Développement de nouvelles applications en analyse d'authenticité des aliments par couplage chromatographie liquide – spectrométrie de masse haute résolution

Katy Dinis

THESE DE DOCTORAT

NNT : 2022UPASB042

université PARIS-SACLAY

Développement de nouvelles applications en analyse d'authenticité des aliments par couplage chromatographie liquide – spectrométrie de masse haute résolution

Development of new applications in food authenticity analysis by liquid chromatography coupled to high resolution mass spectrometry

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°581, Agriculture, alimentation, biologie, environnement, santé (ABIES)
Spécialité de doctorat : Chimie
Graduate School : Biosphera. Référent : AgroParisTech

Thèse préparée dans l'UMR **SayFood** (Université Paris-Saclay, INRAE, AgroParisTech), sous la direction de **Valérie CAMEL**, Professeure et le co-encadrement de **Lucie Tsamba**, Docteure

Thèse soutenue à Paris-Saclay, le 11 juillet 2022, par

Katy DINIS

Composition du Jury

Evelyne VIGNEAU Professeure, ONIRIS	Présidente
Laurent DEBRAUWER Ingénieur de recherche, INRAE (centre Occitanie-Toulouse)	Rapporteur & Examinateur
Gaud DERVILLY-PINEL Directrice scientifique (HDR), Ministère de l'agriculture	Rapporteur & Examinatrice
Carlos AFONSO Professeur, Université Rouen Normandie	Examinateur
Valérie CAMEL Professeure, AgroParisTech (Université Paris-Saclay)	Directrice de thèse
Lucie TSAMBA Docteure, Eurofins Analytics France	Co-encadrante

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Développement de nouvelles applications en analyse d'authenticité des aliments par couplage chromatographie liquide – spectrométrie de masse haute résolution

Mots clés : agroalimentaire, analyse d'authenticité, chimiométrie, LC-HRMS, métabolomique

Résumé : Le contrôle d'authenticité en agroalimentaire permet de garantir la conformité des aliments et ainsi de protéger les consommateurs des fraudes. Bien que les méthodes ciblées actuellement utilisées soient sensibles et spécifiques, elles peuvent faillir à détecter des fraudes sophistiquées. De plus, la complexité des aliments rend leur analyse difficile. Dans le but d'améliorer les contrôles d'authenticité, il semble essentiel de se tourner vers des approches dites non ciblées permettant d'obtenir une empreinte globale de l'échantillon et ainsi d'évaluer rapidement la présence d'éventuelles fraudes.

Dans ce travail, une méthode d'analyse ciblée a été développée pour l'authentification du jus de citron jaune. Puis, une méthodologie d'analyse non ciblée a été développée et testée sur des jus de pommes et des carottes pour contrôler leur authenticité, notamment le mode de production (biologique ou conventionnel).

Title : Development of new applications in food authenticity analysis by liquid chromatography coupled to high resolution mass spectrometry

Keywords : agri-food, authenticity analysis, chemometrics, LC-HRMS, metabolomics

Abstract : Authenticity testing in the food industry is used to ensure food compliance and thus protect consumers from frauds. Although the targeted methods currently used are sensitive and specific, they can fail to detect sophisticated frauds. In addition, the complexity of foods makes them difficult to analyze. In order to improve authenticity controls, it seems essential to turn to untargeted approaches that provide a global fingerprint of the sample in order to quickly assess the presence of possible frauds.

In this work, a targeted analytical method was developed for the authentication of lemon juice. Then, an untargeted analysis method was developed and tested on apple juices and carrots to check their authenticity, including the production method (organic or conventional).

REMERCIEMENTS

J’aimerais dans un premier temps remercier vivement Valérie CAMEL, ma directrice de thèse, pour son encadrement tout au long de cette thèse, malgré la distance physique et le contexte sanitaire lié au COVID-19. Son soutien, ses conseils, les échanges constructifs et ses encouragements m’ont aidé à mener à bien ces trois années de doctorat.

Je remercie également Freddy THOMAS puis Lucie TSAMBA qui ont supervisé cette thèse. Je les remercie pour leur aide et leur disponibilité pour faire avancer les travaux réalisés au cours de cette thèse.

Je voudrais également remercier les membres de mon comité de thèse, Yann GUITTON et Julien BOCCARD. Les échanges constructifs et bienveillants ainsi que leur aide et leurs conseils m’ont été très bénéfiques au cours de cette thèse.

Je remercie également le Dr Laurent DEBRAUWER et la Dr Gaud DERVILLY-PINEL pour avoir accepté d’être les rapporteurs de cette thèse, ainsi que les examinateurs, le Pr Carlos AFONSO et la Pr Evelyne VIGNEAU.

Ces trois années en entreprise m’ont permis de rencontrer de belles personnes. Je m’excuse par avance pour tous les collègues de travail que je pourrais oublier d’évoquer. Tout d’abord, l’équipe chromatographie et arômes : Stéphanie (à qui j’ai envie de dire un grand merci pour son temps et ses conseils pour le passage en prod des méthodes développées), Amélie, Gwen, Emilie, Hélène, Maud, Isaline, Mélina et Erwan ; ainsi que les stagiaires : Sarah, Léa, Cécile et Lucie (merci d’ailleurs d’avoir compati pendant ma période carottes !). Je n’oublie pas les autres équipes : la prépa, la chimie, le SNIF et les isotopes, avec lesquelles j’échangeais plus rarement. Sans oublier la cellule qualité (Kristell et Delphine).

Je souhaite ensuite remercier l’équipe Projets (Jean-François, Hélène, Béatrice (nos rdv à la Cantine d’Albert vont me manquer) et Delphine) pour tous les échanges et leur bonne humeur ; ainsi que l’équipe ASM (Apolline, Emilie, Ellen, Anna, Séverine, Karine, Maylis, Elias et Elisabeth) pour avoir su mettre une bonne ambiance dans l’open space. Et évidemment je remercie également l’équipe R&D pour tous les moments du quotidien : Freddy, Lucie, Sandrine, Vincent (encore félicitations Docteur !), ainsi que les alternantes Raphaëlle et Mégane (bon courage à vous avec le projet TOFoo). Je remercie également Eric, le BUMA de l’authenticité.

Je remercie également mes anciens camarades de fac et mes amis pour tous les moments partagés ainsi que pour leur soutien.

Enfin, je tiens maintenant à remercier ma famille, notamment mes parents et mon frère, pour leur soutien tout au long de mes études, ainsi que Nicolas pour sa présence au quotidien.

VALORISATION DES TRAVAUX DE THESE

Publications dans des revues à comité de lecture

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel (2022). Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics. *Food Control*, (2022), sous presse. https://doi.org/10.1016/j.foodcont.2022.109098

Katy Dinis, Lucie Tsamba, Eric Jamin, Valérie Camel (2022). Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (*Daucus carota* L.) based on geographical origin and production mode. *Food Chemistry*, soumis

Markus Jungen, Nenad Dragićević, Miriam Rodriguez-Werner, Simone Schmidt, **Katy Dinis**, Lucie Tsamba, Eric Jamin, Thorsten Fiedler, Nadine Fischbach, Valérie Camel, Ralf Schweiggert (2022). A pragmatic authenticity assessment of lemon (*Citrus limon* (L.) Burm. f.) juices by its profile of coumarins, psoralens and polymethoxyflavones. *Food Control*, soumission prévue prochainement

Communication en congrès international

Katy Dinis, Freddy Thomas, Lucie Tsamba, Eric Jamin, Valérie Camel (2020). Fruit juice authenticity using untargeted UHPLC-HRMS combined with chemometrics. *Chimiométrie 2020*, Liège, Belgique. **Communication par affiche**

Communication en congrès national

Katy Dinis, Lucie Tsamba, Freddy Thomas, Eric Jamin, Valérie Camel (2020). Assessment of apple juice authenticity using untargeted UHPLC-HRMS and chemometrics. *1ères Journées Sciencitiques Numériques du RFMF (Réseau Français de Métabolomique et de Fluxomique),* en ligne. **Communication orale**

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

ABREVIATIONS

ACN	Acétonitrile
ACP	Analyse en composantes principales
AIJN	Association européenne des jus de fruits
ANOVA	Analyse de variance (*analysis of variance*)
BD	Base de données
CV	Coefficient de variation
EMA	Adultération économiquement motivée (*economically motived adulteration*)
ESI	Ionisation par électrospray (*electrospray ionization*)
FA	Acide formique (*formic acid*)
FC	Facteur multiplicatif (f*old change*)
GC	Chromatographie en phase gazeuse
HILIC	*Hydrophilic interaction liquid chromatography*
HRMS	Spectrométrie de masse à haute résolution
IFU	Association internationale des jus de fruits et légumes *(international fruit and vegetable juice association)*
IR	Infrarouge
LC / UHPLC	Chromatographie en phase liquide (ultra haute pression)
LOD	Limite de détection
LOESS	Régression locale (*locally estimated scatterplot smoother*)
LOQ	Limite de quantification
LV	Variable latente (*latent variable*)
MeOH	Méthanol
MS	Spectrométrie de masse (*mass spectrometry*)
m/z	Rapport masse sur charge
PC	Composante principale (*principal component*)
PLS-(DA) / OPLS-DA	(Analyse de données) des moindres carrés partiels (*projection to latent structure discriminant analysis*) / *Orthogonal projection to latent structure discriminant analysis*
PQN	Normalisation par quotient probabilistique *(probabilistic quotient normalization)*

QC	Contrôle qualité (*quality control*)
RMN	Résonance magnétique nucléaire
ROI	Région d'intérêt (*region of interest*)
RT	Temps de rétention
S/B	Rapport signal sur bruit
SNIF NMR	Fractionnement isotopique naturel spécifique par RMN (*site-specific natural isotope fractionation nuclear magnetic resonance)*
SVM	Machine à support de vecteur (*support vector machine)*
UE	Union européenne
VIP	Importance de la variable dans les modèles (O)PLS(-DA) (*variable importance in the projection*)
W4M	Plateforme *Workflow4Metabolomics*

GLOSSAIRE

Echantillon authentique : un échantillon non fraudé et dont les métadonnées sont connues avec certitude

"Feature" : combinaison *m/z* – RT caractérisant un pic chromatographique

Fichier brut (ou donnée brute) : fichier issu de l'instrument d'analyse

"Pool" : mélange d'échantillons représentatifs

INTRODUCTION GENERALE

Les analyses de contrôle de l'authenticité des denrées alimentaires sont réalisées au quotidien afin de garantir leur conformité (en lien avec leur étiquetage), de protéger les consommateurs de fraudes éventuelles et de préserver leur confiance. En effet, suite à différents scandales très médiatisés (par exemple la viande de cheval dans des lasagnes de bœuf en 2013 ou la présence de fipronil dans les œufs en 2017), les consommateurs cherchent à avoir plus d'informations concernant les aliments qu'ils consomment. Les contrôles d'authenticité visent ainsi principalement à limiter les pratiques frauduleuses qui peuvent avoir lieu tout au long de la chaîne de production et de fabrication des denrées, et ayant pour but de dégager un profit. Afin de limiter ces pratiques et protéger les consommateurs, différents organismes ont ainsi mis en place des normes et des méthodes officielles de contrôle.

Ces méthodes officielles de contrôle de l'authenticité sont dites « ciblées », visant à détecter et quantifier des composés ou des familles de composés connus et marqueurs d'authenticité. Ces méthodes sont rapides, sensibles et spécifiques pour répondre avec une grande certitude sur l'authenticité d'un échantillon. Néanmoins, elles peuvent faillir à la détection de fraudes sophistiquées puisqu'elles sont orientées sur des composés marqueurs bien spécifiques. Par ailleurs, la complexité et la variabilité des denrées alimentaires rendent leur analyse délicate. En effet, la concentration de certains marqueurs peut fluctuer selon différents facteurs naturels (variété, origine géographique, saison, etc.). Ainsi, dans le but d'améliorer les contrôles d'authenticité, il semble essentiel de se tourner vers d'autres approches de type « non ciblées » permettant d'acquérir une empreinte globale de l'échantillon, en vue à terme d'être mieux outillé pour mettre en évidence la présence d'éventuelles fraudes.

Depuis le début des années 2000, les progrès techniques en termes d'instrumentation, en particulier la spectrométrie de masse à haute résolution, et d'informatique, notamment concernant la statistique et le traitement de données, ont permis la mise en place de ce type d'approche non ciblée. Ces méthodologies dites « omiques » ont dans un premier temps été développées dans le cadre d'études cliniques pour identifier les différences de compositions entre plusieurs groupes d'échantillons. Depuis le début des années 2010, ce

type de méthodologie s'est répandu à diverses problématiques, dont notamment le contrôle d'authenticité des denrées alimentaires où différentes études montrent des résultats prometteurs.

C'est dans cette optique de développement d'une approche non ciblée que s'inscrivent les travaux de cette thèse, menés dans le cadre d'une thèse CIFRE entre le laboratoire Eurofins Analytics France (EAF) localisé à Nantes et l'unité mixte de recherche SayFood (Université Paris-Saclay, INRAE, AgroParisTech) basée en Ile-de-France. Les travaux rapportés dans ce manuscrit se concentrent sur le développement d'une méthodologie d'analyse non ciblée par chromatographie en phase liquide couplée à la spectrométrie de masse à haute résolution (LC-HRMS) combinée à des outils chimiométriques pour contrôler l'authenticité des denrées alimentaires, en particulier les jus de fruits.

Les jus de fruits ont été choisis comme principale matrice d'intérêt dans ce travail du fait de leur grande variabilité (variété, origine géographique, procédé de fabrication, mode de production, saison). Cette matrice nécessite de ce fait une analyse fine de l'authenticité. En outre, les jus de fruits sont connus comme faisant partie des matrices agroalimentaires les plus à risque de fraude (en particulier le jus de pommes).

Le premier chapitre de ce manuscrit présente les enjeux analytiques liés aux analyses de contrôle d'authenticité agroalimentaire. Le cas particulier des jus de fruits sera également présenté. Les méthodes d'analyse conventionnelles ainsi que les approches non ciblées seront ensuite présentées.

Le second chapitre présente le développement d'une méthode d'analyse ciblée par LC-HRMS pour l'authentification du citron jaune. Ce développement a eu lieu dans le cadre d'un projet interlaboratoires. Il fait l'objet d'un article en cours de finalisation qui devrait être soumis prochainement à *Food Control*.

Le chapitre 3 présente le développement d'une méthode d'analyse non ciblée par LC-HRMS et son application pour caractériser l'authenticité des jus de pommes selon deux scénarios : discrimination des échantillons selon le mode de fabrication (purs jus ou jus concentrés) ou le mode de production (biologique ou conventionnel). Un premier workflow de traitement

des données a été mis en place en ligne. Ces travaux ont été valorisés par un article accepté dans *Food Control*.

Le chapitre 4 concerne l'authenticité des carottes par analyse non ciblée LC-HRMS. Le changement de matrice a nécessité quelques ajustements dans la méthode non ciblée. En outre, une nouvelle version du workflow de traitement des données a été mise en place en interne au laboratoire. A nouveau, les résultats obtenus ont été valorisés par un article soumis récemment à *Food Chemistry*.

Les résultats marquants de ces travaux de thèse sont ensuite discutés et remis en perspective de quelques références bibliographiques clés en lien avec les matrices étudiées. Certaines limites sont évoquées et des pistes concrètes sont proposées pour poursuivre ce travail.

CHAPITRE 1 ETAT DE L'ART POUR LE CONTROLE D'AUTHENTICITE ALIMENTAIRE

Dans ce premier chapitre, nous allons tout d'abord présenter les différentes notions liées à l'authenticité en agroalimentaire. Puis, les différentes réglementations existantes liées à l'authenticité alimentaire ainsi que l'importance de la mise en place de son contrôle seront également présentées. Les enjeux liés aux nouvelles méthodes analytiques pour le contrôle d'authenticité en alimentaire seront ensuite présentés. Le cas des jus de fruits, principale matrice étudiée dans ce manuscrit, sera enfin détaillé.

Nous allons ensuite discuter des méthodes actuellement utilisées pour le contrôle d'authenticité, aussi appelées méthodes conventionnelles. Nous allons principalement nous focaliser sur deux techniques d'analyse, la LC-HRMS (chromatographie en phase liquide couplée à un spectromètre de masse haute résolution) et la RMN (résonance magnétique nucléaire), et présenter leur principe ainsi que des exemples d'applications.

Enfin, nous allons présenter les nouvelles méthodologies d'analyse pour le contrôle d'authenticité au travers des deux techniques d'analyse précédentes. Les différentes étapes des méthodologies non ciblées pour ces techniques seront détaillées, et en particulier le traitement de ce type de données sera décrit. Nous présenterons en fin de chapitre la méthodologie choisie pour ces travaux de thèse.

1. LE CONTROLE D'AUTHENTICITE EN AGROALIMENTAIRE

1.1. DEFINITIONS

Les questions d'authenticité des denrées alimentaires existent depuis le début de la commercialisation des aliments. Le contrôle d'authenticité des aliments a gagné l'intérêt des consommateurs, notamment après les récents scandales alimentaires fortement médiatisés : mélamine dans le lait en Chine en 2008, viande de cheval dans des lasagnes de bœuf en 2013, et fipronil dans les œufs en 2017.

Les définitions reprises ci-dessous ont récemment été rappelées dans l'article de Spink et al. dans le but d'harmoniser ces notions de plus en plus utilisées par les chercheurs, les consommateurs et les industriels (Spink et al., 2019).

L'**authenticité** d'un aliment consiste à s'assurer de la conformité de celui-ci et à confirmer toute information relative à l'aliment (étiquetage, origine géographique, variété, mode de production, etc.). Il est possible de résumer l'authenticité par cette affirmation simple : la nourriture est ce qu'elle prétend être. Le contrôle d'authenticité de denrées alimentaires permet ainsi de protéger les consommateurs et de préserver leur confiance. Il est d'autant plus important aujourd'hui de garantir l'authenticité des denrées alimentaires aux consommateurs car l'augmentation des échanges internationaux augmente le risque de fraude des aliments.

La **fraude** alimentaire est définie comme « supercherie illégale » délibérée et intentionnelle affectant la qualité d'un aliment pour des raisons économiques. Les principales fraudes alimentaires sont présentées en **Figure 1.1** et concernent :

- L'ajout, la dilution et/ou la substitution de substance par une substance non authentique et moins couteuse,
- La falsification des documents associés à l'aliment ou la modification de son étiquetage, notamment concernant l'origine géographique, la variété, le procédé utilisé ou le mode de production,
- La dissimulation ayant pour but de masquer la faible qualité d'une denrée alimentaire,
- La contrefaçon.

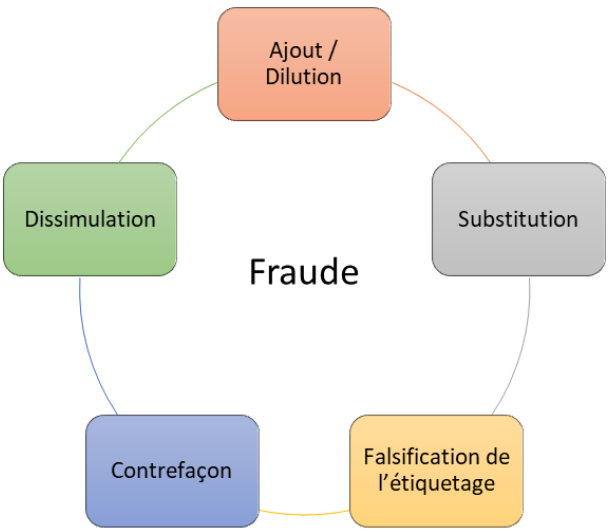

***Figure 1.1** : Les principaux types de fraudes alimentaires*

Il est important de bien distinguer les contrôles d'authenticité des problématiques de sécurité sanitaire des aliments (**Figure 1.2**). Le contrôle d'authenticité a pour but de vérifier l'absence d'adultérant dans les denrées alimentaires et par conséquent l'absence de fraude. Le contrôle de la qualité sanitaire alimentaire vise quant à lui à s'assurer de la conformité de la denrée vis-à-vis de la réglementation pour les éventuels contaminants. Un **adultérant** est une substance ajoutée intentionnellement dans un aliment alors qu'un **contaminant** est une substance présente dans l'aliment résultant de sa production ou fabrication. Le contexte de la sécurité sanitaire n'est pas traité dans ce manuscrit.

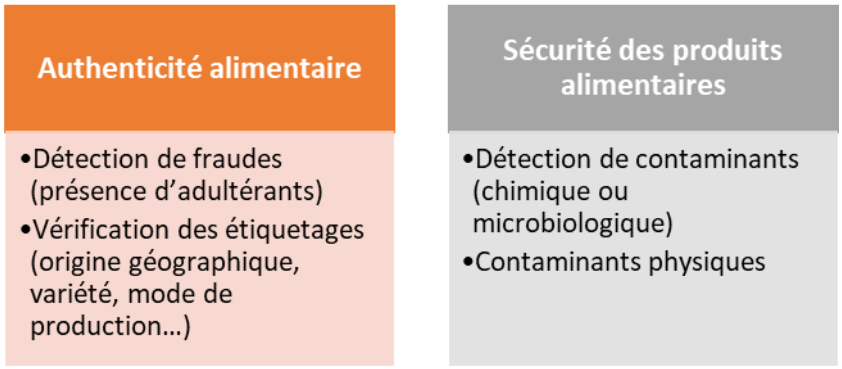

***Figure 1.2** : Principales différences entre l'authenticité et la sécurité sanitaire des aliments*

L'**adultération économiquement motivée** (ou *economically motived adulteration*, EMA) est une fraude alimentaire qui consiste en l'ajout ou la substitution intentionnelle et frauduleuse d'une substance à un aliment pour des raisons économiques. Cette pratique permet d'augmenter la valeur apparente de l'aliment ou d'en réduire son coût de production. Cette manœuvre frauduleuse peut parfois entraîner un risque sanitaire pour le consommateur, comme ce fut le cas lors de la crise du lait adultéré à la mélamine en 2008 en Chine. La mélamine, composé riche en azote, avait été ajoutée dans du lait infantile pour augmenter artificiellement sa teneur en azote et faire croire à un taux élevé de protéines, dans le but d'augmenter sa valeur marchande et ainsi faire du profit. Cela avait causé la mort de 6 bébés et rendu malades près de 300 000 enfants en bas âge car la mélamine s'avère être un composé très toxique (Sharma & Paradakar, 2010 ; Yang et al., 2009).

1.2. LA REGLEMENTATION EXISTANTE

La fraude alimentaire étant bien connue, différentes réglementations existent pour aider à définir les constituants d'un aliment ; elles donnent parfois des indications sur les

concentrations attendues. On peut distinguer différents niveaux de réglementations présentés en **Figure 1.3**.

Il y a tout d'abord au niveau mondial le *Codex Alimentarius* créé en 1963 qui a pour but de protéger la santé des consommateurs, notamment suite au commerce international des aliments toujours plus présent. Celui-ci a permis la création de « normes alimentaires » permettant de garantir aux consommateurs la qualité et la sécurité des denrées alimentaires. Ces normes contiennent également des indications sur les méthodes d'analyse afin de garantir la conformité d'un aliment (FAO/WHO, 2021 ; FAO and WHO, 2019). Ces normes permettent ainsi de faciliter les échanges internationaux des denrées alimentaires.

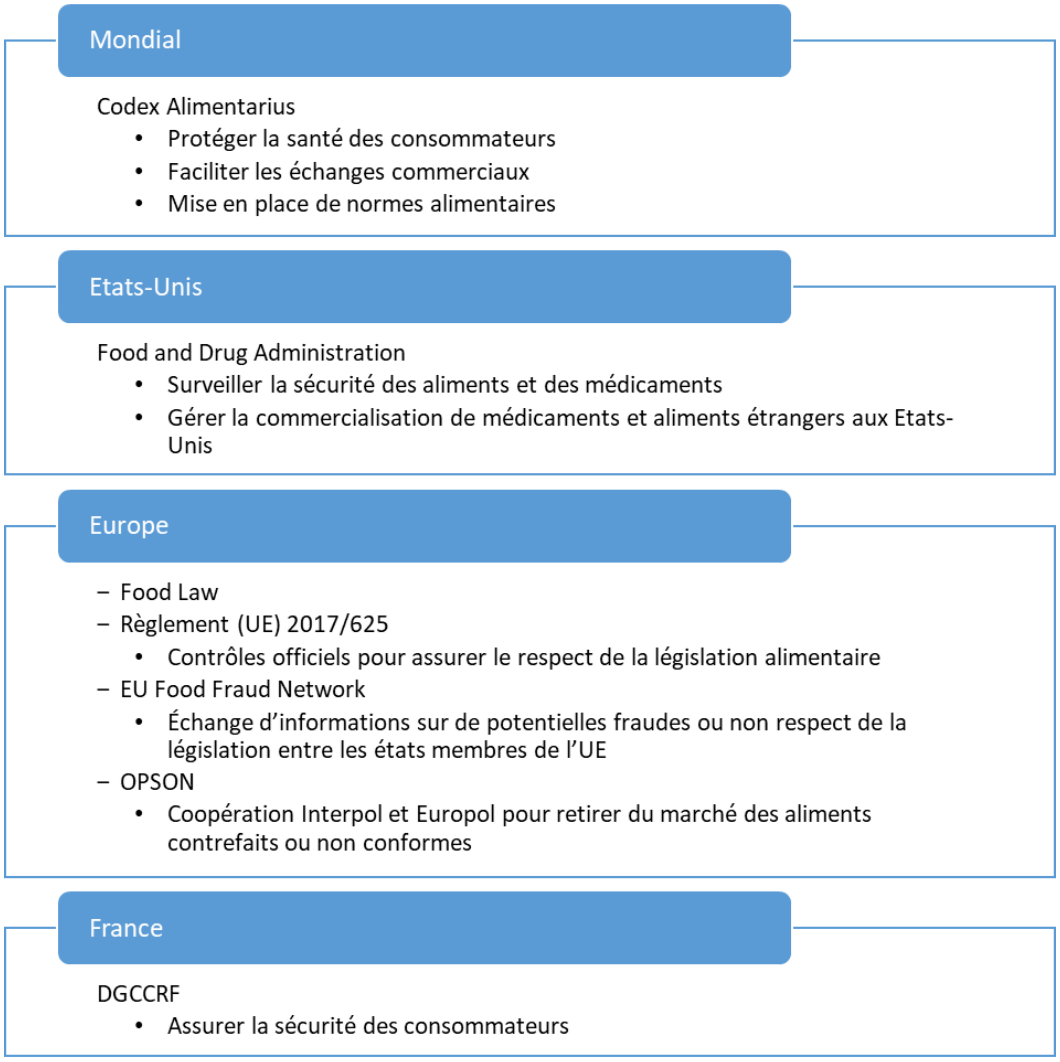

***Figure 1.3** : Présentation des principaux acteurs et réglementations existantes pour le contrôle des fraudes alimentaires*

Il existe également aux Etats-Unis, la FDA (*Food and Drug Administration*, Agence américaine des produits alimentaires et médicamenteux) qui depuis 1848 est chargée de surveiller la sécurité des denrées alimentaires et des médicaments (FDA, 2021). La FDA est notamment en charge de la commercialisation de médicaments ou d'aliments étrangers aux Etats-Unis.

Au niveau européen, il y a tout d'abord eu en 2002 la mise en place de la « *Food Law* » via le règlement CE 178/2002. Cette réglementation établit les principes généraux de la législation alimentaire et concerne tous les acteurs de la filière agroalimentaire depuis la production jusqu'à la commercialisation (« *from farm to fork* » : de la ferme à la fourchette) (Commission Européenne, 2002). Ce règlement est le texte clé du « paquet hygiène » qui est composé de 5 règlements européens et concerne en particulier la sécurité sanitaire des aliments (Ministère de l'agriculture et de l'alimentation, 2020).

On retrouve également au niveau européen le règlement 2017/625 qui concerne les contrôles officiels et les autres activités officielles permettant d'assurer le respect de la législation alimentaire (Règlement (UE) 2017/625, 2021). En 2015, il y a notamment eu la création du réseau européen de lutte contre la fraude alimentaire (*EU Food Fraud Network*) géré par la Commission Européenne. Ce réseau permet aux états membres de l'Union Européenne (UE) d'échanger des informations sur de potentielles fraudes alimentaires ou des non-respects de la législation européenne, et de demander une assistance pour vérifier la conformité de produits suspects (EU Food Fraud Network, 2021). Ce réseau permet ainsi aux pays membres de l'UE ainsi qu'à certains autres pays européens (Suisse, Norvège et Islande) de faire des requêtes concernant des denrées alimentaires suspectes selon 4 critères : (i) la violation de la législation européenne, (ii) la déception du consommateur, (iii) un gain économique direct ou indirect pour l'auteur, et (iv) l'intention de l'auteur. Il existe également l'opération Opson qui est une coopération entre Interpol et Europol avec pour but de retirer du marché tout aliment contrefait ou non conforme (Europol-Interpol, 2020) ; ses opérations ont lieu une fois par an depuis sa création en 2011.

En France, la lutte contre la fraude alimentaire est conduite par la DGCCRF (Direction générale de la concurrence, de la consommation et de la répression des fraudes). Elle a notamment participé à l'opération Opson de 2019 qui s'est principalement intéressée aux produits de la filière biologique.

Il existe également des organismes créés dans le but de donner des recommandations spécifiques selon le type de denrées alimentaires, notamment les aliments à risque élevé de fraudes comme le vin, le miel ou l'huile d'olive. On peut par exemple citer l'IHC (*International Honey Commission*) qui a pour but de contrôler la qualité du miel (IHC, 2021), ou l'OIV (Organisation Internationale de la vigne et du vin) qui vise à harmoniser les pratiques agricoles pour garantir l'authenticité des produits vitivinicoles (OIV, 2021). Ces organisations ont permis le développement et la mise en place de méthodes officielles d'analyse pour ces matrices.

1.3. L'IMPORTANCE DE L'AUTHENTICITE

La fraude alimentaire peut avoir lieu à toutes les étapes de la chaîne de production d'un aliment : de la production primaire jusqu'à sa commercialisation. La fraude est d'autant plus difficile à appréhender que différents intermédiaires sont présents tout au long de la chaîne de production et de fabrication.

Une étude de Moore et ses collègues concernant la fraude alimentaire et l'EMA entre 1980 et 2010 a permis de mettre en évidence les 10 produits ou catégories de produits les plus à risques de fraudes : l'huile d'olive, les produits de la mer, les produits issus de l'agriculture biologique, le lait, les produits céréaliers, le miel et le sirop d'érable, le café et le thé, les épices, le vin et certains jus de fruits (notamment le jus d'orange et le jus de pomme) (Moore et al., 2012). Le rapport annuel 2019 du réseau européen de lutte contre la fraude alimentaire (*Administrative Assistance and Cooperation system – European Food Fraud Network*) présente le top 10 des produits suspectés de fraude (voir **Figure 1.4**) (Commission Européenne, Direction générale de la santé et de la sécurité alimentaire, 2020). Il est intéressant de constater que ce top 10 reste proche de celui établi par Moore et ses collègues en 2012. Cela illustre également le fait que la fraude alimentaire est toujours présente malgré les contrôles.

