

MISE EN PLACE D'UNE APPROCHE NON CIBLEE POUR L'ANALYSE DES BIOTOXINES MARINES ET PREUVE DE CONCEPT

A. Introduction

L'utilisation de la spectrométrie de masse haute résolution pour le criblage non ciblé employant les approches dites « métabolomiques » a connu un essor considérable ces dernières années dans de nombreux domaines notamment en analyse environnementale. Plusieurs travaux sont désormais focalisés sur l'amélioration des connaissances pour la mise en œuvre de cette technique afin d'identifier de nouvelles molécules d'intérêts.

Cependant l'apport de ce type d'approche n'a pas encore été suffisamment étudié dans le domaine des biotoxines marines.

Nous présenterons ici notre démarche et les résultats que nous avons obtenus dans le cadre d'expériences de preuve de concept visant à démontrer le potentiel et les limites des analyses globales selon une approche non ciblée par spectrométrie de masse à haute résolution pour une caractérisation plus fine et plus complète d'échantillons contaminés et l'identification potentielle de composés émergents.

Nous détaillerons dans ce chapitre la démarche entreprise pour l'approche de suspect screening et l'analyse sans a priori. Pour cette dernière, nous mettons l'accent sur les étapes de traitement de données réalisées avec deux logiciels constructeurs (MasterviewTM et MarkerView) et le logiciel open source XCMS disponible sous la plateforme «Workflowformetabolomics».

B. Traitement des données HRMS avec les logiciels constructeurs

I. Matériels et Méthodes

I.1. Produits chimiques et réactifs

Les produits et réactifs chimiques utilisés pour les expériences décrites dans ce chapitre sont les mêmes que ceux présentés dans le chapitre II.B (I.1). La solution étalon de boscalid (à 10 ng/ μ L dans l'ACN) utilisée comme étalon interne a été achetée auprès du fournisseur Dr Ehrenstorfer GmbH (Augsburg, Allemagne).

I.2. Préparation des échantillons et des solutions de travail

Dans le cadre de notre preuve de concept, nous avons choisi de réaliser notre étude sur des échantillons de moules et d'huîtres contrôle que nous avons supplémentés avec des toxines connues dont les étalons sont commercialisés. Les expériences de supplémentation étant très coûteuses, nous avons fait le compromis de sélectionner quelques toxines de familles différentes ; 5 toxines analysables en ESI+ (GYM, SPX1, AZA1, PnTXA, PTX2) et 4 toxines analysables en ESI-. (AO, DTX1, DTX2, AD).

Les échantillons contrôles de moules et d'huîtres ont été extraits selon le mode opératoire standard de l'EURLMB décrit dans le chapitre II.B (I.4). Un contrôle réactif a été également préparé selon le même protocole. Les extraits de moules et d'huîtres ont ensuite été supplémentés par les toxines choisies.

Les solutions d'ajouts ont été préparées à différents niveaux de concentrations à partir d'une solution mère multitoxines (240 ng/mL) tel que décrit dans les

Tableau 22 et Tableau 23

Ne disposant pas de standard interne spécifique aux biotoxines marines, une solution étalon de boscalid (10 ng/mL) a été ajoutée aux solutions de supplémentation comme composé de référence dans le cadre des analyses sans a priori.

Tableau 22. Préparation des solutions d'ajout (SA)

	SA1	SA2	SA3	SA4	SA5	SA6
Concentration en toxines (ng/mL)	20	40	80	120	160	240
<u>ESI +</u> GYM, SPX1, AZA1, PnTXA, PTX2 <u>ESI-</u> AO, DTX1, DTX2, AD						
Volume prélevé de la solution mère (µL)	42	83	167	250	333	470
Volume de la solution étalon de boscalid (µL)				30		
Volume de MeOH (µL)	428	387	303	220	137	0
Volume total (µL)				500		

Tableau 23. Préparation des extraits de moules et d'huîtres supplémentés (N)

	N1	N2	N3	N4	N5	N6
Concentration en toxines (ng/mL)	2	4	8	12	16	24
ESI + GYM, SPX1, AZA1, PnTXA, PTX2 ESI- AO, DTX1, DTX2, AD						
Volume solution d'ajout SA (µL)	50	50	50	50	50	50
Volume d'extrait de moules ou d'huîtres (µL)	450	450	450	450	450	450
Volume total (µL)	500					

I.3. Conditions d'analyse par LC-HRMS

Les analyses ont été menées par LC-HRMS. Les détails de la méthode d'analyse sont décrits dans le chapitre II (I.5).

Les conditions chromatographiques et les paramètres de masse utilisés sont résumés dans les **Tableau 24** et **Tableau 25** respectivement.

Tableau 24. Conditions chromatographiques

Système HPLC	Dionex U3000
Colonne	HSS T3 C18 – 100 x 2,1 mm, 2,6 µm (Waters)
Débit	0,45 mL/min
Température	30°C
Volume d'injection	5 µL
Phases mobile	A : H ₂ O + 2 mM formiate d'ammonium+50 mM acide formique B : 95% ACN + 2 mM formiate d'ammonium+50 mM acide formique
Gradient	<p>Le graphique illustre le gradient de phase mobile utilisé. L'axe des ordonnées représente le pourcentage de phase B (%B) et l'axe des abscisses représente le temps en minutes (min). Le gradient est défini par les points suivants : (0, 0), (10, 100), (15, 100), (16, 0), (20, 0).</p>

Tableau 25. Paramètres de masse

Spectromètre	5600 QTOF (Sciex)
Source	ESI+/ESI-
Gaz 1 (GS1)	35 psi
Gaz 2 (GS2)	45 psi
Température (TEM)	500°C
IonSpray Voltage (ISV)	5,5 kV / -4,5 kV
Gaz rideau (CUR)	30 psi
Potentiel de déclusterisation (DP)	60 V / -100
Ion release delay (IRD)	67 ms
Ion release width (IRW)	25 ms
Energie de collision (CE)	40 eV / -40 eV
Amplitude de l'énergie de collision (CES)	20 V / -20 V
Mode d'acquisition (temps d'accumulation)	MS : TOF MS (0,2 ms) MS/MS : IDA (0,05 ms)
Gamme de masse	TOF MS : 100-1250 Da MS/MS : 50-1250 Da

I.4. Acquisition et traitement des données

L'acquisition des données a été réalisée par le logiciel Analyst® TF 1.7.1 (Sciex, Toronto, ON, Canada).

Pour l'approche suspect screening, les logiciels PeakView® (Sciex, Toronto, ON, Canada) et son algorithme MasterView™ (Sciex, Toronto, ON, Canada) ont été utilisés pour visualiser et traiter les données. Une liste de composés suspects a été créée avec MasterView™ avec 821 molécules incluant biotoxines marines et cyanotoxines. Les seules informations disponibles dans cette liste de suspects sont les noms des molécules, leurs formules brutes et leurs masses exactes. La liste a été utilisée pour les deux modes d'ionisation pour chercher les ions $[M+H]^+$ et $[M-H]^-$ en ESI+ et ESI- respectivement.

Après le traitement, le logiciel permet d'afficher les résultats dans le volet chromatogramme et dans un tableau. Ce tableau affiche des informations pour l'identification des composés en fonction des résultats de recherche de la bibliothèque créée, y compris le temps de rétention, l'erreur de masse (ppm ou Da), la composition élémentaire et le score de pureté. Des paramètres

de confiance pour l'identification des composés sont définis avant de procéder au traitement des données permettant de filtrer les ions détectés selon un code tricolore (vert, orange et rouge) en fonction du niveau de confiance (**Figure 31**).

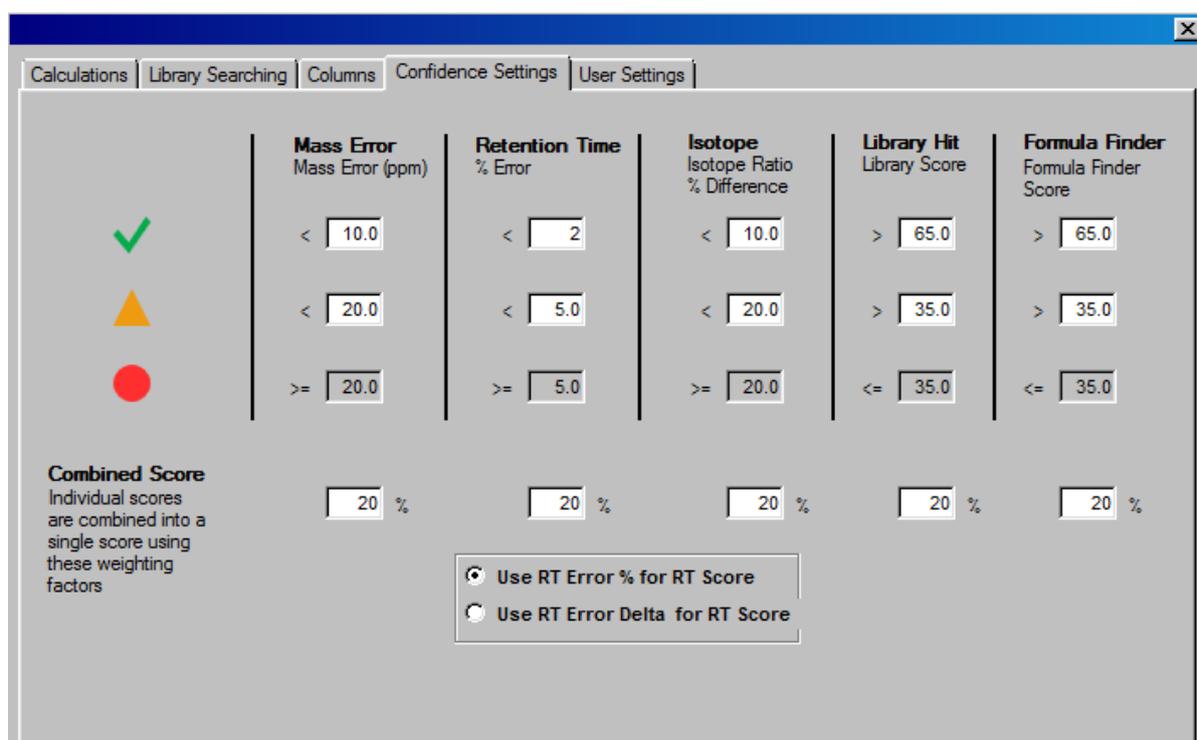


Figure 31. Paramètres de confiance définis pour le suspect screening

Pour l'approche « suspect screening » nous nous sommes intéressés uniquement à l'erreur de masse, le ratio isotopique et le Formula Finder Score. Les composés répondant aux critères fixés sont affichés en vert en haut du tableau et sont considérés comme des potentiels positifs. Cette identification automatique est un premier filtre qui permet de réduire la liste et de prioriser les ions à étudier. Ensuite pour chaque composé de la liste, nous vérifions l'allure du pic correspondant, le rapport S/N, la cohérence du temps de rétention, le profil isotopique et pour confirmation le spectre de fragmentation MS/MS.

Les deux logiciels d'extraction de pics MasterViewTM et MarkerViewTM ont été testés pour l'approche non-ciblée sans a priori :

- Le logiciel MasterViewTM possède un algorithme d'extraction de pics non ciblé « untargeted peak finding » qui permet l'extraction des pics d'intérêt à partir des données brutes. L'extraction des pics a été réalisée en fixant certains paramètres au préalable (**Figure 32**).

Calculations | Library Searching | Columns | Confidence Settings | User Settings

XIC

Do not calculate details for XIC with Intensity < counts or S:N <

Default XIC Width (Da)

Default Retention Time Width (min)

Default Threshold (cps)

Default Threshold (ratio of control)

Non-Targeted Peak Finding

Find new LCMS peaks and add them to the existing XIC list.

Minimum Retention Time (min)

Maximum Retention Time (min)

Find peaks from:

All samples

Selected sample only

Peak Detection Sensitivity

Fast | | | | | | | | | | Exhaustive

Formula Finder

Max Element Mass Tolerance (ppm)

Figure 32. Paramètres d'extraction des pics (Masterview™)

Le logiciel MarkerView™ (version 1.2.1.) a été utilisé pour traiter les données brutes LC-HRMS. Il s'agit d'un progiciel de traitement qui permet la détection des pics, l'alignement et le filtrage des données, générant une matrice de variables dans laquelle sont définis les m/z mesurés, le temps de rétention et l'intensité ionique du signal détecté. L'exploration des données a été effectuée par un algorithme automatisé avec les paramètres d'extraction suivants :

- subtraction offset: 15 scans (cette option réduit les chances de trouver des ions de fond constants sous forme de pics) ;
- subtraction multiplication factor: 1.3 (avant la soustraction, le spectre de fond mentionné ci-dessus est d'abord multiplié par cette valeur, ce qui est utile pour compenser les variations mineures de l'intensité des ions présents dans le bruit de fond) ;
- minimum retention time peak width: 5 scans (les « pics » plus petits que cette valeur sont considérés comme du bruit) ;
- minimum spectral peak width : 10 ppm.

Le logiciel effectue ensuite une correction des temps de rétention et un alignement des pics par rapport à notre composé de référence (le boscalid) avec les paramètres suivants: retention time tolerance 0,2 min; mass tolerance 10 ppm; intensity threshold 50 cps; maximum number of peaks 5000 ; les largeurs minimale et maximale des pics chromatographiques sont fixées respectivement à 0,05 min et 1 min ; les temps de rétention minimal et maximal sont respectivement de 1,2 et 10 min. Pour limiter le nombre de signaux, les pics présents dans moins de 3 échantillons de la séquence d'analyse sont éliminés car considérés comme de potentiels artefacts. De même les pics dont l'intensité dans les échantillons contaminés est moins de 10 fois plus importante que dans le blanc sont éliminés.

Enfin, le logiciel revient aux données brutes pour intégration en utilisant les intervalles de masse m/z et de temps de rétention des pics alignés. En revenant aux données brutes, les pics alignés dans les échantillons analysés sont traités de la même façon, indépendamment de la précision de l'intégration originale. Toutes ces données sont converties en matrice de variables pour procéder ensuite aux analyses statistiques qui vont nous permettre de déterminer les ions d'intérêt avant de procéder à leur identification.

I.5. Analyses statistiques des données

Les données ont été traitées avec différents outils statistiques :

Un t-test qui est un outil d'analyse statistique univariée supervisée pertinent lorsque deux ou plusieurs groupes d'échantillons prédéterminés sont présents. L'outil statistique de MarkerView™ permet une comparaison par paire de tous les groupes ou de comparer un groupe à tous les autres. Nous avons opté pour la comparaison par paire des échantillons contaminés et blancs. Les résultats du t-test indiquent dans quelle mesure chaque signal (ion m/z) distingue les deux groupes. Ceci est rapporté en tant que « p-value » ; plus cette valeur est petite plus le signal est significativement déterminant dans la différence observée entre deux groupes. Les ions ayant une p-value <0.05 ont été retenus comme potentiellement intéressants. Le « fold » est également un paramètre important traduisant la différence d'intensité entre les groupes pour les ions considérés.

Analyse multivariée : une ACP non supervisée a d'abord été réalisée. Avant l'analyse statistique, les réponses aux pics étaient mises à l'échelle avec le modèle pareto selon lequel les données sont centrées autour de la moyenne et divisées par la racine carrée de l'écart type. Cela a pour effet de réduire l'influence des pics très intenses, tout en mettant l'accent sur les pics plus

faibles qui peuvent avoir plus de pertinence. Les données ont également été traitées par analyse supervisée (PCA-DA) qui a un pouvoir plus discriminant que l'ACP pour la détermination des ions responsables des différences significatives entre les échantillons contaminés et blancs.

I.6. Identification des composés d'intérêt

L'identification des composés d'intérêt a été réalisée avec le logiciel Peakview[®] qui peut être programmé pour effectuer automatiquement des calculs empiriques de la formule moléculaire potentielle avec l'application Formula Finder. L'algorithme de Formula Finder utilise les données MS et MS/MS pour trouver les meilleures formules possibles qui correspondent à la masse extraite. Une fois qu'une formule potentielle a été déterminée, elle peut être recherchée dans la base de données ChemSpider. Les structures extraites de la base de données sont liées au spectre MS/MS acquis pour cet ion via les fichiers mol correspondants. Les spectres de fragmentation des biotoxines marines ne sont souvent pas référencés dans les bases de données utilisées (ChemSpider et PubChem). L'identification se fait alors sur la base du spectre MS uniquement dans un premier temps. Pour la confirmation de la structure potentielle, une méthode alternative de fragmentation *in silico* est alors appliquée afin d'élucider les structures. En effet les algorithmes du logiciel sont capable de prédire la fragmentation théorique d'un composé donné utilisant les énergies de liaison et de dissociation (Ruttkies et al., 2016). Les fragments générés *in silico* et ceux acquis sont alors comparés. Le logiciel permet ainsi de calculer un score de probabilité entre les fragments détectés et la structure suspectée. Ce score prend en compte le rapport m/z, l'intensité de chaque fragment coïncidant et l'énergie de liaison.

