluer la fragmentation de l'ion $[M+H]^+$ (ou $[M-H]^-$) avec l'augmentation de l'énergie de collision, ce qui donne graphiquement une sigmoïde. Cette sigmoïde n'est pas la même d'un appareil à l'autre, ce qui permet en effet d'avancer que les énergies de collision affichées ne sont pas équivalentes. Ceci requiert de tester l'influence de différents paramètres des machines, dont certains qui ne sont pas accessibles à l'utilisateur, et il n'a pas été possible de mener le projet à bout. Avec une calibration en énergie entre les appareils, il pourrait être envisagé de joindre à chaque base de données ESI-MS/MS, l'énergie interne pour un calibrant utilisé (la leucine enképhaline ici) à laquelle chaque ion a été fragmenté.

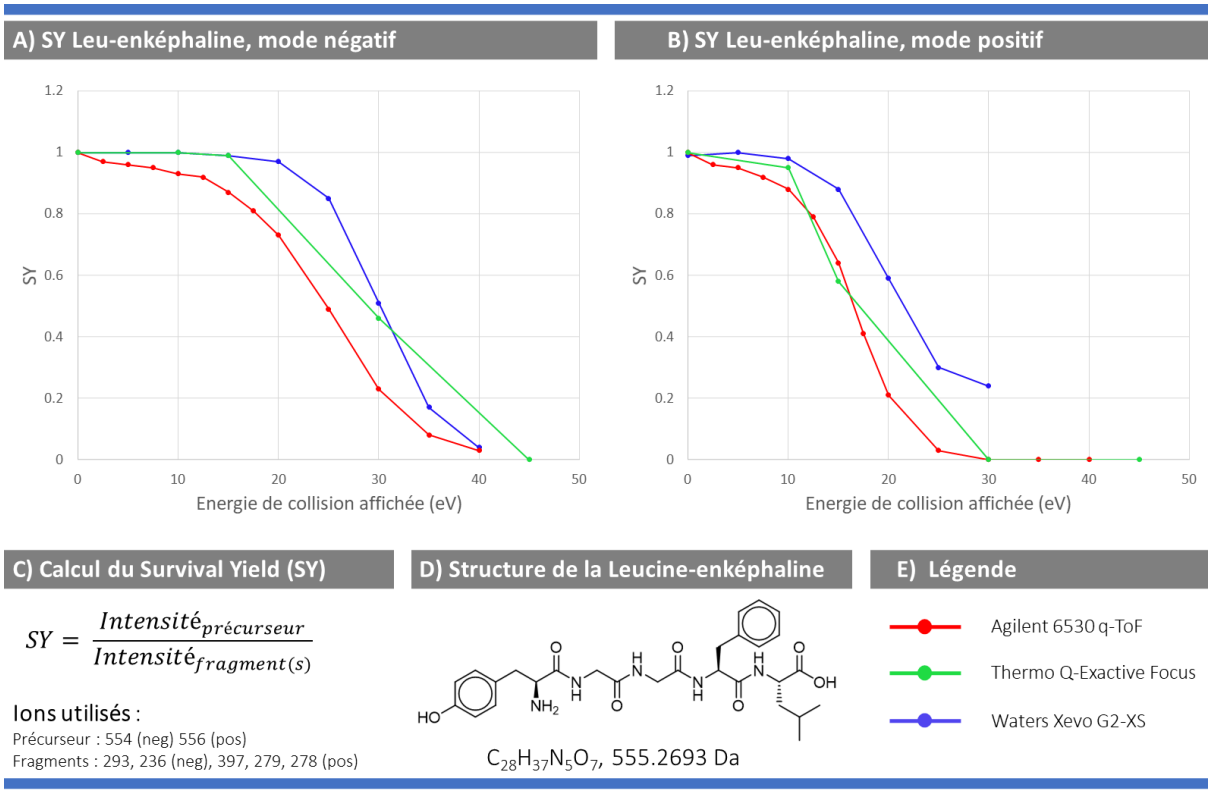


Figure 92 – Impact des énergies de collision de différents instruments LC-MS sur la fragmentation de la leucine-enképhaline, représenté par le rendement de survie SY. A : mode d'ionisation négatif, B : mode

d'ionisation positif, C : formule utilisée pour calculer le rendement de survie de la leucine-enképhaline et les ions utilisés, D : structure de la leucine-enképhaline.

2. Des différences entre CCM et LC-MS.

Compte tenu de l'importance occupée par la CCM dans la lichénologie et le profilage des espèces, il convient d'établir les avantages et les inconvénients de cette technique par rapport à la LC-MS moderne.

En l'état actuel, les outils de métabolomique ne permettent pas de répondre aisément aux questions des lichénologues chimiotaxonomistes, qui peuvent être résumées à cet objectif insidieusement simple : établir la liste des molécules caractéristiques d'une espèce ou variété de lichen. Sans aborder les difficultés liées aux délimitations taxonomiques, ceci est problématique pour la LC-MS sur le plan déréplicatif.

L'antériorité des méthodes de profilage standardisées par CCM permettent d'identifier la plupart des signaux observés grâce aux données accumulées depuis maintenant 50 ans (C. F. Culberson and Kristinsson 1970). Cette antécédence leur confère un avantage déréplicatif par rapport aux techniques LC-ESI-MS/MS pour lesquelles aucune base de données n'existait dans ce domaine.

Ces bases de données ont été produites ici, mais la LC-MS reste désavantagée sur le plan de l'exhaustivité puisque la LDB ne couvre actuellement que 250 métabolites contre les 800-1000 métabolites présentant des données de migration sur CCM dans le LIAS Metabolites (Elix et al. 2012) et Wintabolites (Lafferty and Bungartz 2018).

Dans l'immédiat, ceci pose des problèmes majeurs en termes de profilage : le lichénologue habitué des CCM ne trouvera pas certaines des molécules attendues dans la déréplication automatique puisqu'elles sont absentes de la LDB. De plus, ces absences sont accompagnées par la détection de nombre de molécules insoupçonnées, due notamment à une meilleure sensibilité analytique. Il en résulte que les données de métabolomique sont perçues comme étant en désaccord avec la littérature, voire même comme étant fausses pour les chercheurs les plus sceptiques. Ces différences de résultats contribuent à conforter l'usage de la CCM dans le profilage des lichens.

Un autre point crucial joue en faveur de la CCM : la recherche d'une donnée quantitative. Les lichénologues cherchent les métabolites majoritaires caractérisant l'organisme étudié ; or, l'intensité des signaux LC-ESI-MS/MS dépend de la capacité d'ionisation des composés. Des métabolites abondants comme l'atranorine s'ionisent mal et produisent des pics discrets. De par ses chromophores, celle-ci est facilement détectable par CCM en utilisant une révélation UV, ce qui coïncide avec son abondance dans le lichen. Cet argument est utilisé à tort pour justifier l'emploi de la CCM, avancée comme une méthode quantitative permettant d'établir la composition relative en métabolites caractérisant une espèce. Pourtant, ce caractère quantitatif se résume au mieux à l'intensité de détection UV-visible chromophore-dépendante, au pire, à « la taille des tâches » estimée à l'œil nu après révélation. Comme en LC-MS, certains composés minoritaires seront fortement détectés en UV-visible, tels que les pigments.

La limitation aux composés majoritaires, si usitée dans la chimiotaxonomie des lichens, devrait également être remise en cause. Sur la base de ces métabolites, l'absence ou la présence de certaines voies de biosynthèse est établie et des « races chimiques » sont définies. Les défauts de ce genre d'approche ont déjà été formulés au sujet de l'étude des protéines par électrophorèse sur gel. Dans les deux cas, l'utilisateur se limite aux signaux majoritaires qu'il est capable d'expliquer, avec des techniques qui ne sont résolutives que pour les composés d'une certaine polarité (classiquement, les protéines hydrophiles ou les métabolites hydrophobes). L'intérêt des composés jugés minoritaires est considéré comme négligeable, ce qui va à l'encontre d'une pratique holistique où toutes les molécules doivent être prises en compte (Fiehn 2001).

Un caractère quantitatif reste pourtant désirable dans un contexte multi-omique, mais ce sujet complexe a été volontairement omis dans ces travaux.

Ainsi, bien que la CCM présente un avantage immédiat en termes de déréplication et de coût, les analyses de LC-MS non-ciblées, de par leur sensibilité et la génération de données spectrales, sont plus performantes pour étudier le métabolome des lichens. Alors que quelques dizaines de composés peuvent être détectés à l'œil sur une CCM, plusieurs milliers peuvent être repérés par LC-MS. Les limitations aux composés dits « majoritaires » donnent une image faussement simpliste de la chimie des lichens, alimentant l'idée d'organismes pauvres en métabolites. Le désavantage déréplicatif n'est d'ailleurs pas dû à une infériorité intrinsèque de la LC-MS, mais à une plus grande sensibilité. La quantité de données générée se prête mal aux investigations manuelles telles que pratiquées en CCM, et ceci peut être corrigé avec des bases de données spectrales exhaustives et un traitement approprié.

3. A propos du concept de chimiotype

Le concept de « race chimique » ou chimiotype est fondé sur des bases fragiles. Des espèces sont parfois créées en identifiant des différences chimiques grâce à la CCM ou à des réactions thallines. Bien que cela puisse guider vers une différence taxonomique réelle, le lien avec une chimie caractéristique établie par CCM ne peut être que fortuit, tant cette méthode est peu appropriée pour étudier la complexité du vivant. En guise d'exemple, le cas de *Pseudevernia furfuracea*.

Cette espèce serait composée de trois chimiotypes principaux, souvent désignés comme des espèces à part entière : *Pseudevernia furfuracea* (L.) Zopf, *olivetorina* (Zopf) Zopf et *consocians* (Zopf) Zopf (**Tableau 33**). Elles sont indistingables morphologiquement, ne différant que par leur chimie et leur répartition géographique. *P. furfuracea* se développe en Europe et se caractérise par la production d'acide physodique (et 3-hydroxyphysodique), *P. olivetorina* se développe également en Europe et produit notamment l'acide olivétorique sans l'acide physodique ni le 3-hydroxyphysodique. *P. consocians* se caractérise par l'acide lécanorique et il ne se développe qu'en Amérique du Nord. Il existe cependant une forme rare de *P. olivetorina* qui produit les acides physodique, 3-hydroxyphysodique et olivétorique en même temps, bien qu'ils soient

censés être mutuellement exclusifs (*P. olivetorina-2*) (W. L. Culberson, Culberson, and Johnson 1977; Hale 1968). Il est intéressant de noter que ces auteurs soulignaient déjà que ces délimitations n'étaient pas à considérer de façon dogmatique et qu'elles pourraient bien être causées par des modifications métaboliques mineures, ne justifiant pas la création d'espèces.

Tableau 33 – Composition chimique de quatre chimiotypes de Pseudevernia et la détection de ces molécules dans quatre échantillons analysés ici. « + » signifie la présence de la molécule, « - » signifie son absence, « ± » sa présence occasionnelle, « ! » l'absence de la molécule dans la LDB-Orbitrap mais son identification putative avec la LDB-Lit.

	Atranorine	Acide physodique	Acide 3-hydroxyphysodique	Acide alectoronique	Acide 2' -*O*-méthylphysodique	Acide olivétorique	Acide 4-*O*-déméthylmicrophyllinique	Acide lécanorique
P. furfuracea	+	+	+	+	±	-	-	-
P. olivetorina-1	+	-	-	-	-	+	+	-
P. olivetorina-2	+	+	-	+	±	+	+	-
P consocians	+	-	-	-	-	-	-	+
JB/06/56	!	-	-	-	!	-	!	+
JB/08/99	!	-	+	-	-	-	-	+
JB/95/191	!	+	+	+	!	+	!	+
JB/17/212	!	+	+	+	!	+	!	+

Les analyses LC-MS réalisées ici sur quatre échantillons de *Pseudevernia furfuracea* prélevés en France contredisent déjà cette ségrégation sur la base de composés caractéristiques. Tous contiennent de l'acide lécanorique, normalement associé à chimiotype Nord-Américain, en ne considérant que les données de déréplication, JB/06/56 ne correspond à aucune espèce sur la base de sa chimie, JB/08/99 s'apparenterait soit à *P. consocians* ou à *P. furfuracea*, et JB/95/191 ainsi que JB/17/212 devraient correspondre à la « race chimique » rare de *P. olivetorina* ou à *P. consocians*. En LC-MS, comme en CCM, il est important de garder à l'esprit que l'absence de preuve n'est pas la preuve d'absence, et que ce n'est pas parce qu'un composé n'a pas été détecté, qu'il n'est pas produit par l'organisme. Il ne resterait comme argument que la concentration relative des composés, sur la base de la « taille des tâches » sur CCM, mais il faudrait alors s'interroger sur le sens d'une délimitation d'espèces basée sur le degré d'expression des voies métaboliques. Autrement, ne faudrait-il donc pas créer une nouvelle espèce à chaque pour chaque culture aposymbiotique de lichen, puisque leur chimie diffère singulièrement de celle des spécimens prélevés dans la nature ? Les changements de conditions de culture induisant une modulation de l'expression des gènes seraient-ils donc une méthode de spéciation *in vitro* ? Ces auteurs avaient déjà mis-en-garde quant à

la signification de ces données lors de leurs publications dans les années 70, force est de constater qu'elles ont bel et bien été utilisées de façon dogmatique.

Il reste cependant envisageable d'utiliser la chimie en taxonomie, mais avec des méthodes plus représentatives du métabolome, à faire correspondre avec des données de génomique et de transcriptomique. Du point de vue de la qualité des données, il semble étonnant de comparer des données de génétique à des données de profilage CCM qui sont encore la norme dans ce domaine (Fehrer, Slavíková-Bayerová, and Orange 2008; Miadlikowska and Lutzoni 2000; Matteucci et al. 2017; Leavitt, Johnson, and St. Clair 2011). L'étude de la chimie des lichens doit passer le pas et se mettre au même niveau que l'étude de leur génétique. Les outils permettant ceci commencent déjà à voir le jour, avec des classifications produites à partir de données LC-MS/MS comparables à celles produites à l'aide de données de génétique (**Figure 93**).