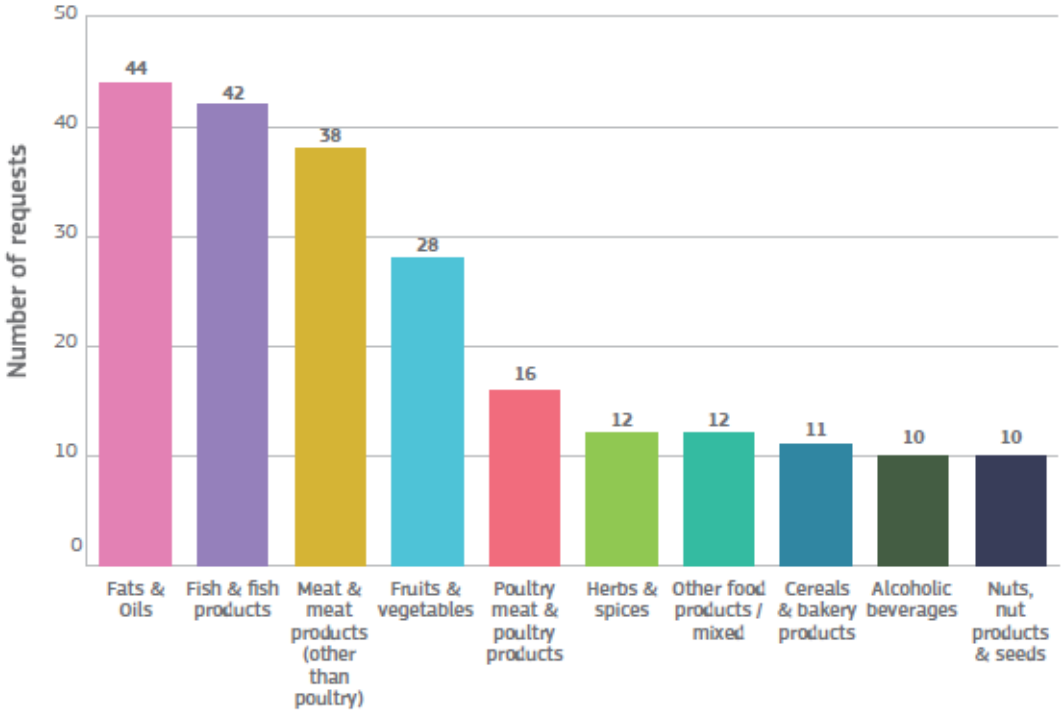

***Figure 1.4** : Le top 10 des produits d'après les requêtes recensées par l'EU Food Fraud Network et l'AAC en 2019. Les jus de fruits sont intégrés dans la catégorie fruits et légumes (Fruits & vegetables).*

Le rapport annuel 2019 cité précédemment indique qu'il y a eu 292 requêtes par les pays membres pour des suspicions de fraudes, un nombre qui a presque doublé en trois ans (cf. 157 requêtes en 2016). Ceci peut s'expliquer grâce à la coopération croissante entre les pays membres de l'UE pour combattre la fraude alimentaire, permettant d'assurer la conformité des denrées alimentaires selon la législation européenne. Cette augmentation montre, qu'au niveau européen, les échanges alimentaires conduisent à des questionnements sur l'authenticité. De ce fait, les efforts se poursuivent pour lutter contre la fraude alimentaire afin de garantir l'authenticité des aliments. Sur les 292 requêtes effectuées par les pays membres en 2019, la Commission Européenne a fait 70 demandes d'enquêtes plus poussées pour vérifier la conformité des produits (Commission Européenne, Direction générale de la santé et de la sécurité alimentaire, 2020).

La fraude et notamment l'EMA ont pour but de diminuer les coûts de production ou d'augmenter la valeur apparente d'un produit pour un agriculteur ou un industriel. Cependant, tous les agriculteurs ou industriels n'utilisent pas ces manœuvres frauduleuses. Ainsi les agriculteurs et les industriels nécessitent également de pouvoir attester de l'authenticité de leurs produits et de leurs bonnes pratiques.

Les consommateurs, quant à eux, s'interrogent de plus en plus sur l'authenticité des aliments qu'ils consomment, notamment après certains scandales alimentaires. Aujourd'hui, c'est notamment avec l'augmentation de la consommation de denrées alimentaires issues de l'agriculture biologique que l'on s'interroge d'autant plus sur l'authenticité (Europol-Interpol, 2020 ; Mihailova et al., 2021). La mise en place de contrôles d'authenticité permet ainsi de rassurer les consommateurs, et de rétablir la confiance entre industriels et consommateurs.

Dans la lutte contre la fraude alimentaire, les laboratoires d'analyses, privés et publics, sont également des acteurs importants. Le contrôle de l'authenticité des aliments existe depuis de nombreuses années : les premiers articles scientifiques datent des années 1960. Depuis, la recherche autour de cette thématique n'a cessé d'être présente et connait même un essor depuis la fin des années 1990 comme observé sur la **Figure 1.5**.

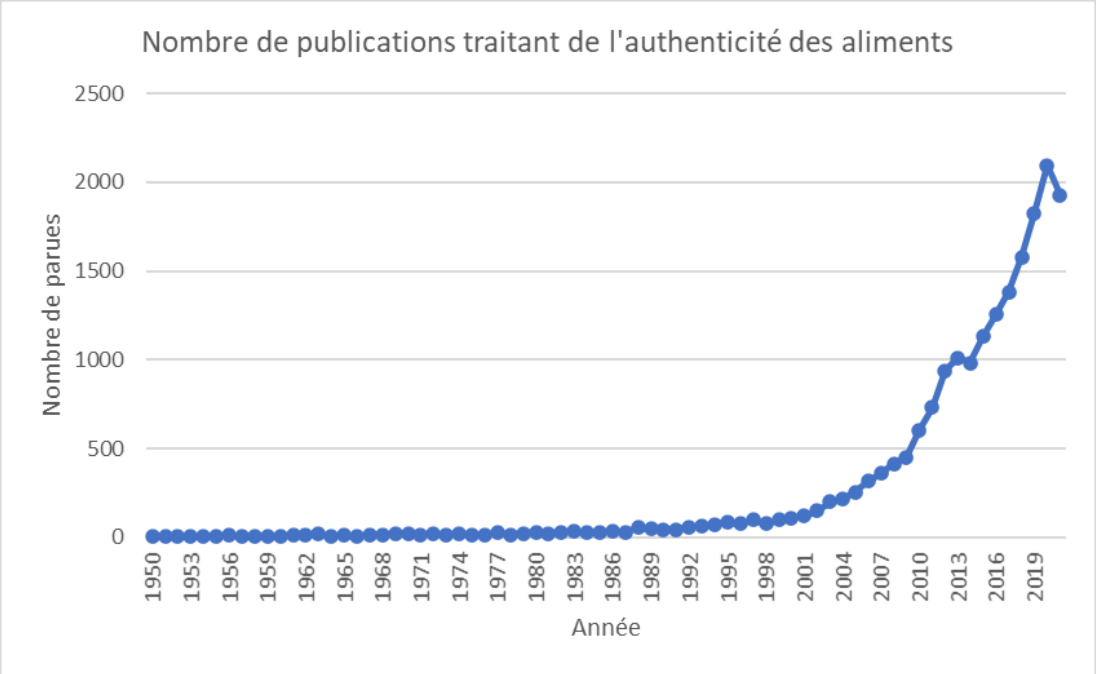

***Figure 1.5** : Nombre de publications parues traitant de l'authenticité des aliments. Base de données utilisée : Science Direct, mot clé : « food authentication »*

1.4. LES ENJEUX ANALYTIQUES

1.4.1. Complexité et variabilité des aliments

Il est souvent dit d'un produit alimentaire, quel qu'il soit, qu'il est complexe. En effet, il est constitué de nombreux constituants chimiques (acides organiques, acides gras, acides

aminés, sucres, vitamines, lipides, glucides, fibres, etc.). Ce grand nombre de constituants peut complexifier l'analyse des denrées alimentaires. De plus, la composition d'un aliment varie en fonction de nombreux facteurs naturels comme l'origine géographique, la variété, le mode de production ou la saison. De ce fait, l'analyse d'authenticité devient d'autant plus complexe.

Si on prend pour exemple le cas des jus de fruits, on observe qu'une grande partie de leur composition est à risque de fraude comme présenté dans la **Figure 1.6** ci-dessous, où tous les composés surlignés en rouge sont susceptibles d'être fraudés, principalement par ajout ou par dilution. Comme la composition d'un jus de fruit varie selon différents facteurs naturels, il peut être difficile de déterminer si une différence de concentration sur ces paramètres à risque de fraude est simplement due à son origine géographique ou sa variété, ou si cela est imputable à une fraude.

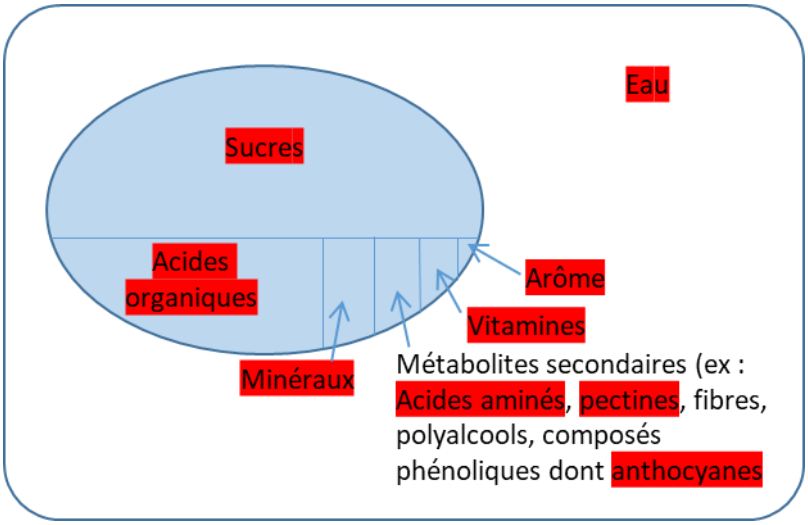

***Figure 1.6** : Composition schématique d'un jus de fruits. Les composés surlignés en rouge peuvent être sujets à des fraudes (principalement par ajout). Figure inspirée de Rinke, 2016*

Par conséquent, les contrôles d'authenticité sont d'autant plus complexes à mettre en œuvre. Cela peut notamment s'avérer plus difficile à déterminer dans les cas de fraudes par ajout ou substitution qui est l'une des fraudes majeures du fait de la grande variabilité naturelle des échantillons. De ce fait, l'analyse doit pouvoir confirmer qu'il n'y a pas eu d'ajout de certains composés comme les acides organiques ou les sucres. Cela nécessite donc de connaitre les concentrations attendues pour ces familles de molécules sujettes aux fraudes. Il est important d'également prendre en compte les différents facteurs naturels qui peuvent impacter les concentrations, rendant ainsi d'autant plus complexe l'analyse

d'authenticité. Pour pallier ce problème, les analyses se reposent sur les réglementations mises en place afin de répondre quant à l'authenticité du produit. Il est également possible de se reposer sur une base de données d'échantillons authentiques, c.à.d. un échantillon non fraudé et dont les métadonnées sont connues avec certitude.

1.4.2. Diversité des marqueurs

Pour aider à contrôler l'authenticité des aliments, il a fallu identifier des composés ou des familles de composés à détecter permettant de garantir une origine géographique, une variété ou un mode de production. Ces études se basent sur l'analyse et la comparaison d'échantillons authentiques et ont permis de mettre en évidence des composés marqueurs.

Par exemple, pour les jus de fruits, l'étude de Willems et Low a permis de mettre en évidence des molécules caractéristiques de la pomme et de la poire, permettant ainsi la détection de l'ajout d'un co-fruit dans les jus de pomme ou de poire (Willems & Low, 2018). Cette pratique est d'ailleurs très répandue dans les jus de fruits. Dans la même idée, Lehnert et Ara ont déterminé des composés discriminant le citron jaune et le citron vert : les flavanones polyméthoxylées (Lehnert, & Ara, 2014) ; ces composés sont donc régulièrement analysés dans les jus de citrus. La recherche de tels marqueurs permet ainsi de mettre en évidence l'une des fraudes les plus répandues dans les jus de fruits : l'ajout d'un co-fruit.

Un autre exemple illustratif est celui du café. Il existe majoritairement deux variétés de café : le café Arabica et le café Robusta. Il est important de pouvoir les discriminer, et en particulier de pouvoir authentifier le café Arabica dont la valeur marchande est plus élevée. Suite à des études, le 16-O-methylcafestol a été identifié comme marqueur de la variété Robusta et est régulièrement recherché dans les échantillons de café pour authentifier la variété Arabica (Monakhova et al., 2015).

Ces marqueurs sont spécifiques d'une origine géographique, d'une variété ou d'un fruit, ce qui montre la diversité de marqueurs d'authenticité à détecter dans les échantillons et par conséquent, une diversité de méthodes à mettre en place pour les analyser.

Bien que de nombreux marqueurs d'authenticité soient connus, des travaux doivent encore être menés car beaucoup restent inconnus. De ce fait, certaines revues scientifiques visent

à indiquer de potentiels composés à utiliser comme marqueurs d'authenticité (Medina et al., 2019a).

1.4.3. Limites des méthodes d'analyse conventionnelles

Comme mentionné au paragraphe 1.2., il existe des méthodes officielles pour garantir l'authenticité des denrées alimentaires. Ces méthodes ont été développées dans le but d'analyser des composés ou familles de composés connus pour être marqueurs d'authenticité. Ces méthodes d'analyse dites conventionnelles sont de ce fait très sensibles et spécifiques face aux composés à analyser, offrant de faibles limites de détection ou de quantification pouvant aller jusqu'au µg/L selon la technique d'analyse utilisée. Ainsi, de par ces méthodes, il est possible de répondre sur l'authenticité des denrées alimentaires avec une grande certitude.

Néanmoins, ces méthodes ayant été développées pour analyser des composés ou familles de composés spécifiques, elles peuvent ne pas être en capacité de détecter des fraudes plus sophistiquées. De plus certains marqueurs d'authenticité étant à ce jour encore inconnus, aucune méthode conventionnelle actuelle n'est donc en capacité de les observer. Par conséquent, un aliment pourra être considéré comme authentique selon les résultats d'analyses obtenus par les méthodes conventionnelles alors qu'il peut être adultéré : les analyses n'ont pas détecté cet adultérant (Everstine et al., 2013). C'est pourquoi il faut mettre en place de nouvelles approches non ciblées pour contrôler l'authenticité.

1.4.4. Les nouvelles approches pour l'analyse d'authenticité

Comme détaillé précédemment, il est intéressant de pouvoir développer de nouvelles méthodologies permettant l'analyse d'un maximum de composés présents dans le but de détecter d'éventuelles fraudes. Ce type de méthodologie a l'avantage de pallier les limites des méthodes actuelles, avec en contrepartie une moindre sensibilité et spécificité. En ce sens, la recherche se poursuit autour de l'authenticité alimentaire comme présenté en Figure 1.5 où le nombre de publications scientifiques traitant de ce sujet a fortement augmenté depuis la fin des années 1990.

Ainsi, afin de garantir l'authenticité des denrées alimentaires, des techniques d'analyse jusqu'alors très peu utilisées dans cette thématique ont montré leur intérêt pour répondre à

cette problématique (**Figure 1.7**). Ces techniques sont de plus en plus utilisées, en particulier les techniques chromatographiques et spectroscopiques sont majoritaires aujourd'hui (Danezis et al., 2016 ; Medina et al., 2019b).

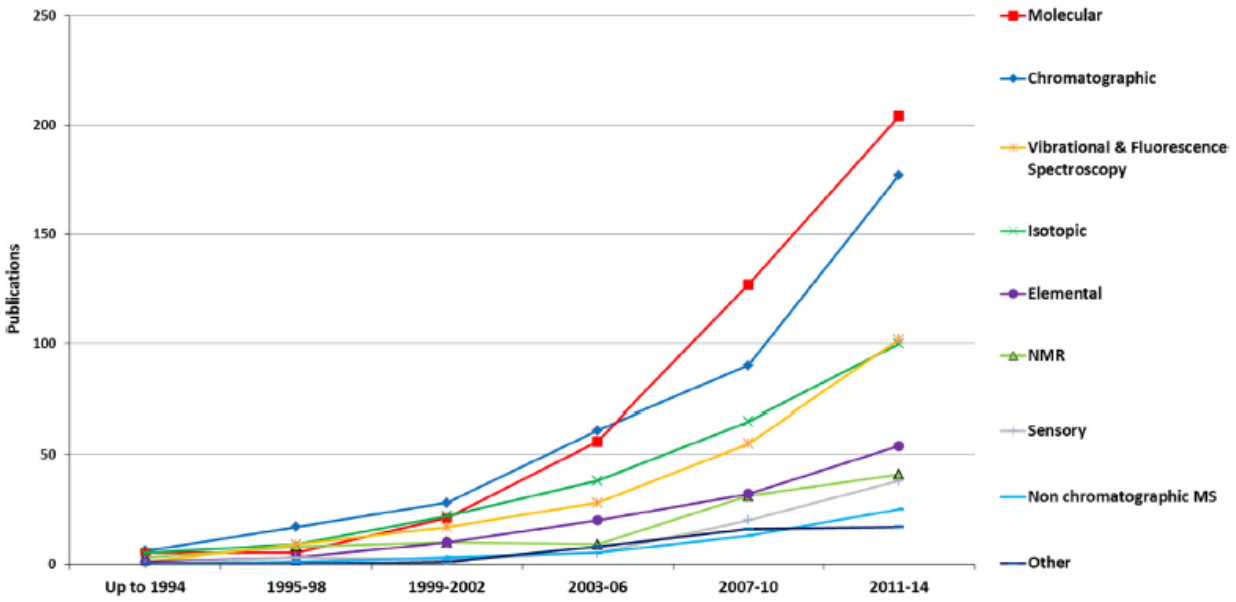

***Figure 1.7** : Evolution de l'utilisation de certaines techniques d'analyse en authenticité alimentaire au cours du temps. Issue de (Danezis et al., 2016). Les techniques « Molecular » concernent la protéomique, la génomique et les méthodes basées sur l'ADN.*

En parallèle de l'émergence de ces 'nouvelles' techniques d'analyse, d'autres approches méthodologiques sont de plus en plus en plus utilisées dans la lutte contre la fraude alimentaire. C'est ainsi que la « *Foodomics* » est née. Celle-ci a pour but de contrôler les denrées alimentaires avec des méthodologies de types 'omiques' (Cifuentes, 2009). Ces méthodes 'omiques' ont d'abord été développées dans le cadre d'études cliniques dans le but de pouvoir détecter des modifications de composition chimique entre plusieurs groupes d'échantillons (par exemple un groupe de contrôle et un groupe pour lequel un traitement a été appliqué). Elles se sont faites aujourd'hui une place dans le contrôle alimentaire aussi bien pour les problématiques d'authenticité que de sécurité sanitaire, notamment les approches métabolomiques et protéomiques (Cubero-Leon et al., 2014 ; Bohme et al., 2019 ; Knolhoff, & Croley, 2016 ; Lopez-Ruiz et al., 2019 ; Ortea et al., 2016).

Les approches métabolomiques permettent de mettre en évidence les signaux discriminants entre deux (ou plusieurs) groupes d'échantillons à l'aide d'une méthode d'analyse la plus exhaustive possible. Ces signaux sont par la suite identifiés (c.à.d. un composé est assigné

avec certitude au signal) ou a minima annotés (c.à.d. qu'une hypothèse quant à la structure du composé correspondant au signal est émise) (Dunn et al., 2011). Il est nécessaire dans ce type de méthodologie d'analyser un grand nombre d'échantillons afin de caractériser la variabilité interne aux échantillons. De ce fait, des grands jeux de données sont obtenus : plusieurs centaines voire des milliers de composés sont détectés selon la technique d'analyse utilisée. Des outils chimiométriques sont alors utilisés afin de permettre l'analyse de ces jeux de données (Dunn et al., 2011) ; ces outils permettent ainsi l'extraction de l'information pertinente pour répondre à une problématique.

Le développement de ce type d'approche a permis de discriminer des carottes selon le mode de production (agriculture biologique ou conventionnelle) (Cubero-Leon et al., 2018). D'autres études ont permis l'authentification de différents jus de fruits selon leur origine géographique et/ou leur variété ainsi que le type de fruits (Diaz et al., 2014 ; Vaclavik et al., 2012 ; Jandric et al, 2014 ; Jandric et al., 2017). Des résultats probants ont aussi montré leur application à l'analyse l'authenticité du safran selon son origine géographique (Rubert et al., 2016), ainsi que l'authenticité du blé dur (Cavanna et al., 2020).

Les approches non ciblées semblent donc très prometteuses pour répondre à la problématique d'authenticité des denrées alimentaires. Néanmoins, pour être utilisables en contrôle de routine, celles-ci nécessitent l'analyse d'un grand nombre d'échantillons authentiques afin de prendre en compte la variabilité des matrices alimentaires. Cela permet de mettre en place une base de données (BD) d'échantillons authentiques permettant de comparer un échantillon suspect à cette BD (Cubero-Leon et al., 2014 ; Danezis et al., 2016). La construction d'une telle BD peut s'avérer être délicate en authenticité car cela requiert d'avoir une bonne connaissance des métadonnées liées aux échantillons. Or certaines informations ne sont pas toujours connues, par exemple dans le cas d'un jus de fruits la variété des fruits n'est pas toujours indiquée.

1.5. LE CAS PARTICULIER DES JUS DE FRUITS

Les jus de fruits sont des matrices particulièrement intéressantes à étudier dans le cas des problématiques d'authenticité du fait de leur grande variabilité. De plus, il s'agit d'une matrice identifiée dans le top 10 des produits à risques de fraude comme précédemment exposé (Moore et al, 2012 ; Commission Européenne, Direction générale de la santé et de

la sécurité alimentaire, 2020). Enfin les jus de fruits sont très consommés à travers le monde, notamment pour leur valeur nutritionnelle (Rinke and Jamin, 2018 ; Dasenaki et al., 2019).

La recherche scientifique concernant l'authenticité des jus de fruits a fortement augmenté depuis les années 1990 comme le montre la **Figure 1.8**.

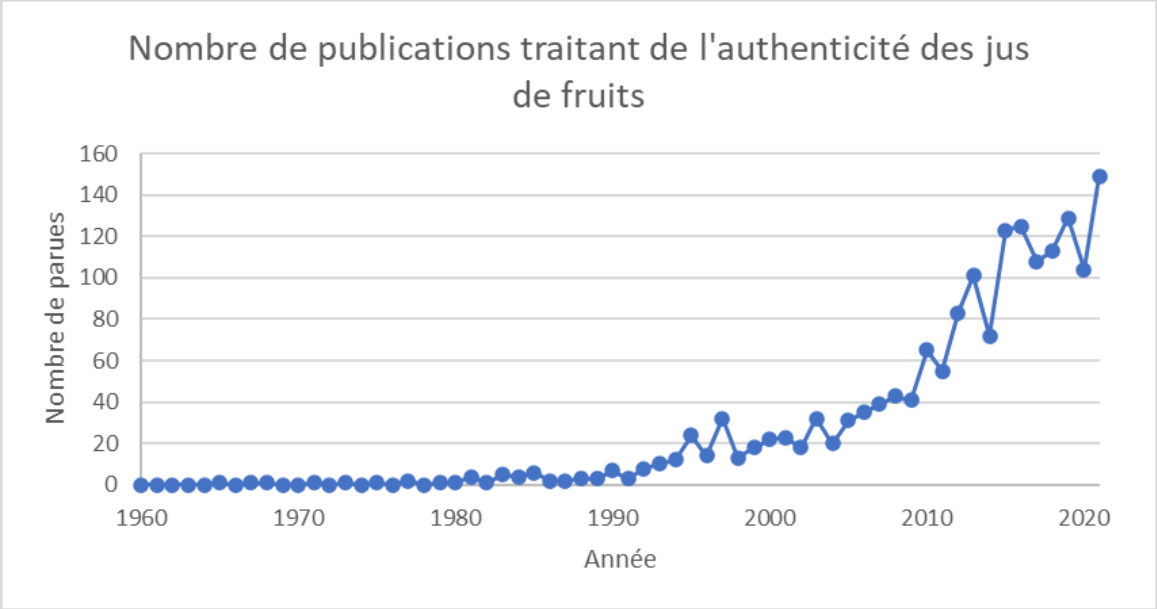

Figure 1.8 *: Nombre de publications parues traitant de l'authenticité des jus de fruits. Base de données utilisée : Science Direct, mots clés : « fruit juice authentication »*

1.5.1. La réglementation

Il existe des réglementations spécifiques selon le type d'aliment. Pour les jus de fruits, la directive européenne 2012/12/EC a permis d'établir des recommandations quant à leur constitution (Directive 2012/12/EC, 2012). Celle-ci détaille notamment des règles concernant la production, la composition et l'étiquetage des jus de fruits. Cette directive a pour but d'améliorer la circulation de ce type de denrée alimentaire au sein de l'UE et de protéger les consommateurs.

De plus, toujours dans le but de garantir la conformité des produits alimentaires, des associations ont été créées donnant ainsi des recommandations ou donnant accès à des méthodes d'analyse de référence. Pour les jus de fruits, on peut citer l'IFU (*International Fruit and Vegetable Juice Association*, association internationale des jus de fruits et légumes), reconnue mondialement (Codex Alimentarius, FDA) : elle agit en tant que

référence dans le secteur des jus (IFU, 2021). L'IFU est divisée en différentes commissions visant à assurer une harmonisation concernant les pratiques agricoles et la législation dans le secteur des jus de fruits, ainsi que la mise en place et la standardisation de méthodes d'analyse. Les méthodes IFU sont notamment reconnues comme méthodes officielles par le Codex Alimentarius.

On retrouve également au niveau européen l'AIJN (*European Fruit Juice Association*) qui a établi des codes de bonnes pratiques (AIJN code of practice, 2021). Ces codes sont des lignes directrices et indiquent la composition naturelle et caractéristique des fruits. Ils apportent une base pour l'évaluation de l'authenticité, de la qualité et de l'identité des jus. Ces codes indiquent pour certains composés soit leurs teneurs minimales ou maximales, soit une gamme de valeurs de concentration habituellement présente dans un jus de fruit typique.

1.5.2. La fraude concernant les jus de fruits

Les principaux types de fraudes ont été présentés en **Figure 1.1**. En ce qui concerne les jus de fruits, les principales fraudes rencontrées sont les suivantes (Rinke and Jamin, 2018) :

- Ajout d'eau (dilution du jus),
- Ajout de sucres, de vitamines (notamment la vitamine C), d'acides organiques (en particulier l'acide citrique et l'acide malique), d'arôme (naturel ou synthétique),
- Substitution du jus par du jus à base de concentré,
- Ajout d'un co-fruit (souvent moins couteux à produire),
- Falsification de l'origine géographique et/ou de la variété.

De ce fait, une grande partie des constituants d'un jus de fruits sont à risques de fraudes (cf. **Figure 1.6**). Il est donc essentiel d'avoir des méthodes d'analyse permettant de garantir l'authenticité des jus de fruits.

1.6. EN RESUME

Le contrôle d'authenticité consiste à vérifier la conformité d'une denrée alimentaire. Ainsi, cela permet de s'assurer d'une part de l'absence de fraudes et d'autre part de la présence de marqueurs d'authenticité.

Il existe différentes réglementations définissant les constituants attendus dans un aliment et donnant des indications sur leurs concentrations. Il existe également des organisations qui ont aidé à la mise en place de méthodes d'analyse de contrôle, en particulier pour les denrées les plus à risque de fraudes comme les jus de fruits. Néanmoins, malgré la présence de contrôles, la fraude est toujours présente.

Les aliments sont des matrices complexes car ils contiennent de nombreux constituants chimiques, les rendant difficiles à analyser. De plus, les facteurs naturels influent sur leur composition. Les méthodes de contrôle officielles peuvent donc échouer à la mise en évidence de fraudes. Il faut donc développer de nouvelles méthodologies d'analyse pour contrôler l'authenticité des denrées alimentaires (**Figure 1.9**).

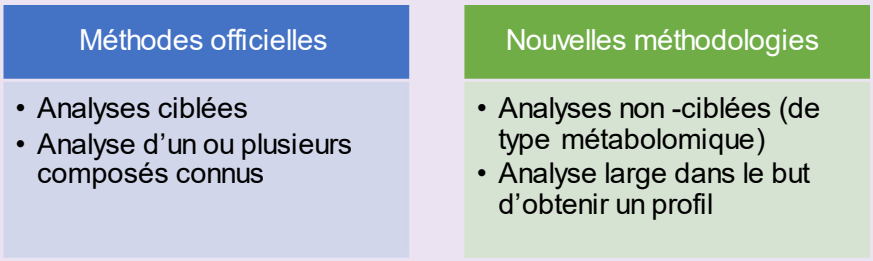

Figure 1.9 *: Différences entre les méthodes officielles et les nouvelles méthodologies utilisées pour le contrôle d'authenticité*

2. LES METHODES D'ANALYSE CLASSIQUEMENT UTILISEES POUR LE CONTROLE D'AUTHENTICITE

De nombreuses techniques analytiques sont mobilisables pour contrôler l'authenticité des aliments, comme précédemment illustré dans la **Figure 1.7**. Les techniques majoritairement utilisées dans l'analyse agro-alimentaire sont les techniques de spectroscopie vibrationnelle (infrarouge (IR), Raman, Fluorescence et UV-visible), la spectrométrie de masse (MS)

utilisée seule ou en couplage avec des techniques chromatographiques (chromatographie en phase liquide (LC) ou gazeuse (GC)) et la résonance magnétique nucléaire (RMN). Chacune de ces techniques présente des avantages et inconvénients dans le contrôle d'authenticité (**Tableau 1.1**) et la plupart d'entre elles sont complémentaires. En effet, la spectroscopie vibrationnelle permet d'observer les liaisons chimiques de la molécule étudiée ; la RMN quant à elle permet d'observer les environnements chimiques des différents atomes de la molécule ; et la MS permet d'observer la fragmentation de la molécule.

Dans ce manuscrit, nous allons principalement nous focaliser sur les méthodes d'analysespar RMN et par LC-MS, car ce sont celles-ci qui sont mises en œuvre au quotidien au sein du laboratoire Eurofins Analytics France en charge du contrôle de l'authenticité des aliments.

***Tableau 1.1** : Avantages et limites des principales techniques d'analyse ciblées (Ellis et al., 2012 ; Luykx & van Ruth, 2008)*

Techniques d'analyse	Avantages	Inconvénients
Spectroscopie vibrationnelle	Rapide, Robuste, Reproductible, Non destructif, Prix modéré de l'instrument	Faible sensibilité (de l'ordre du mg/L), Signal de l'eau très intense
RMN	Rapide, Robuste, Reproductible, Non destructif, Quantitatif, Information structurale	Faible sensibilité (de l'ordre du mg/L), Prix élevé de l'instrument
Couplage LC ou GC – MS	Sensibilité (jusqu'au fg/L), Spécificité, Robuste, Reproductible, Quantitatif (avec calibration), Analyse de mélange complexe	Temps d'analyse plus long, Encrassement du système, Prix élevé de l'instrument

2.1. LES METHODES D'ANALYSE PAR RMN

2.1.1. Les principales étapes d'analyse

L'analyse par RMN permet de déterminer l'environnement chimique d'un atome d'une molécule selon son déplacement chimique. Il existe différentes sondes RMN capables d'étudier différents noyaux atomiques. Les analyses RMN les plus répandues sont la RMN du proton (ou RMN ^{1}H), la RMN du carbone-13 (ou RMN ^{13}C) et la RMN de l'azote-15 (ou RMN ^{15}N).

L'échantillon est soumis à un champ magnétique constant B_0 une fois placé dans le spectromètre RMN. L'aimantation des noyaux se positionne alors selon ce champ. Un champ magnétique B_1 est ensuite appliqué faisant basculer l'aimantation dans le plan transversal. L'aimantation retourne ensuite à sa position initiale selon le phénomène de relaxation tout en émettant un signal qui est réceptionné par la bobine réceptrice.

L'analyse par RMN s'effectue selon les étapes suivantes : (i) la préparation des échantillons, (ii) l'analyse RMN, (iii) le traitement du signal, et si nécessaire (iv) la quantification.

La préparation des échantillons est une des étapes les plus délicates de l'analyse. En effet, dans le cadre des analyses conventionnelles, elle requiert des étapes permettant d'isoler le ou les composés d'intérêt. De ce fait, il est crucial que ces étapes soient bien réalisées afin d'obtenir ces composés avec une pureté suffisante pour éviter des superpositions de signaux. De plus, l'analyte doit être suffisamment concentré (de l'ordre du mg/L) pour être correctement analysé. La méthode d'extraction ou les solvants utilisés lors de la préparation des échantillons peuvent directement impacter les résultats (Sobolev et al., 2016). En effet, dans le cas d'analyses quantitatives, il est important de s'assurer que l'extraction permet d'obtenir un bon rendement pour les composés à analyser. Les étapes d'extraction permettent ainsi d'augmenter la sensibilité de l'analyse (Luykx & van Ruth, 2008 ; Mannina et al., 2012). Différents solvants ou combinaisons de solvants peuvent être utilisés pour les étapes d'extraction (Mannina et al., 2012). Au cours de la préparation des échantillons, il est important d'avoir *a minima* un solvant deutéré (le signal du deutérium est utilisé pour ajuster le « lock » de l'instrument) (Mannina et al., 2012).