II. Résultats et discussions

II.1. Qualités des données acquises

Le développement et la caractérisation de notre méthode en mode ciblée nous a permis de valider les performances chromatographiques (répétabilité des temps de rétention des étalons), et de détection en spectrométrie de masse (précision et répétabilité des mesures de masses précises et répétabilité des intensités des références). Les critères de validation analytique ont été les suivants : mesure de la masse précise des références avec une erreur inférieure à 5 ppm et coefficient de variation des Tr et des intensités, respectivement inférieurs à 5 et 25%.

II.2. Évaluation des performances de l'approche « suspect screening »

Pour évaluer le workflow du suspect screening, nous avons procédé à l'analyse des échantillons de moules et d'huîtres dopés tout en considérant les toxines présentes comme des composés inconnus. La seule information a priori était la masse exacte des ions présents dans la liste des suspects, calculée à partir de la formule brute de chacun des composés candidats.

Les critères d'identification des suspects devaient être choisis avec soin afin de minimiser à la fois le nombre de faux positifs (ions identifiés à tort comme composés d'intérêt) et de faux négatifs (composés réellement présents et pas identifiés).

Les premières étapes du criblage des suspects sont automatisées grâce au logiciel MasterView™ qui permet de filtrer les données et identifier les ions potentiellement présents en se basant sur des critères de confiance que nous fixons à l'avance. Différents paramètres ont alors été étudiés et optimisés pour arriver au workflow optimal. Pour les paramètres relatifs aux pics chromatographiques, une intensité minimale de 1000 cps a été appliquée. Le rapport signal/bruit (S/N) minimum a été fixé à 6 (compris entre la LD (S/N 3) et la LQ (S/N 10)). Les autres critères d'identification ont été déterminés à partir de trois injections d'extraits de moules et d'huîtres supplémentés à la LQ des différentes toxines étudiées. Nous avons vérifié les scores obtenus pour chacune des toxines (erreur de masse, ratio isotopique, Formula Finder score) et les plus mauvais scores obtenus dans ces conditions expérimentales ont été choisis comme paramètres de confiance. Ainsi, nous avons conclu qu'un Formula Finder score de 65 permettrait d'atteindre le nombre minimum de faux négatifs et de réduire le nombre de faux positifs potentiels. Une erreur de masse de 10 ppm et une différence de rapport isotopique de 10% ont été choisies comme meilleurs compromis.

Une fois nos critères choisis, nous avons procédé à l'analyse de nos échantillons à l'aveugle pour évaluer la pertinence du workflow.

Comme des résultats comparables ont été obtenus pour les différentes matrices, seuls les résultats relatifs aux échantillons d'huîtres sont présentés ici. Après la première étape de filtrage automatique en ESI+, 15 composés suspects (allumés en vert) sur une liste de 821 ont d'abord été identifiés comme des toxines potentiellement présentes dans les échantillons analysés. La deuxième étape consistait à vérifier manuellement pour chacun des ions de la liste : la présence d'un ou plusieurs pics chromatographiques, la correspondance des massifs isotopiques, la

présence des pics dans les 3 réplicats et enfin nous avons vérifié que ces pics n'étaient pas également présents dans le blanc réactif.

Cette deuxième étape a conduit à l'élimination de cinq candidats qui n'étaient soit pas présents dans les trois injections (et donc considérés comme des faux positifs) soit n'affichant pas de « vrai » pic chromatographique. Cette étape a abouti à une liste de 10 candidats, parmi lesquels figuraient les 5 toxines présentes dans l'échantillon : PnTX-A, PTX2, GYM, SPX1, AZA1 mais également des analogues isobares de SPX1 (SPX-A, SPX-G) et AZA1 (AZA6, AZA29, AZA40). Pour confirmer l'identité de ces composés, nous avons comparé les spectres MS/MS expérimentaux aux spectres de fragments théoriques extraits des fichiers « mol » obtenus à partir des bases de données ChemSpider ou PubChem. Tous les spectres expérimentaux ont montré une bonne corrélation avec les fragments théoriques (score > 70%). Une exception a été observée pour la GYM, avec seulement 20% de fragments correspondants. Le logiciel a automatiquement attribué le pic le plus intense présent dans le chromatogramme extrait (XIC) à la masse exacte du composé sélectionné dans la liste des suspects. Le XIC de la GYM a ensuite été vérifié visuellement et un deuxième pic moins intense était présent à un temps de rétention différent (5,1 min) dans les réplicats. Ce pic a ensuite été sélectionné manuellement et la corrélation entre les spectres de fragmentation empirique et théorique a été vérifiée à nouveau. Cette fois, nous avons obtenu une correspondance de 100%. Toutes les toxines étudiées ont été identifiées sans équivoque sur la base de leurs spectres de fragmentation à l'exception de la SPX1 et AZA1, qu'il n'a pas été possible de distinguer de leurs analogues isobares.

Ces résultats montrent qu'il est important de garder un esprit critique lors du traitement des résultats générés automatiquement et de toujours procéder à une vérification pour une identification fiable.

Pour cet exercice de preuve de concept, nous avons essayé de simuler les conditions d'analyse d'un échantillon naturellement contaminé pour lequel le laboratoire ne disposerait pas d'un échantillon de contrôle (même échantillon mais non contaminé). Ce cas de figure se présente souvent pour les échantillons impliqués dans les cas de TIAC où seul un échantillon contaminé est envoyé au laboratoire pour investigation.

Dans le cas où un échantillon de contrôle est disponible, une étape supplémentaire de comparaison, contaminé versus blanc, permettrait d'éliminer soit tous les pics communs aux deux échantillons (« blank extraction ») ou d'éliminer les pics avec une intensité moins de 10 fois plus élevée dans le contaminé que dans le blanc, on parlerait ici de « blank reduction ».

Travaillant dans des conditions de preuve de concept, nous avons pu mettre en application cette approche qui nous fait de gagner du temps en réduisant le nombre initial de candidats potentiels et aucun problème n'a été rencontré pour l'identification de la GYM, puisque le pic le plus intense identifié à tort comme la GYM était en fait présent dans le contrôle alors que le second, élué à 5,1 min et correspondant effectivement à la GYM n'était présent que dans le contaminé. Les mêmes étapes ont été suivies pour les données acquises en ionisation négative. Nous avons réussi à identifier sans ambiguïté l'ensemble des toxines supplémentées (AD, AO, DTX1 et DTX2). L'AO et la DTX2 sont deux composés isobares ayant la même formule brute (m/z 803,4587 ; $C_{44}H_{68}O_{13}$), néanmoins les conditions chromatographiques permettent de les séparer. Il y a ainsi deux pics chromatographiques parfaitement résolus attribués à chacune de ces deux toxines (AO première toxine éluée). Sans aucune information sur le temps de rétention de l'AO et de la DTX2, il est difficile de se prononcer sur l'identité du pic lorsque seule une de ces deux toxines est présente, car elles ne peuvent pas être différenciées sur la base de leur spectre de fragmentation puisqu'il est identique.

II.3. Évaluation de l'approche non ciblée sans a priori

II.3.1. MasterView™

Le même jeu de données a été traité selon une approche sans a priori cette fois-ci en employant la fonction « untarget peak finding » de MasterView™. Après la définition de certains paramètres (cf.I.4) le logiciel procède à l'extraction (peak peaking) des signaux présents dans les échantillons traités. Le logiciel va extraire tous les signaux répondant aux critères définis préalablement, il peut générer une liste avec une centaine voire des milliers de signaux comprenant les composantes matricielles, les interférents qui peuvent survenir lors des étapes de préparation d'échantillons et potentiellement nos composés d'intérêts. Il est impossible à ce stade d'investiguer toute la liste de signaux et de réussir à distinguer les contaminants recherchés parmi tous les ions générés. La meilleure approche pour éliminer les composés sans intérêt (composantes matricielles, artéfacts) est d'effectuer un screening comparatif. Le screening comparatif consiste en la comparaison de notre échantillon inconnu (échantillons supplémentés) avec un témoin connu (échantillons blancs non supplémentés). Nous avons appliqué ici un « control ratio » de 10 qui permet d'éliminer tous les signaux dont l'intensité dans les échantillons inconnus ne serait pas 10 fois plus importante que dans le blanc (**Figure 33**).

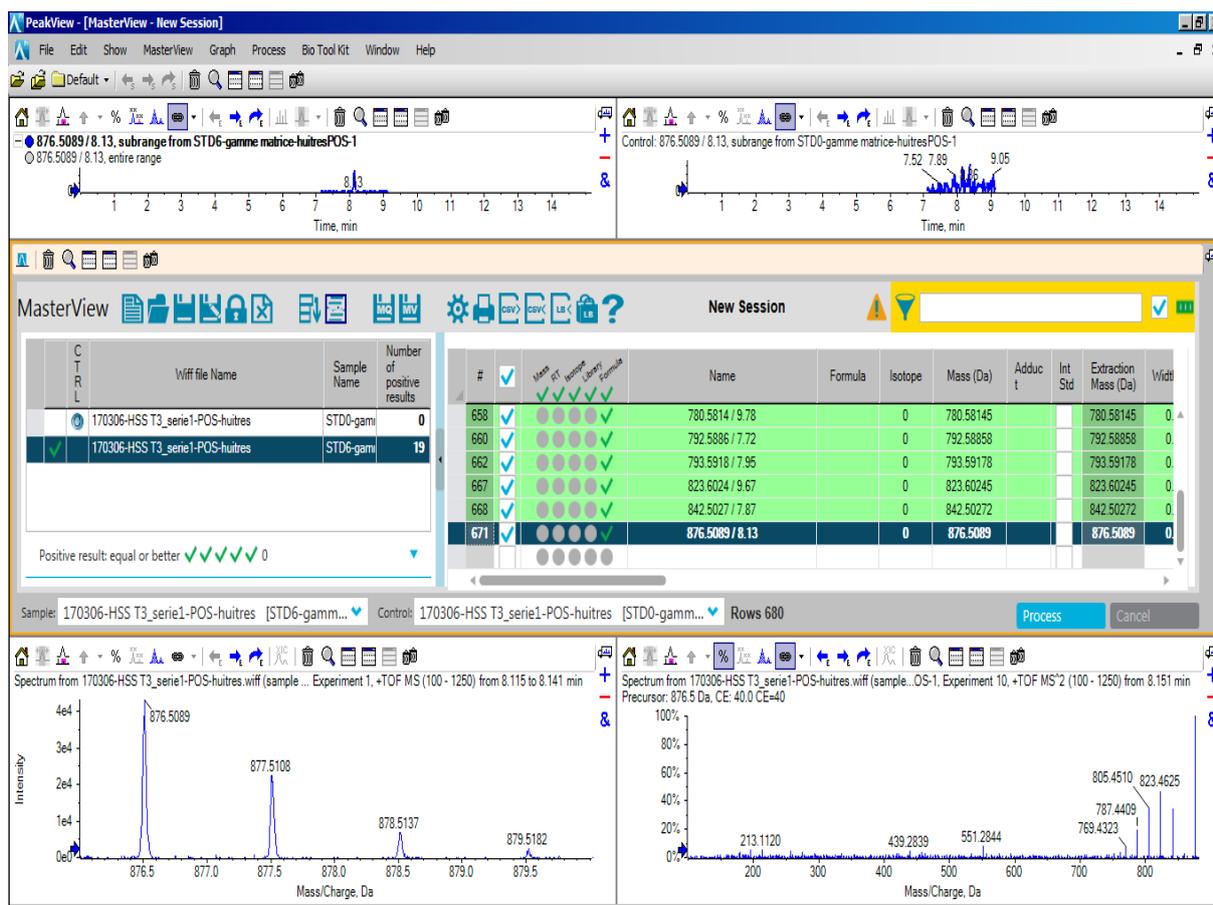


Figure 33. Screening comparatif, échantillon contaminé (chromatogramme en haut à gauche) vs échantillon de contrôle (chromatogramme en haut à droite) avec MasterViewTM

Nous avons d'abord testé différentes vitesses d'extraction « peak detection sensitivity » (**Figure 32**) : « fast » pour un traitement rapide, « exhaustive » pour un traitement plus long et entre les deux pour un temps de traitement moyen. Ce test réalisé sur les répliqués des niveaux les plus contaminés nous a permis de comparer le nombre de signaux générés à chaque fois et de vérifier si nos toxines figuraient dans la liste. Plus la recherche en mode « untarget peak finding » devient exhaustive, plus le nombre de signaux générés augmente, ce qui implique un temps plus long pour l'investigation des ions générés (environ 600). Le traitement rapide n'a pas permis d'extraire les ions correspondants à nos composés d'intérêt, mais avec le temps intermédiaire on arrive à un taux de recouvrement de 90% sur les 5 toxines dans les 6 répliqués testés. L'option « exhaustive » a permis d'extraire 100% des composés d'intérêt. Bien que plus long en termes de temps de traitement et d'investigation des signaux, nous avons opté pour cette dernière option qui permet d'éviter de passer à côté de signaux potentiellement intéressants.

Après le traitement des différents échantillons, nous avons observé que le nombre d'ions extraits entre différentes injections d'un même échantillon peut varier ($CV_{\max} = 3,2\%$).

Aussi, selon le nombre d'échantillons traités, le nombre de signaux générées n'est pas le même. Nous avons donc étudié les échantillons deux par deux (contaminé et contrôle) pour avoir une liste la plus exhaustive possible des ions présents. Les résultats obtenus pour la matrice huître sont présentés dans le **Tableau 26** pour les deux modes d'ionisation. Les résultats relatifs à la matrice moule sont présentés en **Annexe 3**.

Les résultats indiquent qu'après l'application du screening comparatif, le nombre total des signaux initialement générés est réduit d'environ 95 à 99%. C'est à dire que de l'ensemble des signaux détectés, seulement 1 à 5 % sont potentiellement intéressants à investiguer. Les ions retenus après ce premier filtre automatisé sont investigués manuellement ; d'abord visuellement pour vérifier la présence d'un pic chromatographique, ensuite les ions restants sont comparés aux deux autres injections, seuls les composés communs aux trois réplicats sont retenus pour l'étape d'identification. Cette étape permet de réduire le nombre de faux positifs potentiels. Une fois la liste finale des composés établie, nous procédons à l'identification des composés retenus en suivant les étapes décrites dans le paragraphe 0. Un exemple de la procédure d'identification de la SPX1 est présenté en **Figure 34**.

L'analyse sans a priori utilisant l'option « untarget peak finding » de Masterview, nous a permis de retrouver l'ensemble des toxines recherchées sans ambiguïté dans 66% des échantillons analysés à l'aveugle. Cette approche bien qu'efficace, présente certaines limites notamment pour les niveaux de contamination les plus bas. Tel que montré dans le (**Tableau 27**) aucune des toxines n'a été retrouvée pour les deux premiers niveaux de concentration (20 et 40 $\mu\text{g}/\text{kg}$) qui représentent 5 à 10 fois les LQ déterminées par la méthode ciblée. Cela montre qu'il faut bien distinguer sensibilité instrumentale et pouvoir discriminant des logiciels d'extraction de données. Pour les deux toxines AZA1 et PnTX-A au niveau 3, les signaux correspondants ont été éliminés automatiquement de la liste initiale de signaux car présents dans seulement 2 des 3 réplicats investigués. Cette observation montre que bien que l'étape de comparaison des données entre les différents réplicats d'un même échantillon nous ait permis d'éliminer un certain nombre de faux positifs, elle nous a fait également passer à côté de certains composés d'intérêts. Cela montre que malgré la robustesse du système et de l'algorithme utilisé, la détection de pics de faible intensité peut être aléatoire. Pour remédier à ce problème, il faudrait considérer les ions présents dans au moins 2 répétitions plutôt que 3 ou bien augmenter le

nombre de réplicats mais cette dernière option implique un effort supplémentaire dans l'investigation manuelle des signaux.

L'identification des ions déterminés comme composés d'intérêt est l'une des étapes les plus importantes et parmi les plus complexes. En effet, les algorithmes utilisés tel que Formula Finder génèrent un nombre très important de formules possibles pour une même masse exacte. À cette étape intervient l'esprit critique de l'analyste et sa connaissance des compositions élémentaires probables des familles de composés recherchés pour limiter le nombre de formules à confirmer par la comparaison des spectres de fragmentation théoriques (*in silico*) et empiriques. La deuxième difficulté rencontrée lors de l'identification est soit (i) l'indisponibilité des spectres de fragmentations théoriques dans les bases de données pour certains composés, ce qui est notamment le cas pour les biotoxines marines, soit (ii) l'absence d'acquisition d'un spectre de fragmentation exploitable. En effet, bien que l'acquisition en mode IDA soit l'un des meilleurs compromis possibles pour une analyse non ciblée, tous les composés d'intérêt ne sont pas forcément fragmentés à cause de la richesse des matrices analysées. L'une des solutions possibles qui permettrait de résoudre ce problème serait de ré-analyser l'échantillon avec le mode d'acquisition « Product ion » à différentes énergies de collisions en rentrant la masse exacte de l'ion à investiguer. On obtiendrait ainsi des spectres de fragmentation plus riches à étudier.

