III.4.8.

Application de l'approche non ciblée pour l'analyse des composés émergents ou inconnus

IV.1. Principe général

L'approche non-ciblée est principalement employée en métabolomique et a été adaptée pour l'analyse de petites molécules dans divers domaines (pharmaceutique, sécurité alimentaire, analyse environnementale, médicale, etc.). Il s'agit d'une approche multidisciplinaire mettant en jeu des outils d'analyse chimique et chimiométriques ; plus précisément, la préparation des échantillons, la séparation chromatographique et le traitement des données (**Figure 17**).

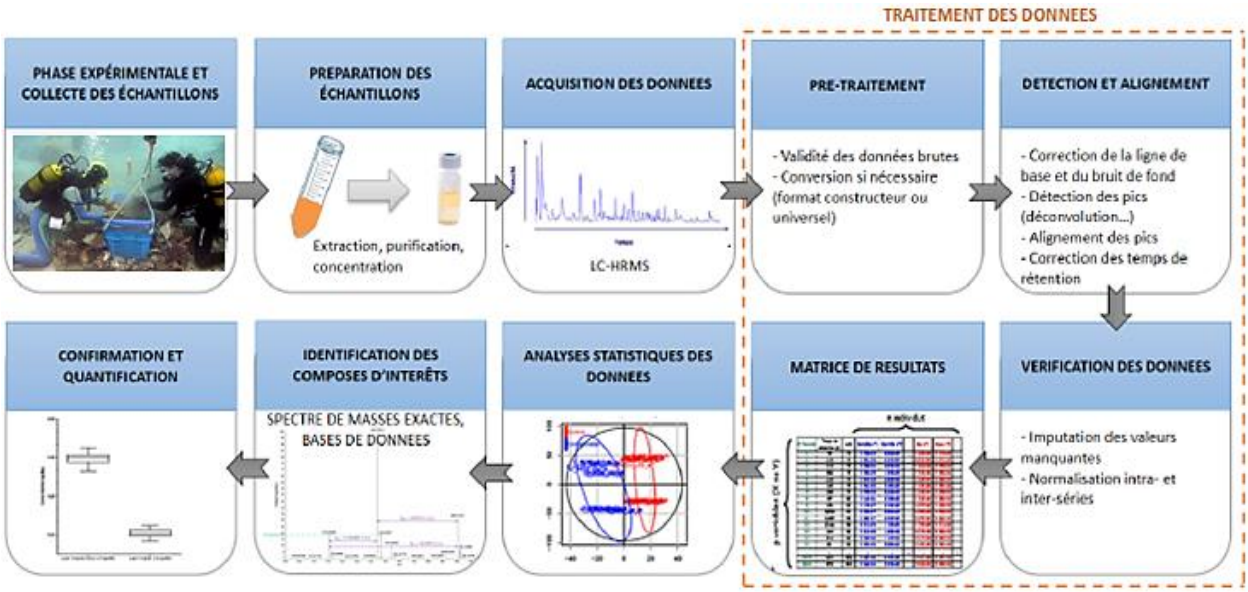

Figure 17. Les différentes étapes d'une analyse non-ciblée par spectrométrie de masse à haute résolution (adapté de la thèse de K. Slimani)

Elles reposent principalement sur l'utilisation des analyseurs à temps de vol (TOF) ou des analyseurs Orbitrap qui ont gagné en popularité en raison de leur grande capacité à fournir des données de qualité exceptionnelles et des informations plus complètes concernant la masse moléculaire exacte, la composition élémentaire et la structure moléculaire détaillée d'un composé donné.

Les données sont acquises en mode full scan ce qui permet la recherche de plusieurs centaines de molécules à la fois, connues et inconnues, dans différentes matrices complexes (animales, végétales, biologiques). L'acquisition d'une empreinte chimique globale de l'échantillon offre la possibilité de réaliser des fouilles de données successives et illimitées en nombre et dans le temps (analyse rétrospective) permettant ainsi l'identification de nouveaux contaminants préoccupant. L'association des analyseurs HRMS avec les quadripôles permet d'enregistrer des spectres de fragmentation (MS^2) indispensables pour l'élucidation structurale des composés détectés. Il convient de noter que les méthodes d'analyse LC-HRMS génèrent une quantité importante de données qui doivent être minutieusement traitées afin d'extraire l'information pertinente. L'identification précise de molécules inconnues est longue et complexe, elle requière de multiples étapes de filtration et de traitement de données

impliquant l'utilisation de différents outils chimiométriques (tests statistiques univariés et multivariés).