Au cours de l'analyse, il est important de contrôler certains paramètres tels que le pH, la température ou la concentration des composés. En effet, ces facteurs peuvent influer sur les spectres RMN obtenus en provoquant des décalages de signaux (Mannina et al., 2012 ; Sobolev et al., 2016). Ainsi, afin de pouvoir comparer les spectres, il est important que ceux-ci aient été réalisés dans les mêmes conditions d'analyse.

Le traitement du signal consiste en différentes étapes : (i) l'application d'une transformée de Fourier pour obtenir le spectre RMN, (ii) le phasage du spectre permettant d'obtenir des pics gaussiens, (iii) la correction de la ligne de base du spectre, et (iv) la correction des déplacements chimiques à l'aide d'une référence.

La quantification par RMN est relativement simple à effectuer du fait que l'aire du signal est proportionnelle à la concentration du composé dans l'échantillon (Balci, 2005).

2.1.2. Exemples d'application

Les édulcorants sont des additifs alimentaires ayant un fort pouvoir sucrant, ce qui permet d'apporter un gout sucré aux aliments. Leurs teneurs étant réglementées par une directive européenne (Directive 2003/115/CE, 2003), il est important de doser ces composés. Certains édulcorants sont ainsi quantifiés par RMN (sucralose, cyclamate et néohespéridine dihydrochalcone) afin de s'assurer que leurs teneurs sont bien inférieures aux limites définies par la directive européenne.

La RMN isotopique développée au début des années 1980 permet de déterminer des rapports isotopiques, principalement $^{2}H/^{1}H$ et $^{13}C/^{12}C$, sur chaque position non équivalente de la molécule analysée. Cette méthode a notamment permis l'authenticité du vin en déterminant son taux de chaptalisation (Martin & Martin, 1981). En effet, en étudiant les rapports isotopiques de chaque site de la molécule d'éthanol, il a été possible de déterminer si des ajouts de sucre avaient été réalisés.

Cette méthode est d'ailleurs à l'origine de la création des laboratoires d'analyse Eurofins et est commercialisée sous le nom de SNIF NMR (*Site-specific Natural Isotope Fractionation Nuclear Magnetic Resonance*). En 1990, cette méthode a été reconnue comme méthode officielle pour l'authenticité des vins. Plus tard, celle-ci a été appliquée sur différentes matrices. Toujours en analysant l'éthanol, elle a été utilisée pour détecter l'ajout de sucre dans les jus de fruits et le sirop d'érable (Jamin et al., 2004). Elle a également été utilisée pour l'analyse de la vanilline, permettant ainsi de déterminer son authenticité et notamment de discriminer la vanilline d'origine naturelle et la vanilline synthétique (Remaud et al., 1997).

2.1.3. Avantages et inconvénients

Les principaux avantages et inconvénients de l'analyse RMN sont présentés en **Figure 1.10**.

L'un des principaux avantages de cette technique est qu'elle permet de faire de l'analyse structurale. De plus, la RMN est non destructive, ce qui est très avantageux dans le cas d'analyse d'échantillons précieux comme les échantillons biologiques. La RMN est

quantitative sur plusieurs décades sans nécessité d'analyser systématiquement un standard. Les analyses RMN ont aussi l'avantage d'être reproductibles sur différents instruments et différents laboratoires, permettant ainsi des comparaisons de résultats facilement.

Cependant, le spectre RMN peut s'avérer difficile à interpréter lors de l'analyse de mélanges complexes (Ellis et al., 2012). Il est important que l'échantillon soit parfaitement homogène afin d'éviter des signaux déformés par la présence de particules dans le tube. Les analyses peuvent être longues afin de permettre d'acquérir suffisamment de scans pour obtenir un rapport signal sur bruit (S/B) correct, essentiel pour les analyses de SNIF NMR. De plus, il est important d'avoir des conditions stables d'analyse pour que celles-ci soient répétables et reproductibles (solvant utilisé, pH, température) (Ellis et al., 2012 ; Mannina et al., 2012 ; Sobolev et al., 2016).

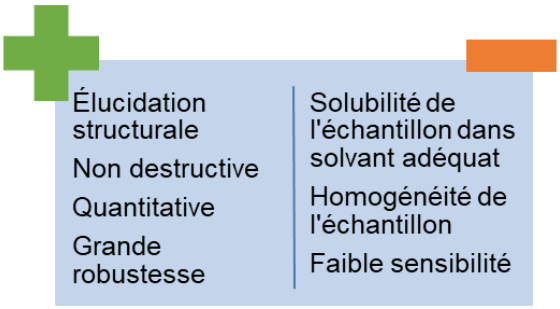

Figure 1.10 *: Avantages et inconvénients des méthodes d'analyse ciblées par RMN, établis à partir de (Ellis et al., 2012)*

2.2. LES METHODES D'ANALYSE PAR LC-MS

2.2.1. Les principales étapes d'analyse

Une fois l'échantillon à analyser préparé et solubilisé, une certaine quantité de l'extrait obtenu (le volume peut varier selon les instruments et les analyses, souvent de l'ordre du µL sur des systèmes UHPLC) est injecté dans le système. Les constituants présents dans l'extrait sont alors séparés par le module de chromatographie liquide (LC) selon leurs propriétés physico-chimiques, la phase stationnaire (c.à.d. la colonne) et la phase mobile (c.à.d. les éluants) utilisées. Les analytes arrivent ensuite dans le spectromètre de masse (MS) où ils sont analysés et détectés. Seuls des analytes sous forme d'ions sont analysés par le module MS. L'ionisation des composés a lieu dans la source d'ionisation du

spectromètre. Les ions ainsi formés sont ensuite séparés selon leur rapport masse sur charge (*m/z*) dans un analyseur de masse et enfin détectés par le détecteur.

Les grandes étapes d'une analyse par LC-MS sont : (i) la préparation des échantillons et du système, (ii) l'analyse, (iii) l'intégration des pics détectés, et si nécessaire (iv) la quantification.

Dans certains cas, les échantillons sont simplement dilués et filtrés préalablement à l'analyse. La dilution permet de diminuer la concentration des analytes dans l'échantillon et ainsi d'éviter toute saturation du signal. La filtration est recommandée pour éviter un bouchage du système, mais peut s'avérer être une étape critique ; en effet, les filtres utilisés peuvent retenir les composés d'intérêt et donc fausser les résultats de quantification. Dans d'autres cas, une préparation des échantillons plus spécifique doit être effectuée. Ainsi les échantillons solides doivent tout d'abord être solubilisés. Il est également parfois nécessaire d'effectuer des extractions permettant ainsi d'isoler les composés d'intérêt. Ces étapes sont également critiques pour l'analyse car les extractions peuvent induire une perte des composés d'intérêt ce qui peut fausser leur quantification.

L'identification d'un composé se fait en fonction de son temps de rétention (RT) et de son rapport masse sur charge (*m/z*) en comparant avec ceux obtenus sur un standard préalablement analysé dans les mêmes conditions. L'utilisation du *m/z* permet très souvent de s'affranchir de co-élutions observées sur le chromatogramme et ainsi d'isoler le pic de la molécule d'intérêt.

Il est parfois nécessaire de quantifier les composés d'intérêt pour confirmer la présence de fraudes. Pour cela, un étalonnage, externe ou interne, est nécessaire. Ainsi, une courbe d'étalonnage peut être établie permettant la quantification des échantillons. L'étalonnage externe est souvent préféré car il n'est pas toujours simple de trouver un étalon interne du fait que des étalons n'existent pas pour toutes les familles de molécules existantes. Il pourrait être intéressant dans le cadre d'analyses LC-MS d'utiliser des étalons internes marqués lorsqu'ils existent, mais leur prix augmente fortement par rapport à un étalon standard classique.

2.2.2. Exemples d'application

Les analyses par LC-MS sont souvent utilisées pour s'assurer de l'absence d'adultérants ou confirmer l'authenticité d'un aliment. Par exemple, au sein de l'unité Chromatographie du laboratoire Eurofins Analytics France, une méthode a pour but de détecter l'arbutine et la phloridzine. Le premier composé est un marqueur de présence de poire, tandis que le second est marqueur pour la pomme (Willems & Low, 2018). Ainsi, il est possible de constater si l'un de ces fruits a été ajouté dans un échantillon de jus de fruits.

Dans le même principe, l'analyse des flavonoïdes permet de détecter l'ajout d'un co-fruit dans les jus d'agrumes. En effet, cette famille de molécules est particulièrement présente dans les agrumes. Certaines molécules sont caractéristiques d'un agrume particulier : l'ériocitrine pour les citrons, le naringine pour les pamplemousses et l'hespéridine pour les oranges (Gattuso et al., 2007).

Dans un autre registre, le 5-hydroxymethylfurfural (5HMF) est également quantifié dans les échantillons de miel ou de jus de fruits car celui-ci permet de mettre en évidence un traitement thermique excessif. En effet, cette molécule se forme dans les aliments contenant des sucres après exposition à la chaleur. Une teneur maximale en 5HMF dans les miels est d'ailleurs indiquée par le Codex Alimentarius (40 mg/kg). Dans les jus de fruits, la teneur maximale est fixée à 20 mg/L par l'AIJN.

2.2.3. Avantages et inconvénients

Les principaux avantages et inconvénients de l'analyse par couplage LC-MS sont présentés en **Figure 1.11**.

L'un des avantages de l'analyse par LC-MS est la possibilité de pouvoir analyser des mélanges complexes (Ellis et al., 2012 ; Luykx & van Ruth, 2008). En effet, le module chromatographique permet la séparation des différents constituants avant d'être détectés par la MS, ce qui peut permettre d'isoler des signaux. De plus, grâce au couplage avec la MS, de nombreux cas de co-élutions peuvent être contournés car il est possible d'extraire le signal pour un *m/z* donné. De ce fait, l'analyse par LC-MS est sensible et spécifique aux molécules d'intérêt. L'analyse LC-MS permet également d'atteindre de faibles limites de détection, de l'ordre du femtogramme.

Néanmoins, l'analyse LC-MS est une méthode destructive. Un encrassement de la source de la MS peut être observé au cours du temps, notamment lors de l'analyse de mélanges complexes comme les matrices alimentaires, ce qui peut impacter directement l'analyse et notamment la quantification. De plus, afin d'avoir des analyses quantitatives, un étalonnage interne ou externe est nécessaire lors de chaque session d'analyse afin de déterminer une courbe d'étalonnage.

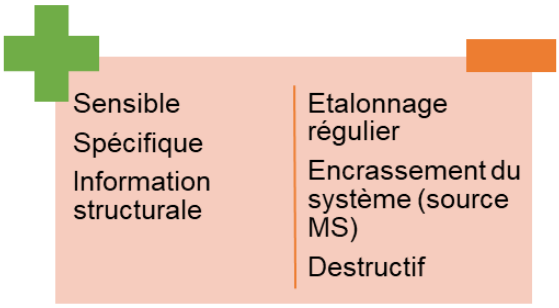

***Figure 1.11** : Avantages et inconvénients des méthodes d'analyse ciblées par LC-MS, établis à partir de (Ellis et al., 2012 ; Luykx & van Ruth, 2008)*

2.3. COMPARAISON DE CES TECHNIQUES

Ces deux techniques permettent l'analyse d'un large panel de matrices. Dans les deux cas, l'analyse se fait sur un extrait liquide et homogène.

2.3.1. Performances

La RMN offre une sensibilité relativement faible. Il est donc difficile avec cette technique de quantifier des composés à l'état de traces, la limite de quantification étant de l'ordre du mg/L. Néanmoins, cette technique est très robuste, et des données provenant de différents instruments et même de différents laboratoires sont très facilement comparables. De plus, du fait de sa grande répétabilité, il est simple de mettre en place une quantification par étalonnage externe.

A l'inverse, la LC-MS est une technique ayant une grande sensibilité et spécificité. La détection et la quantification de composés à l'état de traces sont aisées. Néanmoins, cette technique manque de répétabilité et de reproductibilité lié notamment à l'encrassement de la source, le vieillissement de la colonne et la pureté des éluants utilisés. De ce fait, les intensités varient entre deux sessions d'analyse. Par conséquent, il est nécessaire d'effectuer l'étalonnage externe à chaque session d'analyse.

2.3.2. Coût

Le cout d'un spectromètre RMN et d'une LC-MS sont du même ordre de grandeur, selon les modèles et les options de chaque instrument.

L'analyse RMN étant rapide (notamment la RMN ^{1}H), seulement quelques minutes, et nécessitant peu de consommables, elle est donc peu couteuse. En revanche l'analyse LC-MS dure environ une dizaine de minutes (en particulier avec les systèmes UHPLC), et le prix de l'analyse va également dépendre de la colonne utilisée et des solvants. En effet, ces consommables sont régulièrement changés, induisant de ce fait un cout supplémentaire.

2.4. EN RESUME

Dans cette partie, l'analyse par RMN et par LC-MS ont été présentées dans le cadre du contrôle d'authenticité des denrées alimentaires. Ces techniques d'analyse sont largement utilisées pour répondre aux problématiques d'authenticité des denrées alimentaires, notamment dans plusieurs méthodes officielles de contrôle d'authenticité. Du fait de leurs avantages et inconvénients respectifs, elles sont complémentaires et sont de ce fait souvent toutes deux utilisées dans les laboratoires de contrôle d'authenticité.

Suite à une préparation des échantillons spécifique, il est alors possible de détecter et de quantifier des composés connus marqueurs d'authenticité. L'analyse d'un échantillon est rapide (quelques minutes à quelques dizaines de minutes) et le traitement des données se fait facilement, les composés à détecter étant déjà connus. L'utilisation de ces techniques permet ainsi de répondre rapidement quant à l'authenticité de l'échantillon.

3. LES NOUVELLES APPROCHES NON CIBLEES POUR L'ANALYSE D'AUTHENTICITE

Ces nouvelles approches non ciblées ont pour but d'observer les échantillons dans leur globalité. Les techniques d'analyse sont les mêmes que celles utilisées dans le cas des approches conventionnelles. Chacune de ces techniques présente ses avantages et ses limites, comme présenté dans le **Tableau 1.2**.

***Tableau 1.2** : Principaux avantages et limites des techniques d'analyse utilisées pour les approches non ciblées (Danezis et al., 2016 ; Medina et al, 2019b ; Cubero-Leon et al., 2014)*

Technique d'analyse	Avantages	Inconvénients
Spectroscopie vibrationnelle	Rapide Robuste Traitement des données aisé	Faible sensibilité Identification complexe des signaux
RMN		
Couplage LC ou GC – MS	Identification de composés aisée (notamment avec HRMS) Sensibilité	Temps d'analyse long Traitement complexe des données Peu répétable et peu reproductible

Le principal avantage des techniques spectroscopiques (vibrationnelles et RMN) est leur dimensionnalité. En effet, ces techniques donnent l'information en 2 dimensions puisque l'intensité du signal ne dépend que d'une seule variable (respectivement la longueur d'onde ou le déplacement chimique). De ce fait, le traitement des données appliqué est simplifié en comparaison aux données LC-MS (ou GC-MS). Ces données sont quant à elles en 3 dimensions car dans ce cas l'intensité dépend à la fois du temps de rétention et du rapport masse sur charge.

A l'inverse, le principal avantage des analyses LC (ou GC)-MS est la possibilité d'identifier les composés, notamment lors de l'utilisation de spectromètre de masse à haute résolution (HRMS). En effet, la HRMS permet de détecter avec une grande précision les rapports *m/z* (jusqu'à 4 décimales) ce qui permet ainsi d'obtenir un nombre réduit de propositions de formules brutes. Il est également possible d'avoir des informations sur la fragmentation du composé. En revanche, les spectres obtenus avec les méthodes spectroscopiques montrent souvent des chevauchements de signaux, ce qui rend l'identification des signaux complexe. Dans le cadre de l'analyse d'authenticité, il est intéressant d'être en capacité d'identifier les composés marqueurs.

Ces nouvelles approches d'analyse non ciblées pour le contrôle d'authenticité sont principalement inspirées des études métabolomiques (Cubero-Leon et al., 2014). Les principales techniques utilisées dans les études métabolomiques sont la RMN et la LC-HRMS. En effet, grâce aux récents développements instrumentaux, ces techniques sont aujourd'hui des méthodes de choix dans les études métabolomiques, en particulier,

l'augmentation de la puissance des spectromètres RMN jusqu'au gigahertz, ainsi que les optimisations dans le couplage LC-MS ainsi que les développements dans des MS haute résolution (TOF et Orbitrap®).

Dans ce manuscrit, seules les approches non ciblées par RMN et par LC-HRMS sont présentées.

3.1. LA RMN EN APPROCHE NON CIBLEE

3.1.1. Principe et étapes d'analyse

Le principe de l'analyse non ciblée par RMN est d'acquérir un spectre global d'un échantillon que l'on appelle une empreinte spectrale. Les spectres RMN ainsi obtenus sont ensuite comparés à des échantillons authentiques. Cette comparaison permet de mettre en évidence des signaux aberrants pouvant être dus à une fraude. De plus, cela permet également de s'assurer de la présence de signaux marqueurs d'authenticité. Les analyses non ciblées par RMN peuvent également permettre la quantification de certains composés ; en effet, il est possible de quantifier des signaux isolés et connus.

L'analyse non ciblée par RMN s'effectue selon les étapes suivantes : (i) la préparation des échantillons, (ii) l'analyse RMN, (iii) le traitement du signal, (iv) le traitement des données et (v) les analyses chimiométriques (**Figure 1.12**).

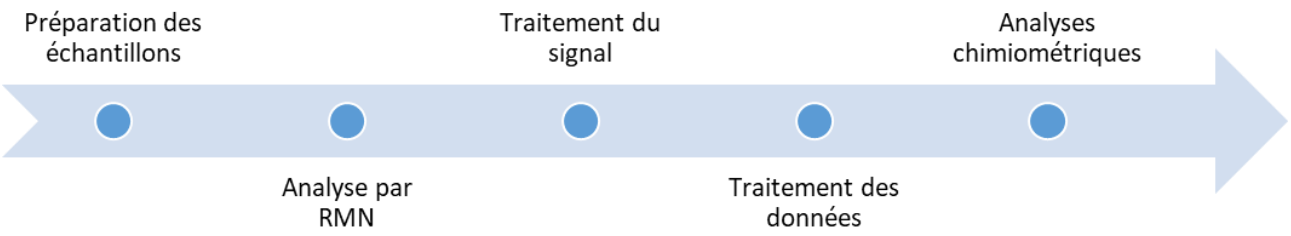

***Figure 1.12** : Principales étapes de l'analyse non ciblée par RMN*

Dans le cadre des analyses non ciblées, la préparation des échantillons doit rester la plus simple possible, permettant ainsi de pouvoir obtenir une empreinte la plus fidèle à l'échantillon. Certaines matrices nécessitent tout de même des extractions, comme par exemples les épices ou les thés (Mannina et al., 2012 ; Laghi et al., 2014 ; Pacholczyk-

Sienicka et al., 2021). De ce fait, les méthodes d'extraction utilisées doivent permettre d'extraire une majorité de composés présents dans l'échantillon afin de garantir un maximum de fidélité. Le choix du solvant peut être une étape délicate car, selon les propriétés des solvants utilisés, ils ne vont pas extraire les mêmes informations sur les échantillons (Mannina et al., 2012).

Lors de l'analyse RMN, comme pour les méthodes ciblées, il est important de bien prendre en compte certains facteurs tels que le pH, la température, la concentration et le solvant utilisé. Ceux-ci peuvent impacter le spectre RMN obtenu, en particulier des décalages de signaux peuvent être observés (Mannina et al., 2012 ; Sobolev et al., 2016 ; Emwas et al., 2018).

Les étapes de traitement du signal permettent de s'assurer de la qualité des spectres RMN. Les traitements appliqués pour les analyses conventionnelles sont également utilisés pour les analyses non-ciblées, c.à.d. correction de la phase, correction de la ligne de base et correction des déplacements chimiques (Emwas et al., 2018). D'autres étapes peuvent être mises en place selon la préparation des échantillons effectuée et selon la problématique étudiée. Par exemple, il peut être nécessaire de vérifier la saturation du signal de l'eau ou du solvant utilisé, ou l'ajustement en pH pour limiter la présence d'éventuels décalages de signaux (Emwas et al., 2018).

Le traitement des données a pour but de préparer les données aux analyses chimiométriques. Cette étape permet ainsi de réduire la taille du jeu de données, par exemple en réduisant le nombre de points du spectre RMN en utilisant les méthodes de « bucketing » (aussi appelées « binning »). Ces méthodes divisent le spectre RMN en différentes zones de largeur égale ; puis, pour chaque zone, les maximums d'intensités sont relevés (Sobolev et al., 2016 ; Emwas et al., 2018 ; Karaman et al., 2018)

Enfin, différentes méthodes chimiométriques peuvent être utilisées afin de s'assurer de l'authenticité de l'échantillon. Ces méthodes permettent de mettre en évidence les signaux discriminants entre plusieurs groupes d'échantillons. Il est également possible d'établir des modèles de prédiction des échantillons. Les analyses RMN étant quantitatives, des modèles de régression peuvent être mis en place pour calculer des concentrations.

3.1.2. Exemples d'application

L'analyse non ciblée par RMN est notamment utilisée dans l'authenticité du miel. En effet, il est possible de discriminer les origines géographiques et les origines botaniques des échantillons à partir du spectre RMN obtenu. Certains signaux sont connus pour être marqueurs de ces origines. A partir de modèles chimiométriques, il est possible de définir une probabilité d'appartenance à une origine en comparant avec la base de données d'échantillons authentiques existante (Spiteri et al., 2017). Cette méthodologie est utilisée en routine dans l'unité Profiling NMR du laboratoire Eurofins Analytics France.

En ce qui concerne l'analyse des jus de fruits, l'analyse non ciblée par RMN permet de s'assurer de différents paramètres pour contrôler l'authenticité. Il est possible de vérifier le mode de production : s'agit-il d'un pur jus ou d'un jus concentré ? L'origine géographique de l'échantillon peut également être confirmée. Différents composés sont également quantifiés afin de s'assurer que les directives établies par l'AIJN sont bien respectées (Spraul et al., 2009). Cette méthodologie a permis la création de l'outil JuiceScreeener™ commercialisé par Bruker. Une méthodologie similaire a été appliquée sur les vins (Godelmann et al., 2013), commercialisée par Bruker sous le nom de WineScreener™. Ces outils permettent à partir d'une acquisition non ciblée RMN ^{1}H de quantifier certains composés (ex : sucres, acides aminés) et d'appliquer des modèles chimiométriques afin de confirmer différentes informations telles que l'origine géographique, le cépage pour les vins et le mode de production des jus de fruits. Ces outils sont d'ailleurs utilisés dans l'unité de Profiling NMR au sein du laboratoire d'Eurofins Analytics France.

Il faut également citer l'implémentation de cette méthodologie pour contrôler l'authenticité des huiles alimentaires. Ainsi, de par l'analyse non ciblée RMN ^{13}C et l'utilisation d'outils chimiométriques, il est possible d'authentifier les huiles alimentaires en détectant l'ajout d'autres types d'huiles. En particulier, cette méthode est capable de détecter jusqu'à 2 % l'ajout d'une huile d'origine végétale, et jusqu'à 5 % l'ajout d'une huile d'origine animale (Guyader et al., 2018). Cette méthode est également utilisée en routine dans l'unité Profiling NMR du laboratoire Eurofins Analytics France.

L'analyse non ciblée par RMN ^{1}H couplée à des outils chimiométriques a également permis de discriminer les deux variétés majeures de cannelle (C. verum et C. cassia). De plus, un

composé, l'eugénol, a été identifié comme étant marqueur de la variété C. verum, et peut donc être utilisé comme marqueur d'authenticité. La variété C. cassia se différencie quant à elle par sa plus forte teneur en acides gras (Farag et al., 2018).

3.1.3. Avantages et inconvénients

Du fait que la RMN soit une méthode répétable, la comparaison des empreintes spectrales peut se faire très facilement. De plus, étant également une analyse reproductible, il est de ce fait possible de comparer des empreintes obtenues sur différents instruments, voire même différents laboratoires d'analyse. Par conséquent, la création d'une base de données d'échantillons authentiques est simple à mettre en œuvre et peut être facilement incrémentée.

L'empreinte spectrale ainsi obtenue contient énormément de signaux qu'il est complexe d'exploiter, rendant ainsi difficile l'identification de composés. Cette étape est pourtant importante car elle permet de donner un sens chimique aux observations et peut être utile pour identifier de nouveaux marqueurs d'authenticité.

3.2. LA LC-HRMS EN APPROCHE NON CIBLEE

3.2.1. Principe et étapes d'analyse

Le principe de l'analyse non ciblée par LC-HRMS est d'acquérir l'empreinte chromatographique globale d'un échantillon. Pour chaque pic chromatographique, il est possible d'extraire les *m/z* majoritaires. Par conséquent, ce type d'analyse permet d'obtenir une information en trois dimensions : le temps de rétention (RT), le *m/z* et l'intensité.

Les principales étapes d'une analyse non ciblée par LC-HRMS sont : (i) la préparation des échantillons, (ii) l'analyse, (iii) le prétraitement des données, (iv) les études chimiométriques et (v) l'annotation des marqueurs discriminants (**Figure 1.13**).

Pour l'acquisition d'une empreinte globale, il est important de prévoir le moins d'étapes possibles de préparation des échantillons. En effet, le but de ces analyses non ciblées est d'observer le maximum de composés présents dans les échantillons afin d'obtenir une empreinte la plus représentative de l'échantillon. De plus, les composés d'intérêt (c.à.d. marqueurs d'authenticité) n'étant pas toujours connus, il est important d'appliquer un

minimum de préparation pour éviter les risques de perte de constituants de l'échantillon. Ainsi, des approches de type « dilution et analyse » (de l'anglais *dilute and shoot*) sont souvent appliquées lorsque que la matrice étudiée le permet.

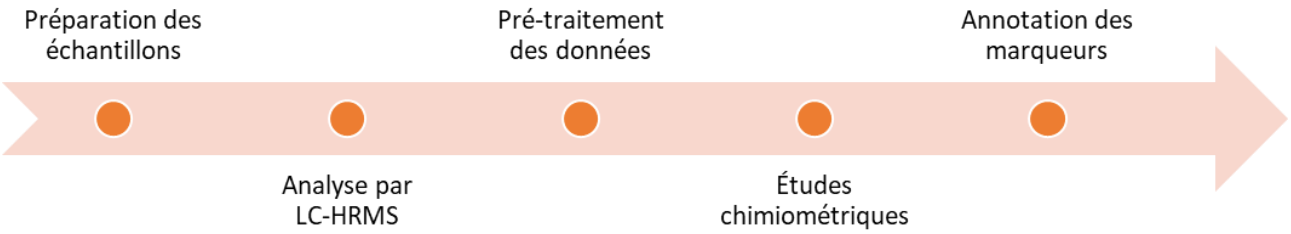

Figure 1.13 *: Principales étapes de l'analyse non ciblée par LC-HRMS*

L'analyse doit être capable de permettre la détection d'un maximum de composés. De ce fait, la méthode doit être la plus générique possible. Par conséquent, les conditions analytiques appliquées sont assez générales, c.à.d. que l'on va se mettre dans des conditions adaptées à la majorité des petites molécules, bien que cela implique qu'elles ne seront pas adaptées à certains composés (notamment les molécules très polaires). En particulier, une colonne avec une phase stationnaire à base de silice et un greffage C18 est souvent utilisée pour ces analyses non ciblées, et le spectromètre de masse fonctionne généralement en mode balayage complet (« *full scan* ») avec une ionisation en mode positif (le mode négatif, nettement moins utilisé, semble donner de moins bons résultats (Vaclavik et al., 2012 ; Diaz et al., 2014)).

Le prétraitement des données est constitué de différentes étapes permettant d'extraire les pics des chromatogrammes obtenus sous la forme d'un tableau. Lors de cette étape d'extraction, les pics sont alors caractérisés par leur combinaison *m/z* – RT, aussi appelée « *feature* ». Puis le jeu de données est nettoyé dans le but de réduire le nombre de variables en supprimant des features instables ou non pertinents. Cette étape permet ainsi de faciliter les analyses chimiométriques. Une sélection des variables peut également être appliquée à la fin du prétraitement.

Les études chimiométriques consistent en la création de modèles permettant de déterminer les features favorisant la discrimination entre les différents groupes d'intérêt. Les outils

utilisés peuvent également permettre la création de modèle de classification et de prédiction. Ils ont également pour but de détecter d'éventuels échantillons ou signaux aberrants afin de les mettre de côté.

Lorsque les features discriminants ont été déterminés, il est important de tenter d'identifier les composés sous-jacents. Cette étape permet ainsi de donner un sens chimique à ces signaux discriminants. Il existe différents niveaux d'identification, définis pour les études métabolomiques, qui sont totalement adaptés aux études agroalimentaires d'authenticité (**Figure 1.14**).

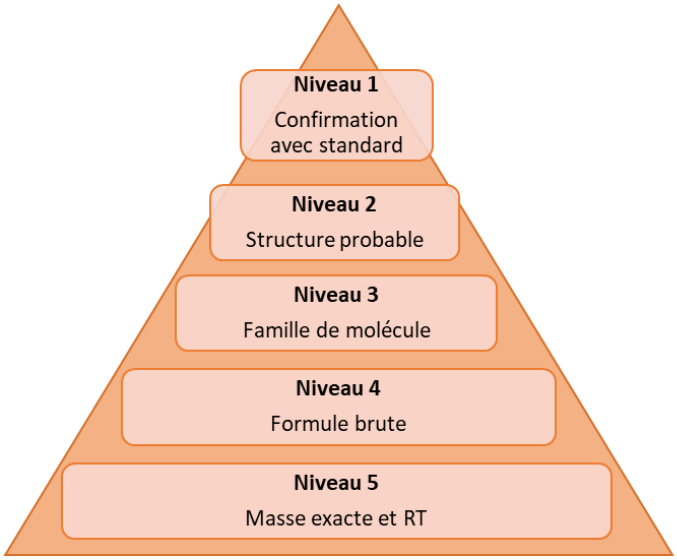

***Figure 1.14** : Les différents niveaux pour l'annotation des composés, inspiré de Schymanski et al., 2014*

Lorsque le signal d'intérêt est uniquement caractérisé par la masse expérimentale exacte obtenue (le *m/z*) et le temps de rétention (RT), il s'agit du niveau 5, le niveau le plus bas, d'identification : le composé est alors inconnu. Pour atteindre le niveau 4, il est nécessaire de connaitre des informations supplémentaires : le type d'adduit du signal observé et le massif isotopique. Grâce à ces informations, il est alors possible de déterminer la formule brute du composé. Puis, le niveau 3 est atteint lorsque l'on a déterminé la classe (ou famille) de composés à laquelle appartient le composé d'intérêt, sans pour autant être capable de définir une structure. La famille d'appartenance du composé est déterminée grâce à l'acquisition de spectres MS/MS, en étudiant la fragmentation du composé. Le niveau 2 correspond à une proposition de structure probable du composé suite à l'analyse plus poussée des spectres MS/MS obtenus, et après des recherches dans les bases de données

spectrales. A ce niveau, on parle alors d'annotation putative du composé (en anglais, *putative annotation*). Enfin, le niveau 1, le cas idéal, correspond à la confirmation du composé *via* l'analyse d'un standard permettant ainsi de vérifier que le RT, le *m/z* et le spectre MS/MS obtenus correspondent au pic d'intérêt (Schymanski et al., 2014 ; De Vijler et al., 2018).

Il existe différentes bases de données en ligne permettant d'aider à l'identification des composés : certaines sont générales (par ex., Metlin, MassBank, NIST, mzCloud) et d'autres sont spécifiques comme par exemple FooDB qui référence des composés présents dans les aliments, ou HMDB qui référence de nombreux composés présents dans le corps humain.

3.2.2. Exemples d'application

Ce type d'approche non ciblée par LC-HRMS inspirée de la métabolomique a été mis en place pour différentes problématiques d'authenticité alimentaire.