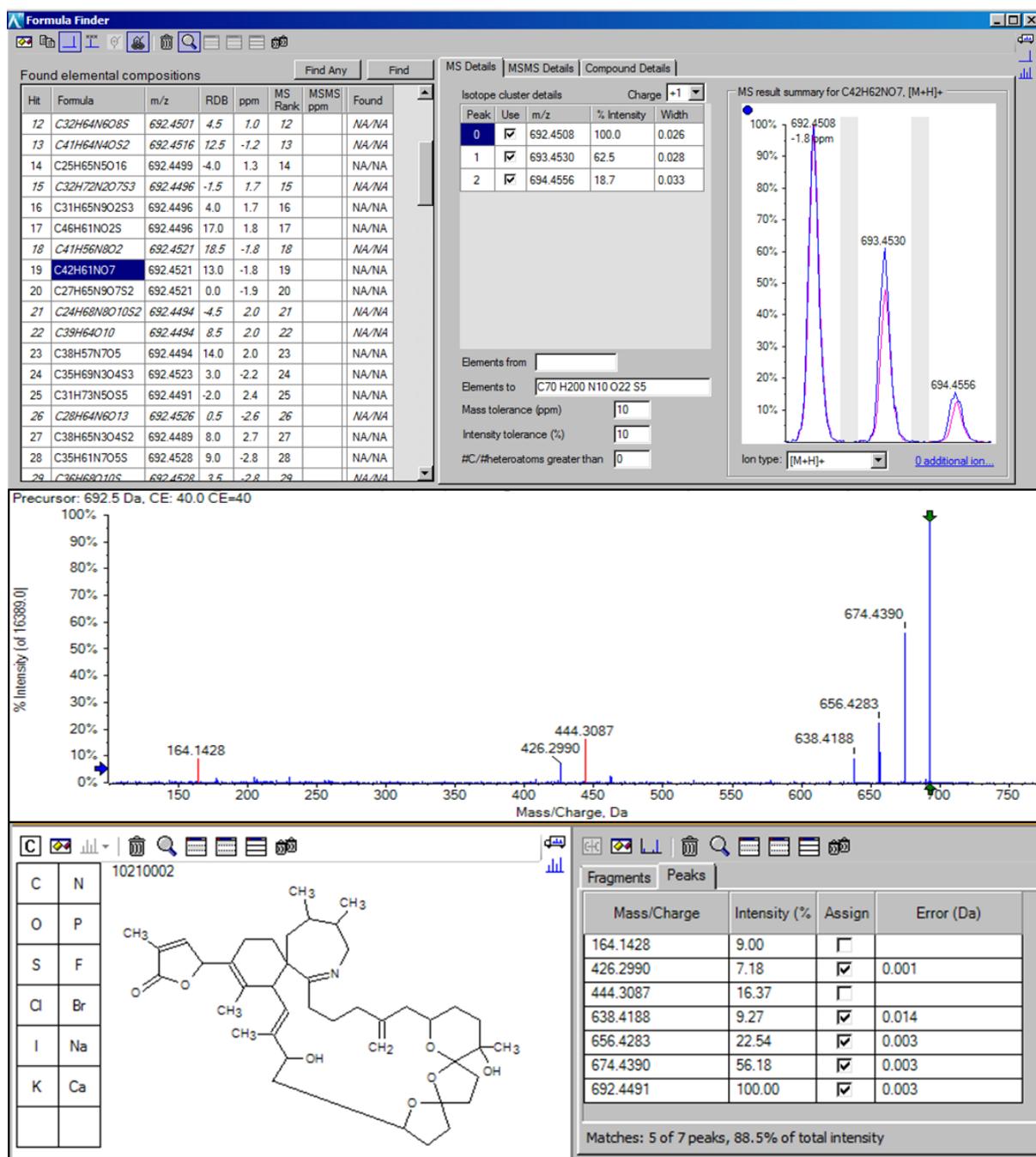


Figure 34. Étapes d'identification d'un composé inconnu (exemple : SPX1) : 1- Recherche de la formule brute dans Chemspider à partir du spectre TOF-MS, comparaison des profils isotopique, 2- Confirmation de la structure par comparaison du spectre de fragmentation acquis avec le spectre de fragmentation théorique (*in silico*)

Tableau 26. Tableaux récapitulatifs des résultats de l'extraction des données avec le logiciel MasterView pour les échantillons d'huître :
(a) ESI+ ET (b) ESI-

(a) ESI+	N1			N2			N3			N4			N5			N6		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Nombre des ions extraits	585	600	590	590	602	582	608	600	590	590	589	603	595	604	607	617	598	589
Nombre d'ion après extraction du contrôle	3	7	5	7	10	6	16	20	13	18	15	18	17	22	19	32	22	27
Nombre d'ions après vérification visuelle	2	3	1	4	2	0	3	4	5	5	5	7	9	6	9	6	5	9
Nombre de faux positifs	2	1	0	2	0	0	1	1	2	0	0	2	3	2	4	1	0	2
nombre de composés Identifiés	0\5	0\5	0\5	0\5	0\5	0\5	3\5	3\5	3\5	5\5	5\5	5\5	5\5	4\5	5\5	5\5	5\5	5\5

(b) ESI-	N1			N2			N3			N4			N5			N6		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Nombre des ions extraits	432	439	412	449	450	437	442	439	430	436	429	438	440	435	439	432	438	437
Nombre d'ion après extraction du contrôle	4	7	4	7	10	5	12	12	9	15	13	13	18	15	18	19	17	20
Nombre d'ions après vérification visuelle	0	2	1	3	2	0	2	1	4	5	6	6	4	5	7	9	6	7
Nombre de faux positifs	0	2	0	1	0	0	1	0	0	0	1	2	0	0	2	1	0	1
nombre de composés Identifiés	0\4	0\4	0\4	0\4	0\4	0\4	3\4	3\4	3\4	4\4	4\4	4\4	4\4	4\4	4\4	4\4	4\4	4\4

Tableau 27. Identification des toxines pour les différents niveaux de contamination par le logiciel MasterView™
✓ : toxine identifiée ; × : toxine non identifiée

	ESI +					ESI -			
	SPX1	GYM	AZA1	PnTX A	PTX 2	AD	AO	DTX	DTX2
N 1	×	×	×	×	×	×	×	×	×
N 2	×	×	×	×	×	×	×	×	×
N 3	✓	✓	✓	×	×	×			
N 4							✓	✓	✓
N 5	✓	✓	✓	✓	✓	✓	✓	✓	✓
N 6	✓	✓	✓	✓	✓	✓			

II.3.2. MarkerView

Le logiciel MarkerView™ a permis d'extraire les données acquises à partir des empreintes LC-HRMS en ESI+ et ESI- de tous les échantillons analysés. Une stratégie de réduction des données a d'abord été appliquée afin de réduire la liste des signaux à investiguer. Les données ont ensuite été analysées par des tests statistiques univariés (t-test) et multivariés (ACP, ACP-DA) afin de déterminer et identifier les composés d'intérêt.

II.3.2.i. Résultats du t-test

L'extraction des données avec le logiciel MarkerView™ a généré 7135 et 6952 signaux pour les échantillons d'huîtres respectivement en ESI+ et ESI- contre 6340 et 6518 signaux pour les échantillons de moules. La stratégie de réduction des données appliquée suivant différentes étapes de préfiltrage basées sur le temps de rétention, et l'élimination des isotopes a permis d'éliminer 55 à 79 % des données initialement extraites (**Tableau 28**). Les données retenues ont ensuite été analysées par un t-test. Le t-test représente une étape supplémentaire de filtrage de données qui permet d'éliminer les signaux redondants dans le set de données et détecter les différences significatives entre les groupes d'échantillons.

Tableau 28. Résumé des résultats de la stratégie de réduction des données pour l'identification manuelle des signaux, après des étapes de préfiltrage et de comparaison par paires en utilisant le t-test avec comme facteurs discriminants la p-value et le log du fold-change

		Huîtres		Moules	
		ESI+	ESI-	ESI+	ESI-
1. Etapes de préfiltrage	nombre total de signaux extraits	7135	6952	6340	6518
	1,2 min < RT <10 min	6290	6175	5567	5841
	Masse monoisotopique	2830	2715	1824	1216
2. Résultats du t-test (comparaison par paire N6 et N0)	P-value < 0,05	147	99	125	133
	Log (fold change) > 0,2	48	28	62	17
	Code tricolore	vert : 20 Orange : 18 Rouge : 10	vert : 9 Orange : 7 Rouge : 12	vert : 19 Orange : 26 Rouge : 17	vert : 15 Orange : 2 Rouge : 0
	Liste finale des signaux pour identification manuelle	20	9	19	15

Les données obtenues à l'issue du t-test ont été classées selon les valeurs croissantes de la p-value. Seuls les ions avec une p-value $< 0,05$ considérés comme signaux significatifs dans la différence entre les deux groupes d'échantillons « dopé » et « contrôle » ont été retenus. Ces signaux ont ensuite été filtrés en fonction du log du « fold change » qui représente le rapport entre l'intensité du signal dans l'échantillon contrôle et contaminé. Seuls les signaux avec un log « fold change » $> 0,2$ ont été retenus à cette étape. La liste des ions pertinents déterminés par le t-test est ensuite importée dans le logiciel PeakView pour visualiser les données. Ces dernières sont alors à nouveau filtrées grâce au code tricolore défini sur la base des critères de confiance déterminés précédemment (II.2). Nous obtenons ainsi une liste finale avec un nombre réduit de composés d'intérêt à identifier (**Figure 35**). On note que le nombre de composés potentiels augmente globalement en fonction du niveau de dopage, une tendance plus perceptible en mode d'ionisation positive puisqu'on passe de 10-12 signaux pour le niveau N1 à 19-20 signaux pour le niveau N6. Le workflow adopté s'est montré très efficace dans la réduction du nombre de signaux à investiguer manuellement ; le nombre de composés d'intérêt représente moins de 1% (entre 0,1 et 0,3%) des signaux totaux extraits.

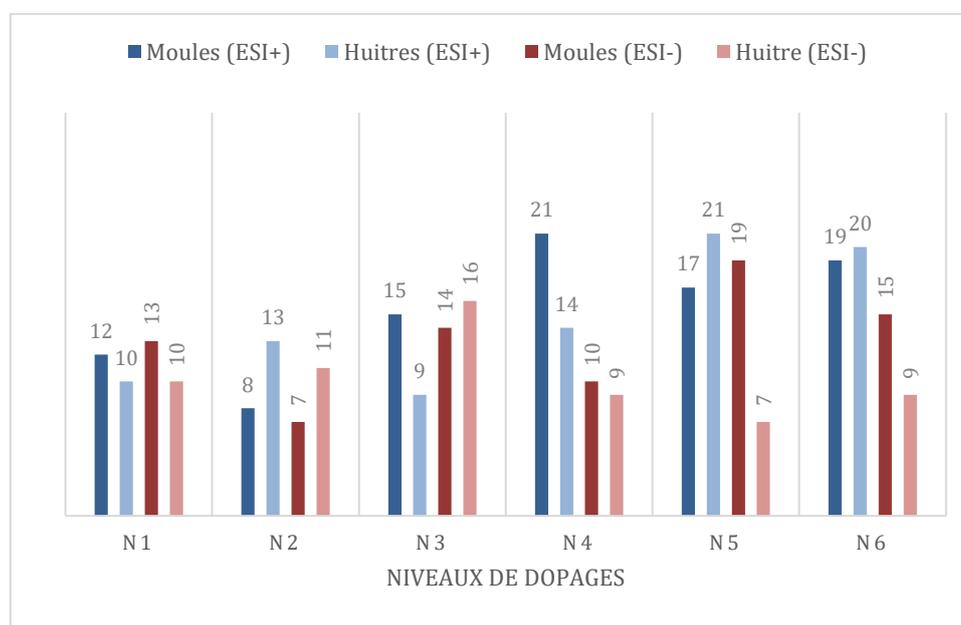


Figure 35. Nombre de signaux restant à identifier manuellement après analyse statistique (t-test) préalable pour les différents niveaux de dopage

Dans cet exercice, le but était d'évaluer l'efficacité de cette approche pour l'identification des toxines dopées aux échantillons de moules et d'huîtres. Nous avons donc simplement vérifié la présence des ions correspondant à ces toxines dans les listes finales d'ions pour chacun des niveaux de contamination. Pour les deux matrices moules et huîtres, les toxines présentes dans nos échantillons ont été identifiées parmi les ions responsables des variabilités observés entre contrôles et contaminés sans ambiguïté pour les niveaux de contamination les plus élevés avec une p-value < 0,01. Pour le niveau de contamination le plus faible, des 5 toxines étudiées seule la SPX 1 a pu être identifiée. Pour la GYM, ce n'est qu'à partir du deuxième niveau de concentration (4 ng/mL) qu'on a pu la retrouver, à partir du niveau 3 (8 ng/mL) pour l'AZA1 et la PnTX-A et à partir du niveau 4 (12 ng/mL) pour la PTX2 (**Tableau 29**).

Tableau 29. Résultats du t-test pour l'identification des toxines en fonction du niveau de concentration étudié, par comparaison avec l'extrait non dopé (niveau N0)

Niveau de contamination comparé à N0	p-value				
	SPX1	GYM	AZA1	PnTX A	PTX 2
N1	✓	×	×	×	×
N2	✓	✓	×	×	×
N3	✓	✓	✓	✓	×
N4	✓	✓	✓	✓	✓
N5	✓	✓	✓	✓	✓
N6	✓	✓	✓	✓	✓

✓ : toxine identifiée (p-value < 0,05) ; × : toxine non identifiée (p-value > 0,05)

II.3.2.i. Résultats des analyses multivariées

Après l'analyse univariée, confirmant la présence de toxines dopées et analysées dans le cadre de l'étude, nous avons voulu tester l'analyse multivariée, également permise par la suite logicielle Sciex.

Initialement, les données extraites par MarkerView™ des différents échantillons, moules, huîtres et solutions standards ont été analysées par une ACP non supervisée. Le graphique des scores plot des deux premières composantes montrent la présence de trois groupes différents correspondant respectivement aux échantillons d'huîtres, de moules et aux solutions multitoxines (étalons) dans le solvant (**Figure 36**).

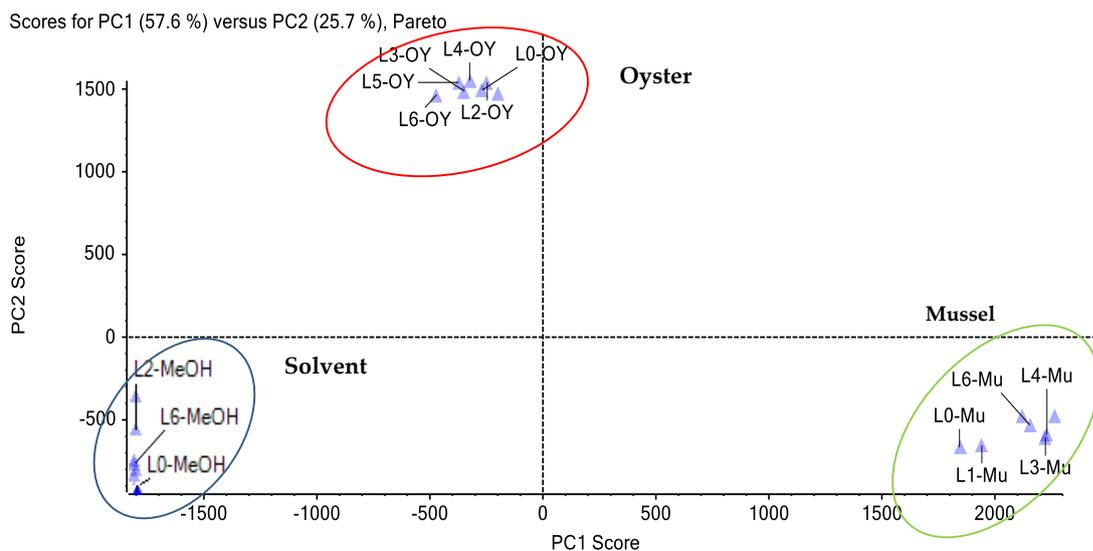


Figure 36. Scores plot obtenu après l'analyse ACP des données générées en ESI+ pour les échantillons de moules, d'huîtres et les solutions multitoxines dans le MeOH

Le score plot montre que les composantes PC1 (57,6 %) et PC2 (25,7%) expliquant à elles deux 83% de la variabilité reflètent en définitive la variabilité liée aux composantes matricielles. Cette distribution n'est pas surprenante dans la mesure où les ions matriciels prédominent par rapport aux ions représentatifs des composés d'intérêt. L'étude des autres composantes (PC3 à PC6) n'a pas révélé non plus de clusterisation basée sur la présence ou non de toxines dans les échantillons. Pour surmonter ou réduire l'impact de la variabilité matricielle qui était prépondérante lors du premier test, nous avons réalisé une analyse supervisée (ACP-DA) plus discriminante, permettant de définir les échantillons de moules et d'huîtres comme appartenant au même groupe. Le scores plot correspondant (**Figure 37**) montre que tous les « L0 » (échantillons non contaminés) sont bien regroupés dans la partie supérieure du graphique et bien séparés des échantillons contaminés. Les autres échantillons sont classés en fonction de leurs niveaux de concentration, du moins contaminé au plus contaminé. Nous observons quatre groupes correspondant aux contrôles (rouge), aux niveaux de contamination faible (vert), moyen (orange) et élevé (bleu), sans distinction des matrices moules (\square) et huîtres (Δ).