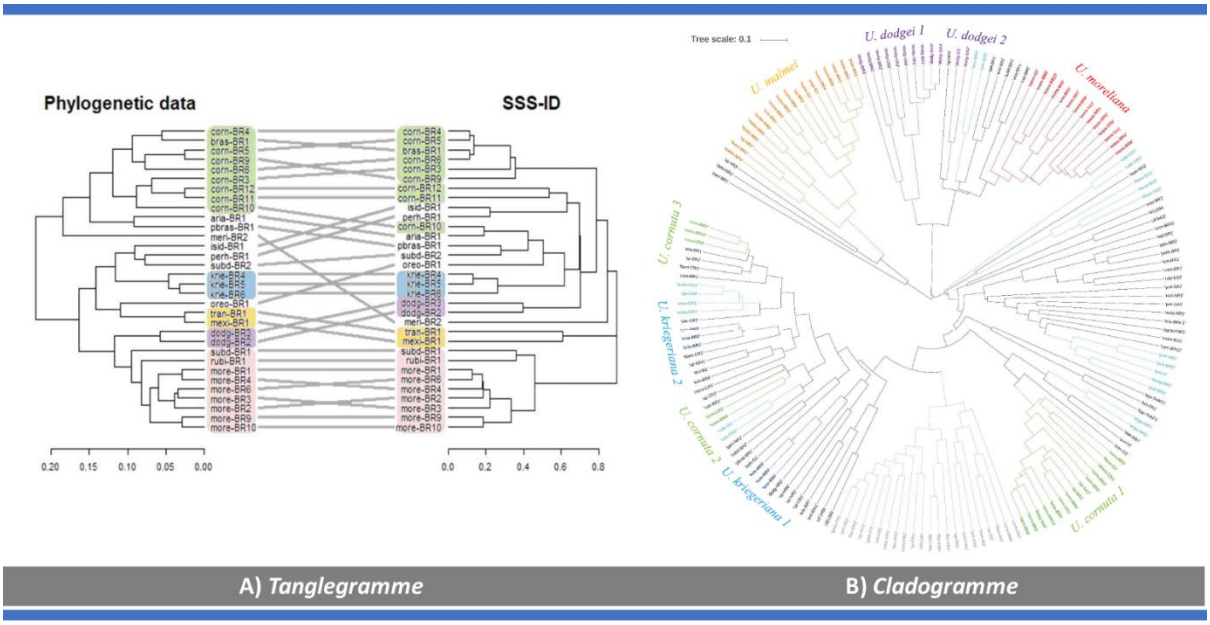


Figure 93 – Exemples d'applications de données LC-MS/MS à la taxonomie. Figures extraites du rapport de stage de M2 de Simon OLLIVIER effectué à Genève en 2019. A) Tanglegramme permettant de comparer une classification de lichens produite par des données de phylogénétique à un équivalent produit par des données LC-MS (SSS-ID). B) Cladogramme généré par des données LC-MS, avec en couleur les coïncidences avec les groupes produits grâce à des données de génétiques.

4. Des solutions apportées par les approches molécule-centrées.

Les approches de réseaux moléculaires basées sur la prédiction de molécules commencent à se développer (Schmid et al. 2020). Elles confèrent des avantages considérables pour l'étude des données LC-MS en simplifiant les données par le regroupement des ions autour du nœud de la molécule qui les a générés. Ceci permet de baser le raisonnement de l'utilisateur directement sur des molécules plutôt que d'interpréter des ions.

Ce genre d'approche existe à présent sur MZmine, XCMS-CAMERA et MS-DIAL avec une visualisation sur des réseaux via le GNPS. Elles ont tout de même été explorées ici indépendamment pour permettre d'avoir un outil plus spécialisé & modifiable à façon

pour les problématiques abordées ici : *Molnotator*. Les différences notables concernent les méthodes de mise en relation molécule-ion, la prédiction de classe structurale, la combinaison de polarité et la déréplication guidée.

Molnotator a été créé en réponse aux problématiques soulevées précédemment, notamment la prise en compte des adduits dans les réseaux et la déréplication. Grâce à la LDB-motifDB, une base de données de motifs créée à partir de la LDB étendue sur MS2LDA, un module de prédiction de classes structurales a été intégré : *Classnotator*. La déréplication par la LDB a été guidée avec la prédiction des adduits et la prédiction des formes neutres a permis d'utiliser la LDB-Lit directement sur les masses des molécules, augmentant considérablement la fiabilité des déréplications.

5. Quant à la diversité chimique des lichens.

Après validation sur la LDB-Orbitrap, *Molnotator* a été utilisé pour étudier de façon globale la chimie des lichens, le sujet principal de ces travaux. Un nombre élevé d'échantillons lichéniques (300) représentant plus de 123 espèces a été analysé par LC-MS (Q-Exactive Focus du Laboratoire de Genève) et soumis à MZmine puis *Molnotator*. Certains de ces organismes étaient déjà bien étudiés, mais la plupart ne l'ont été que sommairement. Ceci a permis de prédire presque 8000 molécules en combinant les deux polarités, même si ce nombre est à considérer avec précaution, beaucoup n'étaient prédites que par l'association d'un faible nombre d'ions (2-3). Un examen manuel complet permettra d'établir le nombre de composés avec plus de certitude. Cependant, même les molécules absentes de la LDB ont pu être détectées sous forme de nœud neutre et dérépliquées grâce à la LDB-Lit. La grande majorité de ces nœuds demeure sans annotation, mais la prise en compte de leur contexte d'ionisation permet d'aider l'utilisateur dans l'interprétation. En considérant également le lichen dans lequel ces molécules ont été détectées, il est possible d'enrichir la LDB pour qu'elle se rapproche du niveau d'exhaustivité de la LDB-Lit et des données utilisées en CCM.

Les données restent encore largement inexplorées, à l'exception de celles d'*Evernia prunastri* et de *Cladonia gracilis*, lichen déjà bien étudié pour le premier et peu étudié pour le second. *C. gracilis* a été sélectionné pour procéder à une étude phytochimique qui a déjà pu confirmer certaines interprétations du réseau, comme l'identité du neutre n439 en tant que l'acide fumarprotocétrarique. Une focalisation particulière sera faite sur les molécules présentes en quantité isolable et n'appartenant pas aux bases de données.

Par ailleurs, certains genres peu réputés pour leur diversité moléculaire (*Peltigera*, *Umbilicaria*) se sont avérés être assez riches en métabolites. Il peut cependant être remarqué que plus une espèce est représentée par des échantillons, plus celle-ci présentera de molécules (comme *Parmelia sulcata*, **Annexe**, **Figure S-190**). La seule chose pouvant être avancée à ce stade est que chaque échantillon apportera de nouvelles molécules. Il est trop tôt à ce pour expliquer de façon rationnelle ce comportement, une liste non-exhaustive de pistes à envisager peut être avancée :

- Contaminations :
    - Par d'autres organismes lors du prélèvement (lichens, substrat, parasites).
    - Par des microbes dans les herbiers.
    - Par d'autres échantillons pendant l'extraction & la préparation.
    - Par d'autres échantillons pendant l'analyse.
- Perte de données :
    - Ponctuelle dans un échantillon qui a été mal préparé (extractions, SPE…)
    - La compétition entre ions en mode DDA.
    - Suppression ionique variable entre échantillons.
- Erreurs d'identification des échantillons :
    - Echantillon mal identifié.
    - Echantillon appartenant à une population présentant des différences chimiques (potentielles crypto-espèces inconnues, ou connues mais non recherchées).
- Variabilité chimique :
    - Différences métaboliques dues à l'âge de l'échantillon.
    - Différences dues aux organes prélevés dans les échantillons.
    - Différences métaboliques dues à des conditions environnementales modifiant ponctuellement le métabolisme.
    - Différences métaboliques dues à une variabilité dans la composition de la symbiose (champignons epi- ou endolichéniques, communauté bactérienne, levures eucaryotes etc…).
    - Différences métaboliques au niveau de l'individu.
- Variabilité inexpliquée dans le traitement informatique.

Finir l'interprétation des réseaux et agrandir la LDB semble cependant être la priorité, et ceci permettrait également de mieux comprendre les variabilités observées. Avec une LDB regroupant les données pour quelques milliers de composés et des méthodes de traitement de données et de déréplication molécule-centrées, l'étude des lichens en métabolomique permet de révéler l'étendue des composés inconnus dans ces organismes. Même si parmi les 8000 molécules prédites certaines sont nécessairement des faux positifs ou des fragments de source, il reste évident que cette simple analyse de 300 échantillons a permis de mettre en évidence plus de molécule qu'il n'y a de composés attribués aux lichens.

L'utilisation des molécules plutôt que des ions serait à privilégier, ou du moins à mettre en parallèle, lors des analyses statistiques. Trop souvent, les ions les plus discriminants entre deux échantillons ne sont en réalité que différents signaux d'une ou deux molécules, induisant l'observateur en erreur quant aux marqueurs à rechercher.

6. Perspectives.

Les méthodes utilisées ici sont appropriées pour l'étude d'herbiers qui ne sont que trop peu exploités dans le domaine de la chimie des produits naturels. Les analyses LC-MS peuvent se contenter de quelques milligrammes d'échantillon, ce qui permet de les

étudier avec un impact qui reste minimal. Ceci peut avoir un rôle dans l'étude de la stabilité dans de temps de certaines molécules, comme ce qui avait été fait précédemment au laboratoire (Chollet-Krugler et al. 2019). Il faut cependant garder à l'esprit que l'évolution rapide des méthodes de métabolomique devrait décourager les prélèvements hâtifs dans des herbiers historiques, puisque ces méthodes pourraient être dépassées avant même la publication des résultats.

Comme établi précédemment, les approches molécule-centrées permettent d'entrer dans un cercle vertueux de déréplication et de mieux comprendre le métabolome des organismes étudiés. Les différents signaux d'une même molécule sont intégrés dans des bases de données, ce qui réduit la part d'inconnues dans les analyses LC-MS et permet à l'utilisateur de se focaliser sur les *unknown unknowns*, ce qui contribue davantage à réduire la portion d'inconnues du métabolome. Ces approches sont en plein développement et sont déjà disponibles publiquement sur le GNPS (Schmid et al. 2020). La vitesse de progression laisserait supposer qu'il soit possible d'identifier la majorité des signaux LC-MS dans un futur proche, ce qui aurait des applications qui dépassent le domaine de la chimie des produits naturels.

Les travaux sur la reproductibilité des spectres MS^2 grâce aux ions thermomètres devraient être continués puisqu'ils permettraient de faire correspondre les analyses LC-MS/MS aux bases de données et réduire le taux de nœuds non-identifiés. Les applications de ces réseaux exploités à leur plein potentiel dépassent le domaine de la chimie des produits naturels et de la recherche de composés bioactifs. Ceci peut trouver une utilité dans la recherche de biomarqueurs, l'écologie chimique, la métabolomique appliquée à la santé et surtout, à faire le lien entre génétique et phénotype, comme la métabolomique avait été originellement conçue.

Les réseaux moléculaires pourraient par ailleurs être comparés aux réseaux métaboliques. Les premiers permettent d'observer les composés détectés dans l'organisme, alors que les second permettent de modéliser de façon rationnelle les composés qui devraient être produits. Le recoupement des deux méthodes pourrait permettre d'établir avec plus de certitude l'identité des molécules détectées. Comme les lichens sont des symbioses, ceci permettrait également de mieux comprendre l'origine des molécules : un composé détecté par LC-MS et non prédit dans le réseau métabolique pourrait provenir d'un organisme qui n'a pas été pris en compte dans la symbiose.

Mot de la fin

J'ai pu m'intéresser au naturalisme très tôt et je dois cela à mes parents qui ont laissé libre cours à mes loisirs dès mon plus jeune âge. Ils m'ont également encouragé à poursuivre dans le domaine scientifique, là où notre regrettable système éducatif n'aura voulu que m'en empêcher. Je leur dois beaucoup, pour cela et pour leur soutien inconditionnel, et ce maigre hommage ne saurait leur rendre justice.

Je retiens de mes études supérieures, l'engagement associatif auprès du CNEN et de l'AMO. Merci à Margot, Zakaria, Corentin et les autres membres de cette association naturaliste étudiante que nous avons dirigé pendant plus de quatre ans, bien que le naturalisme ait été bien mis à mal dans cette université. Merci à René Chéreau, Gilbert Ouvrard, Chantal Maillard et à tous les membres de l'AMO avec qui nous avons fait des sorties et animé les salons mycologiques. C'est surtout grâce à vous que j'ai pu me lancer dans la chimie des produits naturels puisque vous m'avez mis en contact avec Nicolas Ruiz, Catherine Roullier et Yves-François Pouchus au MMS. C'est en passant là que j'ai été mis en relation, suite aux recommandations de Samuel Bertrand, avec le laboratoire de Jean-Luc Wolfender à Genève. J'ai pu y effectuer un stage sous l'encadrement de Pierre-Marie Allard, à qui je dois mon intérêt pour la métabolomique et les réseaux moléculaires.

Suite à un stage à l'ISCR, j'ai pu intégrer cette thèse sur le thème des lichens et la métabolomique, et je dois à ce titre remercier chaleureusement mon encadrante Marylène Chollet-Krugler et mes co-directeurs de thèse David Rondeau et Joël Boustie. Joël, ayant entrepris de me recruter après le départ de Pierre Le Pogam Alluard, avait obtenu un financement me permettant d'outrepasser ces ridicules sélections de candidat qui sont maintenant coutume. Merci Joël, ce n'est pas tout le monde qui m'aurait recruté en thèse.

Grâce aux membres de l'équipe COrInt ainsi qu'à Joël Esnault, Philippe Uriac, les membres de l'AFL et Elise Lebreton, j'ai pu acquérir des bases de lichénologie sans pour autant m'y jeter corps et âme. J'ai pu également être initié à la spectrométrie de masse à l'IETR grâce à David Rondeau et Thomas Delhaye qui m'ont permis de mieux appréhender la complexité de cette discipline. Je remercie également Marc Litaudon et Pierre Le Pogam Alluard pour avoir accepté d'être les membres de mon CSI. Merci aussi à Guillaume Bernadat, Mehdi Beniddir, Victor Turpin et Alexis Pinet pour leur accueil lors de mon passage à BioCis et tout spécialement merci à Pierre Le Pogam pour son aide constante lors de la publication de mon premier article. A nouveau, je dois remercier Pierre-Marie Allard, Jean-Luc Wolfender et les membres des équipe Fatho/Fasie qui nous ont accueillis Simon Ollivier et moi pour un nouveau stage. Ce passage à Genève m'aura permis de faire la transition entre R et Python, de rentrer en contact avec les créateurs de MS2LDA et de produire l'essentiel du travail développé ici.

Je remercie à nouveau mes amis jeunes chercheurs, Simon Ollivier, Julia Mocquard, Elise Lebreton, Zakaria Bouchouireb et Margot Wagner avec qui nous avons pu souvent débattre de questions scientifiques.

L'expérience acquise lors de mon parcours et de cette thèse m'ont habitué à remettre en cause ce qui me semble dogmatique, et je m'excuse déjà auprès de Joël et Marylène d'avoir été une tête de mule à maintes reprises !