Les stratégies pour les analyses LC-HRMS sont très différentes selon les groupes de recherche. Pour identifier des composés inconnus, chaque laboratoire dispose de ses propres « workflow », plusieurs études ont été présentées dans la littérature notamment dans le domaine de l'analyse environnementale (Agüera et al., 2013; Ferrer and Thurman, 2003; Krauss et al., 2010; Schymanski et al., 2015) mais aussi en sécurité alimentaire (Castro-Puyana et al., 2017; Le Boucher et al., 2015; Tengstrand et al., 2013). Malgré la spécificité des protocoles décrits, un schéma global a pu être tiré de ces diverses études englobant le screening ciblé, le screening suspect et le screening non ciblé. Cette stratégie générique décrite par Krauss et al. (2010) (**Figure 18**) peut-être optimisée et adaptée en fonction des instruments utilisés et la finalité de l'étude.

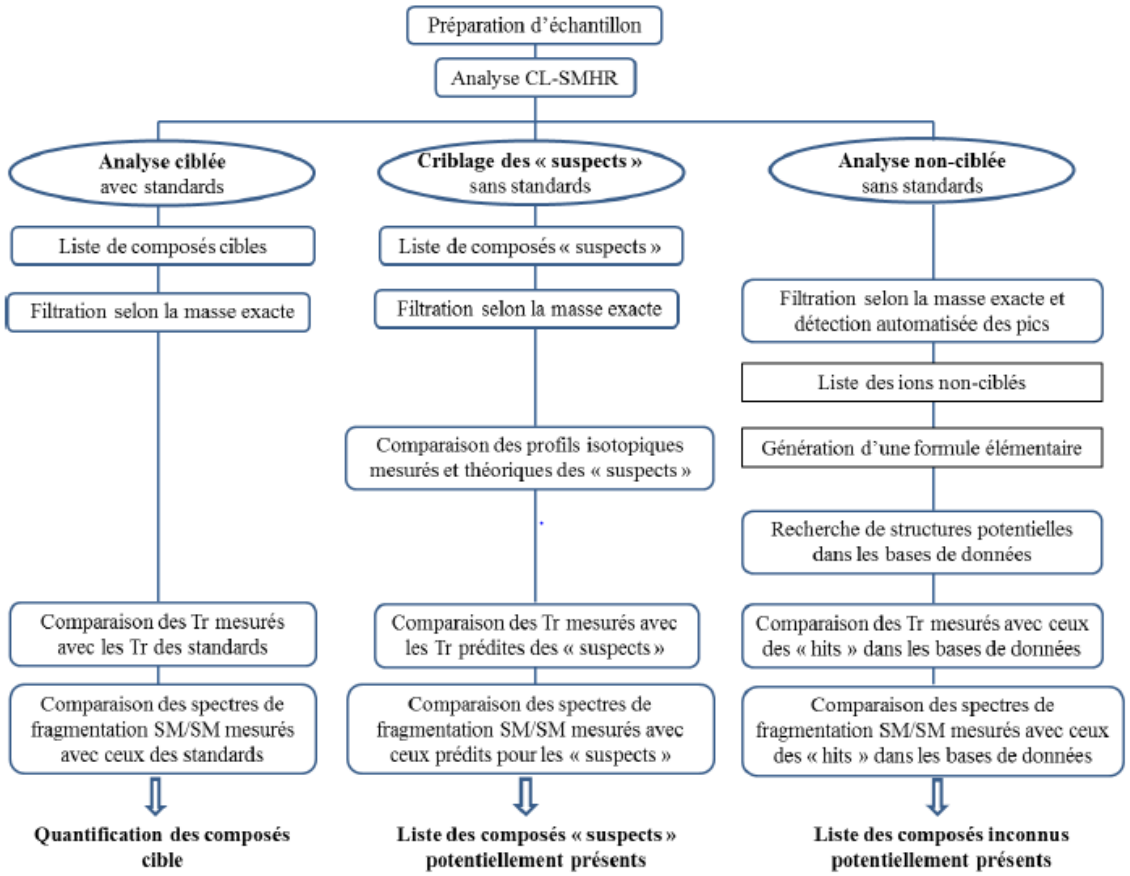

Figure 18. Stratégies d'analyses des données HRMS (adapté de Krauss et al. (2010))