En ce qui concerne les jus de fruits, Diaz et al. ont utilisé cette méthodologie pour authentifier différents jus d'orange selon leur origine géographique. L'utilisation de modèle chimiométrique a permis de différencier les échantillons provenant d'Espagne des échantillons d'autres origines (Brésil, Argentine et Afrique du Sud). Ce modèle a été capable de classer 100 % des échantillons. Un feature a été identifié comme marqueur de l'origine géographique et une proposition d'identification (niveau 2) a été faite (Diaz et al., 2014) : il s'agirait de la citrusine D (à confirmer par analyse du standard). L'approche métabolomique a également été utilisée pour différencier différents types de fruits (orange, pomme et pamplemousse) par LC-HRMS. De plus, de par les modèles chimiométriques établis, il a été possible de déterminer le taux d'adultération de ces jus de fruits jusqu'à 10 % d'adultération (Vaclavik et al., 2012). Une autre étude visant à authentifier le jus de grenade a permis de discriminer différents fruits (grenade, pomme et raisin) grâce à ce type de méthodologie ; les modèles chimiométriques proposés ont permis de détecter la présence d'adultération jusqu'à 1 % dans le jus de grenade (Dasenaki et al., 2019). Ce type d'approche métabolomique par LC-HRMS a également permis de discriminer le jus de cassis du jus d'aronia (Dubin et al., 2017).

Les produits issus de l'agriculture biologique sont également étudiés *via* cette méthodologie. C'est notamment le cas des carottes pour lesquelles la méthode a permis de discriminer des échantillons issus de l'agriculture biologique de ceux issus de l'agriculture conventionnelle. De plus, plusieurs composés marqueurs du caractère bio sur ces échantillons ont été identifiés (niveau 1), dont la sedoheptulose, la L-arginine et l'acide chlorogénique (Cubero-Leon et al., 2018). Ces composés sont liés au métabolisme des sucres ou aux mécanismes de défense de la plante.

D'autres denrées alimentaires ont également été authentifiées *via* ce type d'approche. De par l'analyse LC-HRMS combinée à des traitements statistiques, Rubert et al. ont pu authentifier le safran, notamment dans le but de confirmer l'origine géographique des échantillons. Ainsi ces auteurs ont pu discriminer des échantillons de safran d'origine protégée (AOP) des autres échantillons ; la méthodologie développée permet de classer correctement 100 % des échantillons (Rubert et al., 2016). L'approche non ciblée par LC-HRMS a également été utilisée pour authentifier l'origine géographique du blé dur : sur un jeu de données externe, la méthodologie mise en place a permis de correctement classer 88 % des échantillons (Cavanna et al., 2020).

Ce type d'approche a également été utilisé afin de détecter la présence de contaminants chimiques. En particulier, Delaporte et al. ont pu détecter différents contaminants soumis à des réglementations dans le thé à des niveaux de traces dans les échantillons (Delaporte et al., 2019).

3.2.3. Avantages et inconvénients

Le module de chromatographie en phase liquide permet de séparer les différents constituants de l'échantillon avant leur détection par le spectromètre de masse. Cette séparation est notamment intéressante pour l'identification des signaux discriminants. En effet, lors de l'analyse de standard pour confirmer l'identité du composé, il est important que celui-ci ait le même RT. L'utilisation de la LC-HRMS pour les analyses non-ciblées permet d'atteindre une certaine sensibilité (de l'ordre du µg/L), bien que celle-ci ne soit pas aussi grande que lors d'analyses ciblées. Il est ainsi possible d'observer des composés présents à l'état de traces.

Les données LC-HRMS, de par leur dimensionnalité, sont complexes à traiter. En effet, l'analyse LC-HRMS n'est pas complètement répétable : les RT et les *m/z* doivent être alignés entre les différents chromatogrammes obtenus avant d'être comparés. Les déviations ne sont pas linéaires, ce qui complexifie les prétraitements à réaliser (Karaman et al., 2018 ; Gorrochategui et al., 2016 ; Dunn et al., 2011 ; Pezzatti et al., 2020). Les analyses non-ciblées requièrent d'observer un maximum de composés, ces analyses LC-HRMS sont donc relativement longues en comparaison des analyses RMN (environ 30 min pour une analyse LC-HRMS contre environ 10 min pour une analyse RMN).

3.3. LE TRAITEMENT DES DONNEES D'ANALYSE NON CIBLEES

Au préalable de l'utilisation des méthodes chimiométriques, il est primordial de préparer les données d'analyse obtenues. Le prétraitement des données consiste à extraire les signaux afin d'obtenir un jeu de données. Ce jeu de données est ensuite filtré pour éliminer les variables instables présentes, facilitant ainsi l'analyse chimiométrique.

Le traitement des données est souvent suivi d'étapes visant à sélectionner des variables. Cette sélection doit permettre de réduire le nombre de variables du jeu de données afin de ne conserver que les variables les plus pertinentes pour répondre au problème. La sélection de variables vise notamment à conserver les variables les plus discriminantes.

3.3.1. Traitement des données RMN

Les traitements des données applicables aux données non ciblées RMN sont variés. Leurs utilisations dépendent principalement de la préparation des échantillons effectuée et de la problématique étudiée. En effet, les données RMN sont des données spectrales à 2 dimensions ; elles sont par conséquent simples à analyser avec des modèles chimiométriques. De ce fait, il n'est pas toujours nécessaire d'appliquer un traitement sur ces données. Il est en revanche primordial de vérifier la qualité des spectres RMN obtenus. Si les empreintes spectrales sont de bonne qualité, les modèles chimiométriques mis en place seront efficaces pour répondre à la problématique.

La vérification de la qualité des spectres RMN consiste par exemple à s'assurer que la saturation du signal de H_2O est bien effectuée. En cas d'ajustement du pH lors de la préparation de l'échantillon, il faut s'assurer qu'il a été correctement corrigé. Cela permet

d’aider à la répétabilité des spectres RMN. De plus, cela permet d’éviter certains changements sur le spectre RMN dû au changement de conformation de certaines molécules selon le pH (acides organiques, amino-acides…) (Emwas et al., 2018).

Un traitement des données peut par exemple être appliqué lorsque l’échantillon analysé nécessite une extraction par solvant, car il est possible d’observer un biais selon la proportion de solvant présente dans les échantillons. Il est donc nécessaire de compenser ce biais qui induit des différences d’intensités sur les spectres. Cette compensation est souvent effectuée en divisant l’intensité de chaque point par la somme des intensités du spectre (en omettant les signaux attribués au solvant d’extraction) (Petrakis et al., 2015 ; Farag et al., 2018).

Il est également possible de supprimer des zones du spectre RMN ne contenant aucune information pertinente pour répondre à la problématique, ou contenant des signaux connus et pouvant avoir été affectés par l’analyse. C’est par exemple le cas du signal de l’urée lors de l’analyse d’échantillons d’urine : impacté par la suppression du signal de l’eau, il n’est pas exploitable (Emwas et al., 2018 ; Karaman et al., 2018).

Un autre exemple de traitement appliqué aux données RMN est le « bucketing ». Le spectre RMN pouvant contenir plusieurs dizaines de milliers de points, le « bucketing » permet de réduire cette quantité de données en séparant le spectre en différents « buckets » (de nombre B) de longueur égale. Pour chaque « bucket », la somme des points est effectuée. Le spectre peut alors être résumé en B valeurs (Jellema, 2009 ; Sousa et al., 2013).

De plus, il peut être nécessaire d’appliquer des méthodes de normalisation afin de corriger des différences de concentration entre les échantillons. Cette étape est notamment nécessaire lors de l’analyse d’échantillons biologiques, sensibles à des effets dilutions des échantillons (Emwas et al., 2018).

Si nécessaire, il est également possible d’appliquer des méthodes de mises à l’échelle ou de transformation des données en amont des analyses chimiométriques (Emwas et al., 2018). Ces méthodes permettent de réduire le bruit présent dans le jeu de données tout en gardant intacte l’information provenant des échantillons (van den Berg et al., 2006).

3.3.2. Traitement des données LC-HRMS

Le traitement des données LC-HRMS peut être considéré comme plus complexe que le traitement de données issues d'autres techniques spectroscopiques (RMN, IR…). En effet, les techniques chromatographiques couplées à un spectromètre de masse donnent des données avec 3 dimensions (RT, *m/z* et intensité), contrairement aux autres techniques analytiques qui sont des données avec au maximum 2 dimensions.

Le traitement des données LC-HRMS se compose des différentes étapes suivantes : (i) le prétraitement des données, (ii) le nettoyage du jeu de données, (iii) l'application de méthodes chimiométriques et (iv) l'annotation des variables discriminantes (**Figure 1.15**).

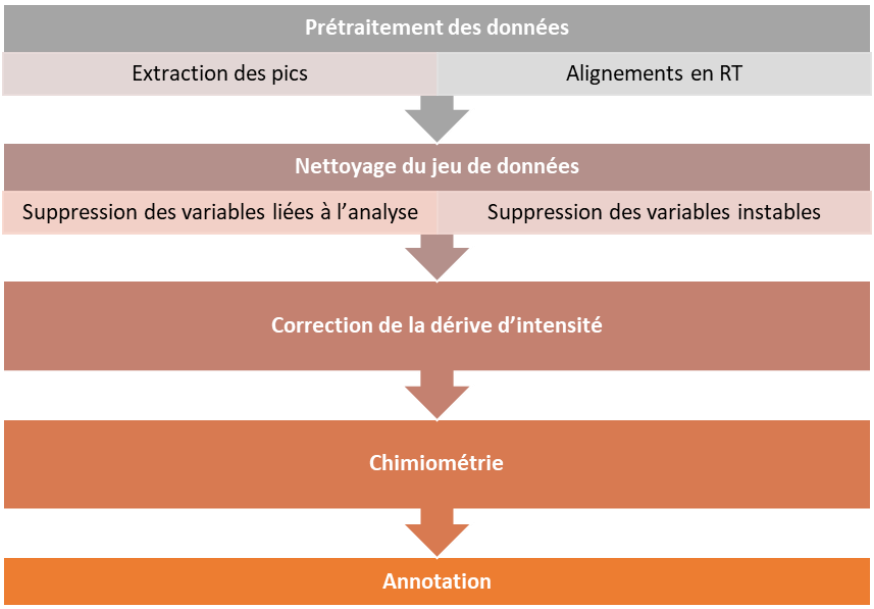

Figure 1.15 *: Principales étapes du traitement des données LC-HRMS*

Différents outils permettent de traiter les données LC-HRMS : les outils propriétaires disponibles auprès des constructeurs d'instrument, et des outils libres. Une comparaison de ces outils est présentée dans le **Tableau 1.3**.

***Tableau 1.3** : Présentation des principaux avantages et limites des outils pour le traitement des données LC-HRMS*

	Avantages	**Inconvénients**
Outils libres	Lecture de fichiers provenant de différents instruments Outils gratuits Souplesse dans le choix des fonctions à appliquer	Nécessite une conversion préalable des fichiers Connaissance du langage de programmation Paramétrage des fonctions utilisées
Outils propriétaires	Simple d'utilisation Lecture des fichiers bruts Automatisation du traitement possible	Logiciels payants Boites noires Peu modulaire

Il existe différents logiciels propriétaires pour le traitement de données LC-HRMS, par exemple Compound Discoverer (ThermoFisher) ou Marker View (Waters). Ces logiciels sont simples d'utilisation pour les utilisateurs et ont l'avantage de pouvoir directement lire les fichiers bruts (c.à.d. les fichiers issus de l'instrument). De plus, le traitement des données peut être effectué automatiquement après l'analyse. Néanmoins, ces logiciels sont payants. De plus, ils sont souvent considérés comme des boites noires car il est difficile de connaitre en détails les calculs effectués des différents modules (notamment pour la partie chimiométrique).

Il existe également d'autres outils pour traiter ces données : il s'agit de logiciels libres (aussi appelés « *open-source* ») comme par exemple MZmine, MSDial ou XCMS (sous R). Ces outils permettent notamment d'avoir plus de souplesse sur les fonctions ou les paramètres appliqués pour le traitement des données. De plus, ces logiciels sont gratuits. Néanmoins, ils nécessitent que l'utilisateur possède des compétences en programmation. Les fonctions présentes sur ces logiciels contiennent différents paramètres qu'il est important de connaitre. De plus, pour être analysés avec ces outils, les fichiers doivent être convertis dans un format libre, généralement au format mzXML.

L'outil le plus utilisé pour traiter des données LC et GC-MS est XCMS (Smith et al., 2006). Ce « *package* » est constitué de différentes fonctions permettant l'extraction des données et l'alignement des temps de rétention. XCMS a connu différentes améliorations depuis son

développement, permettant ainsi l'amélioration de certains modules (Tautenhahn et al., 2008) ainsi que sa mise en place dans des plateformes en ligne telles que *Workflow4Metabolomics* (W4M) (Giacomoni et al., 2014 ; Guitton et al., 2017).

3.3.2.1. *Prétraitement des données LC-HRMS*

Cette première étape d'extraction des données est la partie la plus importante à maitriser car elle peut impacter toute la suite du traitement des données. En effet, c'est au cours de cette étape que la liste de données est établie. Cette liste de données sera ensuite utilisée pour les outils chimiométriques. De ce fait, les résultats des outils chimiométriques peuvent être faussés si la liste de données n'est pas correctement établie : un modèle peut ne pas être concluant alors qu'il devrait l'être ou inversement. L'extraction des features est donc une étape cruciale lors du traitement des données.

L'étape d'extraction des données consiste à extraire les pics chromatographiques. L'extraction permet de définir les pics chromatographiques détectés dans les échantillons selon leur combinaison *m/z* – RT (ou « feature »). Ainsi, la dimensionalité des données LC-HRMS est réduite en données à 2 dimensions plus simples à analyser avec les outils chimiométriques (Gika et al., 2014).

Le but des analyses d'authenticité étant de détecter d'éventuels adultérants, il est important d'extraire un maximum de pics. La méthode d'extraction utilisée peut donc influencer le nombre de pics détectés dans les fichiers (Knolhoff & Croley, 2016). Il existe principalement deux méthodologies d'extraction des pics : le « binning » et la recherche dans des régions d'intérêt (*region of interest* ou ROI) (Gorrochategui et al., 2016).

La méthode de « binning » est l'une des premières méthodologies développées pour l'extraction des features et a notamment été utilisée dans la première version de XCMS (Smith et al., 2006). Cette méthode ressemble au « bucketing » présenté précédemment pour le traitement des données RMN. Le « binning » permet de représenter le fichier brut en une matrice d'intensité avec les RT sur l'axe x et les *m/z* sur l'axe y. Pour cela, l'axe des *m/z* est divisé en différents intervalles (aussi appelées « bin ») de largeur fixe (souvent 0,1 *m/z*). Pour chaque « bin », les maximums d'intensité sont identifiés sur tout le domaine RT (Gorrochategui et al., 2016 ; Smith et al., 2006). L'un des inconvénients majeurs du « binning » est la diminution de la résolution des données. De plus, il est important de bien

définir la taille de chaque « bin » pour éviter d'avoir des co-élutions de pics chromatographiques qui pourraient masquer des pics peu intenses (cas d'un « bin » trop grand) ou à l'inverse un pic non détecté (cas d'un « bin » trop petit) (Gorrochategui et al., 2016 ; Tautenhahn et al., 2008 ; Smith et al., 2006).

Une alternative à la méthode de « binning » est la méthode par détection de ROI. Celle-ci a tout d'abord été présentée par Stolt et al., et se base sur le principe qu'un analyte peut être caractérisé par une région spécifique de points avec une grande densité. Ainsi, un pic est détecté par cette méthode si le *m/z* sur des scans consécutifs reste constant (à une déviation près, définie par l'opérateur selon la précision de l'instrument) et qu'un pic de largeur typique pour la LC est obtenu dans le domaine RT pour ces différents scans (Stolt et al., 2006). Tautenhahn et al. a ensuite repris cette méthodologie pour l'implémenter dans XCMS (algorithme *centWave*) (Tautenhahn et al., 2008). La méthode par détection de ROI est aujourd'hui très largement utilisée, notamment car elle est adaptée aux données de HRMS. De plus, celle-ci permet également de palier les inconvénients de la méthode de « binning » (Tautenhahn et al., 2008 ; Gorrochategui et al., 2016).

Suite à l'extraction des features pour chaque échantillon, il est important d'aligner les temps de rétention entre les échantillons. En effet, les RT ne sont pas stables et reproductibles (Pezzatti et al., 2020 ; Karaman et al., 2018 ; Dunn et al., 2011). L'étape d'extraction précédente a identifié des combinaisons *m/z* – RT : pour comparer les différents features détectés dans les différents échantillons, il faut corriger les déviations en RT. Ces déviations entre échantillons sont non-linéaires (Gorrochategui et al., 2016 ; Pezzatti et al., 2020). Le *package* XCMS contient des fonctions pour aligner les RT entre les différents échantillons permettant ainsi de rendre comparables les features.

A la fin de cette étape, une matrice de données est obtenue contenant les intensités de chaque feature pour chaque échantillon. Il est possible de compléter ce tableau par des tableaux annexes contenant des métadonnées sur les échantillons (groupe, ordre d'injection, etc.) et sur les pics (*m/z*, RT, adduits, etc.).

3.3.2.2. *Nettoyage du jeu de données LC-HRMS*

Le jeu de données obtenu à la fin des étapes de prétraitements peut contenir des dizaines de milliers de combinaisons *m/z* – RT. Il est important d'effectuer différentes étapes de

filtration des données afin de réduire le nombre de features détectés. Ces étapes permettent également de supprimer les features instables et non pertinents.

Une des étapes de filtration consiste à supprimer du jeu de données les features correspondant à l'instrument. Pour cela, cette étape s'appuie sur la présence de blancs analytiques représentatifs des pics résiduels du système. Ces blancs analytiques permettent également de mettre en évidence la présence d'éventuels effets mémoire. Il est également possible d'analyser des blancs d'extractions, c.à.d. des blancs ayant subi la même préparation que les échantillons ; cela permet ainsi d'observer d'éventuels contaminants liés à la préparation des échantillons (Broadhurst et al., 2018 ; Dudzik et al., 2018). Ces pics résiduels peuvent avoir été extraits lors du prétraitement, pour autant ceux-ci n'apportent aucune information pertinente. De plus, ces pics résiduels varient entre différentes sessions d'analyse. Il est donc essentiel de les supprimer pour éviter de les considérer comme discriminants lors de la comparaison de différents jeux de données. Il existe différentes méthodes pour identifier ces pics résiduels liés aux blancs analytiques en vue de les supprimer (Broadhurst et al., 2018). L'une des méthodes considérées comme une des plus représentatives consiste à calculer le « *fold change* » (FC). Pour ce faire, la moyenne des intensités de chaque feature est calculée pour le groupe des échantillons et pour celui des blancs. Le ratio des moyennes est ensuite calculé (échantillons / blancs). Le feature est supprimé si le ratio est inférieur à un seuil défini (Giacomoni et al., 2015). Ce seuil n'est pas toujours simple à définir car il dépend directement des données (Schiffman et al., 2019). Au sein de la communauté de *Workflow4Metabolomics*, une valeur de 4 est recommandée. Ainsi, un échantillon a en moyenne un signal 4 fois plus intense que les blancs. Ce seuil est proche d'un des critères proposés par Dudzik et al. : un blanc a une intensité un tiers plus faible que celle des échantillons (Dudzik et al., 2018).

Une seconde étape de filtration consiste à supprimer les features instables du jeu de données. Cette étape permet ainsi d'éviter d'identifier comme discriminant un feature dont l'intensité varie aléatoirement. Pour cela, il faut s'appuyer sur l'analyse d'échantillons de contrôle qualité (QC) injectés régulièrement tout au long de la session d'analyse. Ces échantillons QC correspondent souvent à un mélange des échantillons analysés au cours de la session (aussi appelé « *pool* » d'échantillons). Les QC sont de ce fait représentatifs des composés analysés et donc des features détectés. Le coefficient de variation (CV) est

calculé sur les intensités de chaque feature et pour chaque groupe d'échantillons. Un feature est considéré comme instable si le CV est supérieur à un seuil défini pour le groupe des échantillons QC. Différents seuils peuvent être appliqués : Dunn et al. recommandent un seuil à 20 % pour des données UHPLC-HRMS, la FDA propose un seuil à 40 % et Gika et al. conseillent un seuil à 30 % (Dunn et al., 2012 ; FDA, 2001 ; Gika et al., 2014). De plus, il est également recommandé d'analyser également des QC dilués régulièrement au cours de la session. Ainsi, il est possible de vérifier que les intensités des features dans les QC dilués diminuent en fonction du facteur de dilution. Si ce n'est pas le cas, les features peuvent être supprimés (Gagnebin et al., 2017 ; Karaman et al., 2018 ; Pezzatti et al., 2020). Il est également possible de filtrer les features selon leur présence dans le jeu de données : par exemple, Gika et al. recommandent de supprimer les features trouvés dans moins de 20 % des échantillons (Gika et al., 2014), tandis que Dunn et al. préconisent de supprimer ceux trouvés dans moins de 50 % des QC (Dunn et al., 2011).

3.3.2.3. *Correction des effets sessions (ou normalisation)*

Il est important de supprimer d'éventuels effets de dérive analytique, bien connus lors de longues sessions d'analyse telles que celles appliquées en analyses non-ciblées par LC-HRMS (Dunn et al., 2011 ; Gika et al., 2014 ; Broadhurst et al., 2018). Deux effets sessions sont observés : l'effet intra-session et l'effet intersession. Le premier effet correspond à une diminution de l'intensité des échantillons au cours de la session d'analyse, liée à l'encrassement de la source de la MS. Le second effet correspond à des différences d'intensité entre les sessions d'analyse principalement liées aux biais du système (sensibilité du détecteur, encrassement de la source, pureté des éluants utilisés, vieillissement de la colonne) (Dunn et al., 2011). Chaque feature ayant un comportement différent face à ces effets sessions, il est donc recommandé d'appliquer une correction spécifique à chaque feature qui sera la même pour tous les échantillons (Dunn et al., 2011 ; Broadhurst et al., 2018).

La correction de ces effets sessions s'effectue principalement en utilisant les QC. Il existe différents types de QC selon les études. Dans le cas où tous les échantillons de l'étude ont été collectés avant de les analyser, il est possible de faire le « *pool* » de tous échantillons qui pourra alors être utilisé pour corriger les effets intra et inter sessions. Il est également possible de faire un « *pool* » à partir d'un certain nombre d'échantillons représentatifs de

l'étude (Dunn et al., 2011 ; Broadhurst et al., 2018). Dans le cas où tous les échantillons n'ont pas été collectés avant le début des analyses, il est possible de faire un QC à partir d'un mélange d'autres échantillons représentatifs de la matrice étudiée (Broadhurst et al., 2018). Dans le cas d'analyse d'échantillons biologiques (sérum, plasma, urine), il existe également des QC commerciaux qui peuvent être utilisés (Dunn et al., 2011 ; Broadhurst et al., 2018). Pour l'analyse d'échantillons alimentaires, il n'existe pas de tels QC commerciaux, c'est pourquoi les QC utilisés sont un mélange des échantillons étudiés : soit un « *pool* » de tous les échantillons, soit un « *pool* » d'une partie des échantillons, soit un mélange d'échantillons extérieurs à l'étude mais représentatifs.

L'effet intra-session peut être corrigé à partir des QC injectés régulièrement au cours de la session. En effet, ce QC étant un mélange d'échantillons, il contient les features détectés au cours de la session (Dunn et al., 2011 ; Wehrens et al., 2016 ; Broadhurst et al., 2018 ; Dudzik et al., 2018 ; Pezzatti et al., 2020). Ainsi, il est possible de modéliser l'évolution de l'intensité d'un feature sur les QC et d'appliquer une correction sur les échantillons. La modélisation la plus utilisée pour corriger cet effet intra-session est la méthode LOESS (*locally estimated scatterplot smoothing*) (Dunn et al., 2011). La régression LOESS est une méthode de régression non paramétrique, c.à.d. qu'elle n'est pas liée à une équation, et permet ainsi d'obtenir une courbe lissée à partir d'un nuage de points. Dans le cas de la correction intra-session, la régression LOESS permet de déterminer une courbe de correction à partir des valeurs obtenues sur les QC. Pour chaque feature, un polynôme de premier degré linéaire localement est déterminé à partir des différentes injections du QC intra-session et selon leur ordre d'injection. Ce polynôme est ensuite lissé avec une régression des moindres carrés pour corriger les échantillons en fonction de leur ordre d'injection. En effet, les échantillons analysés en fin de session sont souvent plus impactés par cette dérive que ceux analysés en début de session. De plus, les métabolites ne subissent pas les mêmes dérives : certains sont peu impactés, d'autres subissent une augmentation de leur intensité ou une diminution (Dunn et al., 2011 ; Broadhurst et al., 2018). La **Figure 1.16** montre un exemple de l'application de la méthode LOESS implémentée dans la plateforme *Workflow4Metabolomics* (Giacomoni et al., 2014).

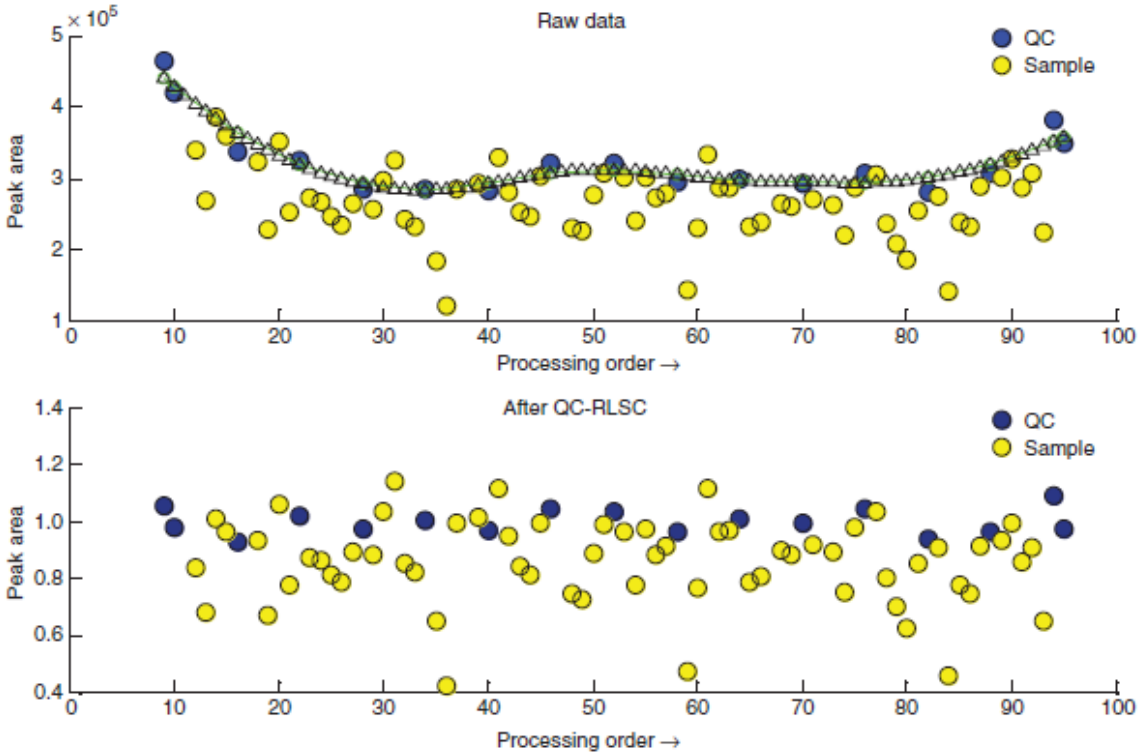

Figure 1.16 *: Normalisation basée sur les QC intra-session pour un feature donné. Les points jaunes représentent les échantillons et les points bleus les différentes injections du QC intra-session. En haut, la courbe LOESS est modélisée (triangle noir) à partir des intensités détectées. En bas, l'intensité de chaque échantillon est corrigée ainsi que la dérive d'intensité observée sur les QC intra-session. Issue de (Dunn et al., 2011)*

L'effet intersession doit ensuite être corrigé lors de l'analyse d'échantillons provenant de différentes sessions analytiques (Dunn et al., 2011 ; Wehrens et al., 2016 ; Broadhurst et al., 2018 ; Dudzik et al., 2018). Il est important de corriger cet effet pour ne pas induire de biais lors de l'analyse statistique des données (c.à.d. en considérant un feature présentant une différence d'intensité entre deux sessions d'analyse comme discriminant lors des analyses statistiques). La correction de l'effet intersession peut se faire de façon identique à la correction intra-session (c.à.d. en utilisant les méthodes précédentes comme la LOESS) à partir du QC analysé lors de chaque session d'analyse et injecté régulièrement pendant chaque session. Selon les études, cet échantillon peut être le même que celui utilisé pour la correction des effets intra-sessions. La **Figure 1.17** montre un exemple de la correction intersession et permet ainsi de visualiser son intérêt.

Cette étape de correction de l'effet intersession peut s'avérer complexe à réaliser. En effet, les features sont caractérisés selon leur *m/z* et leur RT. D'une session à l'autre, du fait d'un manque de répétabilité et de reproductibilité de l'analyse LC-HRMS, un même composé

n'est pas détecté exactement aux mêmes *m/z* et RT (Dunn et al., 2011 ; Dudzik et al., 2018 ; Broadhurst et al., 2018 ; Karaman et al., 2018). De plus, l'utilisation de différents lots de solvants, le vieillissement de la colonne ou les différents lots de colonne utilisés impactent les RT et peuvent donc conduire à des décalages temporels entre les sessions (Dunn et al., 2011 ; Dudzik et al., 2018 ; Broadhurst et al., 2018 ; Karaman et al., 2018).

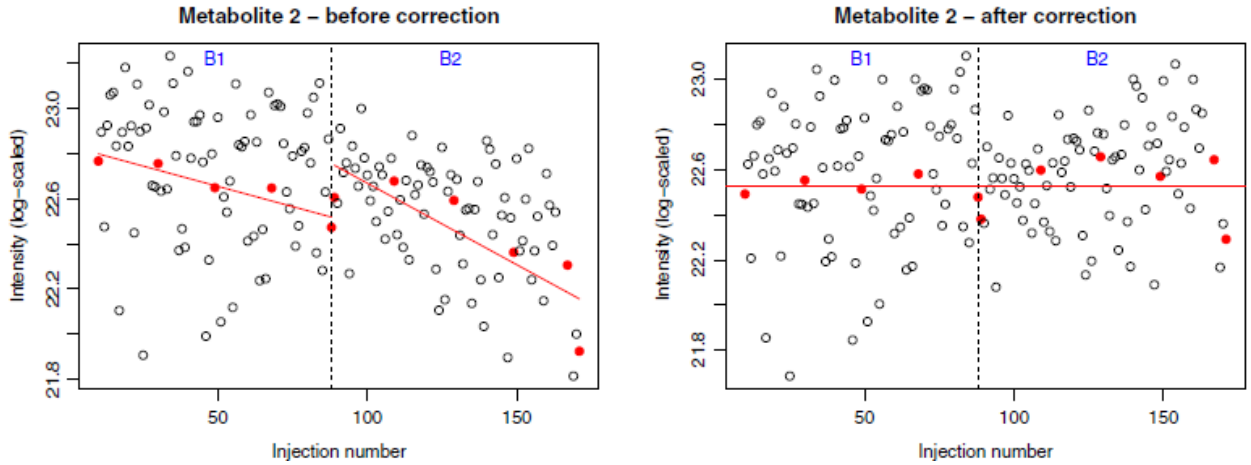

Figure 1.17 *: Représentation d'un feature mesuré dans 2 sessions d'analyse différentes. Les points rouges représentent les QC et les points blancs les échantillons. Les lignes rouges représentent la modélisation de la correction à partir des QC. A gauche, les données non corrigées sont présentées. Les intensités corrigées sont présentées à droite. Issue de (Wehrens et al., 2016)*

Pour pallier ce problème, Dunn et al. proposent une méthodologie utilisant une table de référence. Ainsi, les différentes matrices de données de chaque session d'analyse sont comparées à cette table de référence. Pour cela, des tolérances sont imposées pour le RT et le *m/z* (± 10 secondes et ± 5 ppm) afin de comparer les matrices de données (Dunn et al., 2011). Wehrens et al. proposent également une correction des effets sessions (intra et inter) en utilisant les échantillons et non les QC. En effet, l'utilisation des échantillons permet d'avoir plus de points pour corriger les effets sessions. De plus, dans le cas de métabolites peu concentrés, ils peuvent ne pas être détectés dans les QC, rendant ainsi impossible leur correction ; or, étant détectés dans les échantillons, il est possible de les corriger. Les résultats obtenus en effectuant la correction avec les échantillons sont proches de ceux obtenus en utilisant les QC (Wehrens et al., 2016).