Plusieurs fois pendant cette thèse, l'usage de la LC-MS avec de la bioinformatique m'a été reproché. Ma réponse aura été de ridiculiser les fanatiques de la CCM dans plusieurs chapitres et la conclusion. L'adulation de cette méthode relève au mieux d'un refus de sortir de sa zone de confort, au pire, d'une incompréhension profonde des sciences et de la complexité du vivant. J'ai également souvent entendu dire qu'il n'y avait pas grand-chose à trouver dans les lichens et je pense avoir prouvé le contraire ici. Il reste encore beaucoup à découvrir dans les lichens mais il faudra savoir s'adapter aux méthodes modernes.

Bibliographie

Acharius, Erik. 1798. *Lichenographiae Svecicae Prodromus*. Linköping: D. G. Björn.

———. 1803. *Methodus qua Omnes Detectos Lichenes*. Stockholm.

———. 1810. *Lichenographia Universalis*. Göttingen.

———. 1814. *Synopsis Methodica Lichenum*. Lund.

Adusumilli, Ravali, and Parag Mallick. 2017. "Data Conversion with ProteoWizard MsConvert." In *Proteomics: Methods and Protocols*, edited by Lucio Comai, Jonathan E. Katz, and Parag Mallick, 339–68. Springer New York. https://doi.org/10.1007/978-1-4939-6747-6.

Ahmadjian, Vernon. 1967. "A Guide to the Algae Occurring as Lichen Symbionts: Isolation, Culture, Cultural Physiology, and Identification." *Phycologia* 6 (2–3): 127–60. https://doi.org/10.2216/i0031-8884-6-2-127.1.

Allen, Felicity, Allison Pon, Russ Greiner, and David Wishart. 2016. "Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification." *Analytical Chemistry* 88 (15): 7689–97. https://doi.org/10.1021/acs.analchem.6b01622.

Allen, Felicity, Allison Pon, Michael Wilson, Russ Greiner, and David Wishart. 2014. "CFM-ID: A Web Server for Annotation, Spectrum Prediction and Metabolite Identification from Tandem Mass Spectra." *Nucleic Acids Research* 42 (W1): W94–99. https://doi.org/10.1093/nar/gku436.

Aptroot, André, Narla Mota junior, Viviane Monique dos Santos, and Marcela Eugenia da Silva Cáceres. 2016. "New Tropical Calicioid Lichens from South America." *The Lichenologist* 48 (2): 135–39. https://doi.org/10.1017/S0024282915000547.

Aptroot, Andre, and Marcela Eugenia da Silva Caceres. 2018. "New Lichen Species from Chapada Diamantina, Bahia, Brazil." *The Bryologist* 121 (1): 67–79. https://doi.org/10.1639/0007-2745-121.1.067.

Arnold, A Elizabeth, Jolanta Miadlikowska, K. Lindsay Higgins, Snehal D. Sarvate, Paul Gugger, Amanda Way, Valérie Hofstetter, Frank Kauff, and François Lutzoni. 2009. "A Phylogenetic Estimation of Trophic Transition Networks for Ascomycetous Fungi : Are Lichens Cradles of Symbiotrophic Fungal Diversification ?" *Systematic Biology* 58 (3): 283–97. https://doi.org/10.1093/sysbio/syp001.

Asahina, Yasuhiko. 1951. "Neuere Entwicklungen Auf Dem Gebiete Der Flechtenstoffe." In *Fortschritte Der Chemie Organischer Naturstoffe*, 208–39.

Asahina, Yasuhiko, and Shoji Shibata. 1954. *Chemistry Of Lichen Substances*. University of North Carolina Press.

Association Française de Lichénologie. 2016. "Lexique de l'AFL." Https://Www.Afl-Lichenologie.Fr. 2016. https://www.afl-lichenologie.fr/Lexiq_Lich.htm.

Asta, Juliette, Chantal Van Haluwyn, Michel Bertrand, Jean-Michel Sussey, and Jean-Pierre Gavériaux. 2016. *Guide Des Lichens de France, Lichens Des Roches*. Edited by Belin.

Avalos, Adolfo, and Carlos Vicente. 1987. “The Occurrence of Lichen Phenolics in the Photobiont Cells of *Evernia Prunastri*.” *Plant Cell Reports* 6 (1): 74–76. https://doi.org/10.1007/BF00269744.

Bates, Scott T., Garrett W. G. Cropsey, J. Gregory Caporaso, Rob Knight, and Noah Fierer. 2011. “Bacterial Communities Associated with the Lichen Symbiosis.” *Applied and Environmental Microbiology* 77 (4): 1309–14. https://doi.org/10.1128/AEM.02257-10.

Bawingan, Paulina, Mechell Lardizaval, and John Elix. 2019. “Philippine Lichens with Bulbate Cilia – *Bulbothrix* and *Relicina* (Parmeliaceae).” *Philippine Journal of Science* 148 (May): 637–45.

Bertrand, Robert L., Mona Abdel-hameed, and John L. Sorensen. 2018. “Lichen Biosynthetic Gene Clusters. Part I. Genome Sequencing Reveals a Rich Biosynthetic Potential.” *Journal of Natural Products* 81 (4): 723–31. https://doi.org/10.1021/acs.jnatprod.7b00769.

Białońska, D., and F. E. Dayan. 2005. “Chemistry of the Lichen *Hypogymnia Physodes* Transplanted to an Industrial Region.” *Journal of Chemical Ecology* 31 (12): 2005. https://doi.org/10.1007/s10886-005-8408-x.

Bjerke, Jarle W., Kjetil Lerfall, and Arve Elvebakk. 2002. “Effects of Ultraviolet Radiation and PAR on the Content of Usnic and Divaricatic Acids in Two Arctic-Alpine Lichens.” *Photochemical & Photobiological Sciences* 1 (9): 678–85. https://doi.org/10.1039/b203399b.

Böcker, Sebastian, Matthias C. Letzel, Zsuzsanna Lipták, and Anton Pervukhin. 2009. “SIRIUS: Decomposing Isotope Patterns for Metabolite Identification.” *Bioinformatics* 25 (2): 218–24. https://doi.org/10.1093/bioinformatics/btn603.

Bohman-Lindgren, G. 1972. “Chemical Studies on Lichens—XXXIII: Roccanin, a New Cyclic Tetrapeptide from *Roccella Canariensis*.” *Tetrahedron* 28 (17): 4625–30. https://doi.org/10.1016/0040-4020(72)80043-7.

Boom, Pieter P. G. van den, and A Maarten Brand. 2008. “Some New *Lecanora* Species from Western and Central Europe, Belonging to the *L. Saligna* Group, with Notes on Related Species.” *The Lichenologist* 40 (6): 465–97. https://doi.org/10.1017/S0024282908007299.

Boom, Pieter P. G. van den, and Mireia Giralt. 2011. “The Genus *Buellia* s.l. and Some Additional Genera of Physciaceae in the Canary Islands.” *Nova Hedwigia* 92 (1–2). https://doi.org/10.1127/0029-5035/2011/0092-0029.

Boustie, Joël, Sophie Tomasi, and Martin Grube. 2011. “Bioactive Lichen Metabolites: Alpine Habitats as an Untapped Source.” *Phytochemistry Reviews* 10 (3): 287–307. https://doi.org/10.1007/s11101-010-9201-1.

Briggs, Lindsay H., D. R. Castaing, Alison N. Denyer, E. F. Orgias, and C. W. Small. 1972. “Chemistry of Fungi. Part VIII. Constituents of *Valsaria Rubricosa* and the Identification of Papulosin with Valsarin.” *J. Chem. Soc. Perkin Trans. 1*, no. 0: 1464–66. https://doi.org/10.1039/P19720001464.

Brodo, Irwin M. 1978. “Changing Concepts Regarding Chemical Diversity in Lichens.” *The*

Lichenologist 10 (1): 1–11. https://doi.org/10.1017/S0024282978000031.

———. 1986. "Interpreting Chemical Variation in Lichens for Systematic Purposes." *The Bryologist* 89 (2): 132–38. https://doi.org/10.2307/3242753.

Cansaran-Duman, Demet, Demet Cetin, Husniye Sismek, and Nilay Coplu. 2010. "Antimicrobial Activities of the Lichens *Hypogymnia Vittata*, *Hypogymnia Physodes* and *Hypogymnia Tubulosa* and HPLC Analysis of Their Usnic Acid Content." *Asian Journal of Chemistry* 22 (8): 6125–32.

Cantrell, Thomas P., Christopher J. Freeman, Valerie J. Paul, Vinayak Agarwal, and Neha Garg. 2019. "Mass Spectrometry-Based Integration and Expansion of the Chemical Diversity Harbored within a Marine Sponge." *Journal of the American Society for Mass Spectrometry* 30 (8): 1373–84. https://doi.org/10.1021/jasms.8b06062.

CBS, and Landcare Research (custodians). 2019. "Index Fungorum." CABI. 2019. http://www.indexfungorum.org/names/names.asp.

Chamberlain, Scott, and Eduard Szocs. 2013. "Taxize - Taxonomic Search and Retrieval in R." *F1000Research*. https://doi.org/10.12688/f1000research.2-191.v2.

Chamberlain, Scott, Eduard Szoecs, Zachary Foster, Zebulun Arendsee, Carl Boettiger, Karthik Ram, Ignasi Bartomeus, et al. 2020. "Taxize: Taxonomic Information from around the Web." https://github.com/ropensci/taxize.

Chambers, Matthew C., Brendan MacLean, Robert Burke, Dario Amodei, Daniel L. Ruderman, Steffen Neumann, Laurent Gatto, et al. 2012. "A Cross-Platform Toolkit for Mass Spectrometry and Proteomics." *Nature Biotechnology* 30 (10): 918–20. https://doi.org/10.1038/nbt.2377.

Chambers, Susan, M. Morris, and David C. Smith. 1976. "Lichen Physiology XV. The Effect of Digitonin and Other Treatments on Biotrophic Transport of Glucose from Alga to Fungus in *Peltigera Polydactyla*." *New Phytologist* 76 (3): 485–500. https://doi.org/10.1111/j.1469-8137.1976.tb01485.x.

Chollet-Krugler, Marylène, Thi T. Nguyen, Aurelie Sauvager, Holger Thüs, and Joël Boustie. 2019. "Mycosporine-like Amino Acids (MAAs) in Time-Series of Lichen Specimens from Natural History Collections." *Molecules*. https://doi.org/10.3390/molecules24061070.

Chong, Jasmine, Othman Soufan, Carin Li, Iurie Caraus, Shuzhao Li, Guillaume Bourque, David S. Wishart, and Jianguo Xia. 2018. "MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis." *Nucleic Acids Research* 46 (W1): W486–94. https://doi.org/10.1093/nar/gky310.

Corbett, R. E., and R. A. J. Smith. 1969. "Lichens and Fungi. Part VI. Dehydration Rearrangements of 15-Hydroxyhopanes." *Journal of the Chemical Society C: Organic*, no. 1: 44–47. https://doi.org/10.1039/J39690000044.

Corvec, M. Le, C. Boussard-Plédel, F. Charpentier, N. Fatih, B. Le Dare, F. Massart, F. Rojas, et al. 2016. "Chemotaxonomic Discrimination of Lichen Species Using an Infrared Chalcogenide Fibre Optic Sensor: A Useful Tool for on-Field Biosourcing." *RSC Advances* 6 (110): 108187–95. https://doi.org/10.1039/c6ra17140k.

Crittenden, Peter D., and Neil Porter. 1991. "Lichen-Forming Fungi : Potential Sources of

Novel Metabolites." *Trends in Biotechnology* 9 (1): 409–14. https://doi.org/10.1016/0167-7799(91)90141-4.

Crous, P. W., W. Gams, J. A. Stalpers, V. Robert, and G. Stegehuis. 2004. "MycoBank: An Online Initiative to Launch Mycology into the 21st Century" 50 (1): 19–22. https://edepot.wur.nl/31039.

Culberson, Chicita Frances. 1963. "The Lichen Substances of the Genus *Evernia*." *Phytochemistry* 2 (4): 335–40. https://doi.org/10.1016/S0031-9422(00)84857-8.

———. 1969. *Chemical And Botanical Guide To Lichen Products*. University of North Carolina Press.

Culberson, Chicita Frances, and William Louis Culberson. 1976. "Chemosyndromic Variation in Lichens." *Systematic Botany* 1 (4): 325–39. https://doi.org/10.2307/2418700.

———. 1978. "*Cetrelia Cetrarioides* and *C. Monachorum* (Parmeliaceae) in the New World." *The Bryologist* 81 (4): 517–23. https://doi.org/10.2307/3242338.

———. 2001. "Future Directions in Lichen Chemistry." *The Bryologist* 104 (2): 230–34. https://www.jstor.org/stable/3244888 Accessed:

Culberson, Chicita Frances, William Louis Culberson, and Theodore L. Esslinger. 1977. "Chemosyndromic Variation in the *Parmelia Pulla* Group." *The Bryologist* 80 (1): 125–35. https://doi.org/10.2307/3242518.

Culberson, Chicita Frances, and Hörður Kristinsson. 1969. "Studies on the *Cladonia Chlorophaea* Group: A New Species, a New Meta-Depside, and the Identity of 'Novochlorophaeic Acid.'" *The Bryologist* 72 (4): 431–43. https://doi.org/10.2307/3241383.

———. 1970. "A Standardized Method for the Identification of Lichen Products." *Journal of Chromatography* 46: 85–93. https://doi.org/10.1016/S0021-9673(00)83967-9.

Culberson, William Louis. 1969. "The Use of Chemistry in the Systematics of the Lichens." *Taxon* 18 (2): 152–66. https://doi.org/10.2307/1218673.

Culberson, William Louis, and Chicita Frances Culberson. 1970. "A Phylogenetic View of Chemical Evolution in the Lichens." *The Bryologist* 73 (1): 1–31. https://doi.org/10.2307/3241584.

Culberson, William Louis, Chicita Frances Culberson, and Anita Johnson. 1977. "*Pseudevernia Furfuracea-Olivetorina* Relationships: Chemistry and Ecology." *Mycologia* 69 (3): 604–14. https://doi.org/10.1080/00275514.1977.12020098.

Czeczuga, Bazyli, and Kozo Yoshida. 1991. "Carotenoids in Certain Lichens from the Chichibu Mountains, Central Japan." *Feddes Repertorium* 102 (7-8): 661–66. https://doi.org/10.1002/fedr.19911020718.

Davidson, Robert L., Ralf J. M. Weber, Haoyu Liu, Archana Sharma-Oates, and Mark R. Viant. 2016. "Galaxy-M: A Galaxy Workflow for Processing and Analyzing Direct Infusion and Liquid Chromatography Mass Spectrometry-Based Metabolomics Data." *GigaScience* 5 (1). https://doi.org/10.1186/s13742-016-0115-8.

Dembitsky, Valery M. 2015. "Astonishing Diversity of Natural Peroxides as Potential

Therapeutic Agents." *Journal of Molecular and Genetic Medicine* 9 (1): 1–18. https://doi.org/10.4172/1747-0862.1000163.