Le screening ciblé fait référence au screening basé sur l'utilisation d'étalons de référence mesurés en interne, ce qui permet de confirmer par comparaison du temps de rétention, la concordance de masse exacte, la correspondance du modèle isotopique et enfin la concordance spectrale MS/MS. Les analyses ciblées ont pour objectif final la quantification des composés d'intérêt. Cette approche est la plus communément utilisée dans le cadre d'analyses de routine.

Le screening suspect ne désigne pas nécessairement des composés nouveaux mais peut être utilisé pour des composés qui n'étaient pas initialement recherchés lors de l'analyse. Il peut s'agir de composés connus comme pouvant potentiellement être présents dans les échantillons (analogues, produits de dégradation et/ou de métabolisation par exemple). Contrairement à l'analyse ciblée quantitative, le criblage de composés « suspects » n'est pas tributaire de l'utilisation de standards pour l'identification et la confirmation. Dans le domaine des biotoxines marines, une minorité de standards est disponible à l'heure actuelle. Cependant, des informations spécifiques à chaque molécule existent, telles que la formule moléculaire et la formule développée. Ces informations peuvent être utilisées lors des processus de confirmation de l'identité des « suspects ». La formule moléculaire permet dans un premier temps de calculer le rapport *m/z* exact de l'ion moléculaire recherché. Ce dernier est par la suite extrait du spectre haute résolution sous forme de chromatogramme. En ionisation electrospray, les ions majoritairement formés (à quelques exceptions près pour certaines molécules formant des adduits) sont le $[M+H]^+$ ou le $[M-H]^-$, ce qui facilite les tentatives d'identification (Krauss et al., 2010). Les fragments ou ions isotopiques caractéristiques sont par la suite évalués en comparant les spectres expérimentaux MS ou MS/MS, ou le fragment majoritaire avec ceux reportés dans la littérature (Chemspider, PubChem, MassBank, Metlin *etc*…) (Diaz et al., 2012) ou en se rapportant à la théorie. Idéalement, la disponibilité du standard permettra *apostériori* une identification formelle (grâce aux profils isotopiques, temps de rétention et spectres de fragmentation). Cette approche de criblage des suspects peut aussi se faire grâce aux bases de données « maison » (constituées par chaque laboratoire sur leur(s) matériel(s) et logiciel(s) disponibles) contenant

une liste de composés spécifiques de chaque domaine d'étude (pesticides, composés pharmaceutiques, toxines marines, perturbateurs endocriniens *etc*…).

Le screening non ciblé concerne tous les signaux non attribués par le criblage ciblé ou suspect. Cette approche est utilisée pour identifier des composés d'intérêt sur lesquels on ne dispose d'aucune information préalable. Plusieurs milliers de signaux peuvent être concernés. Chaque signal est caractérisé au minimum par un temps de rétention, une masse exacte (et son massif isotopique) et une intensité.

L'analyse non-ciblée commence généralement par la filtration des signaux en fonction des masses exactes, suivie d'étapes de traitement des données pour éliminer le bruit, les blancs ou les artefacts. Ensuite, une déconvolution automatisée permet d'extraire les pics de tous les composés possibles. Les pics de masse des différents ions d'un composé sont souvent fusionnés en un seul élément (par exemple, [M+H]+, [M+Na]+, [M+NH4]+). L'ensemble de données qui en résulte est ensuite analysé à l'aide de méthodes statistiques afin d'évaluer les caractéristiques les plus pertinentes et sélectionner les ions d'intérêt en comparant différents échantillons et blancs. Pour les ions sélectionnés, la composition élémentaire est calculée et les formules moléculaires les plus probables sont évaluées en faisant correspondre le modèle isotopique. Pour l'identification, les formules moléculaires sont recherchées dans les bases de données ou les bibliothèques MS/MS. Le temps de rétention est souvent utilisé comme critère supplémentaire pour réduire le nombre de « hits ». L'identification est obtenue lorsque la fragmentation MS^2 et le temps de rétention du composé inconnu s'adaptent au spectre de la bibliothèque et au temps de rétention d'un composé de référence.