Pour corriger les effets sessions, il est également possible de retraiter en même temps les différentes sessions analytiques. Dans ce cas l'étape d'extraction des données permet de générer un seul jeu de données pour les différentes sessions, et de ce fait, les features

détectés dans les différentes sessions seront facilement comparables (c.à.d. qu'ils auront les mêmes valeurs de *m/z* et de RT). Néanmoins, pour les études agroalimentaires cette procédure n'est pas idéale. En effet, du fait de la variabilité et de la complexité des matrices alimentaires, il est nécessaire d'analyser un grand nombre d'échantillons pour prendre en compte cette variabilité. Dans le cadre d'analyse d'authenticité, les échantillons nécessaires pour représenter celle-ci étant analysés au fur et à mesure, il n'est donc pas possible d'attendre d'avoir toutes les acquisitions pour réaliser le retraitement, et ainsi répondre à la problématique.

3.3.2.4. *Normalisation et mise à l'échelle du jeu de données*

D'autres étapes existent et peuvent être utilisées ou non pour préparer le jeu de données aux traitements statistiques, en particulier les étapes de normalisation et de mise à l'échelle.

La normalisation a pour but de rendre comparables les concentrations des composés présents dans les échantillons en réduisant les éventuels biais liés à la préparation des échantillons (notamment la dilution) (Gagnebin et al., 2017 ; Dudzik et al., 2018). Il existe différentes méthodes de normalisation. Une des plus utilisées pour la métabolomique par LC-HRMS est la PQN (*probabilistic quotient normalization*) (Dieterle et al., 2006). Cette méthode permet de limiter les effets d'une potentielle dilution des échantillons pouvant affecter les données. Pour cela, la médiane de chaque feature dans les QC est calculée et sert de vecteur de référence : les intensités de chaque feature dans les échantillons sont ensuite divisées par ce vecteur de référence. De ce fait, la médiane des ratios pour chaque échantillon peut être déterminée. Les intensités initiales sont ensuite divisées par la médiane des ratios.

Il est souvent effectué une étape de centrage et de mise à l'échelle des données afin de rendre comparables les intensités des features. Pour cela, il existe deux méthodes largement utilisées : la variance unitaire et la méthode Pareto. La première consiste à diviser chaque variable par son écart-type : cela permet ainsi d'avoir toutes les variables d'importance similaire. La méthode Pareto consiste à diviser chaque variable par la raciné carrée de son écart type ; de ce fait, cette méthode permet de garder le ratio d'intensité en augmentant la contribution des variables peu intenses et en diminuant celle des variables

très intenses. Cette méthode est donc souvent préférée car elle reste plus fidèle au jeu de données d'origine (van den Berg et al., 2006).

Certaines données peuvent également nécessiter une transformation logarithmique (van den Berg et al., 2006). Celle-ci permet de réduire la gamme d'intensités d'un spectre. Il est possible d'utiliser un seuil d'intensité au-dessus duquel la transformation logarithmique doit être appliquée, laissant ainsi intactes les intensités inférieures au seuil.

3.3.3. Méthodes chimiométriques appliquées aux données traitées

Afin d'analyser le jeu de données obtenu suite aux étapes d'extraction des données et de prétraitement, des méthodes chimiométriques sont utilisées : les analyses univariées et les analyses multivariées. La différence entre ces deux approches réside dans le nombre de variables impliquées : une seule pour les analyses univariées et plusieurs pour les analyses multivariées. Il est important de préciser qu'un jeu de données contenant plusieurs variables (comme c'est le cas avec les analyses non ciblées) peut être traité par des analyses univariées et multivariées : les résultats de ces méthodes peuvent être vus comme complémentaires. Toutefois en pratique ce sont surtout les méthodes multivariées qui s'avèrent particulièrement intéressantes dans l'analyse de jeux de données issus d'analyses non ciblées. En effet, une seule et unique variable n'est pas suffisante pour discriminer les différents groupes. De plus, avec les méthodes univariées, le taux de faux positifs augmente lorsqu'un grand nombre de variables est analysé (Vinaixa et al., 2012).

On distingue deux types d'analyses multivariées : les analyses supervisées et les analyses non supervisées. Les méthodes supervisées utilisent également des informations qualitatives ou quantitatives. De ce fait, les calculs sont orientés dans le but de chercher les variables liées à la problématique étudiée. Cela permet ainsi d'être en capacité de prédire, à partir des données obtenues, le groupe auquel appartient l'échantillon (on parle alors de modèle de classification) ou de doser certains composés (il s'agit dans ce cas de modèle de régression) (Kemsley et al., 2019 ; Bevilacqua et al., 2017 ; Pilar Callao & Ruisánchez, 2018). A l'inverse, les méthodes non supervisées n'ont connaissance d'aucune information complémentaire, quelle qu'elle soit, sur les échantillons : il n'y a donc aucune orientation des calculs. Il s'agit alors de méthodes exploratoires (Kemsley et al., 2019 ; Bevilacqua et al., 2017 ; Pilar Callao & Ruisánchez, 2018). Ces méthodes non supervisées sont souvent

appliquées en premier lors des études chimiométriques car elles permettent une première observation des données (Bevilacqua et al., 2017).

Ainsi, les différentes méthodes chimiométriques appliquées aux jeux de données obtenus se séparent en deux grandes familles et sont utilisées selon la problématique étudiée : (i) les approches de discrimination pour lesquelles la différence entre deux ou plusieurs groupes d'échantillons est cherchée, (ii) les modèles de prédiction (ou de classification) qui ont pour but de prédire l'appartenance d'un échantillon à un groupe (ou de classer un échantillon dans un groupe) (Kemsley et al., 2019 ; Pilar Callao, & Ruisánchez, 2018).

3.3.3.1. *L'analyse en composante principale (ACP)*

L'ACP est la méthode la plus utilisée dans les analyses non supervisées (Kemsley et al., 2019 ; Bevilacqua et al., 2017 ; Pilar Callao & Ruisánchez, 2018). Il s'agit d'une technique exploratoire permettant d'avoir une première visualisation de l'information présente dans le jeu de données. De plus, cette technique peut être utilisée pour réduire la dimensionnalité du jeu de données étudié tout en conservant un maximum de variance (Kemsley et al., 2019).

Les composantes principales (PC) sont construites dans le but de maximiser la variance entre les échantillons (aussi appelés individus). Pour cela, les composantes sont créées à partir de combinaisons linéaires des variables. Les PC sont générées par ordre décroissant de variance expliquée : la première PC contient donc le maximum de variance expliquée (Pilar Callao & Ruisánchez, 2018).

Les résultats de l'ACP sont contenus dans deux vecteurs : les scores et les coefficients (ou « *loadings* »). L'analyse du graphique des scores permet d'observer la projection des échantillons dans l'espace formé par les composantes générées. Cette visualisation permet ainsi de détecter si une séparation entre les groupes d'échantillons étudiés est visible. La visualisation des « *loadings* » permet d'observer la contribution des variables à la construction de la PC, et de ce fait, déterminer les variables contribuant à la séparation entre les groupes (si celle-ci est visible). La visualisation de l'ACP permet également de détecter la présence d'échantillons aberrants (Kemsley et al., 2019).

3.3.3.2. *Les modèles PLS*

La régression par les moindres carrés (*Partial Least Squares*) ou régression PLS est la méthode la plus connue de régression (Wold et al., 2001). Elle est notamment souvent utilisée sur les données spectrales (données RMN) dans le but de faire de la quantification. En effet, ces données étant très robustes, il est aisé d'utiliser des modèles PLS pour prédire la concentration de variables quantitatives.

La régression PLS a pour but de maximiser la covariance entre le jeu de données (X) et la classe des échantillons (Y). De ce fait, ce modèle permet de maximiser la description de X et la prédiction de Y. La PLS est un modèle qui peut être difficile à interpréter, en particulier quand une structure forte non-corrélée à Y existe dans X.

Afin d'aider à l'interprétation des modèles PLS, une variante a été développé : l'OPLS (*Orthogonal PLS*). L'OPLS décompose l'information en deux composantes : la première corrélée à Y (donc prédictive) et la seconde non corrélée à Y (et de ce fait orthogonale). Il est important de préciser que l'OPLS donne des prédictions identiques à la PLS (Trygg & Wold, 2002).

Différentes variantes de la régression PLS ont depuis été développées, notamment pour des fins de classification : la PLS-DA (*projection to latent structure discriminant analysis*) (Barker & Rayens, 2003) et l'OPLS-DA (*orthogonal PLS-DA*) (Bylesjo et al., 2006). Ces méthodes font partie des analyses supervisées car elles nécessitent la connaissance des groupes d'appartenance des échantillons pour construire le modèle (c.à.d. la matrice Y). La PLS-DA est la méthode supervisée la plus utilisée (Kemsley et al., 2019 ; Bevilacqua et al., 2017 ; Pilar Callao & Ruisánchez, 2018). Ces techniques sont largement utilisées dans les analyses métabolomiques. En effet, elles permettent la création de modèle de prédiction afin de s'assurer de l'authenticité de l'échantillon. Comme précédemment, ces modèles donnent les mêmes prédictions, mais l'OPLS-DA permet d'obtenir des modèles plus simples à interpréter.

Les modèles obtenus doivent être validés car ils peuvent être sujets à des surapprentissages (ou « *overfitting* »). C'est un phénomène très courant qu'il est important d'éviter sous peine d'obtenir des modèles mauvais en prédiction de futurs échantillons. Il existe deux types de validation : la validation croisée (ou validation interne), et la validation

externe (réalisée sur un nouveau jeu de données). La validation croisée est une première étape de validation simple à mettre en œuvre sur un jeu de données. En effet, celle-ci consiste à séparer le jeu de données aléatoirement en K blocs de taille égale. Le modèle est ensuite calibré sur K-1 blocs (il s'agit du jeu de calibration) et est ensuite testé sur le bloc restant (le jeu de validation). Cette étape est répétée pour que chaque bloc soit utilisé en jeu de validation. De ce fait, K modèles sont ainsi construits et les performances de chacun sont évaluées. Afin d'éliminer un possible effet lié au hasard, cette procédure est répétée N fois. Le modèle est considéré comme performant s'il présente des résultats stables sur les N itérations (Kemsley et al., 2019 ; Bevilacqua et al., 2017 ; Pilar Callao & Ruisánchez, 2018).

Lors de la validation des modèles, une matrice de confusion est générée (voir **Tableau 1.4** ci-dessous) permettant d'avoir un aperçu de la performance des modèles sur de nouveaux échantillons.

***Tableau 1.4** : Illustration d'une matrice de confusion obtenue lors de la validation d'un modèle.*

		Classes réelles	
		Classe A	Classe B
Classes prédites	Classe A	Vrai positif (TP, *true positive*)	Faux négatif (FN, *false negative*)
	Classe B	Faux positif (FP, *false positive*)	Vrai négatif (TN, *true negative*)

A partir de cette matrice de confusion, différents indicateurs peuvent être calculés pour attester de la performance du modèle : la sensibilité (Équation 1), la spécificité (Équation 2), la précision (Équation 3), ainsi que le nombre de mauvaises classifications (NMC, *number of misclassifications*) (Équation 4) (Ballabio & Consonni, 2013 ; Riedl et al., 2015 ; Szymanska et al., 2012).

$$Sensibilité : \frac{TP}{TP + FN} \qquad \textit{(Équation 1)}$$

$$Spécificité : \frac{TN}{TN + FP} \qquad \textit{(Équation 2)}$$

$$Précision : \frac{TP + TN}{TP + FN + FP + FN} \qquad \textit{(Équation 3)}$$

$$NMC = FN + FP \qquad \textit{(Équation 4)}$$

Lors de la validation, sont estimées la somme des carrés résiduelles (SS - *Sum of Squares*) et la somme des carrés des erreurs résiduelles prédites (PRESS - *PRediction Error Sum of Squares*). La performance des modèles (O)PLS-DA est alors caractérisée par d'autres indicateurs :

- R^2X (ou R^2Y) représente la fraction de SS expliquée pour X (ou Y) par une composante (cette grandeur peut être exprimée sous forme cumulative pour rendre compte d'un modèle à plusieurs composantes) ;
- Q^2Y exprime la part des variations exprimée par une composante (ou plusieurs pour son pendant cumulatif) ; $Q^2 = 1 - PRESS/SS$.

En pratique la valeur de R^2Y représente l'ajustement du modèle, et celle de Q^2Y la capacité de prédiction du modèle (Bevilacqua et al., 2017).

3.3.3.3. *L'analyse de variance (ANOVA)*

L'ANOVA (*analysis of variance*) est une méthode univariée classiquement utilisée. En fixant un critère sur la p-value du test de Fisher de l'ANOVA, cette méthode permet de déterminer si la variable est pertinente pour répondre à un problème selon la valeur de p-value obtenue (Stahle & Wold, 1989).

Lors de l'utilisation de l'ANOVA sur un jeu de données issues d'une approche métabolomique, il faut faire attention à bien appliquer une correction pour des tests multiples. En effet, ces jeux de données peuvent contenir des milliers de variables (des dizaines de milliers dans le cas d'analyse LC-HRMS). De ce fait, un risque de faux positifs est important avec un aussi grand nombre de variables. Il est donc nécessaire d'appliquer une correction (Vinaixa et al., 2012).

3.3.3.4. *Sélection de variables : VIP*

Suite à la génération de modèles type PLS (ou variantes), il est possible de sélectionner les variables ayant le plus d'influence dans la création de ces modèles. Cette sélection se fait sur la base du VIP (*variable importance in the projection*) qui est calculé pour chaque variable (Wold et al., 2001).

Pour une variable, plus sa valeur de VIP est importante, plus celle-ci contribue à l'élaboration du modèle. Cette variable peut donc être considérée comme discriminante. Ainsi, il est important de sélectionner les variables ayant de grands VIP. Il est couramment utilisé un seuil de 1 pour le VIP (c.à.d. que les variables ayant une VIP supérieure à 1 sont sélectionnées).

3.3.3.5. *Biosigner*

En 2016, Rinaudo et al. ont mis en place un nouvel outil pour la sélection de variables nommé *biosigner*. Cet outil permet de déterminer le nombre minimal de variables ayant un maximum de contribution dans la performance de différents modèles testés. Le module *biosigner* s'appuie sur 3 modèles chimiométriques pour sélectionner les variables les plus significatives : la PLS-DA, les forêts aléatoires (*random forest*) et la machine à support de vecteur (*support vector machine*, SVM). Ainsi, l'algorithme de *biosigner* réalise différentes itérations pour créer ces différents modèles. A chaque itération, *biosigner* ne garde que les variables contribuant le plus à la performance des modèles. Les itérations s'arrêtent lorsque le nombre de variables significatives est identique à celui obtenu à l'itération précédente (Rinaudo et al., 2016).

3.4. CONCLUSIONS

Cette méthodologie d'analyse combinant acquisition d'une empreinte globale et traitements statistiques des données semble être prometteuse pour les analyses de contrôle alimentaire et notamment pour des problématiques d'authenticité. Elle est d'ailleurs déjà implémentée pour des analyses de routine au sein de l'unité Profiling NMR du laboratoire Eurofins Analytics France.

Les outils chimiométriques appliqués sur des analyses non ciblées permettent d'obtenir différentes informations. Il est possible de définir les variables permettant la discrimination entre plusieurs groupes d'échantillons. De ce fait, des signaux marqueurs d'authenticité peuvent être identifiés. Si l'analyse est réalisée par LC-HRMS, il est même envisageable d'identifier le composé (ou *a minima* de l'annoter). Il est également possible de détecter la présence de signaux suspects parmi un groupe d'échantillons, ce qui peut permettre de détecter d'éventuelles fraudes. Les outils chimiométriques permettent également de mettre en place des modèles permettant la prédiction d'échantillons, comme les modèles (O)PLS-

DA. Si l'analyse est réalisée par RMN, il est possible de quantifier certains signaux *via* la mise en place de modèles de régression (c.à.d. des modèles PLS).

Bien que ces nouvelles approches d'analyse soient prometteuses et permettent d'obtenir des informations supplémentaires, il ne faut pas pour autant oublier les méthodes conventionnelles. Ces méthodes ciblées sont utilisées en routine dans certains laboratoires d'analyses, d'autant qu'il s'agit des méthodes officielles de contrôle d'authenticité. Ces méthodes peuvent aussi être utilisées pour aider à la validation de la nouvelle méthodologie non ciblée présentée ici.

Des travaux doivent encore être menés, notamment sur le traitement des données pour la LC-HRMS. En effet, bien que de nombreuses études montrent l'intérêt de cette technique pour répondre à des problématiques d'authenticité des denrées alimentaires, ces études sont faites sur peu de sessions d'analyse (souvent une seule). De ce fait, les traitements de données peuvent être appliqués directement sur l'intégralité des échantillons analysés. Cependant, les problématiques d'authenticité nécessitent d'analyser un grand nombre d'échantillons (dans différentes sessions analytiques). Le traitement des données doit donc être en capacité de combiner différentes sessions d'analyse, or il est pour le moment impossible d'effectuer cette combinaison. Par conséquent, il est à ce jour impossible d'implémenter cette méthodologie en routine au laboratoire.

3.5. EN RESUME

Dans cette partie, la méthodologie d'analyse non ciblée par RMN et par LC-HRMS a été présentée. Quelques exemples d'application de ces méthodes non ciblées par RMN et LC-HRMS montrent l'intérêt d'appliquer ce type de méthodologie pour le contrôle d'authenticité des denrées alimentaires.

L'application de la RMN non ciblée est déjà mise en place en routine notamment grâce au développement du JuiceScreener™ et du WineScreener™ utilisés par les laboratoires d'Eurofins Analytics France. Néanmoins, cette méthodologie n'est pas encore reconnue comme méthode officielle de contrôle d'authenticité. Il n'y a, à ce jour, pas d'équivalent en LC-HRMS.

Le traitement des données non ciblées obtenues par RMN et par LC-HRMS a été présenté. Les principales étapes du traitement sont similaires pour ces deux techniques d'analyse : (i) préparation des échantillons, (ii) acquisition des données, (iii) prétraitement des données et (iv) analyse chimiométrique.

Les analyses chimiométriques se divisent en deux catégories : les analyses non supervisées et les analyses supervisées. Les premières n'ont aucune connaissance *a priori* des groupes auxquels appartiennent les échantillons, tandis que les dernières nécessitent de connaitre les groupes d'appartenance des échantillons. Les méthodes supervisées permettent de faire de la classification (ou prédiction d'échantillons).

4. CONCLUSION DE L'ETUDE BIBLIOGRAPHIQUE

Dans la première partie de ce chapitre, nous avons vu l'importance des contrôles d'authenticité des denrées alimentaires et leurs enjeux analytiques. Le cas particulier des jus de fruits a également été présenté. Cette matrice est particulièrement intéressante à considérer dans le cadre du contrôle d'authenticité du fait de sa grande variabilité (variété, origine géographique, mode de production, procédé de transformation, composition chimique). De plus, elle est sujette à de nombreux types de fraude (comme détaillé dans le paragraphe 1.5.2.). C'est la raison pour laquelle un volet important de ce travail de thèse a été consacré à l'analyse de jus de fruits (jus de citron jaune, et jus de pommes).

Les analyses de contrôle d'authenticité sont actuellement principalement réalisées par des méthodes conventionnelles dites « ciblées ». Celles-ci visent à détecter et quantifier des composés ou des familles de composés connus et marqueurs d'authenticité. Ces méthodes sont généralement décrites et validées par différents organismes spécifiques (par exemple l'IFU pour les jus de fruits). Les techniques majoritairement utilisées pour ce type de méthodes d'analyse sont la RMN et la LC-MS. Ces deux techniques présentent des avantages et des inconvénients (présentés en **Figure 1.10** et **Figure 1.11** pour la RMN et la LC-MS respectivement) qui les rendent complémentaires. Par conséquent, elles sont souvent toutes deux utilisées dans les laboratoires de contrôle d'authenticité. Ces méthodes ciblées sont sensibles et spécifiques aux composés analysés, permettant ainsi de répondre

avec une grande certitude sur l'authenticité de l'échantillon. Néanmoins, ces méthodes peuvent faillir à la détection de fraudes plus sophistiquées. De plus, il est essentiel de connaître en amont les composés marqueurs ; or, comme présenté dans le paragraphe 1.4.2., beaucoup de marqueurs d'authenticité sont à ce jour encore inconnus. C'est pourquoi il est nécessaire de développer des approches non ciblées permettant d'obtenir une empreinte globale de l'échantillon pour confirmer son authenticité et détecter d'éventuelles fraudes.

Les méthodologies d'analyse non ciblée ont été présentées dans une troisième partie de ce chapitre. Ce type d'approche repose majoritairement sur deux techniques que sont la RMN et la LC-HRMS. Ces méthodologies sont inspirées des études métabolomiques et permettent d'observer un échantillon dans sa globalité, générant ainsi une empreinte. Différentes études utilisant cette méthodologie montrent que celle-ci a su se faire une place dans les analyses de contrôle d'authenticité. Cette approche combine une méthode d'analyse exhaustive et des outils chimiométriques afin de détecter les différences entre plusieurs groupes d'échantillons (par exemple un groupe contrôle et un groupe traité). Les grandes étapes de ces méthodologies non ciblées sont présentées en **Figure 1.12** et en **Figure 1.13** pour la RMN et la LC-HRMS respectivement. Dans les deux cas, l'une des étapes critiques de cette méthodologie est la préparation des échantillons. En effet, cette étape doit rester la plus simple possible pour obtenir une empreinte la plus fidèle possible, et éviter la perte de certains composés d'intérêt. Le traitement des données constitue une étape complexe de cette méthodologie et est également une étape critique. En effet, au cours de cette étape, les données sont généralement filtrées pour ne garder que l'information la plus pertinente. Les données résultantes de cette étape seront ensuite analysées par des outils chimiométriques dans le but de trouver d'éventuelles différences. De ce fait, si les données sont mal traitées, cela va impacter directement le résultat des analyses chimiométriques. En RMN, cette méthodologie a fait ses preuves et certaines méthodes sont aujourd'hui des méthodes de contrôle utilisées en routine (WineScreener™ et JuiceScreener™). Il existe quelques études utilisant une approche non ciblée par LC-HRMS pour répondre à différentes problématiques d'authenticité, prouvant ainsi l'intérêt de cette technique. L'avantage majeur de la LC-HRMS sur la RMN est notamment la séparation des constituants de l'échantillon en amont de leur détection. Cet avantage est notamment

très intéressant pour pouvoir identifier le ou les composés qui permettent de distinguer les différents groupes d'échantillons.

Bien que la méthodologie par RMN soit utilisée en routine pour le contrôle d'authenticité, ce n'est pas encore le cas de la LC-HRMS où il faut encore travailler sur le traitement des données. C'est pourquoi, les travaux de cette thèse sont focalisés sur la LC-HRMS. En particulier, la LC-HRMS souffrant de problèmes de répétabilité (variations des valeurs *m/z* et des RT au cours du temps), la problématique du traitement de données acquises dans différentes sessions d'analyse constitue un verrou scientifique. Dans le cas des contrôles d'authenticité, du fait de la variabilité interne aux échantillons, il est souvent nécessaire d'analyser un grand nombre d'échantillons, et ce sur différentes sessions d'analyse. Or, à ce jour, les traitements de données décrits dans différents articles scientifiques s'intéressent principalement à une seule session d'analyse. Cette problématique du traitement des données acquises en intra- et inter-sessions sera donc abordée dans ce travail de thèse.

5. CHOIX METHODOLOGIQUES

Les travaux de cette thèse se sont découpés en différentes étapes : (i) développement d'une méthode d'analyse ciblée, (ii) développement d'une méthode d'analyse non ciblée, et (iii) automatisation du traitement des données non ciblées acquises par LC-HRMS.

Une méthode d'analyse ciblée pour l'authentification du jus de citron jaune a dans un premier temps été développée. Ce développement s'inscrit dans un projet inter laboratoires visant à vérifier la validité de certains composés comme marqueurs d'espèces de citrus ainsi que définir des valeurs seuils pour garantir l'authenticité des échantillons. Cette méthode a été développée selon la norme NF V03-110 et a été implémentée en routine dans le service Chromatographie du laboratoire d'Eurofins Analytics France. Ce dernier a été accrédité par le COFRAC pour la mise en œuvre de cette méthode dans ce champ d'application.

Une méthode d'analyse non ciblée par LC-HRMS a ensuite été mise au point. Cette méthode a tout d'abord été développée sur des jus de pommes avec une approche *dilute and shoot* (dilution et analyse), permettant ainsi d'avoir une préparation d'échantillons très minime. Le traitement des données obtenues a été réalisé à l'aide d'outils en ligne sur la plateforme W4M. La méthodologie mise en place a été testée sur deux scénarios d'analyse

d'authenticité : la discrimination des purs jus et des jus concentrés, et la discrimination des jus issus de l'agriculture biologique de ceux issus de l'agriculture conventionnelle.

Cette méthode non ciblée a ensuite été appliquée sur des échantillons de carottes. Cela a ainsi permis de tester la méthodologie développée sur un autre type de matrice, qui nécessite notamment des étapes plus complexes de préparation d'échantillons. A nouveau, deux scénarios d'analyse d'authenticité ont été testés : l'authentification de l'origine géographique des échantillons, et l'authentification des carottes issues de l'agriculture biologique. En parallèle, un début d'automatisation du traitement des données a été effectué, en vue d'implémenter en routine au laboratoire ce type de méthodologie. Ainsi, les différentes étapes du traitement des données ont été mises en place sous un logiciel en interne.

6. REFERENCES

AIJN Code of Practice, AIJN European Fruit Juice Association, 2020.

Balci, M., 2005. Basic 1H- and 13C-NMR spectroscopy, 1st ed. ed. Elsevier, Amsterdam.

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods 5*, 3790. https://doi.org/10.1039/c3ay40582f

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. *Journal of Chemometrics 17*, 166–173. https://doi.org/10.1002/cem.785

Bevilacqua, M., Bro, R., Marini, F., Rinnan, Å., Rasmussen, M.A., Skov, T., 2017. Recent chemometrics advances for foodomics. *TrAC Trends in Analytical Chemistry 96*, 42–51. https://doi.org/10.1016/j.trac.2017.08.011

Bohme, K., Calo-Mata, P., Barros-Velázquez, J., Ortea, I., 2019. Recent applications of omics-based technologies to main topics in food authentication. *TrAC Trends in Analytical Chemistry 110*, 221–232. https://doi.org/10.1016/j.trac.2018.11.005

Broadhurst, D., Goodacre, R., Reinke, S.N., Kuligowski, J., Wilson, I.D., Lewis, M.R., Dunn, W.B., 2018. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics 14*, 72. https://doi.org/10.1007/s11306-018-1367-3

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J.K., Holmes, E., Trygg, J., 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics 20*, 341–351. https://doi.org/10.1002/cem.1006

Cavanna, D., Loffi, C., Dall'Asta, C., Suman, M., 2020. A non-targeted high-resolution mass spectrometry approach for the assessment of the geographical origin of durum wheat. *Food Chemistry 317*, 126366. https://doi.org/10.1016/j.foodchem.2020.126366

Cifuentes, A., 2009. Food analysis and Foodomics. *Journal of Chromatography A 1216*, 7109. https://doi.org/10.1016/j.chroma.2009.09.018

Commission Européenne, 2002. RÈGLEMENT (CE) No 178/2002 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 28 janvier 2002 établissant les principes généraux et les prescriptions générales de la législation alimentaire, instituant l'Autorité européenne de sécurité des aliments et fixant des procédures relatives à la sécurité des denrées alimentaires, *Journal Officiel de l'Union Européenne*, 1-42, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2002:031:TOC

Commission Européenne, Direction générale de la santé et de la sécurité alimentaire, 2020. The EU food fraud network and the administrative assistance and cooperation system : 2019 annual report, Publications Office. https://data.europa.eu/doi/10.2875/326318

Cubero-Leon, E., Peñalver, R., Maquet, A., 2014. Review on metabolomics for food authentication. *Food Research International 60*, 95–107. https://doi.org/10.1016/j.foodres.2013.11.041

Cubero-Leon, E., De Rudder, O., Maquet, A., 2018. Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry 239*, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

Danezis, G.P., Tsagkaris, A.S., Camin, F., Brusic, V., Georgiou, C.A., 2016. Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry 85*, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Dasenaki, M.E., Drakopoulou, S.K., Aalizadeh, R., Thomaidis, N.S., 2019. Targeted and Untargeted Metabolomics as an Enhanced Tool for the Detection of Pomegranate Juice Adulteration. *Foods 8*, 212. https://doi.org/10.3390/foods8060212

De Vijlder, T., Valkenborg, D., Lemière, F., Romijn, E.P., Laukens, K., Cuyckens, F., 2018. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrometry Reviews 37*, 607–629. https://doi.org/10.1002/mas.21551

Delaporte, G., Cladière, M., Jouan-Rimbaud Bouveresse, D., Camel, V., 2019. Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study on tea. *Food Chemistry 277*, 54–62. https://doi.org/10.1016/j.foodchem.2018.10.089

Diaz, R., Pozo, O.J., Sancho, J.V., Hernández, F., 2014. Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry 157*, 84–93. https://doi.org/10.1016/j.foodchem.2014.02.009

Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., 2006. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1 H NMR Metabonomics. *Analytical Chemistry 78*, 4281–4290. https://doi.org/10.1021/ac051632c

Directive 2003/115/CE, 2003. Directive 2003/115/CE du Parlement européen et du Conseil du 22 décembre 2003 modifiant la directive 94/35/CE concernant les édulcorants destinés à être employés dans les denrées alimentaires, *Journal Officiel de l'Union Européenne*, 65–71, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32003L0115

Directive 2012/12/CE, 2012. Directive 2012/12/UE du Parlement européen et du Conseil du 19 avril 2012 modifiant la directive 2001/112/CE du Conseil relative aux jus de fruits et à certains produits similaires destinés à l'alimentation humaine, *Journal Officiel de l'Union Européenne*, 1–11, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32012L0012

Dubin, E., Dumas, A.-S., Rebours, A., Jamin, E., Ginet, J., Lees, M., Rutledge, D.N., 2017. Detection of Blackcurrant Adulteration by Aronia Berry Using High Resolution Mass Spectrometry, Variable Selection and Combined PLS Regression Models. *Food Analytical Methods 10*, 683–693. https://doi.org/10.1007/s12161-016-0638-8

Dudzik, D., Barbas-Bernardos, C., García, A., Barbas, C., 2018. Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis 147*, 149–173. https://doi.org/10.1016/j.jpba.2017.07.044

Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N., Nicholls, A.W., Wilson, I.D., Kell, D.B., Goodacre, R., 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols 6*, 1060–1083. https://doi.org/10.1038/nprot.2011.335

Ellis, D.I., Brewster, V.L., Dunn, W.B., Allwood, J.W., Golovanov, A.P., Goodacre, R., 2012. Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chem. Soc. Rev. 41*, 5706. https://doi.org/10.1039/c2cs35138b

Emwas, A.-H., Saccenti, E., Gao, X., McKay, R.T., dos Santos, V.A.P.M., Roy, R., Wishart, D.S., 2018. Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine. *Metabolomics 14*, 31. https://doi.org/10.1007/s11306-018-1321-4

EU Food Fraud Network, 2021. https://ec.europa.eu/food/safety/agri-food-fraud/eu-food-fraud-network_fr#about-the-food-fraud-network consulté le 05/07/2021

Europol-Interpol, 2020 : https://www.europol.europa.eu/activities-services/europol-in-action/operations/operation-opson consulté le 27/07/2021

Everstine, K., Spink, J., Kennedy, S., 2013. Economically Motivated Adulteration (EMA) of Food: Common Characteristics of EMA Incidents. *J. Food Prot. 76*, 13. https://doi.org/10.4315/0362-028X.JFP-12-399

FAO and WHO. 2019. *Codex Alimentarius Commission – Procedural Manual twenty-seventh edition*. Rome. 254 pp.