Deutsch, Eric. 2008. "MzML: A Single, Unifying Data Format for Mass Spectrometer Output." *Proteomics* 8 (14): 2776–77. https://doi.org/10.1002/pmic.200890049.

Díaz-Guerra, Dolores, and Esteban Manrique. 1984. "Sustancias Liquénicas En Taxones de La Provincia de Madrid I. *Evernia Prunastri* (L .) Acb. y *Parmelina Tiliacea* (Hoffm.) Hale." *Lazaroa* 6: 267–68.

Dictionnaire Larousse en Ligne. n.d. "Open Source." Larousse.Fr. Accessed August 6, 2020. https://www.larousse.fr/dictionnaires/francais/open_source/188163.

Dixit, Prateek, Anjali Maurya, Tripti Mishra, Dalip Kumar Upreti, and Mahesh Pal. 2016. "Evaluation of Phytochemical Constituents and Antioxidant Activity of the *Roccella Montagnei*," no. June. https://doi.org/10.21756/cab.v2i01.8610.

Djoumbou-feunang, Yannick, Allison Pon, Naama Karu, Jiamin Zheng, Carin Li, David Arndt, Maheswor Gautam, Felicity Allen, and David S. Wishart. 2019. "CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification." *Metabolites* 9 (72): 1–23. https://doi.org/10.3390/metabo9040072.

Djoumbou Feunang, Yannick, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, et al. 2016. "ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy." *Journal of Cheminformatics* 8 (1): 61. https://doi.org/10.1186/s13321-016-0174-y.

Dole, Malcolm, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson, and M. B. Alice. 1968. "Molecular Beams of Macroions." *The Journal of Chemical Physics* 49 (5): 2240–49. https://doi.org/10.1063/1.1670391.

Drost, Hajk-Georg, Alexander Gabel, Jialin Jiu, Marcel Quint, and Ivo Grosse. 2018. "MyTAI: Evolutionary Transcriptomics with R." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx835.

Duong, Thuc-huy, Mehdi A Beniddir, Grégory Genta-jouve, Thammarat Aree, Marylène Chollet-Krugler, Joël Boustie, Solenn Ferron, et al. 2018. "Tsavoenones A-C: Unprecedented Polyketides with a 1,7-Dioxadispiro[4.0.4.4]Tetradecane Core from the Lichen *Parmotrema Tsavoense*." *Organic & Biomolecular Chemistry* 16 (32): 5913–19. https://doi.org/10.1039/c8ob01280f.

Duong, Thuc-huy, Xuan-phong Ha, Warinthorn Chavasiri, Mehdi A Beniddir, Grégory Genta-Jouve, Joël Boustie, Marylène Chollet-Krugler, et al. 2018. "Sanctis A – C : Three Racemic Procyanidin Analogues from the Lichen *Parmotrema Sancti-Angelii*." *European Journal of Organic Chemistry* 2018 (19): 2247–53. https://doi.org/10.1002/ejoc.201800202.

Duong, Thuc Huy, Bui Linh Chi Huynh, Warinthorn Chavasiri, Marylene Chollet-Krugler, Van Kieu Nguyen, Thi Hoai Thu Nguyen, Poul Erik Hansen, et al. 2017. "New Erythritol Derivatives from the Fertile Form of *Roccella Montagnei*." *Phytochemistry* 137: 156–64. https://doi.org/10.1016/j.phytochem.2017.02.012.

Dutta, Debojyoti, and Ting Chen. 2007. "Speeding up Tandem Mass Spectrometry Database Search: Metric Embeddings and Fast near Neighbor Search." *Bioinformatics*

23 (5): 612–18. https://doi.org/10.1093/bioinformatics/btl645.

Egan, Robert S. 1986. "Correlations and Non-Correlations of Chemical Variation Patterns with Lichen Morphology and Geography." *The Bryologist* 89 (2): 99–110. https://doi.org/10.2307/3242750.

Ekman, Stefan, and Tor Tønsberg. 2019. "*Biatora Alnetorum* (Ramalinaceae, Lecanorales), a New Lichen Species from Western North America." *MycoKeys* 48: 55–65. https://doi.org/10.3897/mycokeys.48.33001.

Elix, John A. 1996. "Biochemistry and Secondary Metabolites." In *Lichen Biology*, 1st ed., 154–80. Nash, Thomas H.

———. 2003. "New Species and New Records of *Xanthoparmelia* (Lichenized Ascomycota, Parmeliaceae) from Western Australia." *The Lichenologist* 35 (4): 291–299. https://doi.org/10.1016/S0024-2829(03)00040-9.

———. 2014. *A Catalogue of Standardized Chromatographic Data and Biosynthetic Relationships for Lichen Substances*. Third Edit. Canberra: Published by the Author.

Elix, John A., and Caroline E. Crook. 1992. "The Joint Occurrence of Chloroxanthones in Lichens , and a Further Thirteen New Lichen Xanthones." *The Bryologist* 95 (1): 52–64. https://doi.org/10.2307/3243785.

Elix, John A., and Kim L. Gaul. 1986. "The Interconversion of the Lichen Depsides Para- and Meta-Scrobiculin, and the Biosynthetic Implications." *Australian Journal of Chemistry* 39 (4): 613–24. https://doi.org/10.1071/CH9860613.

Elix, John A., Umar A. Jenie, and John L. Parker. 1987. "A Novel Synthesis of the Lichen Depsidones Divaronic Acid and Stenosporonic Acid , and the Biosynthetic Implications." *Australian Journal of Chemistry* 40 (8): 1451–64. https://doi.org/10.1071/CH9871451.

Elix, John A., Klaus Kalb, Rupprecht J, and R. Schobert. 2012. "LIAS Metabolites – a Database for the Rapid Identification of Secondary Metabolites of Lichens." 2012. http://liaslight.lias.net/Identification/Navikey/Metabolites/index.html.

Encyclopédie Larousse en ligne. n.d. "Théorie Des Graphes." Larousse.Fr. Accessed August 6, 2020. https://www.larousse.fr/encyclopedie/divers/théorie_des_graphes/183433.

Ernst-Russell, Michael A., John A. Elix, Christina L. L. Chai, David C. R. Hockless, Alanna M. Hurne, and Paul Waring. 1999. "Structure Revision and Cytotoxic Activity of the Scabrosin Esters, Epidithiopiperazinediones from the Lichen *Xanthoparmelia Scabrosa*." *Australian Journal of Chemistry* 52 (4): 279–83. https://doi.org/10.1071/C99019.

Ernst, Madeleine, Kyo B. Kang, Andrés M. Caraballo-Rodríguez, Louis-Felix Nothias, Joe Wandy, Christopher Chen, Mingxun Wang, et al. 2019. "MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools." *Metabolites*. https://doi.org/10.3390/metabo9070144.

Ernst, Madeleine, Louis-Félix Nothias, Justin Johan Jozias van der Hooft, Ricardo R. Silva, C. Haris Saslis-Lagoudakis, Olwen M. Grace, Karen Martinez-Swatson, et al. 2019. "Assessing Specialized Metabolite Diversity in the Cosmopolitan Plant Genus

Euphorbia L." *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2019.00846.

Ernst, Madeleine, Simon Rogers, Ulrik Lausten-Thomsen, Anders Bjorkbom, Susan Svane Laursen, Julie Courraud, Anders Borglum, et al. 2020. "Gestational-Age-Dependent Development of the Neonatal Metabolome." *MedRxiv*, January, 2020.03.27.20045534. https://doi.org/10.1101/2020.03.27.20045534.

Favre-Bonvin, Jean, Noël Arpin, and Christian Brevard. 1976. "Structure de La Mycosporine (P 310)." *Canadian Journal of Chemistry* 54 (7): 1105–13. https://doi.org/10.1139/v76-158.

Fazio, Alejandra T., Mónica T. Adler, Sittiporn Parnmen, Robert Lücking, and Marta S. Maier. 2018. "Production of the Bioactive Pigment Elsinochrome A by a Cultured Mycobiont Strain of the Lichen Graphis Elongata." *Mycological Progress* 17 (4): 479–87. https://doi.org/10.1007/s11557-017-1374-1.

Federhen, Scott. 2012. "The NCBI Taxonomy Database." *Nucleic Acids Research* 40 (D1): D136–43. https://doi.org/10.1093/nar/gkr1178.

———. 2015. "Type Material in the NCBI Taxonomy Database." *Nucleic Acids Research* 43 (D1): D1086–98. https://doi.org/10.1093/nar/gku1127.

Fehrer, Judith, Štěpánka Slavíková-Bayerová, and Alan Orange. 2008. "Large Genetic Divergence of New, Morphologically Similar Species of Sterile Lichens from Europe (*Lepraria*, Stereocaulaceae, Ascomycota): Concordance of DNA Sequence Data with Secondary Metabolites." *Cladistics* 24 (4): 443–58. https://doi.org/10.1111/j.1096-0031.2008.00216.x.

Feige, Guido Benno, Helge Thorsten Lumbsch, Siegfried Huneck, and John A. Elix. 1993. "Short Communication Identification of Lichen Substances by a Standardized Liquid Chromatographic Method." *Journal of Chromatography* 646 (2): 417–27. https://doi.org/10.1016/0021-9673(93)83356-W.

Fiehn Lab. 2016. "MGF Files (MS/MS Container Files)." 2016. https://fiehnlab.ucdavis.edu/projects/lipidblast/mgf-files.

Fiehn, Oliver. 2001. "Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks." *Comparative and Functional Genomics* 2 (3): 155–68. https://doi.org/10.1002/cfg.82.

———. 2002. "Metabolomics — the Link between Genotypes and Phenotypes." In *Functional Genomics*, edited by Chris Town, 155–71. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0448-0_11.

Follmann, Gerhard. 1987. "Vorarbeiten Zu Einer Monographie Der Flechtenfamilie Rocellaceae Chev. VIII: Inhaltsstoffe and Verwandtschaftsverhältnisse von *Roccella Hereroensis* Vain. Und *Roccella Mossamedana* Vain." *Journal of Plant Physiology* 131 (1): 145–51. https://doi.org/10.1016/S0176-1617(87)80275-4.

Fox Ramos, Alexander E., Pierre Le Pogam, Charlotte Fox Alcover, Elvis Otogo N'Nang, Gaëla Cauchie, Hazrina Hazni, Khalijah Awang, et al. 2019. "Collected Mass Spectrometry Data on Monoterpene Indole Alkaloids from Natural Product Chemistry Research." *Scientific Data* 6 (1): 1–6. https://doi.org/10.1038/s41597-019-0028-3.

Franklin, Rosalind Elsie, and Raymond Gosling. 1953a. "The Structure of Sodium Thymonucleate Fibres. II. The Cylindrically Symmetrical Patterson Function." *Acta Crystallographica* 6 (8–9): 678–85. https://doi.org/10.1107/S0365110X53001940.

Franklin, Rosalind Elsie, and Raymond George Gosling. 1953b. "The Structure of Sodium Thymonucleate Fibres. I. The Influence of Water Content." *Acta Crystallographica* 6 (8–9): 673–77. https://doi.org/10.1107/S0365110X53001939.

Gatto, Laurent, and Kathryn S. Lilley. 2012. "MSnbase-an R / Bioconductor Package for Isobaric Tagged Mass Spectrometry Data Visualization , Processing and Quantitation." *Bioinformatics* 28 (2): 288–89. https://doi.org/10.1093/bioinformatics/btr645.

Goga, Michal, Ján Ele, Margaréta Marcin, Dajana Ru, Miriam Ba, and Martin Ba. 2018. "Lichen Metabolites: An Overview of Some Secondary Metabolites and Their Biological Potential." In *Co-Evolution Of Secondary Metabolites*, edited by Jean-Michel Merillon and Kishan Gopal Ramawat, 1–36. Springer International Publishing. https://doi.org/10.1007/978-3-319-76887-8_57-1.

Goloborodko, Anton A., Lev I. Levitsky, Mark V. Ivanov, and Mikhail V. Gorshkov. 2013. "Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics." *Journal of the American Society for Mass Spectrometry* 24 (2): 301–4. https://doi.org/10.1021/jasms.8b04453.

Golubkov, Vladimir V., and Martin Kukwa. 2006. "A Contribution to the Lichen Biota of Belarus." *Acta Mycologica* 41 (1): 155–64. https://doi.org/10.5586/am.2006.019.

Goodacre, Royston. 2005. "Making Sense of the Metabolome Using Evolutionary Computation: Seeing the Wood with the Trees." *Journal of Experimental Botany* 56 (410): 245–54. https://doi.org/10.1093/jxb/eri043.

Gowda, Harsha, Julijana Ivanisevic, Caroline H. Johnson, Michael E. Kurczy, H. Paul Benton, Duane Rinehart, Thomas Nguyen, et al. 2014. "Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses." *Analytical Chemistry* 86 (14): 6931–39. https://doi.org/10.1021/ac500734c.

Green, T. G. A., and David C. Smith. 1974. "Lichen Physiology XIV. Differences between Lichen Algae in Symbiosis and in Isolation." *New Phytologist* 73 (4): 753–66. https://doi.org/10.1111/j.1469-8137.1974.tb01303.x.

Grube, Martin. 2019. "Lichens — a Promising Source of Bioactive Secondary Metabolites." *Plant Genetic Resources* 3 (2): 273–87. https://doi.org/10.1079/PGR200572.

Grube, Martin, and Gabriele Berg. 2009. "Microbial Consortia of Bacteria and Fungi with Focus on the Lichen Symbiosis." *Fungal Biology Reviews* 23 (3): 72–85. https://doi.org/10.1016/j.fbr.2009.10.001.

Grube, Martin, Massimiliano Cardinale, Joao Vieira de Castro Jr, Henry Müller, and Gabriele Berg. 2009. "Species-Specific Structural and Functional Diversity of Bacterial Communities in Lichen Symbioses." *International Society for Microbial Ecology* 3: 1105–15. https://doi.org/10.1038/ismej.2009.63.

Grube, Martin, Paula T. Depriest, Andrea Gargas, and Josef Hafellner. 1995. "DNA Isolation

from Lichen Ascomata." *Mycological Research* 99 (11): 1321–24. https://doi.org/10.1016/S0953-7562(09)81215-X.

Guitton, Yann, Marie Tremblay-Franco, Gildas Le Corguillé, Jean-François Martin, Mélanie Pétéra, Pierrick Roger-Mele, Alexis Delabrière, et al. 2017. "Create, Run, Share, Publish, and Reference Your LC–MS, FIA–MS, GC–MS, and NMR Data Analysis Workflows with the Workflow4Metabolomics 3.0 Galaxy Online Infrastructure for Metabolomics." *The International Journal of Biochemistry & Cell Biology* 93: 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.

Hale, Mason E. 1968. "A Synopsis of the Lichen Genus *Pseudevernia*." *The Bryologist* 71 (1): 1–11. https://doi.org/10.2307/3240645.