Scores for D1 (22.9 %) versus D2 (22.3 %), Pareto (DA)

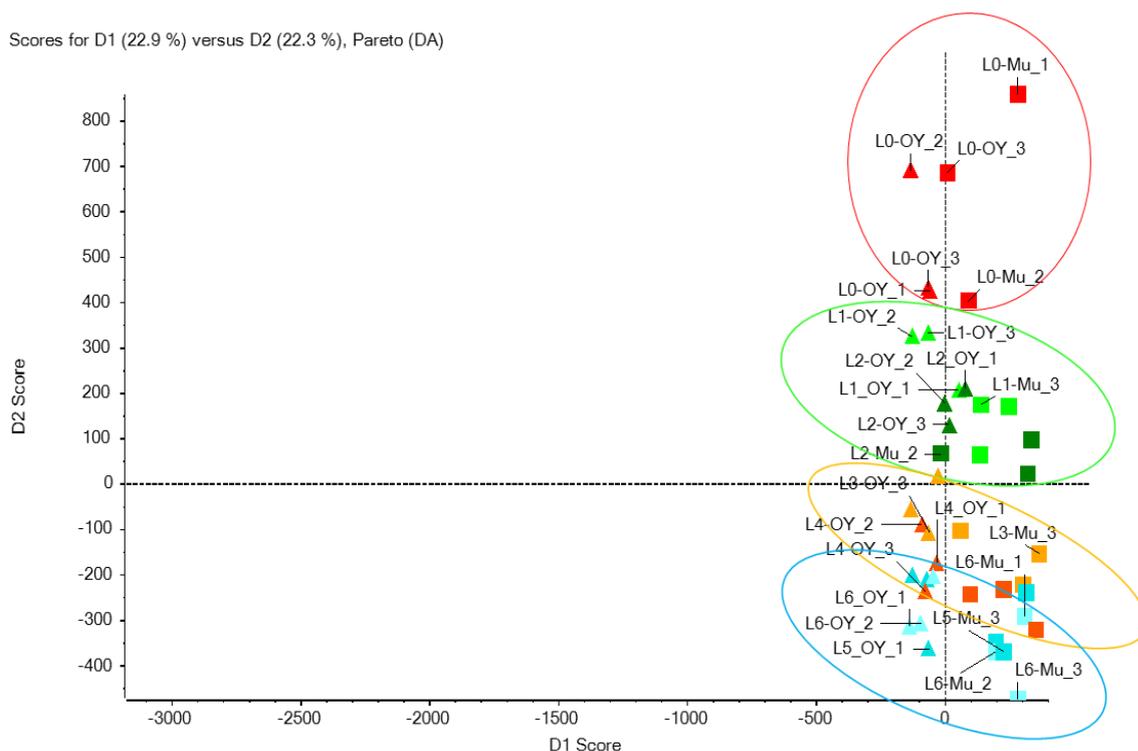


Figure 37. Scores plot obtenu après l'analyse ACP-DA des données générées en ESI+ pour les échantillons de moules et d'huîtres.

□ =moules (Mu), △ = huîtres (OY) ; rouge =niveau 0 ; vert clair = niveau 1 ; vert foncé = niveau 2 ; orange = niveau 3 ; rouge = niveau 4 ; bleu = niveaux 5 et 6.

Pour identifier les ions responsables de la discrimination entre les groupes, nous nous sommes intéressés au graphique des LAOdings plot (**Figure 38**) qui permet de visualiser la distribution des ions en corrélation avec le scores plot. Nous avons donc sélectionné les ions (entourés en bleu) susceptibles de représenter les niveaux les plus contaminés. Nous avons ainsi obtenu une liste de 70 signaux réduites à 55 après l'élimination des isotopes. Les 55 ions monoisotopiques retenus ont été réintroduits dans PeakView® et vérifiés suivant les mêmes étapes que précédemment pour les ions issus du t-test. Les étapes de filtrage manuel ont permis de réduire la liste à 38 et 36 ions pour les matrices d'huîtres et de moules respectivement. Les toxines (GYM, SPX1, AZA1, PnTX-A et PTX2) marquées d'une étoile sur la **Figure 38** ont bien été identifiées parmi la liste finale des signaux responsables de la discrimination observée entre les échantillons contaminés et contrôles. Bien que cette approche ait permis de retrouver les signaux correspondant à toutes les toxines présentes dans la liste des ions d'intérêt, le nombre de faux positifs reste assez important (85 et 87% pour les échantillons de moules et d'huîtres, respectivement). Cela illustre la difficulté qu'il peut y avoir à identifier des composés inconnus par cette approche, du fait du nombre important de signaux qu'il faut investiguer

individuellement pour arriver à identifier le signal d'intérêt, correspondant au composé recherché.

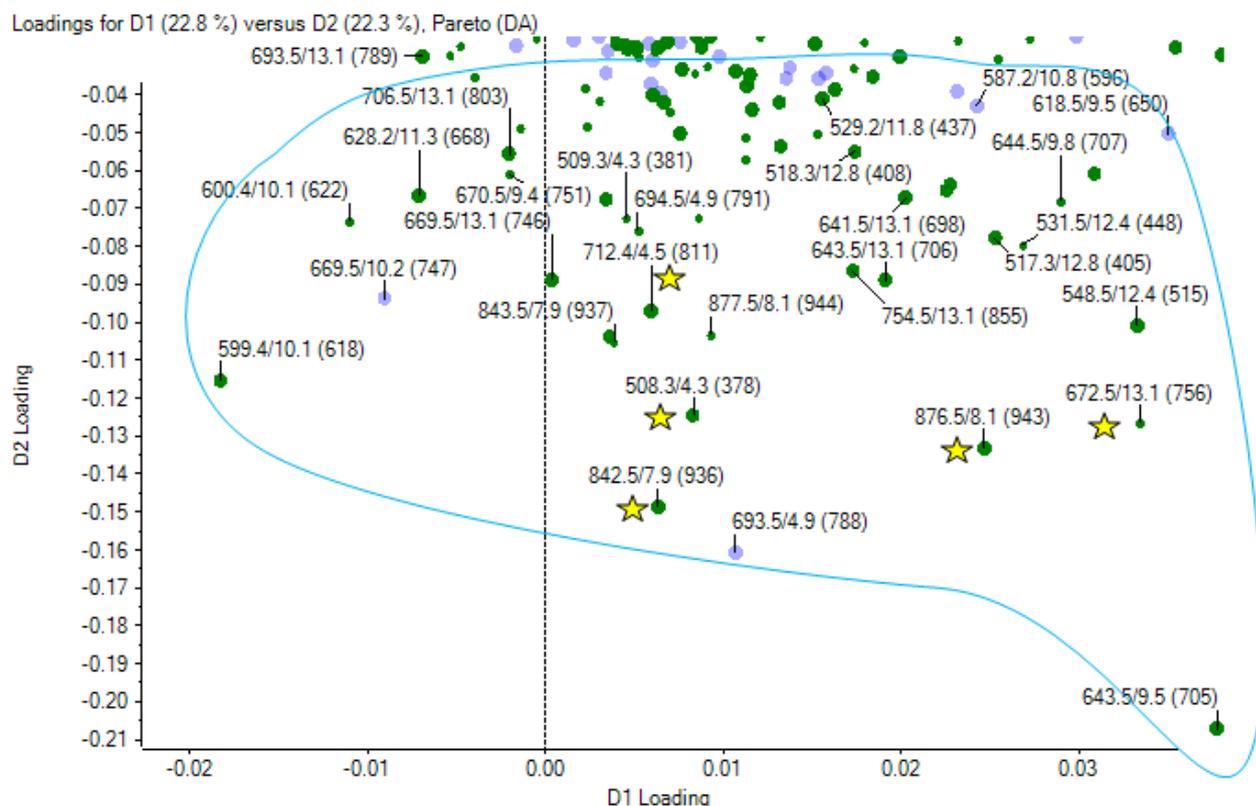


Figure 38. Zoom sur le graphique des loadings plot (ACP-DA): les ions encerclés en bleu correspondent aux variables représentatives (points verts) du niveau le plus contaminé (L6). Les ions correspondant aux toxines d'intérêt sont marqués par des étoiles jaunes. Les trois chiffres indiqués à côté de chaque point vert (tels que 508.3 / 4.3 (378)) représentent respectivement la masse exacte, le temps de rétention et l'aire du pic.

II.1. Conclusions

Dans le cadre d'une preuve de concept, un training test a été mené sur des échantillons de moules et d'huîtres supplémentés avec différentes toxines à différents niveaux analysés par LC-HRMS. Les échantillons analysés ont été traités comme des échantillons inconnus tout au long de l'étude afin d'évaluer les performances des workflows (**Figure 39**) développés pour l'approche « suspect screening » et l'approche non-ciblée sans a priori.

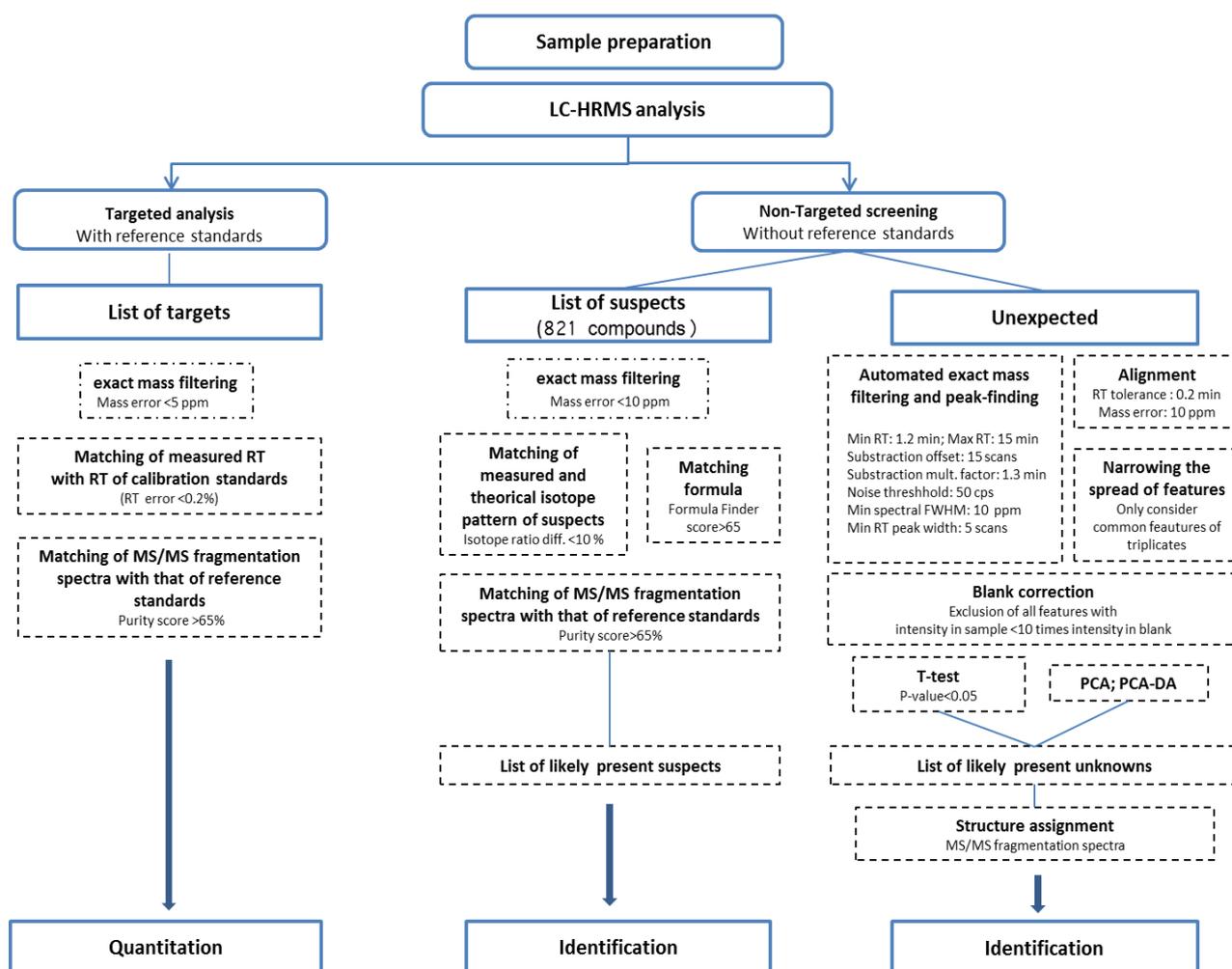


Figure 39. Workflow optimisé pour l'analyse des données LC-HRMS selon les trois approches : (i) screening ciblée, (ii) suspect screening et (iii) screening non ciblée sans a priori

L'optimisation des différentes étapes du workflow développé pour le suspect screening a permis de réduire efficacement le nombre de faux positifs et de faux négatifs. Nous avons réussi à identifier l'ensemble des toxines dans tous les échantillons à tous les niveaux de contamination étudiés. La présence de plusieurs analogues isobares dans la liste des suspects présente une difficulté supplémentaire dans l'identification sans équivoque des composés d'intérêt.

Deux logiciels d'extraction de pics ont été testés pour l'approche non ciblée sans a priori. L'application MasterView™ du logiciel PeakView® s'est avérée très efficace pour l'extraction des composés d'intérêt avec l'approche de screening comparatif avec un pourcentage très faible de faux positifs. Nous avons ainsi réussi à confirmer la présence des toxines étudiées sans difficultés dans l'ensemble des échantillons exceptés pour les 2 niveaux de contamination les plus bas.

Les données extraites avec le logiciel MarkerView™ a permis d'extraire un nombre beaucoup plus important de signaux qu'avec l'application MasterView™. L'analyse univariée des données en appliquant un t-test a permis de recenser les composés responsables des différences significatives entre les échantillons contaminés et contrôle. Les toxines étudiées ont pu être retrouvées et identifiées parmi la liste finale des ions. Bien que le pouvoir discriminant du test statistique se soit montré efficace, il n'était pas possible d'attribuer les différences entre les groupes d'échantillons uniquement aux toxines présentes.

L'analyse multivariée par ACP nous a permis de conclure que ce test n'était pas le plus adapté pour l'analyse des données générées en HRMS. L'analyse supervisée (ACP-DA) s'est avérée plus adéquate et a permis de discriminer les échantillons en différents groupes contaminés et contrôles mais aussi en fonction du niveau de contamination des échantillons. La liste des ions considérés par ce test comme responsables des variabilités entre les échantillons était plus importante qu'avec le t-test. Nos toxines d'intérêt ont bien été retrouvées parmi cette liste, toutefois le nombre de faux positifs restent assez important.

L'identification des ions déterminés comme composés d'intérêt est l'une des étapes les plus importantes et parmi les plus complexes du workflow. En effet, les algorithmes utilisés tel que « Formula Finder » génèrent un nombre très important de formule possibles pour une même masse exacte. À cette étape intervient l'esprit critique de l'analyste et sa connaissance des compositions élémentaires probables des familles de composés recherchés pour limiter le nombre de formules à confirmer par les spectres de fragmentation. La deuxième difficulté rencontrée lors de l'identification est soit (i) l'indisponibilité des spectres de fragmentations théoriques dans les bases de données pour certains composés ce qui est notamment le cas pour les biotoxines marines soit (ii) l'absence d'acquisition d'un spectre de fragmentation exploitable lors de l'analyse LC-HRMS. En effet, bien que l'acquisition en mode IDA soit l'un des meilleurs compromis possibles pour une analyse non ciblée, tous les composés d'intérêt ne sont pas forcément fragmentés à cause de la richesse des matrices analysées. L'une des solutions possibles qui permettrait de résoudre ce problème serait de procéder à une analyse ciblée avec le mode d'acquisition « Product ion » à différentes énergies de collisions en considérant la masse exacte de l'ion à investiguer pour obtenir des spectres de fragmentation plus complets à étudier. Dans le cas d'analyse d'un échantillon réellement contaminé, en la présence d'un nouveau composé, on parlerait plutôt d'élucidation structurale en étudiant les spectres de fragmentation mais l'identification resterait provisoire et devrait être confirmée par d'autres techniques spectrales telles que la RMN.

C. Traitement des données HRMS logiciel open source : XCMS

I. Matériels et Méthodes

I.1. Préparation des échantillons et des contrôles qualité (QC)

Préparation des échantillons

Pour cette étude, nous avons choisi d'élargir la variabilité matricielle de nos échantillons. Nous avons donc sélectionné 3 groupes différents de coquillages (moules *Mytilus edulis*, huîtres *Crassostrea gigas* et coquilles St-Jacques *Pecten maximus*) composés chacun de quatre échantillons différents (**Tableau 30**). L'ensemble des échantillons a été extrait selon le même protocole décrit précédemment (I.4).