Si aucune correspondance dans une base de données MS/MS ou une bibliothèque n'est disponible, des recherches dans les grandes bases de données chimiques telles que PubChem et ChemSpider sont effectuées. Cette recherche se traduit généralement par plusieurs centaines à plusieurs milliers d'occurrences pour une structure possible. La fragmentation peut être utilisée comme critère pour sélectionner les résultats les plus probables. Étant donné que les bases de données chimiques ne contiennent généralement pas de données MS^2, la fragmentation in silico (prédictive) peut être utilisée et les fragments doivent ensuite être

comparés aux fragments MS mesurés. Il en résulte un certain nombre de structures composées proposées. Toutefois, l'identification sans équivoque nécessite encore des informations complémentaires provenant d'autres méthodes d'analyse, comme l'analyse RMN.

IV.2. Traitement des données

Pour une véritable analyse sans a priori, il faudrait, en principe, une identification de tous les éléments détectés dans un échantillon. Compte tenu de la richesse des informations générées l'inspection visuelle de l'ensemble du chromatogramme et le traitement manuel des données spectrales ne sont pas suffisants pour détecter et identifier tous les composés. Le recours au traitement de ces données afin d'extraire l'ensemble de l'information utile est donc nécessaire. Le traitement des données s'effectue principalement par l'utilisation de logiciels commerciaux (MarkerLynx, Waters ; MarkerView, Sciex ; MassHunter, Agilent, etc) ou open source (XCMS (Smith et al., 2006)), MetAlign (Tikunov et al., 2005)) ou MZmine (Katajamaa and Orešič, 2005)). D'autres logiciels « maison » sont développés par certains laboratoires. Ils ont pour but d'éliminer le bruit de fond, de détecter les pics par la mise en œuvre d'algorithmes, et d'aligner les pics entre les différents échantillons analysés. Les données sont converties sous des formats numériques exploitables souvent sous forme matricielle compatible avec les logiciels d'analyse statistiques.

Il est à noter que chaque logiciel dispose d'un algorithme différent assurant les différentes étapes de détection automatisées des signaux. L'avantage des logiciels open source c'est que les détails de leur algorithme sont accessibles à l'utilisateur, ce qui lui offre une grande marge de manœuvre pour optimiser chacun des paramètres à chaque étape du traitement. Quant aux logiciels commerciaux (ou constructeurs) souvent simples d'utilisation, peuvent être qualifié de « boite noire » car les algorithmes sont confidentiels et l'utilisateur n'a pas forcément accès à tous les paramètres de retraitement.

Les étapes de retraitement des données sont logiciel dépendant, nous présenterons ici le principe de quelques étapes clés en prenant comme exemple le logiciel XCMS, l'un des plus utilisés en métabolomique et pour l'analyse HRMS de petites molécules :

Le prétraitement des informations chromatographiques (temps de rétention) et spectrales (rapport m/z et intensité) repose sur la vérification visuelle de la validité qualitative des données d'acquisition brutes LC-HRMS, suivie de leur conversion du format de fichier constructeur vers un format universel exploitables par le logiciel.

La détection automatique des pics qui consiste en la sélection des signaux analytiques pertinents présents dans les données brutes acquises correspondant à l'ensemble des composés détectés dans la totalité des échantillons. Au cours de cette étape, des filtres de bruit tels que le ratio signal sur bruit (S/N) ou la comparaison au niveau de la ligne de base du chromatogramme permettent la suppression d'artefacts électroniques.

L'alignement des empreintes inter-échantillons consiste à associer les signaux identiques provenant d'un même ion détecté dans les différents échantillons, par création de groupes d'ions selon le Tr et le rapport m/z, malgré de légers décalages possibles.