Hall, Robert D. 2006. "Plant Metabolomics: From Holistic Hope, to Hype, to Hot Topic." *New Phytologist* 169 (3): 453–68. https://doi.org/10.1111/j.1469-8137.2005.01632.x.

Haluwyn, Chantal Van, Juliette Asta, Jean-Claude Boissière, Philippe Clerc, and Jean-Pierre Gavériaux. 2012. *Guide Des Lichens de France, Lichens Des Sols*. Edited by Belin.

Haluwyn, Chantal Van, Juliette Asta, and Jean-Pierre Gavériaux. 2013. *Guide Des Lichens de France, Lichens Des Arbres*. Edited by Belin.

Haug, Kenneth, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, et al. 2013. "MetaboLights - an Open-Access General-Purpose Repository for Metabolomics Studies and Associated Meta-Data." *Nucleic Acids Research* 41 (D1): 781–86. https://doi.org/10.1093/nar/gks1004.

Hawksworth, David L. 1988. "A New Name for *Dictyonema Pavonium* (Swartz) Parmasto." *The Lichenologist* 20 (1): 101–101. https://doi.org/10.1017/S0024282988000131.

Heller, Stephen, Alan D. McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. 2013. "InChI - the Worldwide Chemical Structure Identifier Standard." *Journal of Cheminformatics* 5 (1): 7. https://doi.org/10.1186/1758-2946-5-7.

Heller, Stephen R., Alan D. McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. 2015. "InChI, the IUPAC International Chemical Identifier." *Journal of Cheminformatics* 7 (1): 23. https://doi.org/10.1186/s13321-015-0068-4.

Herrero-Yudego, Pilar, Mercedes Martin-Pedrosa, Jesus Norato, and Carlos Vicente. 1989. "Some Features about Usnic Acid Accumulation and Its Movement between the Symbionts of the Lichen, *Evernia Prunastri*." *Journal of Plant Physiology* 135 (2): 170–74. https://doi.org/10.1016/S0176-1617(89)80172-5.

Hesse-Feuerbach, Oswald. 1912. "Die Flechtenstoffe." In *Biochemisches Handlexikon*. Vol. 1. https://doi.org/10.1017/CBO9781107415324.004.

Hill, D J, and David C. Smith. 1972. "Lichen Physiology XII. The 'Inhibition Technique.'" *New Phytologist* 71 (1): 15–30. https://doi.org/10.1111/j.1469-8137.1972.tb04806.x.

Holcapek, Michal, Katerina Volna, and Dana Vanerkova. 2007. "Effects of Functional Groups on the Fragmentation of Dyes in Electrospray and Atmospheric Pressure Chemical Ionization Mass Spectra." *Dyes and Pigments* 75 (1): 156–65. https://doi.org/10.1016/j.dyepig.2006.05.040.

Holzmann, Gerhard, and Christian Leuckert. 1990. “Applications of Negative Fast Atom Bombardment and MS/MS to Screening of Lichen Compounds.” *Phytochemistry* 29 (7): 2277–83. https://doi.org/10.1016/0031-9422(90)83052-3.

Honegger, Rosmarie. 2000. “Simon Schwendener (1829–1919) and the Dual Hypothesis of Lichens.” *The Bryologist* 103 (2): 307–13. https://www.jstor.org/stable/3244159?seq=1.

Hooft, Justin Johan Jozias van der, Hosein Mohimani, Anelize Bauermeister, Pieter C. Dorrestein, Katherine R. Duncan, and Marnix H. Medema. 2020. “Linking Genomics and Metabolomics to Chart Specialized Metabolic Diversity.” *Chemical Society Reviews* 49 (11): 3297–3314. https://doi.org/10.1039/D0CS00162G.

Hooft, Justin Johan Jozias van der, Joe Wandy, Michael P. Barrett, Karl E. V. Burgess, and Simon Rogers. 2016. “Topic Modeling for Untargeted Substructure Exploration in Metabolomics.” *Proceedings of the National Academy of Sciences* 113 (48): 13738 LP – 13743. https://doi.org/10.1073/pnas.1608041113.

Hooft, Justin Johan Jozias van der, Joe Wandy, Francesca Young, Sandosh Padmanabhan, Konstantinos Gerasimidis, Karl E. V. Burgess, Michael P. Barrett, and Simon Rogers. 2017. “Unsupervised Discovery and Comparison of Structural Families across Multiple Samples in Untargeted Metabolomics.” *Analytical Chemistry* 89 (14): 7569–77. https://doi.org/10.1021/acs.analchem.7b01391.

Horai, Hisayuki, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, et al. 2010. “MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences.” *Journal of Mass Spectrometry* 45 (7): 703–14. https://doi.org/10.1002/jms.1777.

Huneck, Siegfried. 1999. “The Significance of Lichens and Their Metabolites.” *Naturwissenschaften* 86 (12): 559–70. https://doi.org/10.1007/s001140050676.

———. 2001. *New Results on the Chemistry of Lichen Substances*. Edited by W. Herz, H. Falk, G. W. Kirby, and R. E. Moore. Springer Vienna. https://doi.org/10.1007/978-3-7091-6196-8_1.

———. 2006. “Progress in the Chemistry of Lichen Substances 2000-2005.” *The Journal of the Hattori Botanical Laboratory* 100: 671–94. https://doi.org/10.18968/jhbl.100.0_671.

Huneck, Siegfried, John A. Elix, R. Naidu, and Gerhard Follmann. 1993. “3-*O*-Demethylschizopeltic Acid, a New Dibenzofuran from the Lichen *Roccella Hypomecha*.” *Australian Journal of Chemistry* 46 (3): 407–10. https://doi.org/10.1071/CH9930407.

Huneck, Siegfried, and Isao Yoshimura. 1996. *Identification Of Lichen Substances*. Springer-Verlag. https://doi.org/10.1007/978-3-642-85243-5_2.

Huovinen, K., R. Hiltunen, and M. Von Schantz. 1982. “A Standardized HPLC Method for Analyses of Lichen Compounds from the Genus *Cladonia*.” *Planta Medica* 45: 152–152. https://doi.org/10.1055/s-2007-971326.

IANA. 2020. “Definition of Tab-Separated-Values (Tsv).” Www.Iana.Org. 2020. https://www.iana.org/assignments/media-types/text/tab-separated-values.

Ito, Tatsuya, and Miyako Masubuchi. 2014. "Dereplication of Microbial Extracts and Related Analytical Technologies." *The Journal of Antibiotics* 67 (5): 353–60. https://doi.org/10.1038/ja.2014.12.

Joulain, Daniel, and Raphaël Tabacchi. 2009a. "Lichen Extracts as Raw Materials in Perfumery. Part 1: Oakmoss." *Flavour and Fragrance Journal* 24 (2): 49–61. https://doi.org/10.1002/ffj.1923.

———. 2009b. "Lichen Extracts as Raw Materials in Perfumery. Part 2: Treemoss." *Flavour and Fragrance Journal* 24 (3): 105–16. https://doi.org/10.1002/ffj.1923.

Kalb, Jutarat, Robert Lücking, and Klaus Kalb. 2018. "The Lichen Genera *Allographa* and *Graphis* (Ascomycota: Ostropales, Graphidaceae) in Thailand—Eleven New Species, Forty-Seven New Records and a Key to All One Hundred and Fifteen Species so Far Recorded for the Country." *Phytotaxa* 377 (1): 1–83. https://doi.org/10.11646/phytotaxa.377.1.1.

Kalb, Klaus, and André Aptroot. 2018. "New Lichen Species from Brazil and Venezuela." *The Bryologist* 121 (1): 56–66. https://doi.org/10.1639/0007-2745-121.1.056.

Kang, Kyo Bin, Madeleine Ernst, Justin Johan Jozias van der Hooft, Ricardo R. da Silva, Junha Park, Marnix H. Medema, Sang Hyun Sung, and Pieter C. Dorrestein. 2019. "Comprehensive Mass Spectrometry-Guided Phenotyping of Plant Specialized Metabolites Reveals Metabolic Diversity in the Cosmopolitan Plant Family Rhamnaceae." *The Plant Journal* 98 (6): 1134–44. https://doi.org/10.1111/tpj.14292.

Kang, Kyo Bin, Sunmin Woo, Madeleine Ernst, Justin Johan Jozias van der Hooft, Louis-Félix Nothias, Ricardo R. da Silva, Pieter C. Dorrestein, Sang Hyun Sung, and Mina Lee. 2020. "Assessing Specialized Metabolite Diversity of *Alnus* Species by a Digitized LC–MS/MS Data Analysis Workflow." *Phytochemistry* 173: 112292. https://doi.org/10.1016/j.phytochem.2020.112292.

Kantivilas, Gintaras. 2012. "The Genus *Menegazzia* (Lecanorales: Parmeliaceae) in Tasmania Revisited." *The Lichenologist* 44 (2): 189–246. https://doi.org/10.1017/S0024282911000685.

Katajamaa, Mikko, Jarkko Miettinen, and Matej Orešič. 2006. "MZmine: Toolbox for Processing and Visualization of Mass Spectrometry Based Molecular Profile Data." *Bioinformatics* 22 (5): 634–36. https://doi.org/10.1093/bioinformatics/btk039.

Kessner, Darren, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. 2008. "ProteoWizard : Open Source Software for Rapid Proteomics Tools Development." *Bioinformatics* 24 (21): 2534–36. https://doi.org/10.1093/bioinformatics/btn323.

Khodosovtsev, Alexander Yevgenovich, Anna Oleksiivna Naumovych, Olga Sergeevna Vondrakova, and Jan Vondrak. 2017. "*Athelium Imperceptum* Nyl. (Thelocarpaceae, Ascomycota), a Scarcely Known Ephemeral Lichen of Biological Soil Crusts, New to Ukraine," no. January. https://doi.org/10.14255/2308-9628/10.63/9.

Kim, Hyun Woo, Mingxun Wang, Christopher A. Leber, and Louis-félix Nothias. 2020. "NPClassifier : Deep Neural Structural Classification Tool for Natural Products." *ChemRxiv*. https://doi.org/10.26434/chemrxiv.12885494.v1.

Kistenich, Sonja, Mika Bendiksby, Stefan Ekman, Marcela E. S. Cáceres, Jesús E. Hernández M., and Einar Timdal. 2019. "Towards an Integrative Taxonomy of *Phyllopsora* (Ramalinaceae)." *The Lichenologist* 51 (4): 323–92. https://doi.org/10.1017/S0024282919000252.

Kok, Annette. 1966. "A Short History of the Orchil Dyes." *The Lichenologist* 3 (2): 248–72. https://doi.org/10.1017/S002428296600029X.

Kosanić, Marijana, Nedeljko Manojlović, Slobodan Janković, Tatjana Stanojković, and Branislav Ranković. 2013. "*Evernia Prunastri* and *Pseudoevernia Furfuraceae* Lichens and Their Major Metabolites as Antioxidant, Antimicrobial and Anticancer Agents." *Food and Chemical Toxicology* 53: 112–18. https://doi.org/10.1016/j.fct.2012.11.034.

Kristinsson, Hörður. 2016. *Íslenskar Fléttur*. Edited by Opna.

la Coba, F. de, J. Aguilera, F. L. Figueroa, M. V. de Gálvez, and E. Herrera. 2009. "Antioxidant Activity of Mycosporine-like Amino Acids Isolated from Three Red Macroalgae and One Marine Lichen." *Journal of Applied Phycology* 21 (2): 161–69. https://doi.org/10.1007/s10811-008-9345-1.

Labuda, Ján, Richard P. Bowater, Miroslav Fojta, Günter Gauglitz, Zdeněk Glatz, Ivan Hapala, Jan Havliš, et al. 2018. "Terminology of Bioanalytical Methods (IUPAC Recommendations 2018)." *Pure and Applied Chemistry* 90 (7): 1121–98. https://doi.org/10.1515/pac-2016-1120.

Lafferty, Daryl, and Frank Bungartz. 2018. "WinTab 64bit, a Revised Version of Wintabolites." Consortium of Lichen Herbaria / Consorcio Herbarios de Líquenes – Help & Resources. https://lichenportal.org/help-resources/index.php/resources-2/resources/.

LaGreca, Scott. 1999. "A Phylogenetic Evaluation of the *Ramalina Americana* Chemotype Complex (Lichenized Ascomycota , Ramalinaceae) Based on RDNA ITS Sequence Data." *The Bryologist* 102 (4): 602–18. https://doi.org/10.2307/3244250.

Laily, B. Din, Zuriati Zakaria, Mohd Wahid Samsudin, and John A. Elix. 2010. "Chemical Profile of Compounds from Lichens of Bukit Larut , Peninsular Malaysia." *Sains Malaysiana* 39 (6): 901–8.

Latkowska, Ewa, Beata Bober, Ewelina Chrapusta, Michal Adamski, Ariel Kaminski, and Jan Bialczyk. 2015. "Phytochemistry Secondary Metabolites of the Lichen *Hypogymnia Physodes* (L.) Nyl. and Their Presence in Spruce (*Picea Abies* (L.) H. Karst.) Bark." *Phytochemistry* 118: 116–23. https://doi.org/10.1016/j.phytochem.2015.08.016.

Lawrey, James D. 2003. "Lichenicolous Fungi: Interactions, Evolution, and Biodiversity." *The Bryologist* 106 (1): 80–120. https://doi.org/10.1639/0007-2745(2003)106[0080:LFIEAB]2.0.CO;2.

Leavitt, Steven D., Leigh Johnson, and Larry L. St. Clair. 2011. "Species Delimitation and Evolution in Morphologically and Chemically Diverse Communities of the Lichen-Forming Genus *Xanthoparmelia* (Parmeliaceae, Ascomycota) in Western North America." *American Journal of Botany* 98 (2): 175–88. https://doi.org/10.3732/ajb.1000230.

Legaz, María Estrella, Adolfo Avalos, Marta de Torres, María Isabel Escribano, Azucena González, Angeles Martin-Falquina, Elena Pérezurria, and Carlos Vicente. 1986. "Annual Variations in Arginine Metabolism and Phenolic Content of *Evernia Prunastri*." *Environmental and Experimental Botany* 26 (4): 385–96. https://doi.org/10.1016/0098-8472(86)90027-4.

Lendemer, James C. 2013. "Two New Sterile Species of *Loxospora* (Sarrameanaceae: Lichenized Ascomycetes) from the Mid-Atlantic Coastal Plain." *Journal of North Carolina Academy of Science* 129 (3): 71–81. https://doi.org/10.7572/2167-5880-129.3.71.

Levitsky, Lev I., Joshua A. Klein, Mark V. Ivanov, and Mikhail V. Gorshkov. 2019. "Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework." *Journal of Proteome Research* 18 (2): 709–14. https://doi.org/10.1021/acs.jproteome.8b00717.