Quatre solutions ont été préparées à partir de chacun des extraits : un échantillon contrôle (N0) et 3 échantillons dopés à 3 niveaux (4, 8 et 16 ng/mL) (**Tableau 31**) par un mix de solutions étalons de toxines (GYM, SPX1, PnTX-A et AZA1) (**Tableau 32**). La PTX2 n'a pas été retenue pour cette étude car il a fallu faire un choix en matière de toxines en raison du coût associé au dopage des matrices plus nombreuses dans cette étude-ci. De plus, le pouvoir discriminant du test statistique a montré que la PTX2 n'est identifiable qu'à partir de 12 ng/mL avec les logiciels constructeurs. Même si un autre logiciel est prévu pour l'exploitation des données générées dans cette partie, nous avons préféré ne pas prendre de risque et passer à côté de deux des trois niveaux de dopage prévus.

Préparation des contrôles qualité (QC)

- **QC 1** est composé de 50 µL de solution étalon de boscalid (10 ng/µL) utilisé comme étalon interne (EI) complété avec 450 µL de MeOH. Il est injecté en début de séquence d'analyse pour vérifier l'état du système analytique.
- **QC 2** est composé de 50 µL de la solution étalon de boscalid et 450 µL de l'extrait correspondant au contrôle réactif (blanc de procédure obtenu en procédant à l'ensemble de la méthode d'extraction mais sans échantillon, l'objectif étant d'identifier les ions apportés par les solvants et autres réactifs utilisés).
- **QC 3** ou QC pool correspond à un extrait représentatif de la totalité des échantillons de la série d'analyse. Pour réaliser le QC pool, il est nécessaire de prélever le même volume (50 µL) de chaque échantillon (« dopés » et « contrôles »). La totalité des volumes prélevés est regroupée, mélangée et répartie dans plusieurs vials d'injection.

Tableau 30. Echantillons utilisés dans le cadre de l'étude

Moules (<i>M. edulis</i>)	11 PHYCO 523 18 BM 008 15 PHYCO 205 15 PHYCO 375
Huîtres (<i>C. gigas</i>)	15 BM 374 15 BM 376 18 BM009 10 PHYCO 013
Coquilles St Jacques (<i>P. maximus</i>)	16 BM 399 17 BM 068 17 BM 059 13 PHYCO 024

Tableau 31. Récapitulatif des niveaux de dopage et des contrôles utilisés

N0	Contrôle (non dopé)
N1	4 ng/mL
N2	8 ng/mL
N3	16 ng/mL
QC 1	MeOH + EI
QC 2	Contrôle réactif + EI
QC 3	QC pool (mélange des extraits)

Tableau 32. Toxines et étalon interne (EI) utilisés pour doper les échantillons

Molécule	Abréviation	Formule brute	Masse exacte (Da)
13-desmethyl spirolide C	SPX 1	C42H61N07	691,4448
Azaspiracid 1	AZA 1	C47H71N012	841,49763
Gymnodimine A	GYM	C32H45N04	507,33486
Pinnatoxine A	PnTX-A	C41H61N09	711,43463
Boscalid*	Boscalid	C18H12Cl2N2O	342,03267

* : utilisé comme EI

I.2. Conditions d'analyse par LC-HRMS

Les analyses ont été menées par LC-HRMS en mode d'ionisation positif utilisant les mêmes conditions chromatographiques et paramètres de masses que ceux décrits pour l'étude précédente (I.5).

Les échantillons ont été analysés de manière aléatoire. Pour cela, le séquençage des échantillons a été randomisé grâce au logiciel R (fonction « rand »). L'ordre d'injection des échantillons est un paramètre important à prendre en compte lors de la préparation de la séquence d'injection pour réduire l'effet de la variabilité du signal sur les résultats. La séquence commence toujours par plusieurs injections de QC standard et de solvant afin de s'assurer de la stabilité du système

analytique, puis de deux injections de QC2 suivie des QC pool (**Figure 40**). Les échantillons sont ensuite insérés dans la séquence en intercalant un QC3 tous les 5 échantillons.

QC 1	15 PHYCO 205-N 1
QC 2	17 BM 059-N2
QC3-1	15 PHYCO 376-N1
13 PHYCO 024-N1	QC3
13 PHYCO 024-N3	[...]
15 PHYCO 375-N2	17 BM 068-N2
18 BM 008- N3	15 PHYCO 375-N3
11 PHYCO 523 -N3	16 BM 399-N0
QC3	QC 3
18 BM 008- N1	QC 2
15 PHYCO 374-N3	QC1

Figure 40. Exemple d'une séquence d'injection

I.3. Traitement des données

I.3.1. Conversion des données

Avant de procéder au traitement des données par les logiciels open source, les données brutes au format constructeur (.raw) ont été converties dans le format mzXML, qui est un format libre, à l'aide de la commande msconvert du logiciel ProteoWizard disponible gratuitement sur Internet (<http://proteowizard.sourceforge.net/>).

I.3.2. Traitement des données brutes

Le traitement des données brutes a été réalisé avec les packages XCMS et CAMERA du logiciel RStudio, tous deux disponibles sur la plateforme Galaxy « Work4metabolomics » (<http://workflow4metabolomics.org/the-galaxy-environment>). Les détails des fonctions (utilisées sous R) ainsi que les paramètres XCMS utilisés sont répertoriés en **Annexe 4**. Les étapes de traitement sont présentées ci-dessous :

- 1- Extraction des pics : Cette étape réalisée avec l'algorithme « Centerwave » consiste en la sélection des signaux analytiques présents dans les données brutes acquises correspondant à l'ensemble des ions détectés dans l'ensemble des échantillons (fonction, « xmsSet »)
- 2- Alignement des empreintes et correction des temps de rétention (Tr) : le logiciel détermine quelles sont les variables qui correspondent à la détection d'un même signal dans différents échantillons. Cela revient à créer des groupes de variables alignées selon un m/z et un temps

de rétention proche (fonctions « group »). Une fois les échantillons alignés, avec la fonction « rector », les légers décalages en temps de rétention sont corrigés.

3- Complétion des données manquantes : il est très fréquent que les variables soient détectées dans certains échantillons et absentes dans d'autres. Pour chaque variable l'étape de complétion des données va déterminer des limites de temps de rétention et de m/z basées sur les différents échantillons dans lesquels celle-ci est présente, puis va intégrer le signal présent dans cette plage de temps de rétention et m/z dans les échantillons où la variable étudiée est absente (fonction « fillpeaks »).

4- Création d'une table de données : avec le package CAMERA une matrice est créée comportant l'ensemble des « variables » avec création pour chacun les EIC (Extracted Ion chromatogram) et boxplots correspondants.

5- Normalisation des variables : la variabilité intra-série du signal des ions présents dans les QC est corrigée en appliquant une régression linéaire pour l'ensemble des échantillons.

6- Filtration des variables : les variables sont filtrées selon leur intensité par rapport aux contrôles analytiques et leur coefficient de variation. Les ions issus des contrôles réactifs sont éliminés. La filtration des variables est réalisée en fonction du rapport des coefficients de variation : les variables ayant un rapport $\frac{CV_{QC\ pool}}{CV_{\text{échantillons}}} > 1$ sont éliminées.

I.4. Analyses statistiques

Les données ont été analysées par des méthodes chimiométriques univariées et multivariées grâce à l'outil « Statistical Analysis » de la plateforme **w4m** regroupant des packages R dont le package « ropls » pour les analyses multivariées.

I.4.1. Analyses univariées

Deux tests paramétriques de Student et non-paramétriques Wilcoxon-Mann-Whitney ont été réalisés avec une erreur α de 5% pour déterminer la significativité (p-value) des variables entre les deux groupes (« dopés » et « contrôles »). Une variable est considérée comme significative dans la discrimination entre deux groupes si sa p-value est inférieure à 0.05.

I.4.2. Analyses multivariées

Une transformation logarithmique des intensités de toutes les variables, puis une normalisation par centrage et mise à l'échelle standard ont été effectuées avant les analyses multivariées.

Dans un premier temps, une analyse en composante principale (ACP) a été réalisée afin de visualiser la répartition des échantillons dans un espace en deux dimensions. Dans un second temps, une analyse supervisée de type PLS-DA (de l'anglais, Partial Least Square Discriminant Analysis) a été appliquées au jeu de données afin d'affiner les résultats et de déterminer quelles variables participent le plus à la différenciation des groupes (dopés /contrôle).

Les modèles de PLS-DA permettent de mettre en évidence les variables les plus impliquées dans la discrimination des groupes « dopés » et « contrôles » qui sont matérialisées par leur score VIP (de l'anglais, Variable Importance in the Projection). Nous considérons ici une variable comme discriminante pour le modèle si son score VIP est supérieur à 5. Des tests de permutations ($k = 1000$) ont été réalisées pour valider les modèles de PLS-DA.

Une comparaison des variables discriminantes obtenus en PLS-DA et en analyses univariées a été réalisée pour conforter ou non les résultats.

II. Résultats et discussions

II.1. Traitement des données brutes

L'ensemble des empreintes acquises par LC-HRMS en ESI+ de tous les échantillons analysés (sans différenciation entre les matrices) a été traité avec le workflow développé. Le logiciel XCMS a permis dans un premier temps d'extraire **9464** signaux. Ces données ont ensuite subi les différentes étapes de prétraitement : alignement (**Figure 41**), complétion des données, etc. pour générer ensuite une matrice des données et une matrice des variables avec le logiciel CAMERA (**Figure 42**). Les données ont été ensuite normalisées (**Annexe 5**) et filtrées en fonction du coefficient de variation des intensités et cette étape a permis d'éliminer 16% des données extraites initialement pour réduire le nombre de signaux à exploiter à **7949**. Pour visualiser et comprendre l'information contenue dans les tableaux de données, deux types d'analyses statistiques univariées et multivariées ont été effectuées.

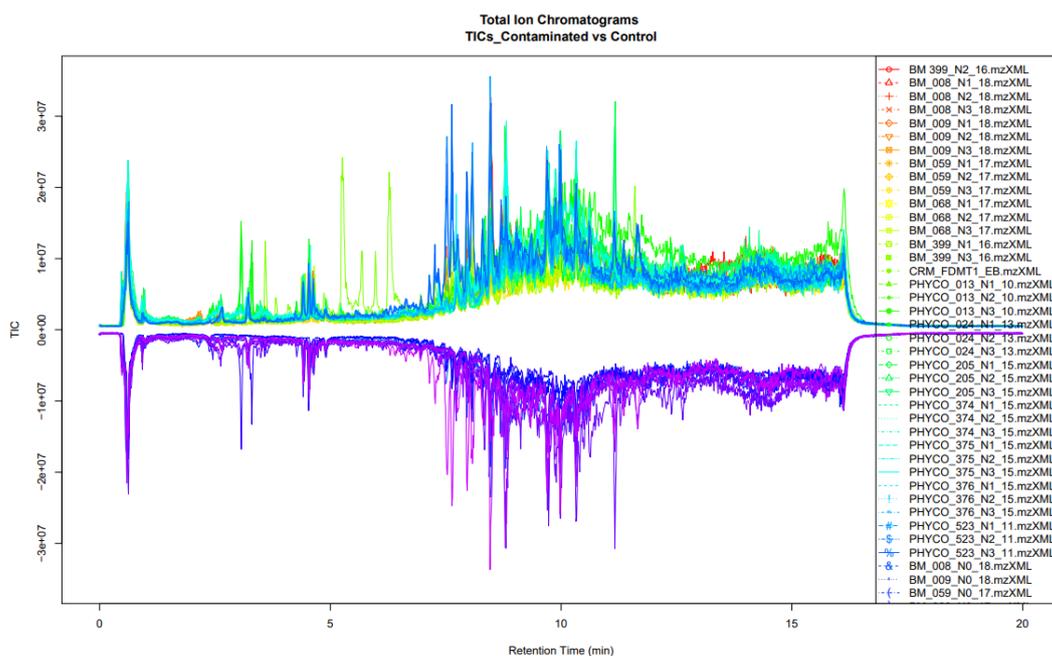


Figure 41. Alignement des pics avec le logiciel XCMS (fonction « rector »)

II.2. Analyses statistiques des données

Les données ont été dans un premier temps analysées en globalité à l'aide d'une analyse non supervisée (ACP) afin de visualiser la répartition « spontanée » des échantillons dans un espace à deux dimensions. Les résultats ont ensuite été affinés à l'aide d'une deuxième analyse multivariée supervisée de type PLS-DA et confirmés par les analyses univariées (t-test et test de Wilcoxon).

Résultats des analyses ACP

La première ACP a été réalisée sur les données relatives à l'ensemble des échantillons (toutes matrices confondues). La carte factorielle du poids des individus, appelée « scores plot », est représentée en **Figure 43a**. Elle montre que les QC pool sont tous regroupés au centre confirmant la qualité du jeu de données et la validité de la séquence analytique. La capacité descriptive du modèle signifiée par la valeur $R^2X = 0,503$ indique que la moitié de la variabilité du jeu de données est prise en compte par ce modèle. La distribution graphique des échantillons ne montre aucune discrimination entre les deux groupes dopés et contrôles et ce, quel que soit le niveau de dopage. La variabilité représentée par les composantes principales 1 (21%) et 2 (8%) traduit majoritairement les différences en fonction de la nature des échantillons et ne prend pas en compte la présence ou absence de toxines. Les projections des autres composantes principales ne permettent pas non plus de mettre en évidence des discriminations en fonction de « l'état

Matrice des variables

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	name	namecustom	mz	mzmin	mzmax	rt	rtmin	rtmax	npeaks	Contaminat	Control	QCpool	isotopes	adduct	pcgroup	
2	M100T153	M100.0751T2	100.0751493	100.0725249	100.0766498	2.557107696	2.523931472	2.846375535	52	32	9	11	[1][M]+			2209
3	M101T84	M101.0226T1	101.0225650	101.0205450	101.0250089	1.392093128	1.321302377	1.482403757	57	35	11	11				2623
4	M101T283	M101.0597A	101.0589671	101.0559372	101.0616328	4.713784418	4.652087093	4.757730385	50	30	10	9				2548
5	M101T232	M101.0597A	101.0589730	101.0564126	101.0629229	3.859034512	3.816899900	3.903985100	45	27	9	9				2338
6	M101T153	M101.0779T2	101.0779250	101.0771727	101.0795163	2.553219829	2.524374090	2.584490676	40	22	7	11	[1][M+]			2210
7	M101T277	M101.0949T4	101.0949166	101.0930078	101.0969674	4.624462584	4.596154644	4.648586713	58	36	12	10				1767
8	M102T158	M102.0451T2	102.0451070	102.0432884	102.0485835	2.636894158	2.597564085	2.648550246	12	7	5	0				1484
9	M102T38	M102.0536T2	102.0536119	102.0515823	102.0580687	0.625832945	0.577615713	0.993719573	54	32	9	10				1317
10	M102T153	M102.0803T2	102.0802984	102.0772933	102.0829835	2.548728397	2.526696031	2.572477900	25	11	5	9	[1][M+]			2211
11	M103T982	M102.9694T1	102.9693648	102.9675901	102.9726437	16.36093789	16.30698074	16.42949676	39	23	9	7				670
12	M103T159	M103.0536T2	103.0535955	103.0516071	103.0566601	2.649602312	2.605442138	2.958418004	56	33	12	10				1485
13	M104T58	M104.0527O	104.0520386	104.0477565	104.0548871	0.972440574	0.649506386	1.030397239	57	34	11	10				2064
14	M104T38	M104.1058T2	104.1057648	104.1038621	104.1082548	0.631096635	0.616035601	0.731134465	57	34	12	11	[2][M]+			1318
15	M105T180	M105.0695T2	105.0695022	105.0651024	105.0710200	2.995518356	2.966543707	3.008756555	40	22	7	11				1416
16	M105T418	M105.0695T2	105.0695442	105.0670807	105.0742795	6.966688855	6.938861314	6.986499439	25	13	7	5				68
17	M105T40	M105.1089T2	105.1088815	105.1062908	105.1108967	0.670803856	0.612658259	0.721475670	33	19	5	9	[2][M+]			1319
18	M106T67	M106.0676T1	106.0675746	106.0658819	106.0702710	1.119104560	1.080527220	1.193147726	20	10	4	6				1978
19	M106T179	M106.0724T2	106.0723626	106.0674750	106.0749249	2.983970366	2.958787062	3.000532893	25	12	5	8				1417
20	M107T159	M107.0484T2	107.0483861	107.0450090	107.0508061	2.646541072	2.615430488	2.799949298	35	25	8	2				33
21	M107T278	M107.0697T4	107.0697074	107.0672182	107.0724291	4.630754765	4.605177798	4.644386734	60	37	12	11				1768
22	M107T523	M107.0701T2	107.0701209	107.0672146	107.0737661	8.723373718	8.376706166	8.771652913	24	15	6	2				944
23	M108T38	M108.0101T2	108.0101422	108.0070872	108.0130624	0.638232029	0.598462789	0.710970731	36	24	6	6				1320
24	M109T61	M109.0397T1	109.0390065	109.0368207	109.0419902	1.012591476	0.990241572	1.054321024	18	7	3	8				1979
25	M109T256	M109.0641T4	109.0640778	109.0598004	109.0661150	4.262664769	4.174942796	4.343067261	57	35	11	11		[M+3H]	3+32	39
26	M109T365	M109.0641T2	109.0641267	109.0613325	109.0663194	6.082034717	6.021618312	6.103418493	53	34	9	10				2788
27	M109T332	M109.0996T5	109.0995630	109.0980496	109.1027658	5.536273308	5.478466689	5.591255378	41	27	9	5				2390
28	M110T976	M110.0079T1	110.0079179	110.0060210	110.0096048	16.27425088	16.21890487	16.30698482	60	37	12	11				10
29	M110T30	M110.0078T2	110.0078455	110.0064853	110.0095168	0.500110237	0.471293839	0.527535743	51	31	10	10				2938
30	M110T37	M110.0262T2	110.0261595	110.0242644	110.0285967	0.613136958	0.587942412	0.639835840	24	17	5	2				1321
31	M111T30	M111.0082T2	111.0081984	111.0036924	111.0108239	0.494828419	0.475422534	0.516136543	36	23	6	7				2939
32	M111T976	M111.0101T1	111.0101228	111.0080715	111.0128577	16.27183416	16.24741657	16.29918670	14	4	5	5				671
33	M111T262	M111.0796T4	111.0796135	111.0778975	111.0816687	4.361797121	4.303826021	4.413939954	60	37	12	11				1652
34	M111T468	M111.1157T7	111.1157481	111.1140305	111.1176893	7.805216050	7.786022282	7.827915568	60	37	12	11	[3][M]+			1159
35	M111T228	M111.1154T7	111.1153621	111.1125236	111.1183270	3.802837185	3.754911948	4.074776350	46	29	10	7				2339
36	M112T30	M112.0065T2	112.0065331	112.0035642	112.0110396	0.491993097	0.468701904	0.520178465	40	27	8	5				2940