FAO/WHO, 2021 : http://www.fao.org/fao-who-codexalimentarius/fr/ consulté le 29/06/2021

Farag, M.A., Labib, R.M., Noleto, C., Porzel, A., Wessjohann, L.A., 2018. NMR approach for the authentication of 10 cinnamon spice accessions analyzed via chemometric tools. *LWT 90*, 491–498. https://doi.org/10.1016/j.lwt.2017.12.069

FDA, US Food and Drug Administration, 2021. https://www.fda.gov/ consulté le 29/06/2021

Gagnebin, Y., Tonoli, D., Lescuyer, P., Ponte, B., de Seigneux, S., Martin, P.-Y., Schappler, J., Boccard, J., Rudaz, S., 2017. Metabolomic analysis of urine samples by UHPLC-QTOF-MS: Impact of normalization strategies. *Analytica Chimica Acta 955*, 27–35. https://doi.org/10.1016/j.aca.2016.12.029

Gattuso, G., Barreca, D., Gargiulli, C., Leuzzi, U., Caristi, C., 2007. Flavonoid composition of citrus juices. *Molecules 12*, 1641–1673. https://doi.org/10.3390/12081641

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... Caron, C. 2015. Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31*(9), 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Gika, H.G., Theodoridis, G.A., Plumb, R.S., Wilson, I.D., 2014. Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis 87*, 12–25. https://doi.org/10.1016/j.jpba.2013.06.032

Godelmann, R., Fang, F., Humpfer, E., Schütz, B., Bansbach, M., Schäfer, H., Spraul, M., 2013. Targeted and Nontargeted Wine Analysis by 1 H NMR Spectroscopy Combined with Multivariate Statistical Analysis. Differentiation of Important Parameters: Grape Variety, Geographical Origin, Year of Vintage. *Journal of Agricultural and Food Chemistry 61*, 5610–5619. https://doi.org/10.1021/jf400800d

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends in Analytical Chemistry 82*, 425–442. https://doi.org/10.1016/j.trac.2016.07.004

Guitton, Y., Tremblay-Franco, M., Le Corguillé, G., Martin, J.-F., Pétéra, M., Roger-Mele, P., Delabrière, A., Goulitquer, S., Monsoor, M., Duperier, C., Canlet, C., Servien, R., Tardivel, P., Caron, C., Giacomoni, F., Thévenot, E.A., 2017. Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *The International Journal of Biochemistry & Cell Biology 93*, 89–101. https://doi.org/10.1016/j.biocel.2017.07.002

Guyader, S., Thomas, F., Portaluri, V., Jamin, E., Akoka, S., Silvestre, V., Remaud, G., 2018. Authentication of edible fats and oils by non-targeted 13 C INEPT NMR spectroscopy. *Food Control 91*, 216–224. https://doi.org/10.1016/j.foodcont.2018.03.046

IFU International Fruit and Vegetable Juice Association, 2021. https://ifu-fruitjuice.com/ consulté le 02/07/2021

IHC International Honey Commission, 2021. https://www.ihc-platform.net/ consulté le 08/07/2021

Jamin, E., Martin, F., & Martin, G. G. 2004. Determination of the 13C/12C ratio of ethanol derived from fruit juices and maple syrup by isotope ratio mass spectrometry: Collaborative study. *Journal of AOAC International, 87*(3), 621–631. https://doi.org/10.1093/jaoac/87.3.621

Jandrić, Z., Roberts, D., Rathor, M.N., Abrahim, A., Islam, M., Cannavan, A., 2014. Assessment of fruit juice authenticity using UPLC–QToF MS: A metabolomics approach. *Food Chemistry 148*, 7–17. https://doi.org/10.1016/j.foodchem.2013.10.014

Jandrić, Z., Cannavan, A., 2017. An investigative study on differentiation of citrus fruit/fruit juices by UPLC-QToF MS and chemometrics. *Food Control 72*, 173–180. https://doi.org/10.1016/j.foodcont.2015.12.031

Jellema, R.H., 2009. Variable Shift and Alignment, in: Comprehensive Chemometrics. Elsevier, pp. 85–108. https://doi.org/10.1016/B978-044452701-1.00104-6

Karaman, I., Climaco Pinto, R., Graça, G., 2018. Metabolomics Data Preprocessing: From Raw Data to Features for Statistical Analysis, in: Comprehensive Analytical Chemistry. Elsevier, pp. 197–225. https://doi.org/10.1016/bs.coac.2018.08.003

Kemsley, E.K., Defernez, M., Marini, F., 2019. Multivariate statistics: Considerations and confidences in food authenticity problems. *Food Control 105*, 102–112. https://doi.org/10.1016/j.foodcont.2019.05.021

Knolhoff, A.M., Croley, T.R., 2016. Non-targeted screening approaches for contaminants and adulterants in food using liquid chromatography hyphenated to high resolution mass spectrometry. *Journal of Chromatography A 1428*, 86–96. https://doi.org/10.1016/j.chroma.2015.08.059

Laghi, L., Picone, G., Capozzi, F., 2014. Nuclear magnetic resonance for foodomics beyond food analysis. *TrAC Trends in Analytical Chemistry 59*, 93–102. https://doi.org/10.1016/j.trac.2014.04.009

Lehnert, N., & Ara, V. 2014. Authenticity analysis of lemon juices concerning the adulteration lime. *Fruit Processing*, 242–248.

Lopez-Ruiz, R., Romero-González, R., Garrido Frenich, A., 2019. Metabolomics approaches for the determination of multiple contaminants in food. *Current Opinion in Food Science 28*, 49–57. https://doi.org/10.1016/j.cofs.2019.08.006

Luykx, D.M.A.M., van Ruth, S.M., 2008. An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry 107*, 897–911. https://doi.org/10.1016/j.foodchem.2007.09.038

Mannina, L., Sobolev, A.P., Viel, S., 2012. Liquid state 1H high field NMR in food analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy 66*, 1–39. https://doi.org/10.1016/j.pnmrs.2012.02.001

Martin, G.J., & Martin, M.L. 1981. Deuterium labelling at the natural abundance level as studied by high field quantitative 2H NMR. *Tetrahedron Letters, 22*(36), 3525-3528. https://doi.org/10.1016/S0040-4039(01)81948-1

Medina, S., Pereira, J.A., Silva, P., Perestrelo, R., Câmara, J.S., 2019a. Food fingerprints – A valuable tool to monitor food authenticity and safety. *Food Chemistry 278*, 144–162. https://doi.org/10.1016/j.foodchem.2018.11.046

Medina, S., Perestrelo, R., Silva, P., Pereira, J.A.M., Câmara, J.S., 2019b. Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology 85*, 163–176. https://doi.org/10.1016/j.tifs.2019.01.017

Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T., Maestroni, B.M., Cannavan, A., 2021. High-resolution mass spectrometry-based metabolomics for the discrimination between organic and conventional crops: A review. *Trends in Food Science & Technology 110*, 142–154. https://doi.org/10.1016/j.tifs.2021.01.071

Ministère de l'agriculture et de l'alimentation, 2020 : https://agriculture.gouv.fr/la-reglementation-sur-lhygiene-des-aliments consulté le 26/07/2021

Monakhova, Y.B., Ruge, W., Kuballa, T., Ilse, M., Winkelmann, O., Diehl, B., Thomas, F., Lachenmeier, D.W., 2015. Rapid approach to identify the presence of Arabica and Robusta species in coffee using 1H NMR spectroscopy. *Food Chemistry 182*, 178–184. https://doi.org/10.1016/j.foodchem.2015.02.132

Moore, J.C., Spink, J., Lipp, M., 2012. Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *Journal of Food Science 77*, R118–R126. https://doi.org/10.1111/j.1750-3841.2012.02657.x

OIV Organisation internationale de la vigne et du vin, 2021. https://www.oiv.int/fr/organisation-internationale-de-la-vigne-et-du-vin Consulté le 08/07/2021

Ortea, I., O'Connor, G., Maquet, A., 2016. Review on proteomics for food authentication. *Journal of Proteomics 147*, 212–225. https://doi.org/10.1016/j.jprot.2016.06.033

Pacholczyk-Sienicka, B., Ciepielowski, G., Albrecht, Ł., 2021. The Application of NMR Spectroscopy and Chemometrics in Authentication of Spices. *Molecules 26*, 382. https://doi.org/10.3390/molecules26020382

Petrakis, E.A., Cagliani, L.R., Polissiou, M.G., Consonni, R., 2015. Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by 1H NMR metabolite fingerprinting. *Food Chemistry 173*, 890–896. https://doi.org/10.1016/j.foodchem.2014.10.107

Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., González-Ruiz, V., Rudaz, S., 2020. Implementation of liquid chromatography–high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: A tutorial. *Analytica Chimica Acta 1105*, 28–44. https://doi.org/10.1016/j.aca.2019.12.062

Pilar Callao, M., Ruisánchez, I., 2018. An overview of multivariate qualitative methods for food fraud detection. *Food Control 86*, 283–293. https://doi.org/10.1016/j.foodcont.2017.11.034

Règlement (UE) 2017/625, 2021. Règlement (UE) 2017/625 du Parlement européen et du Conseil du 15 mars 2017 concernant les contrôles officiels et les autres activités officielles servant à assurer le respect de la législation alimentaire et de la législation relative aux aliments pour animaux ainsi que des règles relatives à la santé et au bien-être des animaux, à la santé des végétaux et aux produits phytopharmaceutiques, *Journal Officiel de l'Union Européenne*, 1–142. https://eur-lex.europa.eu/eli/reg/2017/625/oj

Remaud, G.S., Martin, Y.-L., Martin, G.G., Martin, G.J., 1997. Detection of sophisticated adulterations of natural vanilla flavors and extracts: application of the SNIF-NMR method to vanillin and p-hydroxybenzaldehyde. *Journal of Agricultural and food Chemistry 45*, 859–866. https://doi.org/10.1021/jf960518f

Riedl, J., Esslinger, S., Fauhl-Hassek, C., 2015. Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta 885*, 17–32. https://doi.org/10.1016/j.aca.2015.06.003

Rinaudo, P., Boudah, S., Junot, C., Thévenot, E.A., 2016. biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences 3*. https://doi.org/10.3389/fmolb.2016.00026

Rinke P., 2016. Tradition Meets High Tech for Authenticity Testing of Fruit Juices. In G. Downey (Ed.), Advances in Food Authenticity Testing (pp.625-665). Woodhead Publishing is an imprint of Elsevier

Rinke, P., Jamin, E., 2018. Fruit juices, in: Morin, J.-F., Lees, M. (Eds.), FoodIntegrity Handbook. Eurofins Analytics France, pp. 243–264. https://doi.org/10.32741/fihb.14.juices

Rubert, J., Lacina, O., Zachariasova, M., Hajslova, J., 2016. Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chemistry 204*, 201–209. https://doi.org/10.1016/j.foodchem.2016.01.003

Schiffman, C., Petrick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rappaport, S., Dudoit, S., 2019. Filtering procedures for untargeted LC-MS

metabolomics data. *BMC Bioinformatics 20*, 334. https://doi.org/10.1186/s12859-019-2871-9

Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., Hollender, J., 2014. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology 48*, 2097–2098. https://doi.org/10.1021/es5002105

Sharma, K., Paradakar, M., 2010. The melamine adulteration scandal. *Food Sec. 2*, 97–107. https://doi.org/10.1007/s12571-009-0048-5

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry 78*, 779–787. https://doi.org/10.1021/ac051437y

Sobolev, A.P., Thomas, F., Donarski, J., Ingallina, C., Circi, S., Cesare Marincola, F., Capitani, D., Mannina, L., 2019. Use of NMR applications to tackle future food fraud issues. *Trends in Food Science & Technology 91*, 347–353. https://doi.org/10.1016/j.tifs.2019.07.035

Sousa, S.A.A., Magalhães, A., Ferreira, M.M.C., 2013. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems 122*, 93–102. https://doi.org/10.1016/j.chemolab.2013.01.006

Spink, J., Bedard, B., Keogh, J., Moyer, D.C., Scimeca, J., Vasan, A., 2019. International Survey of Food Fraud and Related Terminology: Preliminary Results and Discussion. *Journal of Food Science 84*, 2705–2718. https://doi.org/10.1111/1750-3841.14705

Spiteri, M., Rogers, K.M., Jamin, E., Thomas, F., Guyader, S., Lees, M., Rutledge, D.N., 2017. Combination of 1H NMR and chemometrics to discriminate manuka honey from other floral honey types from Oceania. *Food Chemistry 217*, 766–772. https://doi.org/10.1016/j.foodchem.2016.09.027

Spraul, M., Schütz, B., Rinke, P., Koswig, S., Humpfer, E., Schäfer, H., Mörtter, M., Fang, F., Marx, U., Minoja, A., 2009. NMR-Based Multi Parametric Quality Control of Fruit Juices: SGF Profiling. *Nutrients 1*, 148–155. https://doi.org/10.3390/nu1020148

Stahle, L., Wold, S., 1989. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems 6*, 259–272. https://doi.org/10.1016/0169-7439(89)80095-4

Stolt, R., Torgrip, R.J.O., Lindberg, J., Csenki, L., Kolmert, J., Schuppe-Koistinen, I., Jacobsson, S.P., 2006. Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data. *Anal. Chem. 78*, 975–983. https://doi.org/10.1021/ac050980b

Szymańska, E., Saccenti, E., Smilde, A.K., Westerhuis, J.A., 2012. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics 8*, 3–16. https://doi.org/10.1007/s11306-011-0330-3

Tautenhahn, R., Böttcher, C., Neumann, S., 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics 9*. https://doi.org/10.1186/1471-2105-9-504

Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics 16*, 119–128. https://doi.org/10.1002/cem.695

Vaclavik, L., Schreiber, A., Lacina, O., Cajka, T., Hajslova, J., 2012. Liquid chromatography–mass spectrometry-based metabolomics for authenticity assessment of fruit juices. *Metabolomics 8*, 793–803. https://doi.org/10.1007/s11306-011-0371-7

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics 7*. https://doi.org/10.1186/1471-2164-7-142

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J.J., Yanes, O., 2012. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites 2*, 775–795. https://doi.org/10.3390/metabo2040775

Wehrens, R., Hageman, J.A., van Eeuwijk, F., Kooke, R., Flood, P.J., Wijnker, E., Keurentjes, J.J.B., Lommen, A., van Eekelen, H.D.L.M., Hall, R.D., Mumm, R., de Vos, R.C.H., 2016. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics 12*. https://doi.org/10.1007/s11306-016-1015-8

Willems, J.L., Low, N.H., 2018. Structural identification of compounds for use in the detection of juice-to-juice debasing between apple and pear juices. *Food Chemistry 241*, 346–352. https://doi.org/10.1016/j.foodchem.2017.08.104

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems 58*, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yang, R., Huang, W., Zhang, L., Thomas, M., Pei, X., 2009. Milk adulteration with melamine in China: crisis and response. *Quality Assurance and Safety of Crops & Foods 1*, 111–116. https://doi.org/10.1111/j.1757-837X.2009.00018.x

CHAPITRE 2 ANALYSE CIBLEE DE COMPOSES MARQUEURS D'AUTHENTICITE DU CITRON JAUNE

1. INTRODUCTION ET RESUME DE L'ARTICLE

La fraude sur le jus de citron jaune est de plus en plus présente depuis le milieu des années 2010. Il s'agit notamment d'une fraude par ajout d'autres espèces de citrus, en particulier le citron vert. Plusieurs composés ont été identifiés comme marqueurs de certaines espèces de citrus : c'est le cas de la 7-méthoxycoumarine, du bergaptène et de l'isopimpinelline pour le citron vert (Lehnert & Ara, 2014 ; Lehnert et al., 2017). Différentes études ont montré que chaque espèce de citrus avait un profil en coumarines, psolarènes et flavonoïdes bien distinct (Gattuso et al., 2007). Le développement de nouvelles méthodes ciblant ces familles de composés est donc nécessaire pour détecter ces fraudes et ainsi améliorer les contrôles d'authenticité.

C'est sur ce constat que l'association internationale SGF a décidé de mener un projet pour l'authentification du citron jaune en s'intéressant à sept composés (7-méthoxycoumarine, bergaptène, isopimpinelline, limettine - aussi connue sous le nom de citropten, nobilétine, sinensétine et tangéritine). Trois laboratoires spécialisés dans les analyses de contrôle des jus de fruits ont participé à ce projet, dont le laboratoire d'Eurofins Analytics France. Dans cette étude, chaque laboratoire a développé et mis en œuvre sa propre méthode d'analyse.

Ainsi, au sein d'Eurofins Analytics France et dans le cadre de ces travaux de thèse, une méthode d'analyse ciblée a été développée pour quantifier ces 7 composés dans les jus de citron. Le développement de cette méthode a suivi la norme NF V03-110, et le laboratoire d'Eurofins Analytics France a par la suite été accrédité par le COFRAC pour la mise en œuvre de cette méthode dans ce champ d'application. Cette méthode est donc désormais utilisée en routine dans le service de chromatographie du laboratoire d'Eurofins Analytics France.

Durant ce projet, des échantillons de jus de citron (fruits pressés, purs jus, jus concentrés et huiles essentielles – 139 échantillons au total) ont été collectés par l'association SGF. Différentes origines géographiques, variétés et procédés de production ont été choisis lors

de la collecte des échantillons. La collecte a permis de recueillir des échantillons produits sur trois années consécutives (2019, 2020 et 2021).

Les résultats d'analyses ont ensuite été mis en commun et comparés. Pour tenir compte des données censurées (i.e. valeurs inférieures à la limite de détection (LOD) ou inférieures à la limite de quantification (LOQ)), des valeurs seuils moyennes ont été appliquées pour comparer les résultats (LOQ/2 pour les huiles essentielles et LOD/2 pour les jus). En raison des performances différentes entre les trois méthodes mises en œuvre, les valeurs de LOD et LOQ ont été différenciées en fonction des méthodes et résultats de chaque laboratoire. Les moyennes et les écarts types ont été calculés pour chaque échantillon. Les échantillons ont été comparés selon leur type (jus ou huile essentielle) et selon leur espèce (citron jaune ou citron vert).

2. CORPS DE L'ARTICLE

A pragmatic authenticity assessment of lemon (*Citrus limon* [L.] Burm. f.) juices by its profile of coumarins, psoralens and polymethoxyflavones

Markus Jungen [1,6,*], Nenad Dragićević [1,*], Miriam Rodriguez-Werner [2], Simone Schmidt [2], Katy Dinis [3,5], Lucie Tsamba [3], Eric Jamin [3], Thorsten Fiedler [4], Nadine Fischbach [4], Valérie Camel [5], Ralf Schweiggert [6]

1: SGF International, Marie-Curie-Ring 10a, 55291 Saulheim, Germany

2: Chelab Dr. V. Ara, Carl-Zeiss-Str. 16, 30966 Hemmingen, Germany

3: Eurofins Analytics France, Rue Pierre Adolphe Bobierre, BP 42301, 44323 Nantes, France

4: GfL Gesellschaft für Lebensmittel-Forschung mbH, Landgrafenstraße 16, 10787 Berlin, Germany

5: UMR SayFood, Université Paris-Saclay, INRAE, AgroParisTech, 91300 Massy, France

6: Geisenheim University, Department of Beverage Research, Chair of Analysis & Technolo-gy of Plant-based Foods, Von-Lade-Str. 1, 65366 Geisenheim, Germany

** Corresponding Authors: Tel.: +49 6732 2779529 - E-mail address: nenad@sgf.org, markus@sgf.org*

Soumis prochainement dans Food Control

Abstract

Coumarins, psoralens and polymethoxyflavones have long been described as authenticity mark-ers in lemon (*Citrus limon* [L.] Burm. f.) and lime (*Citrus* × *aurantifolia* [Christm.] Swingle; *Citrus* × *latifolia* [Yu.Tanaka]) juices. Differing views on what an authentic lemon juice should look like challenge the differentiation between deliberate food fraud and potential GMP lacks. In this study, 139 samples (juice, juice concentrate and peel oil) from eleven countries, covering the usual processing methods on the market, were analysed using three different liquid chroma-tographic methods. First, we were able to confirm that the analytical approaches considered lead to comparable results. Furthermore, analytical differences between lemon juice and peel oil as well as between lemon and lime peel oil were explained. The contents of coumarins, psoralenes and polymethoxyflavones differ both in terms of product type and processed Citrus species. Consequently, maximum levels of investigated marker substances were postulated for the au-thentication of a lemon juice produced under GMP conditions.

Keywords: Citrus; food fraud; chromatography; polyphenols

Chemical compounds studied in this article*:* Herniarin (PubChem CID: 10741), Limettin (PubChem CID: 2775), Isopimpinellin (PubChem CID: 69079), Bergapten (PubChem CID: 2355), Nobiletin (PubChem CID: 72344), Tangeretin (PubChem CID: 68077), Sinensetin (PubChem CID: 145659)

Abbreviations: *AIJN*, European Fruit Juice Association*; AIJN CoP*, AIJN Code of Practice*; EFSA*, European Food Safety Authority*; ESI*, electrospray ionization*; ESI+*, electrospray ionization in positive ion mode operation*; FA*, formic acid*; FAO*, Food and Agriculture Organization of the United Nations*; GMP*, Good Manufacturing Practice*; HESI*, heated electrospray ionization*; HPLC-DAD*, high performance liquid chromatography-diode array detection*; IFU*, International Fruit and Vegetable Juice Association*; ISTD*, internal standard*; LC-MS*, liquid chromatography–mass spectrometry; LoD, limit of detection; LoQ, limit of quantification*; MS*, mass spectrometry*; MS/MS*, tandem mass spectrometry*; mt*, metric tons*; PC*, principal component*; PCA*, principal component analysis*; PDA*, diode-array detector*; SGF*, SGF International e.V.*; SIM*, single ion monitoring*; THF*, tetrahydrofuran*; UHPLC*, ultra-high performance liquid chromatography*; UV*, ultraviolet; *v/v*, volume by volume

2.1. INTRODUCTION

Lemon juices and concentrates are economically important and growing commodities in the food and beverage industry, but especially in the fruit juice industry. In addition to their use as consumer goods (lemon juice or lemon juice from concentrate), their industrial use is primarily as a natural acidifier (in the sense of “clean labelling”) in fruit nectars and soft drinks. While in 2019 the main exporting countries for lemon juice were Brazil, the USA, Mexico, Italy, and Peru (listed in descending order of their market share), the largest producers of lemon juice concentrate were Argentina, Peru, South Africa, Mexico, and Egypt (listed as above). According to FAO data, exports have increased by about 45% for lemon juice concentrate and by about 28% for lemon juice since 2010 (Food and Agriculture Organization of the United Nations).

Rising demand on the one hand, but also difficult harvest and processing situations have led to challenges in the authenticity assessment of lemon juice. Within the framework of routine controls of the voluntary control system of SGF (SGF International e.V.), adulterations in the form of added sugar, added (citric) acid but also additions of foreign species such as lime (*Citrus* × *aurantifolia* [Christm.] Swingle and *Citrus* × *latifolia* [Yu.Tanaka] Tanaka) were detected in lemon juice concentrates, which became more frequent from 2014 onwards compared to previous years. For sugar and acid additions, stable isotope analysis is the analytical method of choice (Jahromi, Pratt, Zhou, Reimann, & Hammon, 2015; Jamin, Martin, Santamaria-Fernandez, & Lees, 2005; Rinke, 2016), but for foreign fruit additions in *Citrus* juices, the analyst is dependent on the use of chromatographic methods. The flavanone glycosides hesperidin, naringin, neohesperidin, eriocitrin and narirutin can be used to detect possible additions of foreign species in *Citrus* juices, as shown by Cautela et. al (2008), but this remains unsuccessful to detect the addition of limes to lemon juice.

The fingerprint of polymethoxylated flavones as measured by HPLC-DAD has played a prominent role in this since the 1990s (Hofsommer, 1999; Ooghe, 2001). Initially used mainly to assess the authenticity of orange juices, it soon became clear that *Citrus* juices have species-specific fingerprint patterns that were discussed to detect possible adulterations. When looking at lemons and limes, the focus is set on coumarins and psoralens in addition to polymethoxylated flavones. The search for suitable marker parameters for the addition of

lime or other *Citrus* species to lemon juice has been occupying the scientific community for some time. While McHale and Sheridan (1989) examined peel oils of key limes, persian limes, bergamot, grapefruit, bitter orange, sweet orange, mandarins, and tangerines, P. Dugo et al. (2009) and G. Dugo & Mondello (2011) were able to present a direct comparison of lime and lemon oils: here, characteristic markers for lime naming herniarin, bergapten and isopimpinellin were shown. Costa et al. (2014) provided insights into composition of coumarins, psoralens and polymethoxylated flavones from Italian lime juice. Lehnert and Ara (2014) established herniarin (referred in their study as 7-methoxycoumarin) as particular lime marker, but Lehnert et al. (2017) added in a further publication that apart from this coumarin the psoralens bergapten and isopimpinellin should be used as marker parameters for lime as well. Jungen et al. (2021) confirmed the usefulness of these previously discussed lime markers bergapten and isopimpinellin adding a further psoralen, 5-geranyloxy-8-methoxypsoralen. By using modified extraction and chromatographic separation techniques, Li et al.(2021) detected in lemon and lime juice the presence of coumarins and methoxylated flavones previously attributed only in other *Citrus* juices. The studies of Lehnert and Ara (2014), Dugrand-Judek et al. (2015) and Lehnert et al. (2017) observed elevated levels of total coumarins and psoralens in peel oil and the outer parts of the fruits compared to the juice from the endocarp. By this, apart from fraudulent addition of foreign *Citrus* species the applied processing technique (and its intensity) could be examined as well and the presence of not permitted processing methods in lemon juice (according to the European Fruit Juice Directive) such as whole-fruit-processing (European Council, 2012) could be detected. Here, Jungen et al. (2021) suggested to combine the targeted analysis of bergapten, isopimpinellin and 5-geranyloxy-8-methoxypsoralen with the phenolic compound phlorin as albedo-marker for over-extraction to assess the influence of applied extraction technology.

With the help of the above-mentioned marker parameters from literature the authentication of market samples is still challenging and final decisions if a sample is to be claimed are often made based on own laboratory-specific databases. In the younger past this situation led to uncertainties in the *Citrus* processing industry, particularly in cases when two laboratories delivered different interpretations while the measured parameters in both laboratories bore practically the same results. The situation is aggravated by the fact that the respective reference guideline for lemon juices of the AIJN Code of Practice does not

have any ranges describing the contents of the analysed coumarins, psoralens and polymethoxyflavones (AIJN European Fruit Juice Association, 2019).

Fruit juices and fruit juice concentrates contain variable amounts of volatile oils, depending on the extraction technique applied. Based on standing European regulation it is allowed to restore once removed flavour from a juice, potentially lost during processing. According to the reference guideline for lemon juice/juice concentrate from the AIJN CoP a product may not contain more than 0.5 mL/L of volatile oils (AIJN European Fruit Juice Association, 2019). Since the highest amounts of coumarins, psoralens and polymethoxyflavones occur especially in flavedo and albedo of lemon fruits, it can be assumed that the contents of coumarins, psoralens and polymethoxyflavones in lemon juices are to a great extent influenced by the oil content of the juice, as noted already by Jungen et al. (2021).For these reasons we conducted a study aiming for a jointly approved evaluation practice of lemon juices and juice concentrates where in total 139 samples (fruits manually squeezed, industrially processed juices and juice concentrates, and peel oils) were analysed in three independent and accredited laboratories. The study was conducted covering products from harvest seasons 2019, 2020 and 2021.

2.2. Materials and methods

2.2.1. Chemicals

The different standards and solvents were purchased from different suppliers by the participating laboratories.

Chelab purchased the standards for 6-methylcoumarin, herniarin, bergapten, limettin, nobiletin, and tangeretin from Sigma Aldrich (Taufkirchen, Germany), isopimpinellin from Phytolab (Vestenbergsgreuth, Germany) and sinensitin from Cayman Chemical (Ann Arbor, Michigan, USA). The solvents toluol and methanol were from VWR (Hannover, Germany), acetonitrile from Carl Roth (Karlsruhe, Germany) and THF was purchased from Thermo Fisher (Kandel, Germany).

Eurofins applied the standards for bergapten, herniarin, and limettin from Sigma Aldrich (St. Quentin Fallavier, France), isopimpinellin, nobiletin, sinensetin, and tangeretin purchased

from Extrasynthèse (Genay, France). The solvents methanol, formic acid and water were all LC-MS grade and were purchased from Thermo Fisher (Les Ulis, France).

GfL purchased standards for herniarin and limettin from Thermo Fisher (Kandel, Germany), phellopterin, nobiletin and sinensetin from PhytoLab (Vestenbergsgreuth, Germany) and bergapten, isopimpinellin, and tangeretin from Sigma Aldrich (Taufkirchen, Germany). The solvents acetonitrile, methanol, and toluol were from Chemsolute (Renningen, Germany) and THF from Merck (Darmstadt, Germany).

All solvents used were at least of analytical or HPLC grade. De-ionized water was used unless stated otherwise.

2.2.2. Samples

To reflect a representative market picture (e.g., extraction technology, dominant cultivars, cloudy and clear products, different filling technologies) in our study, lemon, and lime samples from eleven different countries were sampled during routine inspections and sample requests of the SGF. Among them were typical cultivated lemon varieties of Europe (ICI Business, 2020; Klimek-Szczykutowicz, Szopa, & Ekiert, 2020). These were: Femminello Siracusa, Femminello Trapani, Eureka, Fino, Primofiore and Verna. The varieties Primofiore and Verna were additionally sampled in various times during the harvesting period: early, middle, and late. A part of the samples during presence audits were sampled here in step controls at different points of the extraction process (extractor, finisher/centrifuge, evaporator, decanter, mixing tanks) by independent and SGF-trained and -accredited auditors, guaranteeing the authenticity of the samples. The different commodities are presented in **Table 2.1** with reference to the geographical origin and the dominant extraction technology documented during sampling. Here it was observed that in the industrial processing of lemons and limes into juice, juice concentrate and peel oil, “squeezer” (e.g., JBT/FMC extractors) and “reamer extractors” (e.g., Brown extractors) dominate globally, while “rotary press” technology (e.g., Flli. Indelicato) is mainly used in Italy and Turkey.

***Table 2.1**. Products, predominating Citrus extraction technology and geographical origins of analysed lemon (Citrus limon [L.] Burm. f.) and lime (Citrus × aurantifolia [Christm.] Swingle and Citrus × latifolia [Yu.Tanaka] Tanaka) samples.*

Origin	Extraction technology	Product	*n*
Argentina	reamer-type and squeezer-type	Lemons	1
		Lemon juice	18
		Lemon juice concentrate	13
		Cold-pressed lemon oil	3
Bolivia	n/a	Lemons	1
Brazil, São Paulo	squeezer-type	Lemon juice concentrate	8
		Lime juice	1
		Cold-pressed lemon oil	1
		Cold-pressed lime oil	1
China	squeezer-type	Lemon juice concentrate	1
Israel	squeezer-type	Lemon juice	1
Italy	rotary press extractors	Lemons	5
		Lemon juice	8
		Lemon juice concentrate	6
		Cold-pressed lemon oil	3
Mexico	squeezer-type	Lemon juice concentrate	1
		Cold-pressed lime oil	3
South Africa	reamer-type and squeezer-type	Lemons	1
		Lemon juice	4
		Lemon juice concentrate	4
		Cold-pressed lemon oil	6
Spain	squeezer-type	Lemons	12
		Lemon juice	9
		Lemon juice concentrate	10
		Cold-pressed lemon oil	13
		Cold-pressed lime oil	1
Turkey	rotary press extractors	Lemon juice concentrate	1
Uruguay	reamer-type	Lemon juice	2
		Lemon juice concentrate	1

2.2.3. Determination of coumarins, psoralens and polymethoxyflavones

The compounds were determined with different methodology in three different laboratories. In this way, we were able to combine the answer to the question which components could be detected in the different product types with different analytical methods. By doing so, we hope to show that, regardless of the methodology chosen, lemon juices can be evaluated based on coumarins, psoralens and polymethoxyflavones.

2.2.3.1. HPLC-DAD method A (Chelab)

Fresh fruits were extracted by hand using a *Citrus* press prior to analyses. The processing of the samples was carried out in accordance with Pupin et al. (1998) with some methodological modifications. While it was Pupin et al. (1998) objective to authenticate orange products by means of their contents of polymethoxyflavones we aimed for an optimisation in the analytical differentiation of lemon and lime products. Using a gradient-based methodology we could reach a better separation of the target compounds (herniarin, sinensetin, limettin, isopimpinellin, bergapten, nobiletin, and tangeritin) as shown in the HPLC chromatogram at 330 nm in **Figure 2.1**.

Briefly, 10 mL of the respective sample of juice was transferred to centrifuge tubes, internal standard and 5 mL of toluene were added. The tubes were subsequently sealed, shaken, and centrifuged and supernatants were carefully removed and collected in a conical flask. This extraction was repeated two more times. Hereafter the combined organic phases were concentrated to dryness: the compounds were extracted with 15 mL toluene three times. To concentrate the compounds, the solvent was removed by vacuum evaporation, then the dried extracts were re-dissolved in 0.5 mL methanol. After centrifugation, the solution obtained was used for HPLC analysis. The oil samples were diluted with methanol (0.1 g sample in 5 mL), internal standard was added, centrifuged and the supernatants were measured.