Lin, Simon M., Lihua Zhu, Andrew Q. Winter, Maciek Sasinowski, and Warren A. Kibbe. 2005. "What Is MzXML Good For?" *Expert Review of Proteomics* 2 (6): 839–45. https://doi.org/10.1586/14789450.2.6.839.

Linné, Carl von. 1753. *Species Plantarum*. Edited by Lars Salvius. 1st ed. Stockholm.

Louwhoff, Simone H. J. J. 2005. "The Lichen Genus *Nephroma* in Australia." *Muelleria* 10: 3–10.

Lücking, Robert, Brendan P. Hodkinson, and Steven D. Leavitt. 2017. "The 2016 Classification of Lichenized Fungi in the Ascomycota and Basidiomycota – Approaching One Thousand Genera." *The Bryologist* 119 (4): 361–416. https://doi.org/10.1639/0007-2745-119.4.361.

Lumbsch, Helge Thorsten. 1998a. "Taxonomic Use of Metabolic Data in Lichen-Forming Fungi." In *Chemical Fungal Taxonomy*, edited by Marcel Dekker, 345–87. New York: CRC Press.

———. 1998b. "The Use of Metabolic Data in Lichenology at the Species and Subspecific Levels." *The Lichenologist* 30 (4): 357–67. https://doi.org/10.1006/lich.1998.0147.

Lumbsch, Helge Thorsten, Alan W. Archer, and John A. Elix. 2007. "A New Species of *Loxospora* (Lichenized Ascomycota: Sarrameanaceae) from Australia." *The Lichenologist* 39 (6): 509–17. https://doi.org/10.1017/S0024282907007153.

Lumbsch, Helge Thorsten, and Steven D. Leavitt. 2011. "Goodbye Morphology ? A Paradigm Shift in the Delimitation of Species in Lichenized Fungi." *Fungal Diversity* 50 (1): 59–72. https://doi.org/10.1007/s13225-011-0123-z.

Mack, L. L., P. Kralik, A. Rheude, and M. Dole. 1970. "Molecular Beams of Macroions. II." *The Journal of Chemical Physics* 52 (10): 4977–86. https://doi.org/10.1063/1.1672733.

Mann, Matthias. 1990. "Electrospray: Its Potential and Limitations as an Ionization Method for Biomolecules." *Organic Mass Spectrometry* 25 (11): 575–87.

Mark, Kristiina, Tiina Randlane, Göran Thor, Jae-seoun Hur, Walter Obermayer, and Andres Saag. 2019. "Lichen Chemistry Is Concordant with Multilocus Gene Genealogy in the Genus Cetrelia (Parmeliaceae, Ascomycota)." *Fungal Biology* 123

(2): 125–39. https://doi.org/10.1016/j.funbio.2018.11.013.

Mathey, Annick, Peter Spiteller, and Wolfgang Steglich. 2002. "Draculone, a New Anthraquinone Pigment from the Tropical Lichen *Melanotheca Cruenta*." *Zeitschrift Für Naturforschung C* 57 (7–8): 565–67. https://doi.org/10.1515/znc-2002-7-801.

Matrixscience. 2019. "Mascot Database Search, Data File Format." Matrixscience.Com. 2019. http://www.matrixscience.com/help/data_file_help.html.

Matteucci, Enrica, Andrea Occhipinti, Rosanna Piervittori, Massimo E. Maffei, and Sergio E. Favero-Longo. 2017. "Morphological, Secondary Metabolite and ITS (RDNA) Variability within Usnic Acid-Containing Lichen Thalli of *Xanthoparmelia* Explored at the Local Scale of Rock Outcrop in W-Alps." *Chemistry & Biodiversity* 14 (6): e1600483. https://doi.org/10.1002/cbdv.201600483.

May, Philip F. 1997. "*Ophioparma Lapponica*—a Misunderstood Species." *Harvard Papers in Botany* 2 (2): 213–28. http://www.jstor.org/stable/41761547.

McAvoy, Andrew C., Olakunle Jaiyesimi, Paxton H. Threatt, Tyler Seladi, Joanna B. Goldberg, Ricardo R. da Silva, and Neha Garg. 2020. "Differences in Cystic Fibrosis-Associated *Burkholderia Spp.* Bacteria Metabolomes after Exposure to the Antibiotic Trimethoprim." *ACS Infectious Diseases* 6 (5): 1154–68. https://doi.org/10.1021/acsinfecdis.9b00513.

McCarthy, Patrick M., and John A. Elix. 1996. "*Myeloconis*, a New Genus of Pyrenocarpous Lichens from the Tropics." *The Lichenologist* 28 (5): 401–414. https://doi.org/10.1006/lich.1996.0038.

McCarthy, Patrick M., G. Kantvilas, and John A. Elix. 2001. "*Amphorothecium*, a New Pyrenocarpous Lichen Genus from New South Wales, Australia." *The Lichenologist* 33 (4): 291–96. https://doi.org/10.1006/lich.2001.0330.

McLuskey, Karen, Joe Wandy, Isabel Vincent, Justin Johan Jozias van der Hooft, Simon Rogers, Karl Burgess, and Rónán Daly. 2020. "Decomposing Metabolite Set Activity Levels with PALS." *BioRxiv*, January, 2020.06.07.138974. https://doi.org/10.1101/2020.06.07.138974.

McNaught, Alan D. 2006. "The IUPAC International Chemical Identifier: InChl — a New Standard for Molecular Informatics." *Chemistry International* 28 (6). http://publications.iupac.org/ci/2006/2806/4_tools.html.

McNaught, Alan D., A. Wilkinson, and S. J. Chalk. 2019. "IUPAC. Compendium of Chemical Terminology, Online Version." Blackwell Scientific Publications. 2019. https://doi.org/10.1351/goldbook.

Mendes, Pedro. 2002. "Emerging Bioinformatics for the Metabolome." *Briefings in Bioinformatics* 3 (2): 134–45. https://doi.org/10.1093/bib/3.2.134.

Miadlikowska, Jolanta, and François Lutzoni. 2000. "Phylogenetic Revision of the Genus *Peltigera* (Lichen-forming Ascomycota) Based on Morphological, Chemical, and Large Subunit Nuclear Ribosomal DNA Data." *International Journal of Plant Sciences* 161 (6): 925–58. https://doi.org/10.1086/317568.

Mietzsch, E., Helge Thorsten Lumbsch, and John A. Elix. 1992. "Wintabolites (Mactabolites for Windows)." In .

———. 1993. “A New Computer Program for the Identification on Lichen Substances.” *Mycotaxon* 47: 475–79.

Millot, Marion, and Lengo Mambu. 2019. “Champignons Endolichéniques et Épilichéniques : Les Habitants Cachés Des Lichens.” *Bulletin de l’Association Française de Lichénologie* 44: 71–76.

Millot, Marion, Sophie Tomasi, Elisabeth Studzinska, Isabelle Rouaud, and Joël Boustie. 2009. “Cytotoxic Constituents of the Lichen *Diploicia Canescens*.” *Journal of Natural Products* 72 (12): 2177–80. https://doi.org/10.1021/np9003728.

Misra, Biswapriya B., Johannes F. Fahrmann, and Dmitry Grapov. 2017. “Review of Emerging Metabolomic Tools and Resources: 2015–2016.” *Electrophoresis* 38 (18): 2257–74. https://doi.org/10.1002/elps.201700110.

Mitchell, Tom. 1997. *Machine Learning*. Edited by McGraw-Hill.

Mohanty, Ipsita, Sheila Podell, Jason S. Biggs, Neha Garg, Eric E. Allen, and Vinayak Agarwal. 2020. “Multi-Omic Profiling of *Melophlus* Sponges Reveals Diverse Metabolomic and Microbiome Architectures That Are Non-Overlapping with Ecological Neighbors.” *Marine Drugs*. https://doi.org/10.3390/md18020124.

Molnár, Katalin, and Edit Farkas. 2010. “Current Results on Biological Activities of Lichen Secondary Metabolites: A Review.” *Zeitschrift Für Naturforschung C* 65 (3–4): 157–73. https://doi.org/10.1515/znc-2010-3-401.

———. 2011. “Depsides and Depsidones in Populations of the Lichen *Hypogymnia Physodes* and Its Genetic Diversity.” *Annales Botanici Fennici* 48 (6): 473–82. https://doi.org/10.5735/085.048.0605.

Moreno-Ulloa, Aldo, Victoria Sicairos Diaz, Javier A. Tejeda-Mora, Marla I. Macias Contreras, Fernando Díaz Castillo, Abraham Guerrero, Ricardo Gonzales Sanchez, Rafael Vazquez Duhalt, and Alexei Licea-Navarro. 2019. “Metabolic and Metagenomic Profiling of Hydrocarbon-Degrading Microorganisms Obtained from the Deep Biosphere of the Gulf of México.” *BioRxiv*, January, 606806. https://doi.org/10.1101/606806.

Moriyasu, Yuchiko, Hisashi Miyagawa, Nobuo Hamada, Hiromi Miyawaki, and Tamio Ueno. 2005. “Total Synthesis of Four Naturally Occurring 2-Azaanthraquinone Antibiotics, 6-Deoxy-8-Methylbostrycoidin, 6-Deoxybostrycoidin, 7-*O*-Demethyl-6-Deoxybostrycoidin and Scorpinone.” *Tetrahedron* 61 (9): 2295–2300. https://doi.org/10.1016/j.tet.2005.01.035.

Mullis, Kary B. 1990. “The Unusual Origin of the Polymerase Chain Reaction.” *Scientific American* 262 (4): 56–65. https://doi.org/10.1038/scientificamerican0490-56.

Murray, Kermit K., Robert K. Boyd, Marcos N. Eberlin, G. John Langley, Liang Li, and Yasuhide Naito. 2013. “Definitions of Terms Relating to Mass Spectrometry (IUPAC Recommendations 2013).” *Pure and Applied Chemistry* 85 (7): 1515–1609. https://doi.org/10.1351/PAC-REC-06-04-06.

Myers, Owen D., Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. 2017. “One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for

Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks." *Analytical Chemistry* 89 (17): 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947.

Narui, Takao, Keiko Sawada, Satoshi Takatsuki, Toru Okuyama, Chicita Frances Culberson, William Louis Culberson, and Shoji Shibata. 1998. "NMR Assignments of Depsides and Tridepsides of the Lichen Family Umbilicariaceae." *Phytochemistry* 48 (5): 815–22. https://doi.org/10.1016/S0031-9422(97)00958-8.

Narui, Takao, Satoshi Takatsuki, Keiko Sawada, Toru Okuyama, Chicita Frances Culberson, William Louis Culberson, and Shoji Shibata. 1996. "Lasallic Acid, a Tridepside from the Lichen, *Lasallia Asiae-Orientalis*." *Phytochemistry* 42 (3): 839–42. https://doi.org/10.1016/0031-9422(95)00960-4.

Nguyen, Don Duy, Cheng-Hsuan Wu, Wilna J. Moree, Anne Lamsa, Marnix H. Medema, Xiling Zhao, Ronnie G. Gavilan, et al. 2013. "MS/MS Networking Guided Analysis of Molecule and Gene Cluster Families." *Proceedings of the National Academy of Sciences* 110 (28): E2611 LP-E2620. https://doi.org/10.1073/pnas.1303471110.

Nothias-Esposito, Mélissa, Louis Felix Nothias, Ricardo R. Da Silva, Pascal Retailleau, Zheng Zhang, Pieter Leyssen, Fanny Roussi, et al. 2019. "Investigation of Premyrsinane and Myrsinane Esters in *Euphorbia Cupanii* and *Euphobia Pithyusa* with MS2LDA and Combinatorial Molecular Network Annotation Propagation." *Journal of Natural Products* 82 (6): 1459–70. https://doi.org/10.1021/acs.jnatprod.8b00916.

Nothias, Louis Felix, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, et al. 2019. "Feature-Based Molecular Networking in the GNPS Analysis Environment." *BioRxiv*, January, 812404. https://doi.org/10.1101/812404.

Nylander, Wilhelm. 1866. "Hypochlorite of Lime and Hydrate of Potash, Two New Criteria in the Study of Lichens." *Botanical Journal of the Linnean Society* 9 (38): 358–65. https://doi.org/10.1111/j.1095-8339.1866.tb01301.x.

O'Boyle, Noel M., Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. 2011. "Open Babel: An Open Chemical Toolbox." *Journal of Cheminformatics* 3 (1): 33. https://doi.org/10.1186/1758-2946-3-33.

Oliver, Stephen G., Michael K. Winson, Douglas B. Kell, and Frank Baganz. 1998. "Systematic Functional Analysis of the Yeast Genome." *Trends in Biotechnology* 16 (9): 373–78. https://doi.org/10.1016/S0167-7799(98)01214-1.

Olivier-Jimenez, Damien, Marylène Chollet-Krugler, David Rondeau, Mehdi A. Beniddir, Solenn Ferron, Thomas Delhaye, Pierre-Marie Allard, et al. 2019. "A Database of High-Resolution MS/MS Spectra for Lichen Metabolites." *Scientific Data* 6 (1): 294. https://doi.org/10.1038/s41597-019-0305-1.

Olivier-Jimenez, Damien, Marylène Chollet-Krugler, David Rondeau, Mehdi A. Beniddir, Solenn Ferron, Thomas Delhaye, Harrie J. M. Sipman, Robert Lücking, Joël Boustie, and Pierre Le Pogam. 2019. "A Database of High-Resolution MS/MS Spectra for Lichen Metabolites (Repository)." MetaboLights. 2019. https://www.ebi.ac.uk/metabolights/MTBLS999.

Olivon, Florent, Nicolas Elie, Gwendal Grelier, Fanny Roussi, Marc Litaudon, and David Touboul. 2018. "MetGem Software for the Generation of Molecular Networks Based on the T-SNE Algorithm." *Analytical Chemistry* 90 (23): 13900–908. https://doi.org/10.1021/acs.analchem.8b03099.

Olivon, Florent, Gwendal Grelier, Fanny Roussi, Marc Litaudon, and David Touboul. 2017. "MZmine 2 Data-Preprocessing to Enhance Molecular Networking Reliability." *Analytical Chemistry* 89 (15): 7836–40. https://doi.org/10.1021/acs.analchem.7b01563.

Pedrioli, Patrick G. A., Jimmy K. Eng, Robert Hubley, Mathijs Vogelzang, Eric W. Deutsch, Brian Raught, Brian Pratt, et al. 2004. "A Common Open Representation of Mass Spectrometry Data and Its Application to Proteomics Research." *Nature Biotechnology* 22 (11): 1459–66. https://doi.org/10.1038/nbt1031.

Perru, Olivier, and Armand Colin. 2006. "Aux Origines Des Recherches Sur La Symbiose Vers 1868-1883." *Revue d'histoire Des Sciences* 59 (1): 5–27. https://doi.org/10.3917/rhs.591.0005.