Matrice des données

name	BM_399_N2	BM_008_N1	BM_008_N2	BM_008_N3	BM_009_N1	BM_009_N2	BM_009_N3	BM_059_N1	BM_059_N2	BM_059_N3	BM_059_N4
M100T153	3417.787687	317429.4413	316912.9786	313386.2365	6563.1342	5429.0175	4583.956866	3251.075257	3979.520574	2916.109761	2
M101T84	2914.62	7212.560409	6888.191421	3906.138901	6376.314230	7439.358411	7083.982317	1102.739205	1841.435016	1222.179895	1
M101T283	16290.90606	13985.78857	8778.944954	9489.723571	13887.54759	16565.22197	9206.129310	9949.675437	9129.601538	5647.439118	1
M101T232	12014.25970	12539.13313	5673.45725	9756.960447	9250.776488	12927.42646	13323.71486	11568.32592	13443.21325	8791.691072	8
M101T153	1998.788455	29202.85714	24290.7904	25188.51977	2303.181991	2252.165924	2622.283954	2182.254499	2777.946260	2334.697655	2
M101T277	11793.84431	10701.67635	14073.19107	18560.31767	14448.37771	12240.006	15916.44846	12162.10931	16166.51255	15356.20978	1
M102T158	806.7200750	1169.151960	2026.521615	2105.508937	859.936235	775.6458056	798.0776576	629.1453675	671.9386176	647.8366150	9
M102T38	18462.3543	13215.98068	10861.52100	9836.328705	11521.92	13842.28613	13377.81087	25193.95875	27623.50297	27753.87537	3
M102T153	926.3442694	2327.188880	2539.250297	1045.310142	984.3785442	1084.766732	1197.483987	947.4686056	964.9805511	1109.895324	9
M103T982	2186.347333	2026.072612	1890.643530	2008.859523	2094.431379	1849.386272	1965.500945	2251.318914	1541.426554	1829.812363	1
M103T159	2319.910588	3625.884285	2934.429473	4970.066617	2515.812391	2799.605122	4013.009	1405.591058	604.4396249	1323.073079	7
M104T58	2137.161142	3251.28	4073.246307	3636.8316	4969.530085	1123.235466	2121.5115	1660.125	1469.208888	1264.1578	1
M104T38	132209.8008	47112.53433	42522.25785	34947.51959	92525.40670	85675.28114	96854.38252	19285.82712	24542.291	19331.95387	2
M105T180	2040.287975	2682.693090	2665.505045	3309.265387	4960.030537	4546.32768	4151.3235	2581.489827	2670.190061	2019.897146	2
M105T418	2309.322157	4408.253577	1817.694923	5680.262506	6242.242128	4956.829230	5723.770461	4754.53775	4216.346001	3389.778088	3
M105T40	3273.69394	5538.124533	5177.241641	5057.825162	7073.140376	9100.393778	8641.130084	2885.170197	3838.118068	3182.302856	4
M106T67	1351.962106	1004.829343	919.2020555	1215.961992	1202.513395	1075.358601	1819.239069	1053.164978	1415.864576	1045.420388	1
M106T179	789.4928185	1068.413053	686.6546739	835.9996234	645.4572769	995.9769502	934.8793218	821.3269774	798.4172649	619.987728	6
M107T159	1154.692208	1557.735739	1208.41875	979.2648571	1448.572101	1743.757306	1551.760573	1087.017091	1170.902944	1434.066995	8
M107T278	26682.43349	26529.85636	35587.00326	22574.89096	38482.152	27945.78660	29309.06314	30110.58014	32906.58193	31427.69784	2
M107T523	9019.211731	11473.47405	10879.36333	9590.413759	10656.33100	9665.826542	9766.806020	1125.059000	10579.40323	8175.943548	9
M108T38	3555.566527	8797.086315	11785.40764	6479.475714	7374.984085	9650.924043	12861.77765	8732.377284	2715.852615	6595.040897	1
M109T61	1930.350143	3936.933	2550.9792	4452.684857	871.8793197	727.7776141	1225.203400	1625.055598	1816.193637	1230.350710	1
M109T256	17229.69282	15977.28214	12930.75202	15384.92663	19343.49035	18560.212	26232.04942	12870.22762	22395.70069	17660.84916	1
M109T365	13282.76434	7947.491416	1167								

de contamination » des échantillons. Ces résultats sont cohérents avec les observations faites lors de la première étude avec MarkerView™ et expliquées par la prédominance des composantes matricielles.

Les analyses ACP réalisées à partir des données issues des échantillons (contrôle N0 et dopés au niveau N3) et traitées individuellement pour chacune des trois matrices (moules, huîtres et CSJ) ont été effectuées et ont abouti aux mêmes conclusions, à savoir une absence de discrimination entre échantillon dopé et contrôle (non dopé). Le graphique des scores plot issu de l'ACP correspondant à la matrice moule est représentée en **Figure 43b**. Les ACP correspondant aux autres matrices sont présentées en **Annexe 6**.

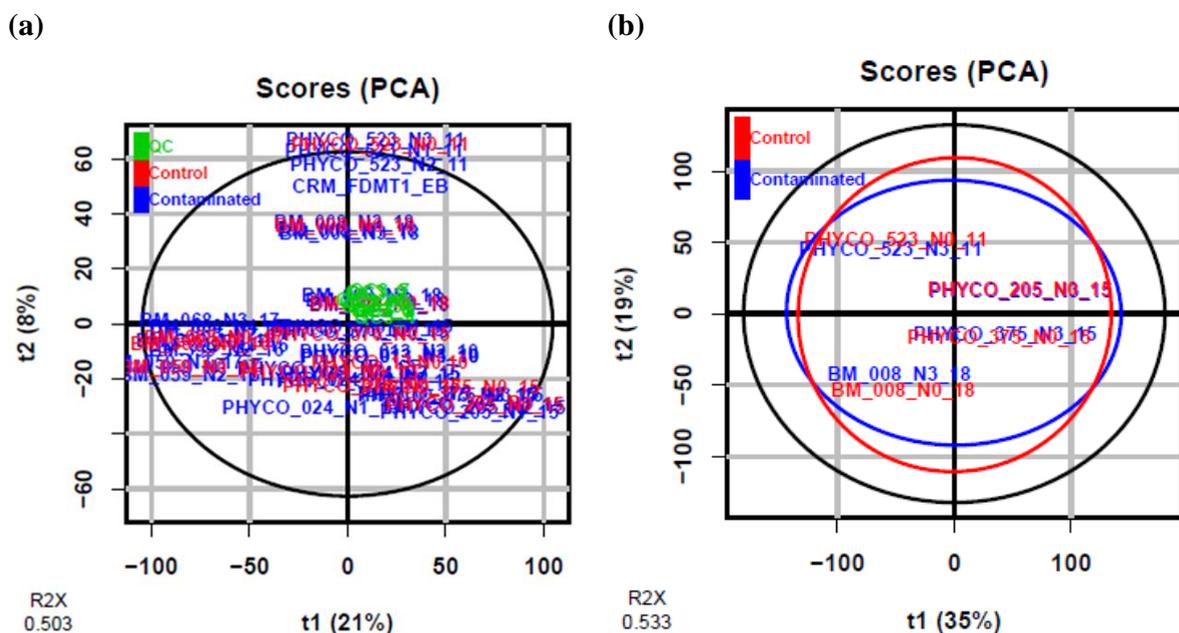


Figure 43. Représentation « scores plot » de l'ACP réalisée (a) pour l'ensemble des échantillons (moules, huîtres et CSJ) et (b) pour les échantillons de la matrice moule

Résultats des analyses PLS-DA et des tests univariés

La PLS-DA, autre modèle dit cette fois « supervisé », a été réalisée sur le même jeu de données utilisées pour l'analyse ACP. Les deux PLS-DA réalisées d'abord sur l'ensemble des échantillons (**Figure 44a**) puis uniquement sur les échantillons de la matrice moule (**Figure 44b**) ont permis d'obtenir des résultats différents de ceux observés avec les ACP. Les graphiques des scores montrent une discrimination assez nette des échantillons en deux groupes correspondant respectivement aux échantillons dopés (en bleu à gauche du graphique) et contrôles (en rouge à droite).

La qualité des modèles ainsi construits peut être évaluée par trois paramètres : $R^2(Y)$, qui correspond au pourcentage de la variabilité qui peut être expliquée par la variable Y, et $Q^2(Y)$ la capacité à restituer (prédire) la variable Y. R^2X correspond à la proportion de la variabilité représentée par le modèle.

Les graphiques correspondant respectivement à l'ensemble des échantillons (PLS-DA 1) et à la matrice moule (PLS-DA 2) montrent qu'une faible proportion de la variabilité ($R^2X=0,286$ et $0,422$) est représentée par le modèle. Cette proportion bien qu'inférieure à la moitié de la variabilité totale a permis une bonne discrimination des échantillons en fonction de leur état de contamination. Le pourcentage de variabilité (R^2Y) expliqué par le modèle est de 98,3% et 99,6% pour la PLS-DA 1 et 2 respectivement. Le pouvoir prédictif (Q^2Y) du modèle est de 68,8% et 78,7% pour les PLS-DA 1 et 2 respectivement. Ces données montrent que le modèle est excellent d'un point de vue descriptif et satisfaisant pour expliquer et prédire l'appartenance d'un échantillon à un groupe donné.

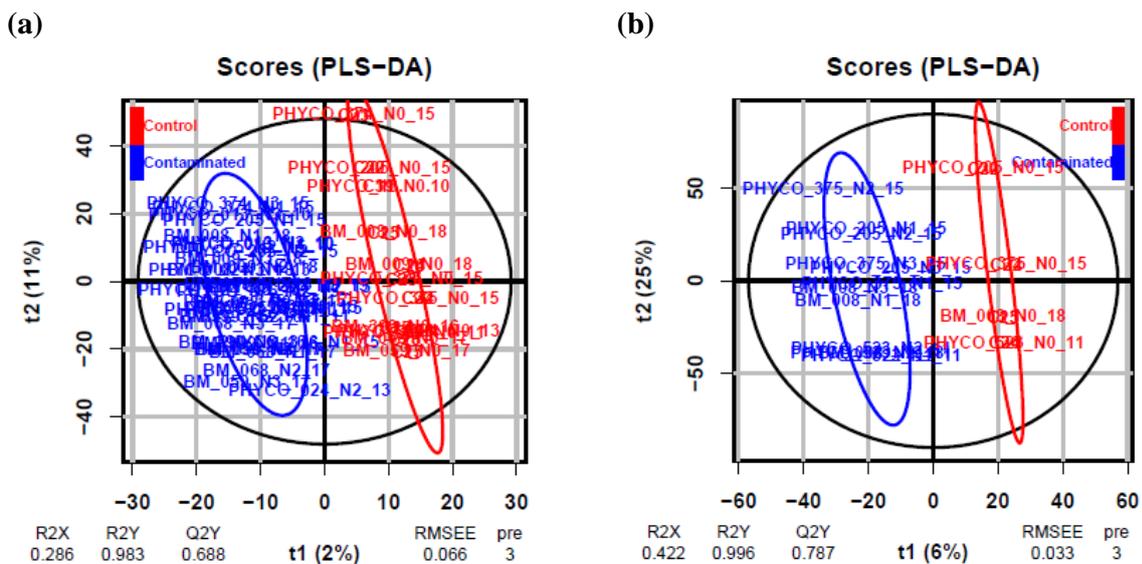


Figure 44. PLS-DA 1 (a) sur l'ensemble des échantillons (moules, huîtres, CSJ) et PLS-DA 2 (b) sur les échantillons de la matrice moule

Le nombre d'échantillons analysés étant nettement inférieur au nombre de variables exploitées, il est important de vérifier que la discrimination observée n'est pas liée à des variables non significatives et par conséquent à un facteur chance.

Des tests de permutations des étiquettes de données ($k=1000$) ont alors été effectués pour évaluer la significativité des critères diagnostiques R^2Y et Q^2Y . Le test de permutation consiste en la permutation aléatoire des étiquettes de données Y des échantillons de façon à obtenir un

nouveau jeu de données. À chaque permutation des valeurs de Y, un nouveau modèle statistique est construit auquel sont associés de nouveaux $pR2Y$ et $pQ2Y$.

Si le modèle obtenu est aussi robuste statistiquement qu'avec les données réellement observées, alors la validité du modèle peut être remise en question. On obtient ainsi des valeurs de p (ou p-value), $pR2Y$ et $pQ2$, qui sont des indicateurs de performances du modèle. Les valeurs obtenues de $pR2Y$ inférieure à 0.01 et $pQ2$ inférieure à 0.05 traduisent la performance du modèle construit et confirme sa validité (**Figure 45**).

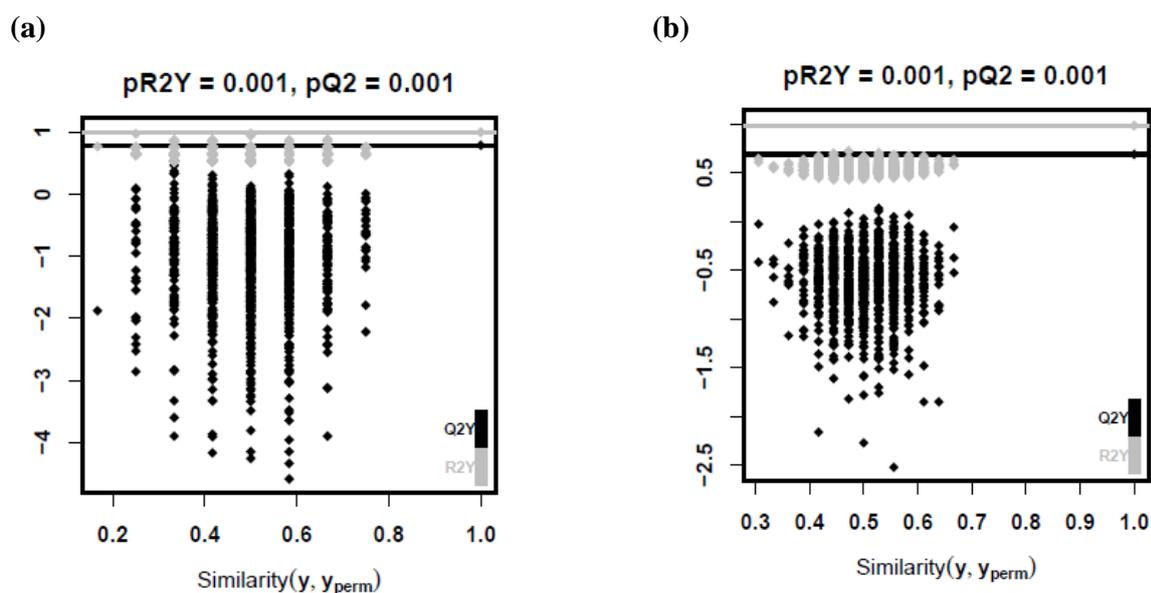


Figure 45. Représentation de la validation du modèle PLS-DA par des tests de permutations ($k=1000$) : (a) PLS-DA 1 (moules, huîtres et CSJ ensemble) et (b) PLS-DA 2 (moules)

Les variables issues de la PLS-DA, ont été classées par ordre croissant selon le critère VIP qui prend en compte l'importance de chaque variable dans la création de la composante et dans la prédiction de Y (variables significatives). Nous avons dans un premier temps sélectionné par défaut les variables pour lesquelles le critère VIP était supérieur à 2 ; 428 variables ont été retenues parmi les 7949 initialement générées après l'analyse LC-HRMS. Dans cette liste, nous avons observé que 12 variables se détachaient du reste avec un VIP supérieur à 5. Ces variables sont définies par le test statistique comme étant les plus significatives dans la discrimination entre les deux groupes « dopés » et « contrôle ».