Quantitative analyses were performed using an HPLC-DAD System (Agilent 1260 HPLC-UV/DAD Infinity II) and using a Reprosil 100 C18, 250 x 4.6 mm, 5 µm column with a C18 precolumn (Dr. Maisch, Ammerbuch-Entringen, Germany). A gradient was used, based on the ternary eluent water/acetonitrile/THF (eluent A: 88/8/4, [v/v/v]) and the binary eluent acetonitrile/THF (eluent B: 96/4, [v/v]). The gradient program was as follow: isocratic at 30% B (30 min), 30 to 80% B (10min), 80 to 30% B (5min), isocratic at 30% B (10 min). Flow rate was 0.7 mL/min, total run time was 55 min and injection volume 20 µL. Column oven temperature was set at 30 °C. The detection was carried out by means of a PDA in the range of 190 to 400 nm with a detection wavelength for sinensetin, limettin, nobiletin, tangeretin at 330 nm, while the detection wavelength for herniarin, isopimpinellin, bergapten was 320 nm. Using external calibration curves with external standards of herniarin, sinensetin, limettin, isopimpinellin, bergapten, nobiletin, and tangeritin the quantification was carried out at

concentrations ranging from 0.03 to 45 mg/L. The applied ISTD 6-methylcoumarin was used for quality control and monitoring of diverging retention times, and not for quantitation.

2.2.3.2. HPLC-DAD method B (GfL)

Lemons were extracted with a heavy cast iron lever *Citrus* press to obtain the juice. Concentrates were pre-diluted with double-distilled water in a ratio of 1:5, juices were used directly, and *Citrus* oils were diluted 1:10 with methanol. Afterwards the sample were membrane filtered and used for HPLC measurement. 5 mL of juice, diluted concentrate or oil were pipetted into a 15 mL centrifuge tube, 0.05 mL of phellopterin ISTD solution (100 mg/L) and 5 ml of toluene were added and thoroughly mixed for 10 s on a vortex-mixer (Model Heidolph Reax top; top speed).

Subsequently, the upper organic phase was transferred to a conical flask. The toluene was completely removed on a rotary evaporator (40°C) and any solvent residues were removed under nitrogen stream. The sample was redissolved in 0.5 mL methanol. The sample transferred to an HPLC vial is used directly for the analysis.

Analysis was performed by HPLC-DAD System (Agilent 1100 HPLC, Agilent Technologies, Santa Clara, USA), using a reversed phase C18 column (Hypersil ODS, 250 x 4.6 mm, 3µm, VDS optilab; Berlin, Germany) operated at 50 °C. Two mobile phases were used: eluent A (water) and eluent B (binary mixture of acetonitrile and THF [60:40, v/v]). The gradient program was as follows: 0 to 15 min from 75% A to 65% A, 15 to 20 min from 65% A to 40% A, 20 to 22 min from 40% A to 20% A, 22 to 28 min from 20% A to 20% A, 28 to 28.1 min from 20% A to 75% A and 28.1 to 35 min from 75% A to 75% A.

Flow rate was 0.8 mL/min, total run time 35 min, and the injection volume 20 µL. Detection wavelength was 330 nm. UV spectra were recorded in the range of 200-450 nm. The contents of the individual substances were calculated according to the generally applicable procedure of the internal standard method.

2.2.3.3. UPLC-MS^n (Eurofins)

Lemon juices were extracted from the fresh fruit samples using an electric *Citrus* squeezer. Juice concentrates were rediluted with de-ionized water to single strength level (8 °Bx). Aliquots (5 mL) of the samples were centrifuged during 5 min at 4,000 rpm. The supernatants

of juices and rediluted juice concentrates were collected and 4-fold diluted with water directly in a vial before analysis. The supernatants of the oils were diluted 1:1000 with methanol and aliquots of them were added 1:4 with de-ionised water directly to a sample vial for analysis. Standards were prepared at different concentration level and a calibration curve was constructed for quantification of samples.

Analyses were performed on a ThermoFisher Vanquish Flex UHPLC system, composed of a binary pump, a refrigerated sampler, and a column oven, connected to a ThermoFisher high resolution Orbitrap mass spectrometer QExactive Plus with a heated electrospray ion source (HESI). The UHPLC separation was achieved using a C18 Hypersil Gold column (50 x 2.1 mm, 1.9 µm) at a 0.4 mL/min flowrate. The column temperature was set to 30°C. The mobile phases were water acidified with 0.1% FA (A), and methanol acidified with 0.1% FA (B), with the following linear gradient elution: 0-5 min, B: 35%; 5-8 min, B: 35-70%; 8-8.5 min, B: 70-98%; 8.5-11.5 min: B: 98%; 11.5-11.6 min, B: 98-35%; 11.6-15 min, B: 35%. The injection volume was 1 µL.

MS data were acquired using positive ion mode (ESI+) operating in SIM (Single Ion Monitoring) with a resolution of 70,000. An inclusion list containing the mass-to-charge ratio (*m/z*) for all the coumarins, psolarens, and polymethoxyflavones analysed was used to perform MS/MS analysis on these precursor ions. Precursor ions were used for quantification whereas product ions confirmed the compounds identities.

2.2.4. Descriptive statistics, data evaluation and visualisation

The differences in analytical methods between the laboratories generated different LoDs and LoQs. To be able to calculate descriptive statistics from continuous, left-censored quantitative data ("<LoQ"), the medium bound values were applied (LoQ/2 in case of oils and LoD/2 in case of juices), as described by EFSA (2010) and proposed by Antweiler and Taylor (2008), and George et al. (2021).

Arithmetic means and standard deviations were calculated on measurements performed on individual samples. Descriptive statistics were expressed from arithmetic means per sample.

Statistical evaluations and graphical presentations were created using Python 3.5 software (Python Software Foundation) and the NumPy (Harris et al., 2020), pandas (Stéfan van der

Walt & Jarrod Millman, 2010; Wes McKinney, 2010), Matplotlib (J. D. Hunter, 2007) and Seaborn libraries (Waskom, 2021). For the principal component analysis, the measured values were also normalised in Python, the procedure used corresponded to the calculation of a *z*-value by removing the mean and scaling to unit variance with the StandardScaler from the class *sklearn.preprocessing* and *sklearn.decomposition* (Pedregosa et al., 2011).

2.3. Results and discussion

The authentication of lemon juices by means of coumarins, psoralens and polymethoxyflavones is a sensitive topic and it must be differentiated between lacking GMP e.g., due to carry-overs of other *Citrus* species or higher contents of peel oil and the intentional addition/co-processing of further *Citrus* species, namely lime. For this reason, different analytical methods for quantifying the target substances are considered in our study regarding their comparability. A second step examines the contents of the target substances in lemon juices and juice concentrates with the contents in lemon peel oils. As marker substances for lime juices were already described elsewhere (Hofsommer, 1999; Lehnert N. & Ara V., 2014) it was decided to compare the coumarins, psoralens and polymethoxyflavones in lemon and lime oils in a third step. Based on the previously obtained analytical results on juice products and peel oils, and considering industry-recognised standards, pragmatic assessment criteria and subsequent recommendations of actions are then proposed in the authentication of lemon juices and juice concentrates.

2.3.1. Comparability of different analytical approaches

Of the samples examined, 65 samples were analysed by all three methods of analysis described above, 73 samples were analysed by two of the methods mentioned, of which 59 were analysed by methods HPLC-DAD method A and UPLC-MS^n and 14 by methods HPLC-DAD method A and HPLC-DAD method B. All the three methods applied yield comparable results – not only in the elution order of the lime marker parameters of interest for the authentication of lemon juice, but also about the quantifications resulting from the recorded chromatograms. As an example, **Figure 2.1** shows the chromatograms of a typical lemon juice and a typical lime juice measured with the analytical methods described above. In these exemplary chromatograms, quantifiable contents are only found for the coumarins citropten/limettin (lemon: 1.34±0.60 mg/L, lime: 11.92±2.47 mg/L) and herniarin (lemon:

<LoQ, lime: 3.99±1.34 mg/L) as well as for the psoralens bergapten (lemon: <LoQ, lime: 4.74±1.27 mg/L) and isopimpinellin (lemon: <LoQ, lime: 2.89±0.87 mg/L).

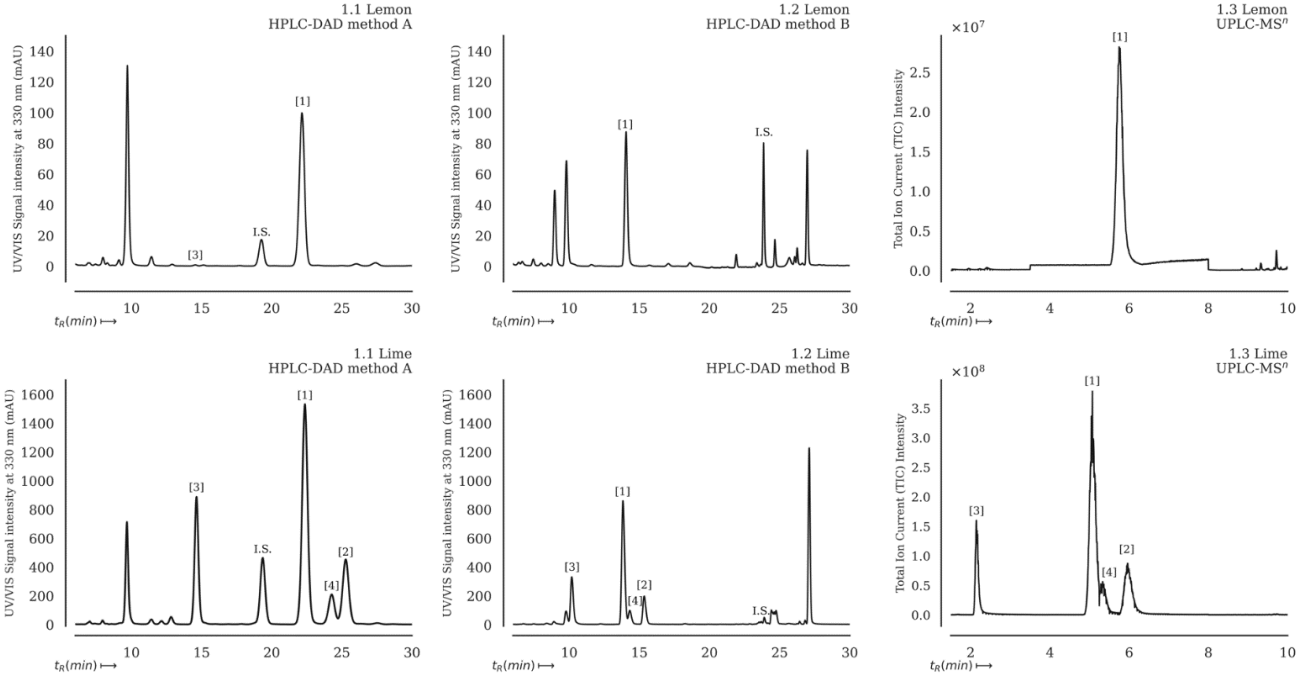

Figure 2.1. *Representative chromatograms of typical lemon (Citrus limon [L.] Burm. f.) and lime (Citrus × aurantifolia [Christm.] Swingle) juice measured with HPLC-DAD method A (1.1), HPLC-DAD method B (1.2), and UPLC-MSⁿ (1.3) with the coumarins limettin [1], herniarin [3], and the psoralens isopimpinellin [4] and bergapten [2] including different internal standards [I.S.].*

A certain lack of reproducibility in the results of the lemon fruits is caused by the different hand extraction techniques (manual hand press, electric hand press, manual iron lever press) used to de-juice the fruits, especially for the coumarins limettin/citropten and herniarin. For the psoralenes bergapten and isopimpinellin, which are discussed as lime markers, as well as for the polymethoxyflavones nobiletin, sinensetin and tangeritin, there are no abnormalities in the examined fruits in agreement with the work of Lehnert et al. (2017).

In the examination of fruit juices and concentrates, the contents of coumarins, psoralens and polymethoxyflavones determined by means of the three different analytical techniques are very comparable. This can be explained by the presence of a homogeneous analyte pattern in all three laboratories (compared to the individual de-juicing of lemons) as well as by the fact that practically only the coumarin citropten/limettin is quantified to any appreciable extent in the lemon juices and juice concentrates presented here and

quantifiable contents of the psoralens bergapten and isopimpinellin are the exception. It is remarkable that the presence of the coumarin herniarin is limited to only a few exceptions in lemon juices.

The analysis of the lemon and lime peel oils, on the other hand, reveals partly major differences in the quantifications. Possible reasons here could be the different sample preparations in the individual laboratories.

2.3.2. Lemon juice products vs. lemon peel oils

We confirmed previous results reported from Dugrand et al. (2013) and Zhao et al. (2017) by observing generally higher contents of coumarins, psoralens, and polymethoxyflavones in lemon peel oils compared to lemon juices (**Table 2.2**). In the different product types limettin represents the dominant component (1.04 ± 0.83 mg/L in lemons, 0.89 ± 0.64 mg/L in juice, 1.07 ± 1.06 mg/L in juice concentrates, and 907.84 ± 213.55 mg/L in peel oils) with a comparably high range of variation, which can be attributed to different origins, lemon varieties and processing methods. The other components are quantified in lower contents or in traces (< LoQ) in lemons, juices, and juice concentrates while their contents in peel oils are significantly higher. It is noteworthy that the psoralens bergapten and isopimpinellin, which are discussed as lime markers in the scientific community (Hofsommer, 1999; Lehnert N. et al., 2017; Lehnert N. & Ara V., 2014), were quantifiable (with a high degree of variation) in the lemon oils investigated, while in lemons, juices and juice concentrates only traces were detectable (isopimpinellin < LoQ) or very low concentrations were quantifiable (bergapten in lemon juices at 0.02 ± 0.05 mg/L). The findings of isopimpinellin in lemon oil are in accordance with the previous observations from Dugrand et al. (2013). The polymethoxyflavone nobiletin was quantified in lemon peel oils (5.65 ± 1.48 mg/L) but could only be found in traces (< LoQ) in lemons, lemon juices and juice concentrates. Sinensetin and tangeritin could not be quantified in all lemon products. According to Pupin et al. (1998) and Dugo et al. (2009) higher quantities of nobiletin, tangeritin, and sinensetin are commonly found in oranges, grapefruits, and tangerines. According to Lehnert et al. (2017) these components are not found in lemons, but neither in oranges, tangerines or pomelo from China. On the contrary, according to Li et al. (2021), the three compounds were detected as well in lemon fruits from China (compounds not quantified), and according to Xi et al. (2017) higher contents of nobiletin and sinensetin were quantified in lemon peel oil from China than

reported in this study. In the two last mentioned studies, the possibility of the presence of lemon hybrids was not considered.

A PCA calculated on the base of our results in lemon juice products (black circles, **Figure 2.2**) and peel oils (grey triangles) showed a clear separation of juice products vs. peel oils as expected from the analytical results. A total variance of 95.59% was explained by the first two principal components (PCs), where the first PC already explained 83.55%. Bergapten, herniarin and isopimpinellin showed positive loadings on both PCs while the coumarin limettin and the polymethoxyflavones sinensetin, nobiletin and tangeritin are characterised by positive loadings on PC1 and negative loadings on PC2. By interpretation of the loading plot, we conclude that PC1 is influenced mainly by the quantitative contents of the measured components that are significantly higher in peel oils while PC2 describes qualitative contents related to the occurrence of the considered parameters in the diverse products. In addition to the clear separation between juiced products and peel oils, the principal component analysis also shows differences in the examined peel oils: in the score plot, a larger group of oils (in the quadrant of positive PC1 and negative PC2 direction) is located, while only five peel oils are in the opposite quadrant in positive PC2 direction. The latter five samples mentioned are four peel oils from South Africa (sampling year 2020) and one oil from Spain (sampling 2021). These samples have in common that they all show high contents of herniarin and bergapten as well as significant contents (> 10 mg/L) of isopimpinellin compared to the other measured lemon oils. While Dugo et al. (2009) and Russo et al. (2021) did not report the presence of isopimpinellin in lemon peel oils, Dugrand-Judek et al. (2015) detected 0.82 ± 0.29 mg/kg in peel extracts of Eureka lemons. Jungen et al. (2021) found with 0.1 ± 0.1 mg/kg and 0.2 ± 0.0 mg/kg lower contents in whole-processed and tincture-pressed lemons, respectively. These papers also point out that isopimpinellin could not be detected in the endocarp of the lemon. Even though a peel oil is much more concentrated compared to a peel extract, whole-processed or tincture-pressed lemons, the levels of isopimpinellin in the five lemon peel oils mentioned appear very high. Considering the loading plot and the substances located in the quadrant of positive PC1 and PC2 directions – bergapten, herniarin and isopimpinellin – this shows that the group of lemon oils is not as homogeneous as the group of “juiced products” in comparison. The reasons for these differences, even within the group of lemon oils, could be due to the processing of

different lemon cultivars or to different technological processes for oil extraction and requires further investigation in the future.

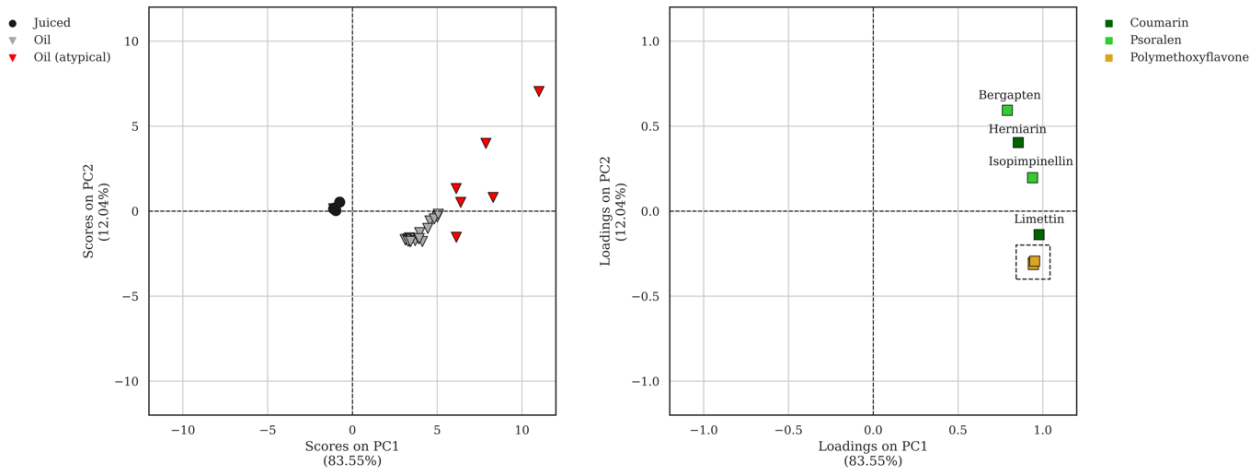

Figure 2.2. *Score and corresponding loading plots of the principal component analyses (PCAs) on lemon (Citrus limon [L.] Burm. f.) samples, calculated based on coumarins, psoralens and polymethoxyflavons.*

2.3.3. Lemon and lime peel oils

In contrast to lemon peel oils, the lime peel oils demonstrated generally higher contents (especially bergapten was higher with a factor of 50, isopimpinellin was higher by more than a factor of 100 compared to lemon peel oil) of the investigated parameters and the polymethoxyflavone tangeritin was detected above the limit of quantification (**Table 2.2**).

Table 2.2. *Contents of coumarins, psoralens, and polymethoxylated flavons in lemon (Citrus limon [L.] Burm. f.) and in lime (Citrus × aurantifolia [Christm.] Swingle and Citrus × latifolia [Yu.Tanaka] Tanaka) samples.*

	Lemon				**Lime**
	fruits [a] (n=20)	juice [a] (n=42)	juice concentrates [b] (n=45)	peel oil [c] (n=26)	peel oil [c] (n=5)
Limettin [1]	1.04 ± 0.83	0.89 ± 0.64	1.07 ± 1.06	907.84 ± 213.55	2,571.25 ± 1,617.52
Bergapten [2]	< LoQ	0.02 ± 0.05	< LoQ	22.6 ± 31.45	1,128.34 ± 959.6
Herniarin [3]	0.02 ± 0.05	0.03 ± 0.08	< LoQ	13.85 ± 13.29	2,058.89 ± 1,472.24
Isopimpinellin [4]	< LoQ	< LoQ	< LoQ	8.74 ± 5.71	1,138.72 ± 443.06
Nobiletin [5]	< LoQ	< LoQ	< LoQ	5.65 ± 1.48	8.50 ± 9.82
Sinensetin [6]	< LoQ	< LoQ	< LoQ	< LoQ	< LoQ
Tangeritin [7]	< LoQ	< LoQ	< LoQ	< LoQ	6.57 ± 5.31

Values represent means ± standard deviation of at least two analytical replicates (n = 2) per sample. Expressed as mg/L.
[a]: LoQ (0.02 mg/L) for fruits and juice single strength
[b]: LoQ (0.15 mg/L) for juice concentrates
[c]: LoQ (5.00 mg/L) for peel oils

The first two PCs of the PCA comparing peel oils explained a total variance of 92.75% (**Figure 2.3**). The lime peel oils (grey triangles) are separated from most lemon peel oils (black circles). Two lemon peel oils, unlike the others, are shown in the lower right quadrant of the coordinate system (red circles), indicating positive scores for PC1 and negative scores for PC2. In contrast to the other lemon oils analysed, these two lemon oils are characterised by clear contents of isopimpinellin (> 13 mg/L) and at the same time quantifiable contents of nobiletin and tangeritin. Apart from these two samples, nobiletin and tangeritin could not be quantified in any lemon oil, so that a carry-over of other *Citrus* species (other than lemon or lime) can be assumed from the content of both polymethoxyflavones.

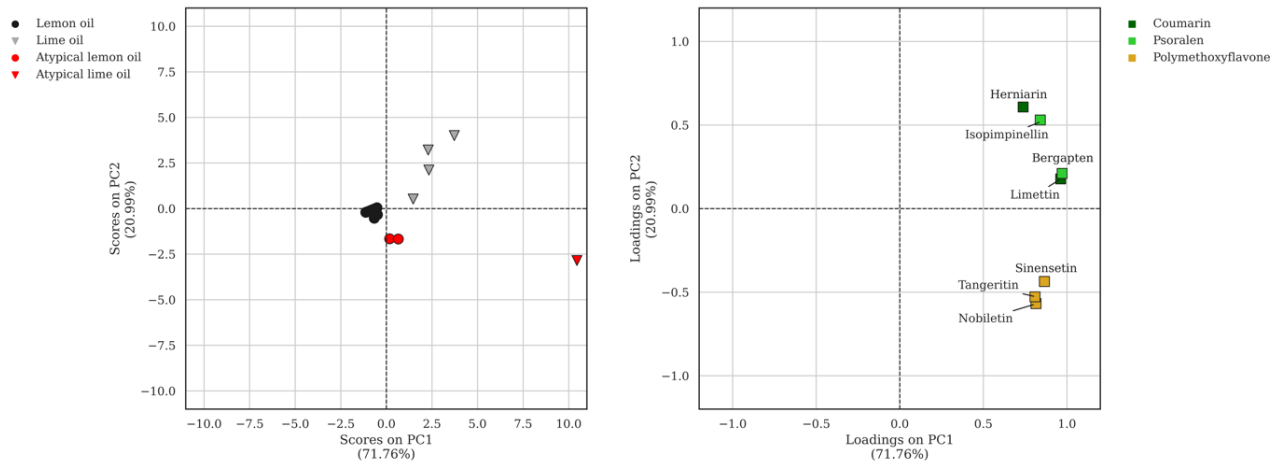

Figure 2.3. *Score and corresponding loading plots of the principal component analyses (PCAs) on lemon (Citrus limon [L.] Burm. f.) and lime (Citrus × aurantifolia [Christm.] Swingle and Citrus × latifolia [Yu.Tanaka] Tanaka) peel oils, calculated on the basis of coumarins, psoralens and polymethoxyflavons.*

Except for one lime peel oil (red triangle in **Figure 2.3**, with quantifiable contents of nobiletin and tangeritin, carry-over of orange or mandarin assumed) with positive scores at PC1 and negative scores at PC2, the others with positive scores at PC1 and PC2 are in the upper right quadrant of the coordinate system. These three deviating peel oils (two lemons and one lime) deviate from the expectation based on the non-targeted PCA about their species-related allocation. By looking at the disposition of the samples, we could observe a narrow distribution of lemon samples, whilst the lime samples exhibited a wide distribution with big standard deviations in their contents of target substances. Possible reasons for this could be found maybe in the definition of these species: a lime is defined by the presence of two subspecies: *Citrus* × *aurantifolia* [Christm.] Swingle and *Citrus* × *latifolia* [Yu.Tanaka] Tanaka, while the lemon is defined only by *Citrus limon* [L.] Burm. f. While all considered

parameters showed positive loadings on PC1 coumarins and psoralens feature positive loadings on PC2 as well, but polymethoxyflavones reveal negative loadings on PC2.

The findings of nobiletin and tangeritin in one of the lime oils and two of the lemon oils influenced its position on positive side of PC1 and on negative side of PC2 as pictured in the scores plot. Considering the number of lime peel oils studied in this work (n=5), further data collection could provide further insight. For this study, however, the differences in coumarin, psoralen and polymethoxyflavone contents between these fruit types are sufficient to draw initial conclusions and to make a proposal for pragmatic assessment practice.

2.3.4. Influence of the oil content in lemon juices and juice concentrates

As described in our study, lemons as well as their juice products (fruit juice and fruit juice concentrate) can be well distinguished from lemon peel oil mainly by the total concentration of coumarins, psoralenes and polymethoxyflavones (see **Figure 2.2**). In addition, atypical samples could also be identified within the products declared as lemon peel oil, which show particularly high contents of the coumarin herniarin and the psoralens bergapten and isopimpinellin, and contain significant contents of the polymethoxyflavones sinensetin, nobiletin and tangeritin.

Lime peel oils are clearly separated from lemon peel oils by their composition and their content of coumarins and psoralens (see **Figure 2.3**); two atypical lemon peel oils and one suspicious lime peel oil could also be identified. In all three products, the polymethoxyflavone contents are above expectation.

In **Figure 2.4**, we continue this multivariate approach and consider all the lemons, juices, juice concentrates and peel oils examined together with the lime peel oils. PC1 explains 73.65% of the observed variance, PC2 20.92%. The differentiations of lemon fruits, juices, and juice concentrates, to the lemon peel oils and the lime peel oils are very evident. While lime peel oils are separated from lemons, lemon juices and lemon juice concentrates mainly by their remarkably high contents of bergapten, isopimpinellin and herniarin, the contents of the psoralenes are significantly lower in lemon peel oils. The polymethoxyflavones nobiletin, sinensetin and tangeritin can be described as atypical for lime peel oils. This is illustrated in **Figure 2.4** for one of the lime peel oils far right in positive PC1 direction. In the case of lemon

peel oils, we can observe at least two lemon peel oils in our study that are also to be judged as atypical here due to quantified polymethoxyflavones.

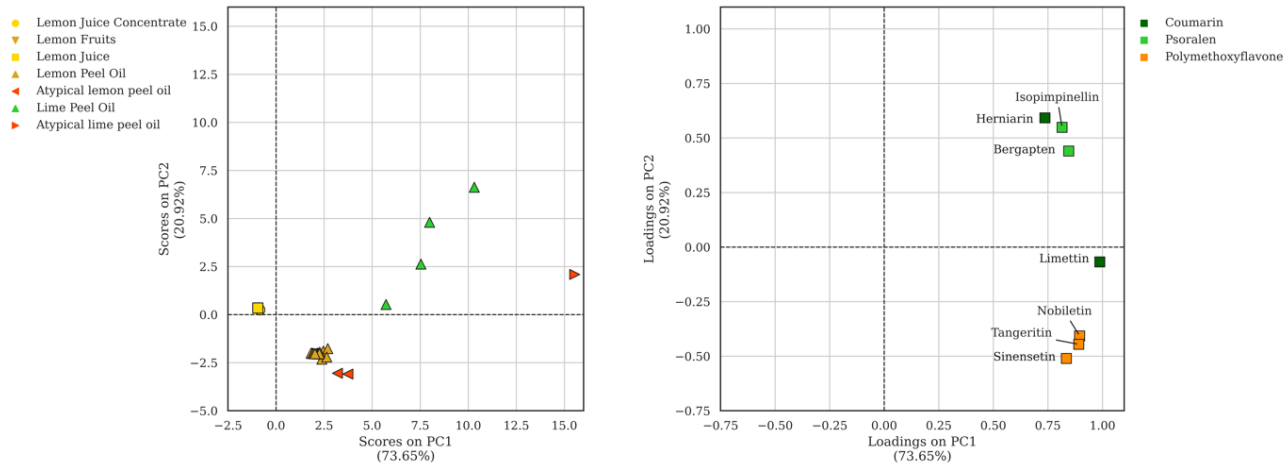

Figure 2.4. *Score and corresponding loading plots of the principal component analyses (PCAs) on lemon (Citrus limon [L.] Burm. f.) fruits, juices, juice concentrates, and peel oils and lime (Citrus × aurantifolia [Christm.] Swingle and Citrus × latifolia [Yu.Tanaka] Tanaka) peel oils, calculated on the basis of coumarins, psoralens and polymethoxyflavones.*

Based on our results, an interim conclusion can be drawn that the coumarin limettin is not a suitable marker parameter for the addition of lime to lemon due to its occurrence in all samples examined. The situation is similar with coumarin herniarin, which could be quantified in all diverse products apart from lemon juice concentrate. Its use as stand-alone lime marker is not suitable, as previous work by Lehnert et al. (2017) and Jungen et al. (2021) has shown. The polymethoxyflavones sinensetin, nobiletin and tangeritin indicate addition/carry-over of other *Citrus* species or hybrids such as orange, mandarin, and grapefruit, as they are practically absent in the lemon products investigated.

We can confirm the previously discussed lime markers bergapten and isopimpinellin, even though we were able to quantify comparatively low contents of both psoralens in lemon peel oil (**Table 2.2** - 6 out of 26 samples in the case of isopimpinellin, and 11 out of 26 in the case of herniarin, quantifiable in minimum two laboratories) in agreement with Dugrand et al. (2013) and solely bergapten in agreement with Frerot et al. (2004). Therefore, it can be said that a consideration of the theoretical input of the previously mentioned components by oil contents in the lemon juice and lemon juice concentrate must be done to assess the contents especially of the psoralens bergapten and isopimpinellin as well as the other substances.

Considering the maximum content of volatile oils of the Reference Guideline for Lemon of the AIJN Code of Practice (0.5 mL/L) (AIJN European Fruit Juice Association, 2019) and by means of the maximum contents of coumarins, psoralens and polymethoxyflavones in lemon peel oils (**Table 2.3**) determined in our study, we carried out a "worst-case" calculation by means of the rule of three of the sources of entry of industry-wide accepted oil contents in lemon juices and juice concentrates. The theoretical influence of the lemon oils to the possible intake of target substances to a juice or juice concentrate is small if the maximum content of 0.5 mL/L volatile oil is not exceeded. When in doubt if the content of volatile oils in the juice is conform with the good industrial practice as laid down in the CoP, oil content should be measured with the official Scott method, IFU Method No. 45 (2005).

Table 2.3. *Maximum contents of coumarins, psoralens, and polymethoxylated flavons in lemon (Citrus limon [L.] Burm. f.) oils, calculated values assuming a volatile oil content of 0.5 mL per litre of lemon juice, and proposed maximum levels in lemon juices.*

	MAX. CONTENTS IN LEMON OILS[A]	CALC. CONTENTS IN LEMON JUICE	PROPOSED MAXIMUM LEVELS
Limettin [1]	1219.60	0.61	n/a[c]
Bergapten [2]	75.30	0.04	< 0.20
Herniarin [3]	41.00	0.02	< 0.25
Isopimpinellin [4]	27.20	0.01	< 0.03
Nobiletin [5]	9.20	0.00	< 0.03
Sinensetin [6]	tr.[b]	0.00	< 0.02
Tangeritin [7]	tr.	0.00	< 0.02

[a]: Values represent maximum values of three analytical replicates (n = 3) per sample with exception of nobiletin (two analytical replicates per sample, n = 2). Expressed as mg/L
[b]: tr. (traces): < LoQ (5.00 mg/L) for peel oils
[c]: n/a, Limettin is not suitable as marker parameter for foreign species.