Pittayakhajonwut, Pattama, Veera Sri-indrasutdhi, Aibrohim Dramae, Sanisa Lapanun, Rapheephat Suvannakad, and Morakot Tantichareon. 2009. "Graphisins A and B from the Lichen *Graphis Tetralocularis*." *Australian Journal of Chemistry* 62 (4): 389–91. https://doi.org/10.1071/CH08313.

Pletnev, Igor, Andrey Erin, Alan D. McNaught, Kirill Blinov, Dmitrii Tchekhovskoi, and Steve Heller. 2012. "InChIKey Collision Resistance: An Experimental Testing." *Journal of Cheminformatics* 4 (1): 39. https://doi.org/10.1186/1758-2946-4-39.

Pluskal, Tomáš, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. 2010. "MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data." *BMC Bioinformatics* 11 (1): 11. https://doi.org/10.1186/1471-2105-11-395.

Pogam, Pierre Le, and Joël Boustie. 2016. "Xanthones of Lichen Source: A 2016 Update." *Molecules* 21 (3): 30. https://doi.org/10.3390/molecules21030294.

Pogam, Pierre Le, Gaëtan Herbette, and Joël Boustie. 2015. "Analysis of Lichen Metabolites, a Variety of Approaches - Recent Advances in Lichenology: Modern Methods and Approaches in Biomonitoring and Bioprospection, Volume 1." In , edited by Dalip Kumar Upreti, Pradeep K Divakar, Vertika Shukla, and Rajesh Bajpai, 229–61. New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2181-4_11.

Pogam, Pierre Le, Anne-cécile Le Lamer, Béatrice Legouin, Joël Boustie, and David Rondeau. 2016. "In Situ DART-MS as a Versatile and Rapid Dereplication Tool in Lichenology : Chemical Fingerprinting of *Ophioparma Ventosa*." *Phytochemical Analysis* 27 (6): 354–63. https://doi.org/10.1002/pca.2635.

Pogam, Pierre Le, Anne-Cécile Le Lamer, Bandi Siva, Béatrice Legouin, Arnaud Bondon, Jérôme Graton, Denis Jacquemin, et al. 2016. "Minor Pyranonaphthoquinones from the Apothecia of the Lichen *Ophioparma Ventosa*." *Journal of Natural Products* 79 (4): 1005–11. https://doi.org/10.1021/acs.jnatprod.5b01073.

Pogam, Pierre Le, Andreas Schinkovitz, Béatrice Legouin, Anne-Cécile Le Lamer, Joël

Boustie, and Pascal Richomme. 2015. “Matrix-Free UV-Laser Desorption Ionization Mass Spectrometry as a Versatile Approach for Accelerating Dereplication Studies on Lichens.” *Analytical Chemistry* 87 (20): 10421–28. https://doi.org/10.1021/acs.analchem.5b02531.

R Development Core Team. 2008. “R: A Language and Environment for Statistical Computing.” Vienna, Austria. http://www.r-project.org.

Rabinowitz, J. D. R., J. G. P. Urdy, L. V. Astag, T. S. Henk, and E. K. Oyuncu. 2011. “Metabolomics in Drug Target Discovery.” *Cold Spring Harbor Laboratory Press* LXXVI. https://doi.org/10.1101/sqb.2011.76.010694.

Rafaëlly, L., Sylvie Héron, Witold Nowik, and Alain Tchapla. 2008. “Optimisation of ESI-MS Detection for the HPLC of Anthraquinone Dyes.” *Dyes and Pigments* 77 (1): 191–203. https://doi.org/10.1016/j.dyepig.2007.05.007.

Rambold, Gerhard. 1996. “LIAS–The Concept of an Identification System for Lichenized and Lichenicolous Fungi.” In *The Third Symposium IAL*.

Rambold, Gerhard, John A. Elix, Bärbel Heindl-tenhunen, Thomas Köhler, Thomas H Nash Iii, Dieter Neubacher, Wolfgang Reichert, Luciana Zedda, and Dagmar Triebel. 2014. “LIAS Light – towards the Ten Thousand Species Milestone.” *MycoKeys* 8: 11–16. https://doi.org/10.3897/mycokeys.8.6605.

Rambold, Gerhard, and Dagmar Triebel. 2007. “Genera of Lichenized and Lichenicolous Ascomycetes – LIAS: A Global Information System for Lichenized and Non-Lichenized Ascomycetes.” In .

Rambold, Gerhard, Luciana Zedda, Jessica R. Coyle, Derek Persoh, Thomas Köhler, and Dagmar Triebel. 2016. “Geographic Heat Maps of Lichen Traits Derived by Combining LIAS Light Description and GBIF Occurrence Data, Provided on a New Platform.” *Biodiversity and Conservation* 25 (13): 2743–51. https://doi.org/10.1007/s10531-016-1199-2.

Ranković, Branislav, Marijana Kosanić, Nedeljko Manojlović, Aleksandar Rančić, and Tatjana Stanojković. 2014. “Chemical Composition of *Hypogymnia Physodes* Lichen and Biological Activities of Some Its Major Metabolites.” *Medical Chemistry Research* 23 (1): 408–16. https://doi.org/10.1007/s00044-013-0644-y.

Ranković, Branislav, M. Mišić, and S. Sukdolak. 2007. “Evaluation of Antimicrobial Activity of the Lichens *Lasallia Pustulata*, *Parmelia Sulcata*, *Umbilicaria Crustulosa*, and *Umbilicaria Cylindrica*.” *Microbiology* 76 (6): 723–27. https://doi.org/10.1134/S0026261707060112.

Řezanka, Tomáš, and Irene A. Guschina. 1999. “Brominated Depsidones from *Acarospora Gobiensis*, a Lichen of Central Asia.” *Journal of Natural Products* 62 (12): 1675–77. https://doi.org/10.1021/np990114s.

Řezanka, Tomáš, Jitka Jáchymová, and Valery M. Dembitsky. 2003. “Prenylated Xanthone Glucosides from Ural’s Lichen *Umbilicaria Proboscidea*.” *Phytochemistry* 62 (4): 607–12. https://doi.org/10.1016/S0031-9422(02)00539-3.

Richardson, David. H. S., D. Jackson Hill, and David C. Smith. 1968. “Lichen Physiology XI. The Role of the Alga in Determining the Pattern of Carbohydrate Movement between

Lichen Symbionts." *New Phytologist* 67 (3): 469–86. https://doi.org/10.1111/j.1469-8137.1968.tb05476.x.

Richardson, David. H. S., and David C. Smith. 1968a. "Lichen Physiology IX. Carbohydrate Movement from the *Trebouxia* Symbiont of *Xanthoria Aureola* to the Fungus." *New Phytologist* 67 (1): 61–68. https://doi.org/10.1111/j.1469-8137.1968.tb05454.x.

———. 1968b. "Lichen Physiology X. The Isolated Algal and Fungal Symbionts of *Xanthoria Aureola*." *New Phytologist* 67 (1): 69–77. https://doi.org/10.1111/j.1469-8137.1968.tb05455.x.

Richardson, David. H. S., David C. Smith, and D. H. Lewis. 1967. "Carbohydrate Movement between the Symbionts of Lichens." *Nature* 214 (5091): 879–82. https://doi.org/10.1038/214879a0.

Richardson, Leonard. 2007. "Beautiful Soup Documentation." *April*.

Robert, Vincent, Duong Vu, Ammar Ben Hadj Amor, Nathalie van de Wiele, Carlo Brouwer, Bernard Jabas, Szaniszlo Szoke, et al. 2013. "MycoBank Gearing up for New Horizons." *IMA Fungus* 4 (2): 371–79. https://doi.org/10.5598/imafungus.2013.04.02.16.

Rogers, Roderick W. 1989. "Chemical Variation and the Species Concept in Lichenized Ascomycetes." *Botanical Journal of the Linnean Society* 101 (2): 229–39. https://doi.org/10.1111/j.1095-8339.1989.tb00156.x.

Rogers, Simon, Cher Wei Ong, Joe Wandy, Madeleine Ernst, Lars Ridder, and Justin Johan Jozias van der Hooft. 2019. "Deciphering Complex Metabolite Mixtures by Unsupervised and Supervised Substructure Discovery and Semi-Automated Annotation from MS/MS Spectra." *Faraday Discussions* 218 (0): 284–302. https://doi.org/10.1039/C8FD00235E.

Rossum, Guido Van, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Röst, Hannes L., Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, et al. 2016. "OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis." *Nature Methods* 13 (9): 741–48. https://doi.org/10.1038/nmeth.3959.

Roullier, Catherine, Marylène Chollet-Krugler, Aurélie Bernard, and Joël Boustie. 2009. "Multiple Dual-Mode Centrifugal Partition Chromatography as an Efficient Method for the Purification of a Mycosporine from a Crude Methanolic Extract of *Lichina Pygmaea*." *Journal of Chromatography B* 877 (22): 2067–73. https://doi.org/10.1016/j.jchromb.2009.05.040.

Roullier, Catherine, Marylène Chollet-Krugler, Eva-Maria Pferschy-Wenzig, Anne Maillard, Gerald N. Rechberger, Béatrice Legouin-Gargadennec, Rudolf Bauer, and Joël Boustie. 2011. "Characterization and Identification of Mycosporines-like Compounds in Cyanolichens. Isolation of Mycosporine Hydroxyglutamicol from *Nephroma Laevigatum* Ach." *Phytochemistry* 72 (11): 1348–57. https://doi.org/10.1016/j.phytochem.2011.04.002.

Roullier, Catherine, Marylène Chollet-Krugler, Pierre van de Weghe, Françoise Lohézic-

Le Devehat, and Joël Boustie. 2010. “A Novel Aryl-Hydrazide from the Marine Lichen *Lichina Pygmaea*: Isolation, Synthesis of Derivatives, and Cytotoxicity Assays.” *Bioorganic & Medicinal Chemistry Letters* 20 (15): 4582–86. https://doi.org/10.1016/j.bmcl.2010.06.013.

Safe, Stephen, Lorna M. Safe, and Wolfgang S. G. Maass. 1975. “Sterols of Three Lichen Species: *Lobaria Pulmonaria*, *Lobaria Scrobiculata* and *Usnea Longissima*.” *Phytochemistry* 14 (8): 1821–23. https://doi.org/10.1016/0031-9422(75)85302-7.

Santiago, Krystle Angelique A., Jayne Nicholei C. Borricano, Joecela N. Canal, Denisse Marie A. Marcelo, Myleen Claire P. Perez, and Thomas Edison E. dela Cruz. 2010. “Antibacterial Activities of Fruticose Lichens Collected from Selected Sites in Luzon Island, Philippines.” *Philippine Science Letters* 3 (2): 18–29.

Schinkovitz, Andreas, Pierre Le Pogam, Séverine Derbré, Emilie Roy-vessieres, Patricia Blanchard, Sangeetha-laura Thirumaran, Dimitri Breard, et al. 2018. “Secondary Metabolites from Lichen as Potent Inhibitors of Advanced Glycation End Products and Vasodilative Agents.” *Fitoterapia* 131: 182–88. https://doi.org/10.1016/j.fitote.2018.10.015.

Schmid, Robin, Daniel Petras, Louis-Félix Nothias, Mingxun Wang, Allegra T. Aron, Annika Jagels, Hiroshi Tsugawa, et al. 2020. “Ion Identity Molecular Networking in the GNPS Environment.” *BioRxiv*, January, 2020.05.11.088948. https://doi.org/10.1101/2020.05.11.088948.

Schumm, Felix, and John A. Elix. 2015. *Atlas of Images of Thin Layer Chromatograms of Lichen Substances*. Norderstedt: Herstellung und Verlag: Books on Demand GmbH.

Schymanski, Emma L., Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ru, Heinz P. Singer, and Juliane Hollender. 2014. “Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence.” *Environmental Science and Technology* 48 (4): 2097–98. https://doi.org/10.1021/es5002105.

Seriña, Estela, Rosario Arroyo, Esteban Manrique, and Leopoldo G. Sancho. 1996. “Lichen Substances and Their Intraspecific Variability within Eleven *Umbilicaria* Species in Spain.” *The Bryologist* 99 (3): 335–42. https://doi.org/10.2307/3244307.

Shafranovich, Y., and Network Working Group. 2005. “Common Format and MIME Type for Comma-Separated Values (CSV) Files.” Https://Tools.Ietf.Org/. 2005. https://tools.ietf.org/html/rfc4180.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Res.*, no. Karp 2001: 2498–2504. https://doi.org/10.1101/gr.1239303.metabolite.

Shibata, Shoji. 2000. “Yasuhiko Asahina (1880-1975) and His Studies on Lichenology and Chemistry of Lichen Metabolites.” *The Bryologist* 103 (4): 710–19.

Silva, Ricardo R. da, Pieter C. Dorrestein, and Robert A. Quinn. 2015. “Illuminating the Dark Matter in Metabolomics.” *Proceedings of the National Academy of Sciences* 112 (41): 12549 LP – 12550. https://doi.org/10.1073/pnas.1516878112.

Sipman, Harrie J. M., John A. Elix, and Thomas H. Nash. 2009. "*Hypotrachyna* (Parmeliaceae, Lichenized Fungi)." *Flora Neotropica* 104 (June): 1–176. http://www.jstor.org/stable/25660972.

Skult, Henrik. 1997. "Notes on the Chemical and Morphological Variation of the Lichen *Ophioparma Ventosa* in East Fennoscandia." *Annales Botanici Fennici* 34 (4): 291–97. http://www.jstor.org/stable/23726487.

Smith, Colin A., Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. 2006. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification." *Analytical Chemistry* 78 (3): 779–87. https://doi.org/10.1021/ac051437y.

Smith, David C., and E. A. Drew. 1965. "Studies in the Physiology of Lichens V. Translocation from the Algal Layer to the Medulla in *Peltigera Polydactyla*." *New Phytologist* 64 (2): 195–200. https://doi.org/10.1111/j.1469-8137.1965.tb05390.x.

Smith, David C., and Susan Molesworth. 1973. "Lichen Physiology XIII. Effects of Rewetting Dry Lichens." *New Phytologist* 72 (3): 525–33. https://doi.org/10.1111/j.1469-8137.1973.tb04403.x.

Smith, David C., and Mark R. D. Seaward. 2013. "A Tribute to Vernon Ahmadjian (19 May 1930–13 March 2012)." *The Lichenologist* 45 (2): 133–36. https://doi.org/10.1017/S0024282912000795.

Søchting, Ulrik. 2016. "Chemosyndromes with Chlorinated Anthraquinones in the Lichen Genus *Caloplaca*," no. January 2001.

Solhaug, Knut Asbjørn, Marius Lind, Line Nybakken, and Yngvar Gauslaa. 2009. "Possible Functional Roles of Cortical Depsides and Medullary Depsidones in the Foliose Lichen *Hypogymnia Physodes*." *Flora* 204 (1): 40–48. https://doi.org/10.1016/j.flora.2007.12.002.