Les résultats des analyses univariées (t-test et Wilcoxon) pour une p-value < 0,05 ont permis d'identifier les mêmes 12 variables issues de la PLS-DA confirmant ainsi la significativité de

ces variables dans la discrimination des deux groupes. Ces variables ont alors été retenues pour l'annotation (**Tableau 33**).

Dans la liste des 12 variables retenues, certaines correspondent simplement à différents isotopes d'un même ion (M844T534, M509T303, M693T339) et d'autres à un même ion détecté avec un décalage négligeable dans la précision en masse ou en temps de rétention mais identifiés comme variables différentes par le logiciel. Il ne s'agit finalement que de 5 ions responsables des différences observées. Pour chacune des variables nous avons recherché les chromatogrammes (XIC) correspondants et représenté leur présence dans les différents groupes d'échantillons par des boxplots (**Figure 46**). Ces représentations montrent bien que les cinq ions identifiés comme significatifs pour expliquer les différences entre les deux groupes d'échantillons (dopés et contrôle) ne sont présents que dans les échantillons dopés. Quatre des cinq ions jugés pertinents par le test statistique ont été identifiés comme correspondant aux quatre toxines avec lesquelles nous avons dopé nos échantillons (GYM, PnTX-A, SPX1, AZA1). L'identité du 5^{ème} ion (m/z 227,1752, Tr 3,33 min) reste inconnue. Le fait qu'il soit seulement présent dans les échantillons dopés est probablement due à une contamination apportée lors des expériences de dopage. Quoi qu'il en soit, le traitement des données a permis de mettre en évidence un ion que nous ne nous attendions pas à retrouver, prouvant ainsi l'efficacité de l'approche utilisée via XCMS.

Ces résultats confirment le pouvoir discriminant du modèle créé qui a permis d'attribuer les différences observées entre les groupes à la présence des toxines qui ont été ajoutées aux échantillons testés.

Tableau 33. Extraction de la matrice des données issues de l'analyse PLS-DA
Variables classées par ordre croissant selon le critère VIP

variableMetadata	m/z	Tr	npeaks	Contaminated	Control	QCpool	class_PLSDA_VIP
M758T966_1	757,5371	16,1048	14	7	5	1	2,001
M229T92	229,1196	1,5364	19	9	5	4	2,002
M703T819	703,2102	13,6556	32	18	6	8	2,938
M185T545	185,1167	9,0751	37	24	7	6	2,940
M743T620_2	742,5597	10,3311	29	18	2	4	2,945
M895T701	894,7663	11,6766	27	9	7	5	2,959
M391T824	391,3019	13,7399	8	1	5	0	2,961
M677T600	677,4987	9,9952	10	3	5	0	2,982
M340T267_2	340,2400	4,4450	19	15	1	3	2,983
M280T553_2	280,1814	9,2240	33	23	4	6	2,985
[...]							
M819T620	818,5633	10,3383	20	8	5	4	2,986
M795T604	794,6179	10,0617	28	16	2	3	3,002
M794T701_1	793,5906	11,6881	20	11	0	6	3,002
M762T904	761,5882	15,0649	20	9	0	5	3,007
M724T819	724,1397	13,6554	10	4	1	5	3,130
M762T921	761,5715	15,3578	21	15	1	2	3,141
M774T621	773,5294	10,3462	20	5	7	4	3,153
M804T620_2	803,5304	10,3334	17	5	7	3	3,156
M784T966_4	783,5755	16,0976	19	12	2	5	3,169
M735T547	734,5763	9,1163	20	8	8	2	3,178
[...]							
M536T867_1	536,1477	14,4467	16	7	7	2	3,246
M432T418_2	432,2386	6,9690	36	24	3	9	3,254
M785T965_2	784,5899	16,0758	28	19	2	3	3,262
M702T819_2	702,2388	13,6517	25	16	4	5	3,271
M789T967	788,5618	16,1124	21	13	1	7	3,283
M769T965_2	768,6060	16,0805	29	17	4	3	3,305
M721T740	720,5730	12,3331	8	3	5	0	3,310
M447T724	447,3550	12,0625	28	21	4	2	3,312
M769T965_1	768,5590	16,0818	25	13	3	6	3,320
M735T773	734,5383	12,8894	17	7	7	3	3,324
M844T534_1	843,5015	8,9071	20	15	0	5	5,289
M844T534_2	843,5114	8,9079	17	10	1	6	5,621
M227T199	227,1752	3,3239	58	35	12	10	5,795
M508T303_1	508,3185	5,0479	19	16	0	3	5,859
M843T534	842,5026	8,9070	21	16	0	5	5,957
M509T303	509,3436	5,0452	31	24	0	7	6,852
M508T303_2	508,3427	5,0479	33	26	0	7	6,879
M712T318	712,4394	5,3005	27	20	0	7	6,913
M693T339	693,4511	5,6460	23	19	0	4	6,976
M692T339_2	692,4512	5,6487	28	23	0	5	7,158
M692T339_3	692,4571	5,6536	10	5	0	5	7,220
M692T339_1	692,4454	5,6489	9	4	0	5	7,223

Les variables significatives avec un score VIP > 5 (cf. colonne « class_PLSDA_VIP ») sont surlignées en orange

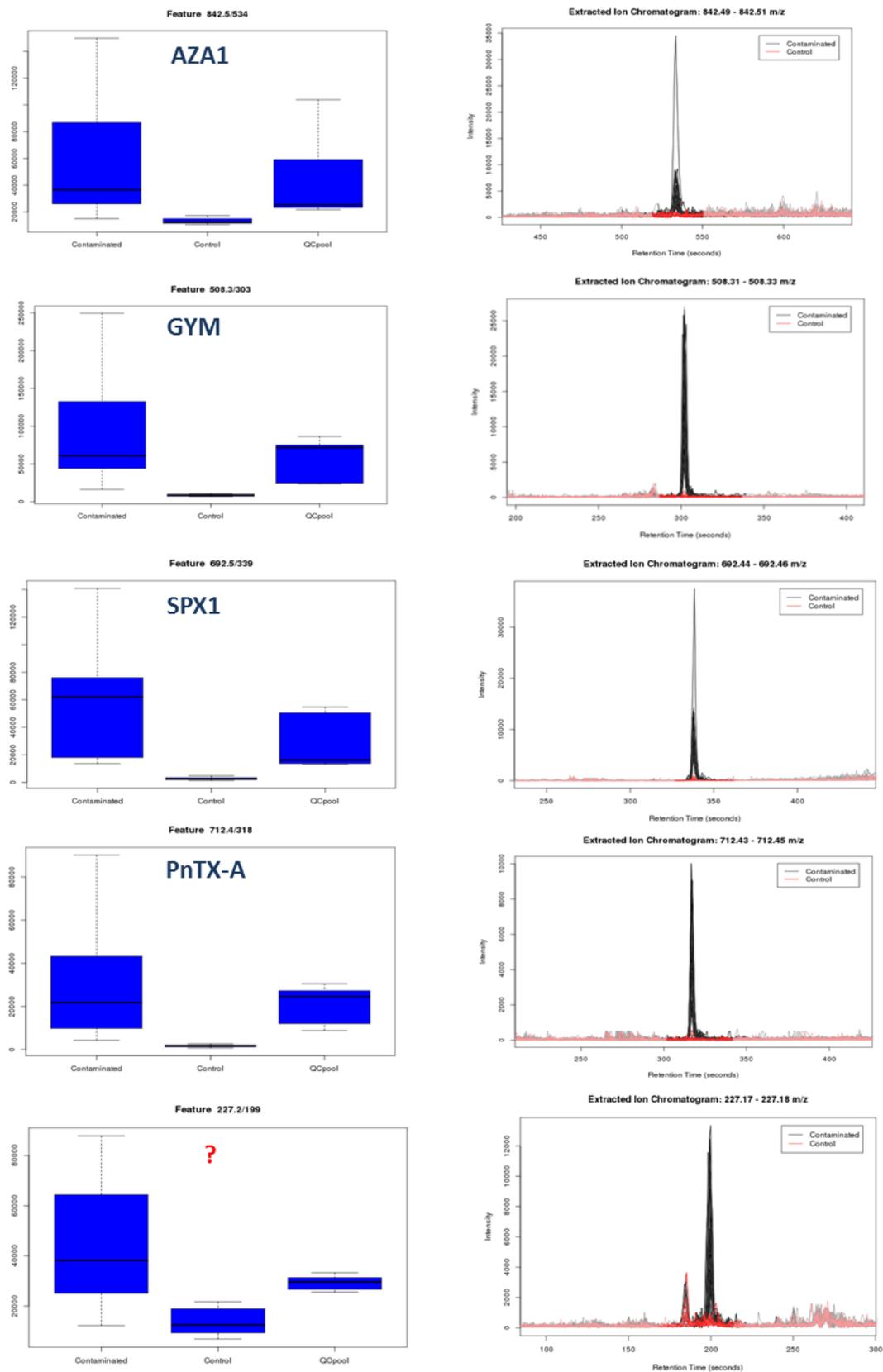


Figure 46. Boxplot (de gauche à droite : contaminé, contrôle et QC pool) et EIC des cinq ions déterminés par la PLS-DA comme étant les plus significatifs dans la discrimination des échantillons dopés et contrôles

II.3. Conclusions

Le logiciel XC/MS dispose d'un algorithme très puissant qui a permis une extraction assez exhaustive de pics à partir des empreintes acquises par LC-HRMS des différents échantillons étudiés. Les étapes de prétraitement des données a permis d'éliminer 16% des pics extraits initialement. Les données ont ensuite été analysées utilisant deux approches statistiques univariées et multivariées afin de détecter les toxines supplémentées dans les échantillons et traitées comme des composés « inconnus ».

Les résultats des analyses statistiques multivariées non supervisées de type ACP ont montré que ce test statistique n'était pas le plus adapté dans le cadre de notre étude car ne permettant pas de mettre en évidence une discrimination entre les échantillons «contaminés » et les échantillons de « contrôles ». Les répartitions observées étaient principalement reliées aux composantes matricielles.

Le test supervisé PLS-DA a démontré un pouvoir discriminant plus intéressant. Le modèle créé a permis de discriminer les échantillons en deux groupes en fonction de leur état de contamination (supplémentés vs blancs). L'étude des données utilisant le score VIP a permis d'attribuer les différences significatives entre les deux groupe essentiellement à la présence/absence de toxines. En effet, les ions relatifs aux toxines étudiées ont pu être identifiés comme étant les variables les plus pertinentes du modèle qui a été validé par les tests de permutation.

Les résultats obtenus par PLS-DA ont été confirmés par les tests statistiques univariés (t-test et Wilcoxon) qui ont également permis d'identifier nos toxines sans ambiguïté dans les échantillons supplémentés.

Cette approche s'est montrée très efficace pour la détermination des toxines à l'aveugle et ceux quel que soit la matrice étudiée.

D. Investigation de cas de TIAC liés à la consommation de violets du genre *Microcosmus*

I. Introduction

Les violets, également appelés bijoux, figures de mer ou encore patates de mer, sont des ascidies comestibles présentes en Méditerranée notamment (**Figure 47**). Ces organismes marins se nourrissent en filtrant l'eau de mer et sont très prisés pour leur goût fortement iodé. Si tous les violets du genre *Microcosmus* sont comestibles, l'espèce *M. sabatieri* est la plus commercialisée.

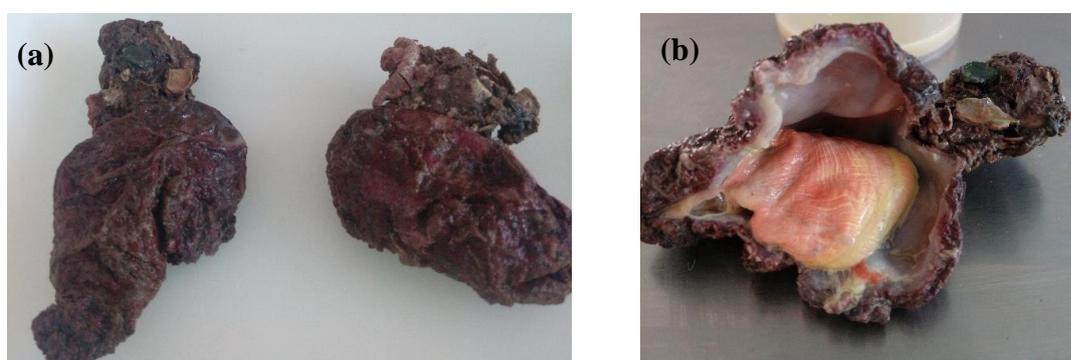


Figure 47 : Violets du genre *Microcosmus* (a) entiers et (b) ouverts longitudinalement

Entre janvier 2011 et mars 2014, 10 cas de toxi-infections alimentaires (collectives ou non ; TIAC/TIA), liés à la consommation de violets ont été recensés par le centre antipoison et de toxicovigilance de Marseille (CAPTVM) et l'Agence régionale de santé (ARS). Au total 15 personnes (7 femmes et 8 hommes) âgées de 23 à 80 ans résidant dans les régions Provence-Alpes-Côte d'Azur et Occitanie ont été intoxiquées (**Tableau 34**). Les symptômes présentés par les patients concernés sont apparus entre 10 min et 1h30 après ingestion des violets ; ils étaient d'ordre neurologique avec comme manifestations principales des troubles de la vue (diplopie), une ataxie et des vertiges. L'ensemble des symptômes rapportés est présenté dans la **Figure 48**. Début 2018, une femme de 56 ans a contacté le CAPTVM pour des troubles neurologiques d'apparition brutale (1 h) de type diplopie et vertiges survenus suite à la consommation de violets. Dans tous les cas rapportés, les personnes intoxiquées se sont remises rapidement (généralement en moins de 24 h).

Le(s) composé(s) à l'origine de ces intoxications n'a(ont) toujours pas été identifié(s) et dans le cadre de cette partie, nous avons souhaité appliquer le workflow développé à l'investigation de cas de TIAC survenus en 2014 et 2018, en analysant les échantillons concernés.

Tableau 34 : Cas de TIA en région Provence-Alpes-Côte d'Azur et en Occitanie entre janvier 2011 et mars 2014, associés à la consommation de violets

	Date, Lieu	Nombre d'intoxiqués	Age
Cas n°1	Janvier 2011, Aude (11)	2 hommes	30 et 52 ans
Cas n°2	Décembre 2011, Bouches du Rhône (13)	2 femmes	70 et 78 ans
Cas n°3	Janvier 2012, Bouches du Rhône (13)	1 femme	55 ans
Cas n°4	Mars 2012, Bouches du Rhône (13)	2 hommes	33 et 40 ans
Cas n°5	Aout 2012, Hérault (34)	1 femme	42 ans
Cas n°6	Décembre 2012, Bouches du Rhône (13)	1 homme	60 ans
Cas n°7	Octobre 2013, Bouches du Rhône (13)	1 homme	80 ans
Cas n°8	Novembre 2013, Bouches du Rhône (13)	1 femme et 1 homme	23 et 29 ans
Cas n°9	Février 2014, Bouches du Rhône (13)	1 femme et 1 homme	33 et 57 ans
Cas n°10	Mars 2014, Bouches du Rhône (13)	1 femme	51 ans

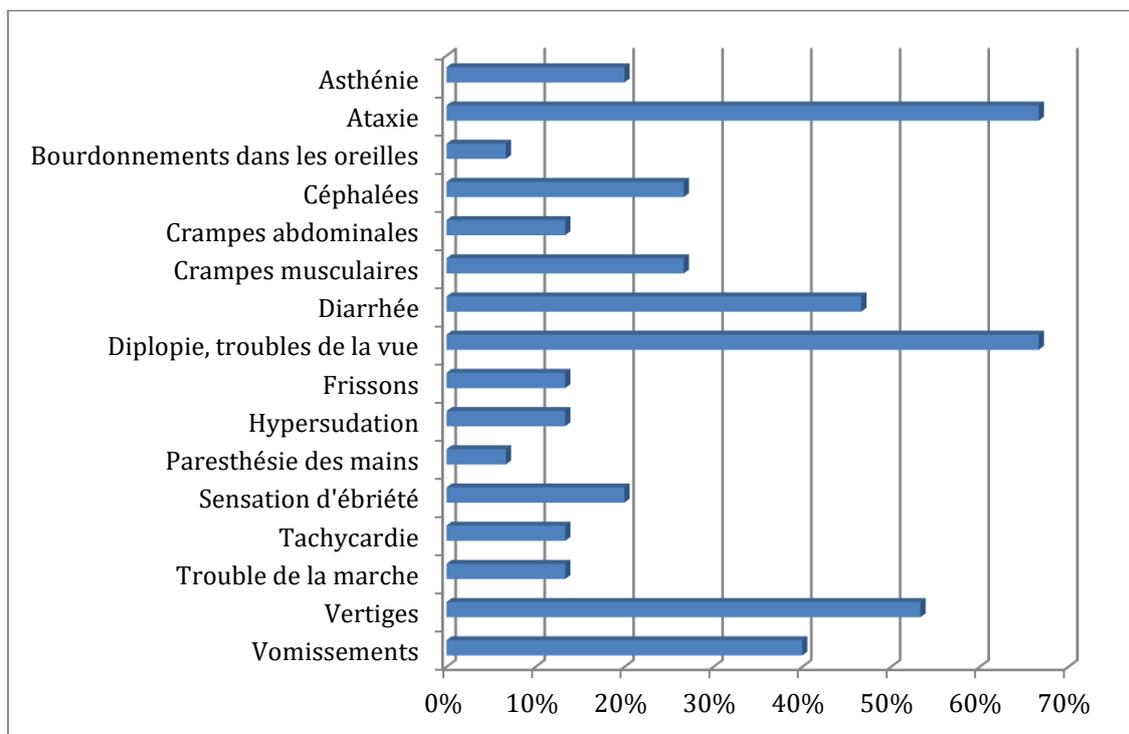


Figure 48 : Symptômes associés aux TIA liées à la consommation de violets du genre *Microcosmus*. Cas recensés entre 2011 et 2014 par le CAPTVM et l'ARS.