La correction des temps de rétention est une étape complémentaire à la précédente, elle permet de recaler les temps de rétentions et corriger les dérives éventuelles pouvant survenir au cours de la séquence d'acquisition.

La complétion des données manquantes permet de rattraper des pics non détectés en raison du bruit de fond trop important ou de la forme du pic chromatographique atypique mais aussi de remplacer la donnée manquante par l'intégration du bruit de fond local. Cette étape est très importante pour l'utilisation de statistiques notamment les tests de significativité.

La normalisation permet de maîtriser et corriger les fluctuations de rendement d'ionisation dues notamment aux effets de suppression afin qu'elles n'entravent pas l'interprétation des données. Pour être en mesure de corriger cette dérive analytique,

la solution la plus commune est l'utilisation d'échantillons de contrôle qualité (*Quality control*, QC). Ces QCs sont généralement composés d'un mélange de l'ensemble des échantillons analysés et sont injectés régulièrement tout au long de la séquence d'analyse. Le biais analytique pourra ainsi être modélisé et corrigé par régression.

Tout au long du processus de traitement des données, différents paramètres peuvent être optimisés par l'utilisateur afin d'aboutir à une matrice de données à la sortie du logiciel. Ces données contiennent les informations sur les variables détectées. Chaque variable est caractérisée par un couple rapport *m/z* et temps de rétention et présente l'information sur l'aire du pic chromatographique dans chaque échantillon. Cette matrice de données à deux dimensions est à son tour traitée par des algorithmes ou des programmes d'exploration de données. Ces aspects seront abordés plus en détail dans le chapitre III dédié à l'analyse non ciblée.

IV.3. Analyses statistiques des données

L'analyse statistique des données permet de synthétiser et de structurer l'information contenue dans les données mesurées en spectrométrie de masse. L'analyse d'empreintes chimiques globales peut être réalisée selon deux approches statistiques univariées et multivariées (Gorrochategui et al., 2016).

IV.3.1. Analyses univariées

Les analyses statistiques univariées sont des techniques classiquement utilisées en biologie. Elles permettent d'analyser une à une les variables explicatives en fonction d'une métadonnée (concentration, ordre de passe, origine géographique…) sans tenir compte des interactions existant entre les variables. Le type de test utilisé dépend de la nature de la variable et de la nature de la métadonnée à étudier : des calculs de corrélation sont réalisés si les deux sont quantitatives alors que des analyses de variance sont utilisées si l'une des deux

est qualitative. La distribution de probabilité de la variable conditionne aussi le choix du test. Une distribution selon la loi normale autorise l'utilisation de tests paramétriques (test de Student ou corrélation de Spearman), alors que les variables dont la distribution au sein de la population étudiée ne suit pas une loi statistique sont étudiées à l'aide de tests non-paramétriques tels que le test de Wilcoxon ou la corrélation de Pearson.

IV.3.2. Analyses multivariées

Les analyses statistiques multivariées sont utilisées afin d'étudier ou de décrire un ensemble de données. Elles permettent de synthétiser et de visualiser rapidement une grande quantité d'informations, ceci en projetant les données initiales dans un espace de dimensions réduites ce qui permet une visualisation aisée des données. Les analyses statistiques multivariées peuvent être divisées en deux groupes :

Les analyses descriptives ou non supervisées qui ne nécessitent pas d'information « a priori » sur la nature des échantillons. Leur but est de décrire des données et de visualiser la répartition des échantillons. L'analyse en composante principale (ACP) est la plus représentative de ce groupe (Hotelling, 1933; Wold et al., 1987).

Les analyses statistiques explicatives ou supervisées qui visent à expliquer une réponse (variable qualitative Y). Parmi ces méthodes, les régressions PLS (Partial Least Square ;(Joreskog, 1982)) l'avantage d'être insensibles aux multicolinéarités c'est-à-dire à la présence de variables très corrélées. Ceci est particulièrement intéressant pour les études menées en spectrométrie de masse où les rapports m/z des fragments issus d'un même ion sont très corrélés.