Spribille, Toby, Veera Tuovinen, Philipp Resl, Dan Vanderpool, Heimo Wolinski, M. Catherine Aime, Kevin Schneider, et al. 2016. "Basidiomycete Yeasts in the Cortex of Ascomycete Macrolichens." *Science* 353 (6298): 488–92. https://doi.org/10.1126/science.aaf8287.

Stocker-Wörgötter, Elfie. 2004. "Secondary Chemistry of Lichen-Forming Fungi : Chemosyndromic Variation and DNA- Analyses of Cultures and Chemotypes in the *Ramalina Farinacea* Complex." *The Bryologist* 107 (2): 152–62. https://doi.org/10.1639/0007-2745(2004)107[0152:SCOLFC]2.0.CO;2.

———. 2008. "Metabolic Diversity of Lichen-Forming Ascomycetous Fungi: Culturing, Polyketide and Shikimate Metabolite Production, and PKS Genes." *Natural Product Reports* 25 (1): 188–200. https://doi.org/10.1039/b606983p.

Sturm, Marc, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, et al. 2008. "OpenMS – An Open-Source Software Framework for Mass Spectrometry." *BMC Bioinformatics* 9 (1): 163. https://doi.org/10.1186/1471-2105-9-163.

Swartz, Michael E. 2005. "UPLC™: An Introduction and Review." *Journal of Liquid Chromatography & Related Technologies* 28 (7–8): 1253–63.

https://doi.org/10.1081/JLC-200053046.

Takenaka, Yukiko, Takao Tanahashi, Naotaka Nagakura, and Nobuo Hamada. 2000. "Production of Xanthones with Free Radical Scavenging Properties, Emodin and Sclerotiorin by the Cultured Lichen Mycobionts of *Pyrenula Japonica*." *Zeitschrift Für Naturforschung C* 55 (11–12). https://doi.org/10.1515/znc-2000-11-1211.

———. 2003. "Phenyl Ethers from Cultured Lichen Mycobionts of *Graphis Scripta* Var. *Serpentina* and *G. Rikuzensis*." *Chemical and Pharmaceutical Bulletin* 51 (7): 794–97. https://doi.org/10.1248/cpb.51.794.

Tarasova, Viktoria N., Angella V. Sonina, Vera I. Androsova, and Irina S. Stepanchikova. 2016. "The Lichens of Forest Rocky Communities of the Hill Muroigora (Arkhangelsk Region, Northwest Russia)." *Folia Cryptogamica Estonica* 53 (0 SE-Articles): 111–21. https://doi.org/10.12697/fce.2016.53.13.

Tautenhahn, Ralf, Gary J. Patti, Duane Rinehart, and Gary Siuzdak. 2012. "XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data." *Analytical Chemistry* 84 (11): 5035–39. https://doi.org/10.1021/ac300698c.

Tehler, Anders, Martin Irestedt, Mats Wedin, and Damien Ertz. 2010. "The Old World *Roccella* Species Outside Europe and Macaronesia: Taxonomy, Evolution and Phylogeny." *Systematics and Biodiversity* 8 (2): 223–46. https://doi.org/10.1080/14772001003789554.

Tsurykau, Andrei, Volha Khramchankova, and Jurga Motiejūnaitė. 2012. "*Pycnora Sorophora* (Lecanoraceae) – Lichen Species New to Belarus." *Botanica* 18 (1): 80–82. https://doi.org/10.2478/v10279-012-0010-x.

Varol, Mehmet. 2018. *Lichens as a Promising Source of Unique and Functional Small Molecules for Human Health and Well-Being*. *Studies In Natural Products Chemistry*. 1st ed. Vol. 60. Elsevier B.V. https://doi.org/10.1016/B978-0-444-64181-6.00012-7.

Vicente, Carlos, and Elena Pérez-Urria. 1988. "Tolbutamide, a Urea Derivative, Impedes Phenolic Accumulation in the Lichen *Evernia Prunastri*." *Journal of Plant Physiology* 132 (5): 580–83. https://doi.org/10.1016/S0176-1617(88)80257-8.

Vitikainen, Orvo. 2001. "William Nylander (1822–1899) and Lichen Chemotaxonomy." *The Bryologist* 104 (2): 263–67. https://doi.org/10.1639/0007-2745(2001)104[0263:WNALC]2.0.CO;2.

Wandy, Joe, Yunfeng Zhu, Justin Johan Jozias van der Hooft, Rónán Daly, Michael P. Barrett, and Simon Rogers. 2017. "Ms2lda.Org: Web-Based Topic Modelling for Substructure Discovery in Mass Spectrometry." *Bioinformatics* 34 (2): 317–18. https://doi.org/10.1093/bioinformatics/btx582.

Wang, Mingxun, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, et al. 2016. "Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking." *Nature Biotechnology* 34: 828–37. https://doi.org/10.1038/nbt.3597.

Watrous, Jeramie, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, et al. 2012. "Mass Spectral Molecular

Networking of Living Microbial Colonies." *Proceedings of the National Academy of Sciences* 109 (26): E1743 LP-E1752. https://doi.org/10.1073/pnas.1203689109.

Watson, James Dewey, and Francis Harry Compton Crick. 1953. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

Wegrzyn, Michal Hubert, Paulina Wietrzyk-Pełka, Agnieszka Galanty, Beata Cykowska-Marzencka, and Monica Alterskjær Sundset. 2019. "Incomplete Degradation of Lichen Usnic Acid and Atranorin in Svalbard Reindeer (*Rangifer Tarandus Platyrhynchus*)." *Polar Research* 38: 1–11. https://doi.org/10.33265/polar.v38.33751.

Weininger, David. 1988. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. https://doi.org/10.1021/ci00057a005.

———. 1990. "SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures." *Journal of Chemical Information and Computer Sciences* 30 (3): 237–43. https://doi.org/10.1021/ci00067a005.

Weininger, David, Arthur Weininger, and Joseph L. Weininger. 1989. "SMILES. 2. Algorithm for Generation of Unique SMILES Notation." *Journal of Chemical Information and Computer Sciences* 29 (2): 97–101. https://doi.org/10.1021/ci00062a008.

Westring, Johan Peter. 1805. *Svenska Lafvarnas Färghistoria*. Stockholm.

Whitehouse, Craig M., Robert N. Dreyer, Masamichi Yamashita, and John B. Fenn. 1985. "Electrospray Interface for Liquid Chromatographs and Mass Spectrometers." *Analytical Chemistry* 57 (3): 675–79. https://doi.org/10.1021/ac00280a023.

Wikipedia. 2019. "Bibliothèque Logicielle." Wikipedia. 2019. https://fr.wikipedia.org/wiki/Bibliothèque_logicielle.

———. 2020a. "Comma-Separated Values." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Comma-separated_values.

———. 2020b. "Expression Régulière." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Expression_régulière.

———. 2020c. "Extensible Markup Language." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Extensible_Markup_Language.

———. 2020d. "Format Propriétaire." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Format_propriétaire.

———. 2020e. "Taxonomie." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Taxonomie#:~:text=La taxinomie ou taxonomie est,décrire%2C les nommer et les.

———. 2020f. "Théorie Des Graphes." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Théorie_des_graphes.

———. 2020g. "Théorie Des Réseaux." Wikipedia. 2020. https://fr.wikipedia.org/wiki/Théorie_des_réseaux.

Wishart, David S., Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, et al. 2007. "HMDB: The Human Metabolome Database." *Nucleic Acids Research* 35 (suppl_1): D521–26. https://doi.org/10.1093/nar/gkl923.

Wisniak, Jaime. 2013. "Pierre-Jean Robiquet." *Educación Química* 24: 139–49. https://doi.org/10.1016/S0187-893X(13)72507-2.

Wong, S. F., C. K. Meng, and J. B. Fenn. 1988. "Multiple Charging in Electrospray Ionization of Poly(Ethylene Glycols)." *The Journal of Physical Chemistry* 92 (2): 546–50. https://doi.org/10.1021/j100313a058.

Xia, Jianguo, Rupasri Mandal, Igor V. Sinelnikov, David Broadhurst, and David S. Wishart. 2012. "MetaboAnalyst 2.0—a Comprehensive Server for Metabolomic Data Analysis." *Nucleic Acids Research* 40 (W1): W127–33. https://doi.org/10.1093/nar/gks374.

Xia, Jianguo, Nick Psychogios, Nelson Young, and David S. Wishart. 2009. "MetaboAnalyst: A Web Server for Metabolomic Data Analysis and Interpretation." *Nucleic Acids Research* 37 (suppl_2): W652–60. https://doi.org/10.1093/nar/gkp356.

Xia, Jianguo, Igor V. Sinelnikov, Beomsoo Han, and David S. Wishart. 2015. "MetaboAnalyst 3.0—Making Metabolomics More Meaningful." *Nucleic Acids Research* 43 (W1): W251–57. https://doi.org/10.1093/nar/gkv380.

Yamamoto, Yoshikazu, Hideki Matsubara, Yasuhiro Kinoshita, Kaoru Kinoshita, Kiyotaka Koyama, Kunio Takahashi, Vernon Ahmadjiam, Teiko Kurokawa, and Isao Yoshimura. 1996. "Naphthazarin Derivatives from Cultures of the Lichen *Cladonia Cristatella*." *Phytochemistry* 43 (6): 1239–42. https://doi.org/10.1016/S0031-9422(96)00495-5.

Yang, Jane Y., Laura M. Sanchez, Christopher M. Rath, Xueting Liu, Paul D. Boudreau, Nicole Bruns, Evgenia Glukhov, et al. 2013. "Molecular Networking as a Dereplication Strategy." *Journal of Natural Products* 76 (9): 1686–99. https://doi.org/10.1021/np400413s.

Yoshihiro, Nishikawa, Michishita Kazuhiko, and Kurono Goichi. 1973. "Studies on the Water Soluble Constituents of Lichens. I. Gas Chromatographic Analysis of Low Molecular Weight Carbohydrates." *Chemical & Pharmaceutical Bulletin* 21 (5): 1014–19. https://doi.org/10.1248/cpb.21.1014.

Yoshimura, Isao, Yasuhiro Kinoshita, Yoshikazu Yamamoto, Siegfried Huneck, and Yasuyuki Yamada. 1994. "Analysis of Secondary Metabolites from Lichen by High Performance Liquid Chromatography with a Photodiode Array Detector." *Phytochemical Analysis* 5 (4): 197–205. https://doi.org/10.1002/pca.2800050405.

Yosioka, Itiro, Hiroshi Yamauchi, and Isao Kitagawa. 1972. "Lichen Triterpenoids. V. On the Neutral Triterpenoids of *Pyxine Endochrysina* Nyl." *Chemical & Pharmaceutical Bulletin* 20 (3): 502–13. https://doi.org/10.1248/cpb.20.502.

Yuan, Xunlai, Shuhai Xiao, and T. N. Taylor. 2005. "Lichen-like Symbiosis 600 Million Years Ago." *Science* 308 (5724): 1017 LP – 1020. https://doi.org/10.1126/science.1111347.

Zopf, Wilhelm. 1895a. "Zur Kenntniss Der Flechtenstoffe, Zweite Mittheilung."

Mittheilungen Aus Dem Chemischen Institut Der Universitat Halle, 38–74.

———. 1895b. "Zur Kenntniss Der Flechtenstoffe." In *Mittheilungen Aus Dem Chemischen Institut Der Universitat Halle*, 107–32.

———. 1897. "Zur Kenntniss Der Flechtenstoffe, Dritte Abhandlung." *Mittheilungen Aus Dem Chemischen Institut Der Universitat Halle* 111: 79–80.

———. 1907. *Die Flechtenstoffe In Chemischer, Botanischer, Pharmakologischer Und Technischer Beziehung*. Jena: Gustav Fischer Verlag.

UNIVERSITÉ DE RENNES 1

Titre : Etude de la diversité chimique des lichens par LC-MSn : acquisition et optimisation du traitement des données métabolomiques.

Mots clés : Lichens, Métabolomique, LC-MS, Déréplication, Bioinformatique, Chimie des Produits naturels.

Résumé : Les lichens sont des champignons symbiotiques dont la chimie est exploitée par l'Homme depuis l'antiquité. Ils n'ont cependant pas été intégrés aux études de métabolomique récentes ce qui a installé l'idée que les lichens sont pauvres en molécules. 1050 molécules leur sont classiquement attribuées, bien que ce décompte date et qu'il semble éloigné de ce qui pourrait être attendu pour un mode de vie concernant 19 387 espèces. En métabolomique, LC-MS et la déréplication à l'aide de bases de données sont régulièrement usitées pour permettre le profilage des échantillons. Ces bases de données ne sont cependant pas adaptées à l'étude des lichens, qui produisent principalement des molécules qui leur sont uniques. Dans cette optique, plusieurs bases de données spécifiques aux lichens ont été créées ici, en utilisant des données de la littérature ainsi qu'en produisant des données spectrales. Des outils ont été crées pour améliorer la déréplication par la prédiction des molécules contenues dans les extraits à partir des ions qu'elles produisent. Tout ceci a été appliqué à l'analyse de 300 échantillons de lichens pour mettre en évidence la diversité chimique de ces champignons à l'aide de techniques modernes. Ceci a permis de prédire quelques 8000 molécules avec des degrés de certitude variables. L'étude détaillée des résultats pour mettre à jour les conaissances sur les lichens reste à faire, mais ceux-ci permettent déjà d'avancer que ces organismes sont à l'origine d'une chimie sous-estimée et qui reste encore à explorer.

Title : Study of the chemical diversity of lichens in a metabolomics context : production of appropriate databases, dereplication and prediction of molecules in LC-MS^n analyses.

Keywords : Lichens, Metabolomics, LC-MS, Dereplication, Bioinformatics, Naturel Products Chemistry.

Abstract : Lichens are symbiotic fungi, the chemistry of which has been used by humans since antiquity. They were however not studied with modern tools in metabolomics, which contributed in cementing the idea that lichens have poor chemical diversity. They are traditionally associated with 1050 molecules, although this figure is dated and does not seem to match the number of lichen species (19 387). In the field of metabolomics, LC-MS and dereplication using databases are frequently used for chemical profiling. These databases are nevertheless not adapted to the study of lichens, since they produce unique metabolites. Thus, several specific databases were produced herein, using the literature and pure standards analysed by LC-MS. Additional tools were developped to better the dereplication efficiency by predicting the molecules from the ions detected. These were all applied to the analysis of 300 lichen samples to highlight the chemical diversity of lichens using modern techniques. This resulted in the prediction of 8000 molecules with varying precision. The detailed investigation of the results has yet to be completed, but these already indicate that lichens are host to an underestimated chemical diversity that remains to be explored.