II. Matériels et Méthodes

II.1. Produits chimiques et réactifs

Les produits et réactifs chimiques utilisés pour les expériences décrites dans cette partie sont les mêmes que ceux présentés précédemment dans ce chapitre (I.1).

II.2. Echantillons de violets du genre *Microcosmus*

Nous avons pu récupérer des échantillons de 2014 et 2018, impliqués dans des cas de TIAC. Il s'agit de violets originaires de Croatie (14PHYCO096 et 14PHYCO097) et de France (18BM141), respectivement. Les échantillons de Croatie nous ont été fournis sous la forme d'homogénats de violets alors que l'échantillon français a été récupéré au domicile de la patiente de 56 ans ayant été intoxiquée début 2018. Cet échantillon est constitué de cinq violets qui ont été broyés et extraits individuellement ; de même, les exsudats récupérés à l'ouverture des violets ont été conservés et extraits au même titre que la chair des ascidies.

Afin de pouvoir disposer d'un échantillon de référence non contaminé, nous avons utilisé un échantillon originaire de l'étang de Thau (14PHYCO484), acheté dans le commerce en décembre 2014 et constitué de six violets. Les ascidies ont été traitées individuellement, mais sans distinction de chair ni d'exsudat.

Les échantillons de violets (contaminés et référence) ont été extraits selon le mode opératoire standard du LRUE pour les biotoxines marines décrit dans le chapitre II (I.4). La prise d'essai utilisée était de 2 g de chair ou d'exsudat. La méthode d'extraction utilisée est la même quelle que soit la nature de la matrice.

Un contrôle réactif a été également préparé selon le même protocole.

II.3. Conditions d'analyse par LC-HRMS

Les analyses ont été menées par LC-HRMS. Les détails de la méthode d'analyse sont décrits dans le chapitre II (I.5). Chacun des échantillons a été injecté en triplicate et pour éviter tout biais lié à la séquence d'injection, cette dernière a été randomisée.

II.4. Acquisition et traitement des données

Les modalités d'acquisition et de traitement des données sont les mêmes que celles décrites précédemment (I.4). La suite logicielle Sciex a été utilisée pour les approches de suspect screening et d'analyse sans a priori. De même, les données ont été analysées par le biais de tests univariés (t-test) et multivariés (ACP-DA).

III. Résultats et discussion

Les échantillons de violets, contaminés ou non ont été analysés selon l'approche suspect screening. En mode négatif très peu de feux tricolores ont été allumés dans la table contenant les 821 toxines. Certaines saxitoxines comme les dcCTX1 à 4 ou encore la *Gymnodinium catenatum* toxine 6 (GC6) ont été détectées dans quelques échantillons mais il s'agit de faux positifs car ces toxines ne sont pas retenues dans les conditions d'analyses utilisées, en raison de leur caractère très hydrophile. Si les spectres MS2 avaient été acquis durant l'analyse, il aurait été possible de le confirmer par comparaison avec le spectre *in silico* de ces toxines.

En mode positif, davantage de toxines potentielles ont été détectées dans les échantillons de violets contaminés ou non. Il s'agit notamment de cyanotoxines hydrophiles (anatoxine-a (ATX-a), homo-anatoxine-a (hATX-a)), de GC3a et de portimine. La présence d'ATX-a a été confirmée dans un seul des triplicats d'injection de l'échantillon 14 PHYCO 097 par comparaison avec le spectre de fragmentation *in silico* (8 fragments confirmés sur 8). Ce résultat unique remet en question le fait qu'il s'agisse effectivement d'ATX-a, d'autant qu'il pourrait s'agir de phénylalanine, un acide aminé qui est un composé isobare de l'ATX-a présentant des similitudes au niveau du spectre de fragmentation avec cette dernière (Furey et al., 2005). La présence d'hATX-a et de portimine a également été écartée après comparaison des spectres empiriques et *in silico*.

A l'issue de l'étape de suspect screening, aucune toxine n'a été identifiée dans les échantillons de violets originaires de France ou de Croatie et responsables de TIAC. Nous avons ensuite procédé à une analyse sans a priori.

La **Figure 49** présente les résultats de l'analyse multivariée (ACP-DA) réalisée sur les échantillons de violets à partir des données générées en ESI+. On note que les échantillons contaminés sont séparés de l'échantillon non contaminé par la composante principale D1 qui explique à elle seule 34,1% de la variabilité observée. On distingue quatre clusters correspondant respectivement aux violets non contaminés (14PHYCO484) et contaminés, originaires de France (18 BM 141) et de Croatie (14 PHYCO 096 et 097). L'Analyse de la chair et de l'exsudat des violets contaminés originaires de France montre une différence en matière de composition puisqu'on distingue deux clusters différents séparés par la composante principale D2. En revanche, les échantillons

originaires de Croatie sont assez proches du point de vue de leur composition comme en atteste leur proximité au niveau de la **Figure 49**. On note également que les exsudats des violets français (18 BM 141) sont plus proches des violets croates qu'ils ne le sont de la chair des violets dont ils sont issus. Ceci illustre l'importance du choix de la référence lorsque l'on souhaite procéder à des analyses sans a priori car la variabilité matricielle est telle que deux matrices de même nature (ex. moules) peuvent être plus distantes l'une de l'autre que deux matrices différentes (ex. moule et huître). Si l'on souhaite être en mesure d'identifier les composés responsables d'une TIAC, il est donc nécessaire d'avoir une matrice non contaminée de composition similaire. Or dans les cas de TIAC, nous ne disposons pas toujours de matrice témoin et sommes obligés de faire avec ce que l'on trouve. C'est ainsi qu'on en arrive à utiliser des violets de l'étang de Thau comme référence pour l'analyse de violets de Croatie. Une solution pour pallier ce problème de référence serait de constituer des bibliothèques matricielles permettant d'appréhender la différence qu'il peut y avoir du fait de paramètres tels que la saisonnalité, la localisation, le genre et l'espèce des organismes considérés, leur âge etc. C'est l'approche choisie par le LRUE fruits et végétaux qui a constitué des bibliothèques pour différentes matrices entrant dans son périmètre d'action, appartenant à huit groupes définis par la DG SANTE (Anonymous, 2018).

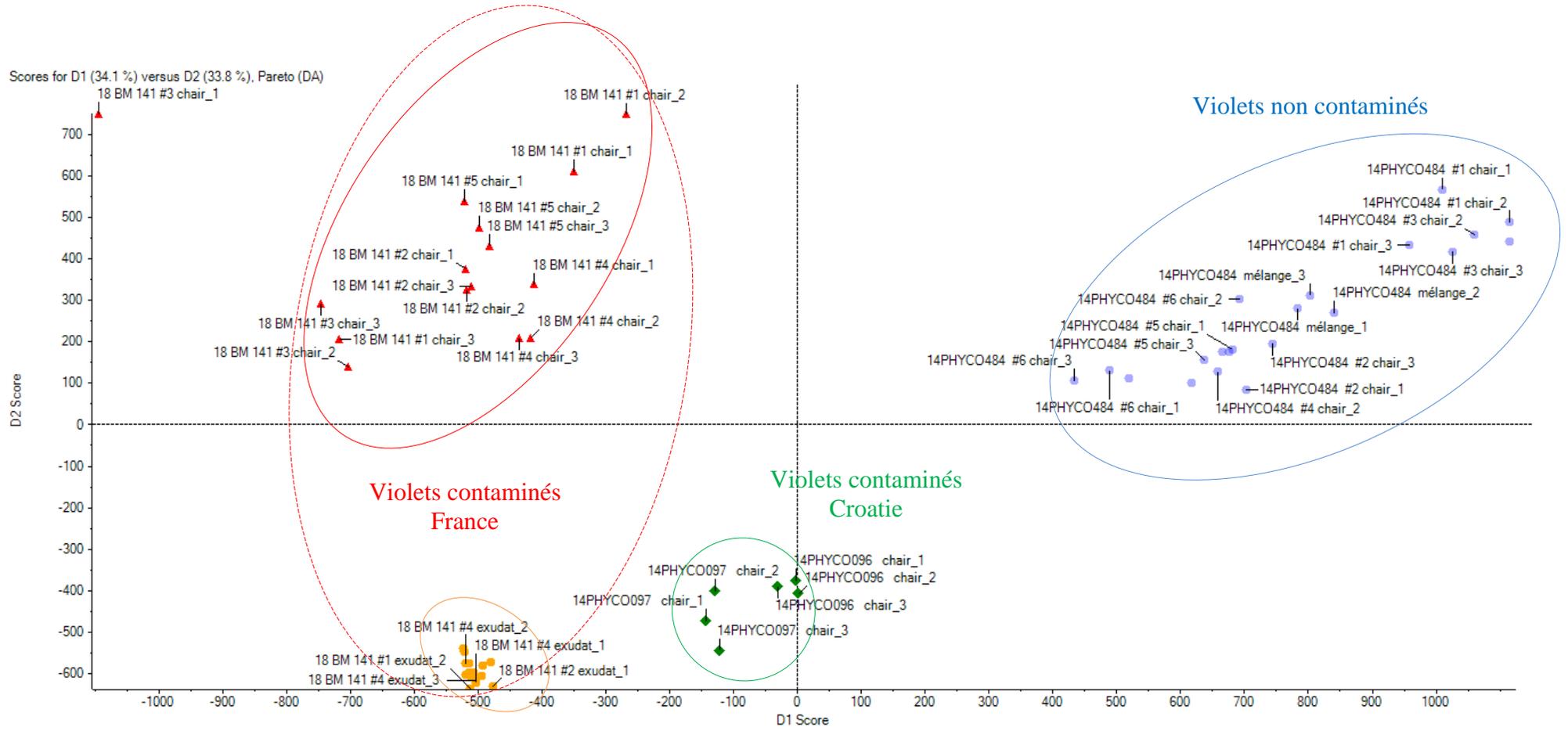


Figure 49 : Scores plot obtenu après l'analyse ACP-DA des données générées en ESI+ pour les échantillons de violets.

Pour identifier les signaux à l'origine des différences de composition entre violets contaminés et non contaminés et espérer identifier les composés responsables des TIAC, nous avons procédé à une analyse univariée entre échantillons contaminé et non contaminé. Le **Tableau 35** présente la stratégie adoptée pour réduire le nombre de signaux. Cette stratégie a permis de réduire de plus de 99% les données initialement générées avec au final entre 7 et 56 signaux potentiels, selon le mode d'ionisation et les violets concernés. A l'issue de cette étape, les signaux restants ont été analysés via l'application Masterview™ afin de vérifier qu'il s'agissait bien de pics chromatographiques, caractéristiques des violets contaminés.

Tableau 35. Résumé des résultats de la stratégie de réduction des données pour l'identification manuelle des signaux, après des étapes de préfiltrage et de comparaison par paires en utilisant le t-test avec comme facteurs discriminants la p-value et le fold-change.

		Violets Croatie		Violets France	
		ESI+	ESI-	ESI+	ESI-
1. Etapes de préfiltrage	nombre total de signaux extraits	9894	9922	9894	9922
	1,2 min < RT < 10 min	5690	4289	5690	4289
2. Résultats du t-test	P-value < 0,05	1565	1489	2113	2485
	Fold change > 10	40	108	210	446
	Signaux ayant une intensité > 100	11	7	56	19

Ainsi, 14 ions ont été identifiés dans les violets originaires de France et 9 dans ceux de Croatie (**Tableau 36**). On note qu'aucun de ces ions n'est commun aux deux origines. Il peut y avoir plusieurs explications à cela, la première étant que bien que la symptomatologie associée aux TIAC à violets soit similaire, les composés impliqués peuvent être différents, comme le sont les analogues d'une famille de toxines. Dans ce cas, il pourrait être intéressant d'utiliser l'approche des réseaux moléculaires basée sur la fragmentation des ions car elle permet d'affilier les composés inconnus au groupe dont elles sont le plus proche du point de vue spectral et ainsi d'identifier d'éventuels nouveaux analogues. Une autre explication à l'absence de signaux communs peut venir du fait que lors du traitement des données et en particulier durant les étapes de filtrage, nous soyons passés à côté d'un signal important. Il est également possible que le problème soit bien en amont du traitement et soit lié à l'étape d'extraction qui ne serait pas appropriée. Ne sachant par définition pas ce que l'on cherche lors d'analyses non ciblées, il est difficile de savoir si on est passé à côté de quelque chose.

Des propositions de formules brutes ont été faites automatiquement par l'application Masterview™ mais ne correspondent pas nécessairement à des composés réels. Il nous faut poursuivre le travail d'annotation des signaux identifiés et vérifier leur potentiel toxique, sur lignées cellulaires par exemple afin de vérifier

si le(s) composé(s) responsable(s) des TIAC figurent bien dans la liste des signaux identifiés. Dans le cas d'investigations comme celle que nous avons entreprise, il pourrait être intéressant de combiner l'analyse dirigée par l'effet et l'analyse non ciblée afin d'identifier les fractions chromatographiques toxiques sur lignées cellulaires et rechercher dedans les composés responsables de cette cytotoxicité.

Tableau 36. Liste des signaux retenus et formules brutes proposées par l'application Masterview™. Les ions issus des violets de France sont en bleu et ceux des violets de Croatie sont en noir. Les ions en gras ne sont présents que dans les violets contaminés

Ionisation	Masse (Da)	TR (min)	Formules proposées par Masterview™	
ESI+	252,0243	4,6	C11H8O7	
	268,0393	4,7	C6H6N8Na2O2 C6H13NaO10 C16H4N4O	
	277,2398	6,5	C18H31NO	
	281,2435	9,7	C14H32N3NaO	
	287,9846	5,1	C7H9N2NaO5S2	
	299,2541	9,7	C12H29N9	
	316,2801	9,7	C12H32N10	
	356,0486	4,9	C15H10N4Na2O4	
	399,3651	4,9	C23H49N3S	
	413,3801	5,0	C24H51N3S	
	426,3911	5,6	C20H46N10 C25H52N2Na2	
	456,2623	9,9	C22H45N2NaS3 C24H42Na2O3S C26H36N2O5 C27H32N6O C27H40N2S2	
	562,3444	8,2	C23H46N8O8 C26H47N6NaO6 C28H52Na2O8	
	585,3778	8,5	C26H51N9O4S C30H55N3O6S C32H48N7NaO2 C34H47N7O2	
	ESI -	129,0431	1,2	C5H7NO3
		197,9183	2,6	C4HNa3OS2
		324,9060	8,2	C7H7NNa4S4
343,9488		6,3	C7H12N2Na4S4 C9H9Na5O3S2 C9HN4Na5O4 C10H3N2Na7O2 C13H7Na3O3S2	
358,8644		8,9	C8HN3Na4S4 C8H2NNaO6S4	
410,9662		6,6	C8H12NNa7O6S C10H10N7NaO2S4	

Ionisation	Masse (Da)	TR (min)	Formules proposées par Masterview™
			C10H16NNa3O4S4
			C12H12NNa7O5S2
			C13H4N5Na3O5S
			C14H7NNa2O9S
	443,3059	9,5	C24H45NO4S
	503,2471	6,6	C16H38N9Na5S
			C22H33N9O3S
			C23H43NNa6S
	629,9846	9,9	C17H16N2Na2O19S
			C21H18Na8O9S
			C16H23Na9O9S2
			C16H14N2Na2O22
			C17H29Na5O6S6
			C18H21N4NaO10S5
			C18H33Na5OS8

IV. Conclusions

Le workflow développé dans le cadre des travaux de thèse a été utilisé pour tenter d'élucider les TIAC associées à la consommation de violets du genre *Microcosmus*. L'approche suspect screening a permis d'écartier l'ensemble des 821 toxines de la liste des composés potentiellement impliqués. L'approche sans a priori a permis d'identifier une liste de signaux absents de l'échantillon de référence (violets de l'étang de Thau) ou alors présents en quantité bien moindre. Néanmoins, nous touchons là aux limites de l'exercice avec la suite logicielle Sciex. Il serait intéressant d'utiliser XCMS pour confirmer les résultats obtenus en matière d'identification et éventuellement aller plus loin. De même, la combinaison de l'analyse dirigée par l'effet et de l'analyse non ciblée pourrait s'avérer intéressante et aurait l'avantage d'avoir comme indicateur une mesure de la toxicité cellulaire confirmant le caractère toxique des composés identifiés.