Based on this pragmatic calculation and on our analytical results, we propose the maximum contents listed in **Table 2.3** for industrial practice in the assessment of lemon juices: Limettin cannot be used as a marker parameter for foreign fruit additives, as we were able to detect this coumarin in all lemon products regardless of processing intensity. The coumarin herniarin is a parameter that occurs in lemons but is strongly influenced by processing intensity, as previous work has shown (Jungen et al., 2021). Therefore, we recommend a content of < 0.25 mg/L as an assessment limit. Regarding the psoralens bergapten and isopimpinellin, we propose a limit of < 0.20 and < 0.03 mg/L, respectively. For the polymethoxyflavones, which are considered marker substances for species such as orange

or mandarin, among others, we propose threshold levels of < 0.03 mg/L for nobiletin and < 0.02 mg/L for sinensetin and tangeritin.

For contents of the critical psoralens and polymethoxyflavones above the maximum levels suggested in **Table 2.3**, further analysis of the volatile oil content is recommended. If the volatile oil content is below or equal to 0.5 mL/L, a carry-over or an addition of other *Citrus* species must be suspected and a detailed root cause analysis must be carried out.

2.4. CONCLUSIONS AND OUTLOOK

Comparing the analytical profiles of the samples examined in our study, it can be summarised that existing lime markers could be confirmed. Furthermore, significant differences in the contents of the coumarins herniarin and limettin, the psoralens bergapten, isopimpinellin and the polymethoxyflavones nobiletin, sinensetin and tangeritin could be resolved in the comparison of fruit juices and concentrates to peel oils. Our comparison of lemon peel oils with lime peel oils just provided clear distinguishing features.

The elevated presence of bergapten, herniarin and isopimpinellin in lime peel oils in contrast to lemon juice and oil clearly illustrates the analytical difference between the two species. In accordance with literature values, low contents (in the range of the LoQ) are possible although rare, based on the data we have presented. Higher contents, especially of isopimpinellin in lemon peel oil, should be critically questioned.

Given that the amounts of sinensetin, nobiletin and tangeritin, found in lemon juices and oils were in trace region, this could be explained by presence of foreign *Citrus* fruits or a carry-over issue. Presence of foreign *Citrus* fruits in lemon juice could be a sign of adulteration but could as well point to GMP issues during production, such as inadequate sorting of the fruits in juice production. and lack of a defined specification in the procurement of fresh lemons, where the presence of lemon hybrids in the orchards of the suppliers is not considered. Furthermore, it has been proven that the contents of the observed substances in the endocarp of the fruits, in the juices and juice concentrates produced from the endocarp according to the European Fruit Juice Directive, must be lower than in the flavedo from which the peel oil is extracted. Regarding the carry-over, this point is clearly GMP-related since it concerns unsuitable cleaning process between the different batches of *Citrus* juices.

With our pragmatic proposal for possible assessment thresholds, we would like to stimulate further discussion in the scientific community. By employing the limits set out in this work we hope to provide further guidelines for evaluation of the juices marketed on the European market, and further define the line between unintentional and intentional deviations in the profile of lemon juice. Further developments in instrumental analysis as well as a larger data basis should have an impact on the maximum levels proposed by us in the future. Therefore, our current proposals only represent a status quo.

2.5. REFERENCES

AIJN European Fruit Juice Association (2019). *6.6 Reference Guideline for Lemon Juice.*

Antweiler, R. C., & Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. *Environmental Science & Technology*, *42*(10), 3732–3738. https://doi.org/10.1021/es071301c

Cautela, D., Laratta, B., Santelli, F., Trifirò, A., Servillo, L., & Castaldo, D. (2008). Estimating Bergamot Juice Adulteration of Lemon Juice by High-Performance Liquid Chromatography (HPLC) Analysis of Flavanone Glycosides. *Journal of Agricultural and Food Chemistry*, *56*(13), 5407–5414. https://doi.org/10.1021/jf8006823

Costa, R., Russo, M., Grazia, S. de, Grasso, E., Dugo, P., & Mondello, L. (2014). Thorough investigation of the oxygen heterocyclic fraction of lime (Citrus aurantifolia (Christm.) Swingle) juice. *Journal of Separation Science*, *37*(7), 792–797. https://doi.org/10.1002/jssc.201300986

Dugo, G., & Mondello, L. (Eds.) (2011). *Medicinal and aromatic plants - industrial profiles: Vol. 49*. *Citrus oils: Composition, advanced analytical techniques, contaminants, and biological activity*. Boca Raton, Fla.: CRC Press.

Dugo, P., Piperno, A., Romeo, R., Cambria, M., Russo, M., Carnovale, C., & Mondello, L. (2009). Determination of oxygen heterocyclic components in citrus products by HPLC with UV detection. *Journal of Agricultural and Food Chemistry*, *57*(15), 6543–6551. https://doi.org/10.1021/jf901209r

Dugrand, A., Olry, A., Duval, T., Hehn, A., Froelicher, Y., & Bourgaud, F. (2013). Coumarin and Furanocoumarin Quantitation in Citrus Peel via Ultraperformance Liquid Chromatography Coupled with Mass Spectrometry (UPLC-MS). *Journal of Agricultural and Food Chemistry*, *61*(45), 10677–10684. https://doi.org/10.1021/jf402763t

Dugrand-Judek, A., Olry, A., Hehn, A., Costantino, G., Ollitrault, P., Froelicher, Y., & Bourgaud, F. (2015). The Distribution of Coumarins and Furanocoumarins in Citrus Species Closely Matches Citrus Phylogeny and Reflects the Organization of Biosynthetic Pathways. *PLoS ONE*, *10*(11). https://doi.org/10.1371/journal.pone.0142757

European Council (2012). *Fruit juice Directive 2012/12/EU* (OJ L 115, 27.4.2012, p. 1–11). Fruit juice Directive 2012/12/EU.

European Food Safety Authority (2010). Management of left-censored data in dietary exposure assessment of chemical substances. *EFSA Journal*, *8*(3), 1557. https://doi.org/10.2903/j.efsa.2010.1557

Food and Agriculture Organization of the United Nations. FAOSTAT - Crops 2019. Retrieved from http://www.fao.org/faostat/en/#data/QC

Frérot, E., & Decorzant, E. (2004). Quantification of Total Furocoumarins in Citrus Oils by HPLC Coupled with UV, Fluorescence, and Mass Detection. *Journal of Agricultural and Food Chemistry*, *52*, 6879–6886. https://doi.org/10.1021/jf040164p

George, B. J., Gains-Germain, L., Broms, K., Black, K., Furman, M., Hays, M. D., . . . Simmons, J. E. (2021). Censoring Trace-Level Environmental Data: Statistical Analysis Considerations to Limit Bias. *Environmental Science & Technology*, *55*(6), 3786–3795. https://doi.org/10.1021/acs.est.0c02256

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hofsommer, H.-J. (1999). New Analytical Techniques for Judging the Authenticity of Fruit Juices. *Fruit Processing*, *No. 12*, page 471-479.

CBI - Centre of for the promotion of Imports from developing countries (2020, March 10). *The European market potential for fresh lemons* [Press release]. Retrieved from https://www.cbi.eu/market-information/fresh-fruit-vegetables/lemons/market-potential

IFU - International Fruit and Vegetable Juice Association (2005). *IFU Method 45 - Determination of essential oils [Scott method].* (IFU Method 45). Zug, Switzerland: IFU - International Fruit and Vegetable Juice Association: IFU - International Fruit and Vegetable Juice Association.

J. D. Hunter (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.5281/zenodo.592536

Jahromi, R., Pratt, H., Zhou, Y., Reimann, L., & Hammon, A. D. (2015). *Recent Developments to Detect Lemon Juice Adulteration.* AOAC. Annual AOAC meeting, Los Angeles, CA, USA. Retrieved from https://www.eurofinsus.com/media/447792/lemon-juice-aoac-2015.pdf

Jamin, E., Martin, F., Santamaria-Fernandez, R., & Lees, M. (2005). Detection of exogenous citric acid in fruit juices by stable isotope ratio analysis. *Journal of Agricultural and Food Chemistry*, *53*(13), 5130–5133. https://doi.org/10.1021/jf050400b

Jungen, M., Lotz, P., Patz, C.-D., Steingass, C. B., & Schweiggert, R. (2021). Coumarins, psoralens, and quantitative 1H-NMR spectroscopy for authentication of lemon (Citrus limon [L.] Burm.f.) and Persian lime (Citrus × latifolia [Yu.Tanaka] Tanaka) juices. *Food Chemistry*, *359*, 129804. https://doi.org/10.1016/j.foodchem.2021.129804

Klimek-Szczykutowicz, M., Szopa, A., & Ekiert, H. (2020). Citrus limon (Lemon) Phenomenon-A Review of the Chemistry, Pharmacological Properties, Applications in the Modern Pharmaceutical, Food, and Cosmetics Industries, and Biotechnological Studies. *Plants (Basel, Switzerland)*, *9*(1). https://doi.org/10.3390/plants9010119

Lehnert N., & Ara V. (2014). Authenticity analysis of lemon juices concerning the adulteration lime. *Fruit Processing*. (Nov/Dec), 242–248.

Lehnert N., Schmidt M., & Ara V. (2017). Authenticity proof of lemon juices by means of fingerprint methods. *Fruit Processing*, 314–318.

Li, G., Rouseff, R., Cheng, Y., Zhou, Q., & Wu, H. (2021). Comprehensive identification and distribution pattern of 37 oxygenated heterocyclic compounds in commercially important citrus juices. *LWT*, *152*, 112351. https://doi.org/10.1016/j.lwt.2021.112351

McHale, D., & Sheridan, J. B. (1989). The oxygen heterocyclic compounds of Citrus peel oils. *Journal of Essential Oil Research*, *1*(4), 139–149. https://doi.org/10.1080/10412905.1989.9697775

Ooghe, W. (2001). HPLC analysis of polymethoxyflavones: a collaborative study. Biologically-active phytochemicals in food. *Royal Society of Chemistry, Cambridge (UK)*, 7 p.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, *12*, 2825–2830.

Pupin, A. M., Dennis, M. J., & Toledo, M. (1998). Polymethoxylated flavones in Brazilian orange juice. *Food Chemistry*, *63*, 513–518. https://doi.org/10.1016/S0308-8146(98)00033-8

Python Software Foundation. Python (Version 3.5) [Computer software]. Retrieved from http://www.python.org

Rinke, P. (2016). *Tradition Meets High Tech for Authenticity Testing of Fruit Juices*. Woodhead Publishing is an imprint of Elsevier. https://doi.org/10.1016/B978-0-08-100220-9.00023-0

Russo, M., Rigano, F., Arigò, A., Dugo, P., & Mondello, L. (2021). Coumarins, Psoralens and Polymethoxyflavones in Cold-pressed Citrus Essential Oils: a Review. *Journal of Essential Oil Research*, *33*(3), 221–239. https://doi.org/10.1080/10412905.2020.1857855

SGF International e.V. Official website of SGF International e.V. Retrieved from https://www.sgf.org/

Stéfan van der Walt, & Jarrod Millman (Eds.) (2010). *Proceedings of the 9th Python in Science Conference.*

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Chairs), *Proceedings of the 9th Python in Science Conference*.

Xi, W., Lu, J., Qun, J., & Jiao, B. (2017). Characterization of phenolic profile and antioxidant capacity of different fruit part from lemon (Citrus limon Burm.) cultivars. *Journal of Food Science and Technology*, *54.* https://doi.org/10.1007/s13197-017-2544-5

Zhao, Z., He, S., Hu, Y., Yang, Y., Jiao, B., Fang, Q., & Zhou, Z. (2017). Fruit flavonoid variation between and within four cultivated Citrus species evaluated by UPLC-PDA system. *Scientia Horticulturae*, *224*, 93–101. https://doi.org/10.1016/j.scienta.2017.05.038

3. CONCLUSION

Les résultats de quantification obtenus par les trois laboratoires sont très similaires, bien que les méthodes d'analyse soient différentes. Des différences plus importantes ont été

observées pour l'analyse des huiles essentielles ; ces différences sont probablement liées à la préparation des échantillons qui est différente pour chaque laboratoire.

Cette étude a permis de confirmer trois composés comme marqueurs du citron vert (7-méthoxycoumarine, bergaptène et isopimpinelline). D'après les résultats d'analyse, la limettine ne peut être considérée seule comme un marqueur car elle a été retrouvée dans tous les échantillons étudiés. Les trois autres composés d'intérêt (nobilétine, sinensétine et tangéritine) quant à eux indiquent la présence d'autres espèces de citrus (comme les oranges ou les pamplemousses).

Les échantillons d'huile essentielle se différencient des échantillons de jus de par les hautes teneurs obtenues dans les différents composés étudiés. Bien que le bergaptène et l'isopimpinelline soient des marqueurs du citron vert, ces composés sont aussi détectés et quantifiés dans les huiles essentielles de citron jaune. D'après les résultats d'analyse, les teneurs de ces 7 composés varient en fonction de l'espèce de citrus et du procédé utilisé.

Au vu des résultats de quantification, et en accord avec les trois laboratoires participants, des valeurs seuils ont été proposées afin de garantir l'authenticité du jus de citron jaune. Les valeurs hautes suivantes sont ainsi recommandées : 0,25 mg/L pour la 7-méthoxycoumarine ; 0,2 mg/L pour le bergaptène ; 0,03 mg/L pour l'isompimpinelline et la nobilétine ; 0,02 mg/L pour la sinensétine et la tangéritine. Ces valeurs peuvent de ce fait servir de guide pour contrôler l'authenticité du citron jaune. Elles pourraient également être reprises dans des réglementations officielles (comme l'AIJN) pour garantir la conformité du citron jaune.

CHAPITRE 3 ANALYSE NON CIBLEE POUR L'AUTHENTIFICATION DU JUS DE POMMES

1. INTRODUCTION ET RESUME DE L'ARTICLE

Le but de cet article est de mettre en place une méthode d'analyse non ciblée par LC-HRMS pour l'authentification des jus de pommes selon deux scénarios : la discrimination des purs jus et des jus concentrés, et la discrimination des jus issus de l'agriculture biologique et des jus issus de l'agriculture conventionnelle.

Il s'agit ici d'une étude préliminaire sur l'authentification de jus de pommes par analyse non ciblée LC-HRMS de type métabolomique. Les approches métabolomiques semblent prometteuses pour répondre aux problématiques de contrôle de l'authenticité en agroalimentaire. En effet, l'utilisation d'une méthode générique couplée à des outils chimiométriques permet la mise en évidence de signaux caractéristiques capables de discriminer des groupes d'échantillons. De premières études ont montré l'intérêt d'une telle méthodologie pour l'authentification de l'origine géographique ou la détection d'adultération.

Il est important de développer de telles méthodologies non ciblées car les méthodes actuelles de contrôle d'authenticité peuvent faillir à la mise en évidence de fraudes. En effet, ces méthodes conventionnelles utilisées actuellement pour s'assurer de l'authenticité sont des méthodes dites « ciblées » et visent à détecter des composés ou des familles de composés connus. Ces méthodes sont généralement décrites et validées par différents organismes comme l'IFU pour le cas des jus de fruits.

Pour cette étude préliminaire, il a été choisi de travailler sur les jus de pommes du fait de la grande variabilité existante pour ce type de matrice : les variétés, les modes de productions, l'origine géographique, la maturité du fruit, les conditions de stockage et les procédés de transformation.

Des échantillons de jus de pommes provenant de différents modes de production et modes d'agriculture ont été analysés par LC-HRMS. La méthode d'analyse non ciblée mise en place a été la plus générique possible : gradient d'élution large (de 3 à 98 % de phase

organique), colonne de silice greffée C18, et paramètres de sources standards permettant l'ionisation d'un maximum de composés.

Le traitement des données a été réalisé à partir d'outils en ligne, sur la plateforme Workflow4Metabolomics. Les différentes étapes du traitement des données brutes acquises par LC-HRMS sont présentées en **Figure 3.1**. Les pics chromatographiques ont été dans un premier temps extraits à l'aide des fonctions de XCMS. Puis, différentes étapes de filtration des données ont été appliquées permettant de supprimer environ 50 % des features détectés. Des corrections intra et intersessions ont été ensuite effectuées dans le but de rendre comparables les différents échantillons. Enfin, différents outils chimiométriques (ACP, PLS-DA, OPLS-DA, ANOVA) ont été utilisés pour identifier les features discriminants.

Des analyses MS/MS ont également été réalisées sur de nouveaux échantillons afin d'identifier les composés marqueurs d'authenticité.

2. CORPS DE L'ARTICLE

Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics

Katy Dinis [a,b], Lucie Tsamba [a]*, Freddy Thomas [a], Eric Jamin [a], Valérie Camel [b]

[a] Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323, Nantes Cedex 3, France

[b] UMR SayFood, Université Paris-Saclay, INRAE, AgroParisTech, 91300 Massy, France

***** Corresponding author**:** Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323 NANTES Cedex 3, France, tel.: +33 2 51 82 55 39, fax: +33 2 51 83 21 11. E-mail address: LucieTsamba@eurofins.com

Food Control (2022), sous presse. https://doi.org/10.1016/j.foodcont.2022.109098

Abstract

In this work, apple juice samples from different farming and production processes (direct and concentrated juices; organic and conventional juices) were analyzed by ultra-high

performance liquid chromatography coupled to high resolution mass spectrometry (UHPLC-HRMS). A workflow was developed and implemented for data processing using the Workflow4Metabolomics (W4M) platform. First, features were detected using XCMS, and next data filtration steps were applied leading to the removal of nearly 50% of the detected features. Intra- and inter-batch correction was then performed, followed by chemometric tools (PCA, PLS-DA, OPLS-DA, ANOVA). The developed approach successfully discriminated apple juice samples in two distinct scenarios simultaneously (direct *vs.* concentrated juices and organic *vs.* conventional juices). PCA highlighted the reproducibility of the method and confirmed the efficiency of batch corrections. OPLS-DA models showed good quality metrics, particularly after feature selection for organic *vs.* conventional juices discrimination (almost 80% of predictive ability). Based on ANOVA and OPLS-DA results, 24 features were retained as significantly discriminant. Among them, some compounds were identified as amino-acids and derivatives, using additional MS/MS experiments and online databases. An independent data set was used to evaluate their potential as marker compounds, with promising results obtained. Further investigation is needed to validate such an untargeted method and its routine application to detect apple juice adulteration and confirm its authenticity.

Keywords

Food authenticity, High resolution mass spectrometry, Liquid chromatography, Metabolomics, PCA, PLS-DA, OPLS-DA

Highlights

- Development of an UHPLC-HRMS metabolomics approach with great potential in juice authentication
- Discrimination between sample groups in two distinct authentication applications
- Relevant markers selected by OPLS-DA and ANOVA were tentatively identified
- Main discriminant compounds were identified as amino-acids and derivatives

2.1. INTRODUCTION

Food fraud is a worldwide issue, and recent crises (like the horse meat scandal in 2013) have sparked interest on food authentication among consumers and food industries (Brooks et al., 2017). Concerning food fraud and Economically Motivated Adulteration between 1980 and 2010, fruit juices are one of the top ten products most at risk, particularly apple and orange juices (Moore et al., 2012). Typical frauds on fruit juices include (1) dilution with water, (2) addition of sugars or organic acids, (3) addition of foreign fruits (mostly cheaper ones) and (4) false labeling of the product (cultivar or geographical origin, as well as production mode such as organic) (Vaclavik et al., 2011).

To facilitate the detection of food fraud, the Codex Alimentarius, the Association of the Industry of Juices and Nectars of the European Union (AIJN) and the European Commission have established guidelines and standards to define permitted practices and evaluate the quality and authenticity of juices (Directive 2012/12/EC; CODEX STAN 247-2005; AIJN Code of practice). However, fruit juices authentication may be challenging due to their complex chemical composition influenced by several factors such as variety, geographical origin, stage of maturity, storage conditions and processing techniques (Jandric et al., 2014; Cubero-Leon et al., 2018; Dasenaki et al., 2019).

Juice authentication is routinely performed by conventional analytical methods (called targeted methods) that are usually described and validated by the IFU (International Fruit and Vegetable Juice Association) (IFU website). For example, sugars, organic acids, minerals, phenolic compounds and several volatile compounds are analyzed to authenticate direct apple juice samples (AIJN Code of practice, Wolter et al., 2008). These methods are sensitive and usually provide low limits of detection and quantification as they have been developed to detect specific compounds or classes of compounds (e.g., molecular markers of foreign fruits or low fruit content). However, these targeted approaches generally focus on a specific fraud and may fail to reveal more sophisticated frauds such as false organic claims (Knolhoff and Croley, 2016; Dasenaki et al., 2019). The illegal addition of vegetable water (such as water obtained during the grape juice concentration process) to orange juice concentrate, with the false claim of “orange juice not from concentrate”, is another illustrative example since conventional $^{18}O/^{16}O$ isotope ratio analysis fails to detect this fraud; in that

case, there is an additional health concern related to the presence of allergenic sulphur dioxide (Rinke and Jamin, 2018).

Therefore, it is important to move toward untargeted methods to detect adulteration and confirm authenticity (Dasenaki et al., 2019; Rinke, 2016). Untargeted methods allow to have an overview of the sample, also called a fingerprint (Medina et al., 2019). Thousands of compounds can be detected, making them more holistic than the conventional methods (Dasenaki et al., 2019). These untargeted methods have emerged with the improvements of analytical techniques (e.g., the development of high resolution mass spectrometers) and the use of advanced statistical methods. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are widely used in the assessment of food authentication using untargeted methodology, in particular liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS) (Cubero-Leon et al., 2014; Sobolev et al., 2019; Esteki et al., 2018; Danezis et al., 2016).

Metabolomics-based approaches using LC-HRMS have already been used in food safety assessment (Knolhoff et al., 2016; Delaporte et al., 2019) and revealed their potential. In the field of food authentication, untargeted LC-HRMS analysis coupled to chemometrics has been used to attest the geographical origin of saffron (100% of the investigated samples were correctly classified) (Rubert et al., 2016). Using a similar methodology, Cavanna and co-workers have assessed the authentication of durum wheat based on geographical origin, with approximately 90% of samples correctly classified (Cavanna et al., 2020). Moreover, metabolomics-based methodology using LC-HRMS has already been successfully implemented for juice authentication regarding geographical origin (Diaz et al., 2014) or for adulteration detection and classification of juices types and varieties (Vaclavik et al., 2012; Jandric et al., 2014; Jandric et al., 2017). Similarly, Dubin et al. used this methodology to authenticate blackcurrant, specifically to detect adulteration with aronia, with a detection limit of 5% aronia concentrate in blackcurrant concentrate (Dubin et al., 2017). Furthermore, the untargeted methodology was also used for pomegranate juice authentication allowing detection of 1% adulteration (Dasenaki et al., 2019).

Thereby, the metabolomics-based methodology appears to be a method of choice for juice authentication. However, this trend deserves confirmation and further methodological development. In particular, studies with large data sets and/or with models that offer broad

applicability are needed to validate the potential of this methodology for food authentication (Cubero-Leon et al., 2018). Moreover, the authentication of organic food also requires further work due to limited number of studies regarding this topic, especially in the juice sector (Cuevas et al., 2017; Mihailova et al., 2021), and the lack of reliable analytical techniques to confirm the organic production of a sample (Cuevas et al., 2019).

In this work, apple juice samples from different farming and production processes (organic and conventional, direct and from concentrate juices) were analyzed using untargeted UHPLC-HRMS analysis. A data processing workflow was developed to select relevant features after peak detection. Based on these features, models were built for sample groups discrimination using chemometric tools in two distinct scenarios (direct juice *vs*. concentrated juice, and organic juice *vs*. conventional juice). Chemical markers allowing the discrimination were then tentatively identified, using online and in-house databases as well as UHPLC-HRMS/MS analyses. The discriminant potential of these marker compounds was evaluated using an independent set of samples.

2.2. MATERIALS AND METHODS

2.2.1. Reagents and chemicals

Methanol (MeOH), water and formic acid (FA), all LC-MS grade, were purchased from Fisher Scientific.

Some compounds known to be present in apple juice, and routinely analyzed by targeted methods, were purchased from Sigma-Aldrich: alpha-terpineol (purity: >99%), hexyl acetate (purity: 99%), ethyl 2-methylbutyrate (purity: > 98%), limonene (purity: 97%), phloridzin (purity: > 99%) and 2-methylbutyl acetate (purity: > 99%). Hydroxymethylfurfural (purity: 100%) was purchased from ACROS. These commercial standards were considered to assess the ability of our untargeted method to detect them. In addition, they may be considered as possible candidates for markers responsible of the discrimination between our sample groups. With the aim to develop a real untargeted method, these target compounds were not included in our inclusion list in our MS/MS experiments. Individual standard solutions were prepared in methanol with a concentration of 0.2 mg/L for most compounds, and of 0.5 mg/L for hydroxymethylfurfural and phloridzin. These solutions were analyzed using the UHPLC-HRMS analytical conditions described in section 2.3, in order to

determine the *m/z* and retention time (RT) of the compounds which will be used to highlight their potential presence in the analyzed samples.

2.2.2. Samples description and preparation

One hundred and ten apple juice samples from several geographical origins and farming processes were collected (organic and non-organic juices; direct juices, concentrated juices and juices from concentrate). Samples were stored in the freezer until analysis. After thawing, aliquots (5 ml) of samples were centrifuged for 10 min at 4,500 rpm. The supernatant was collected and diluted with water directly into a vial before analysis. Three replicates per sample were prepared. Sample vials were randomized in the analytical sequence. Quality Control (QC) samples (pool of apple juice samples) and diluted QC samples were also prepared and analyzed every 10 injections. The repeated injections of QC samples were used to evaluate analytical performance. Also, analytical blanks were analyzed regularly to check for carry over (every 20 samples). Moreover, these blanks were useful to detect residual peaks corresponding to the mobile phases used.

Samples were analyzed in different batches. The first one contained 24 samples (12 organic and 12 conventional apple juices). The second batch contained 30 samples (15 direct apple juices and 15 concentrated apple juices). The third batch contained 26 samples including organic and conventional juices as well as direct and concentrated juices. Another set of samples coming from a different harvest year was also analyzed in a fourth batch in which MS/MS acquisition was performed; this batch contained 30 samples (10 concentrated juice samples, 10 conventional direct juice samples and 10 organic direct juice samples). A detailed list of the samples is presented in **Table 3.A.1** (Supplementary material).

2.2.3. Analytical method

Analyses were performed on a ThermoFisher® Vanquish Flex UHPLC system, composed of a binary pump, refrigerated sampler and column oven, connected to a ThermoFisher® QExactive Plus Orbitrap® high resolution mass spectrometer (version 2.9) with a heated electrospray ion source (HESI). The UHPLC separation was achieved using a C18 Hypersil Gold column (150 x 2.1 mm, 1.9 µm) at a 0.3 mL/min flow-rate. The column temperature was set to 30°C. The mobile phases were water acidified with 0.1% FA (A), and MeOH acidified with 0.1% FA (B), with the following linear gradient elution: 0-2 min, B: 3%; 2-20

min, B: 3-98%; 20-24 min: B: 98%; 24-24.1 min, B: 98-3%; 24.1-32 min, B: 3%. The injection volume was 1 µL.

Raw data were acquired using TraceFinder software (version 3.1, ThermoFisher®). MS data were acquired in positive ion mode (ESI+) with a mass range set at *m/z* 120-1000 in full scan mode and with a resolution of 70,000. The parameters applied on the electrospray ion source are presented in **Table 3.A.2** (Supplementary material); MS data was acquired in centroid mode. The MS detector was weekly calibrated using the Pierce™ positive and negative ion calibration solution purchased from Thermo Fisher Scientific.

For MS/MS acquisition, full scan data-dependent analyses were carried out using an inclusion list. This inclusion list was established after the data processing of the first three batches where several features were identified as discriminant (24 features for both studies). The resolution was set at 17,500. An isolation window of ± 1 uma was used to select the *m/z* of interest at the expected retention time of the features (± 1 min). Three normalized collision energies were applied (10; 30 and 60 eV) for the MS/MS spectrum acquisition.

2.2.4. Data processing

Raw data files were analyzed using the Workflow4Metabolomics (W4M) platform (version 3.0) (Giacomoni et al., 2015) after conversion of the data files to mzXML format using ProteoWizard (Chambers et al., 2012). The main steps of data processing are: (1) peak detection; (2) retention time alignment; (3) peaks grouping; (4) peak annotation; (5) data filtration and normalization, and (6) chemometric analysis. The first four steps were performed using functions of the XCMS (an acronym for various forms (X) of chromatography mass spectrometry) package (Smith et al., 2006) on the W4M platform as illustrated in **Figure 3.1**.

The features, defined by their *m/z* and retention time, and their intensities in different samples were used for the statistical analysis as commonly reported (Cavanna et al., 2018). The chemometric methods used were principal component analysis (PCA) for exploratory purpose, as well as partial least squares - discriminant analysis (PLS-DA), and orthogonal partial least squares - discriminant analysis (OPLS-DA) in order to build models for discrimination and classification of samples groups. Also, analysis of variance (ANOVA) and

biosigner were used to reduce the number of features selected for models building, which may improve models quality.

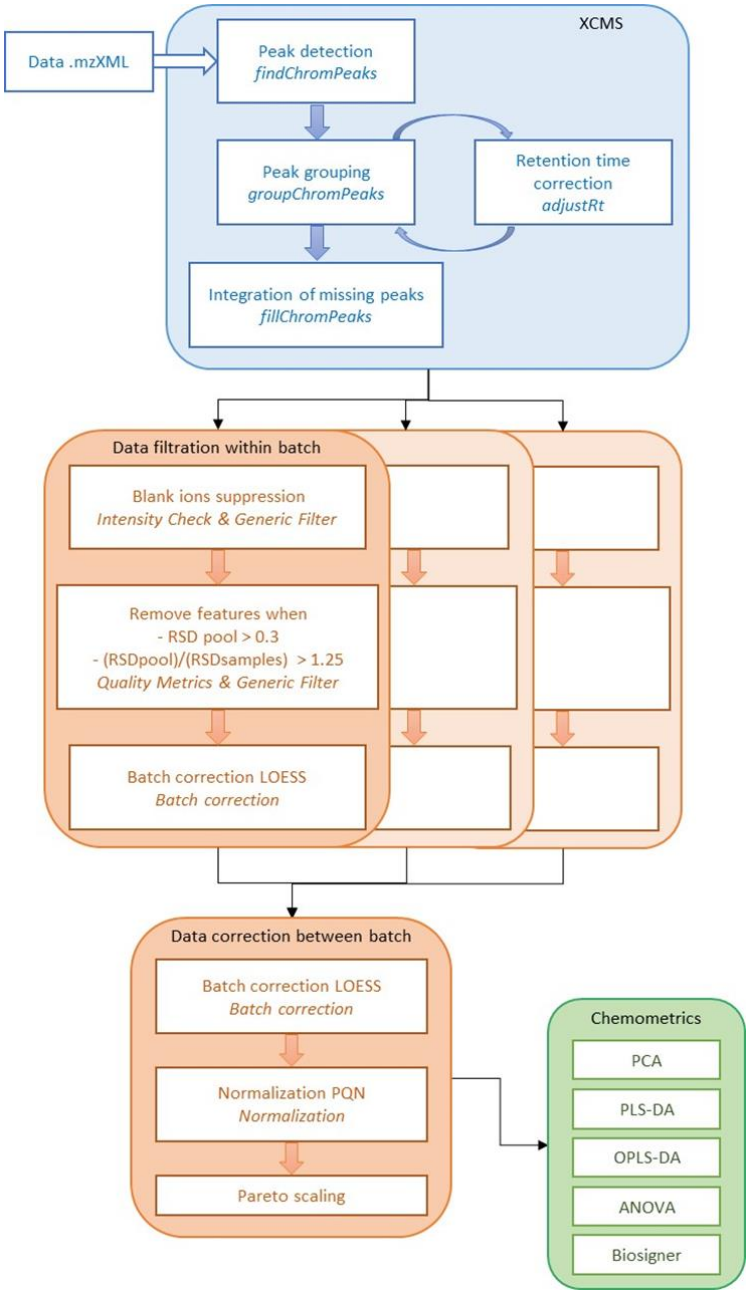

Figure 3.1. *Workflow of the data treatment using W4M* (RSD: relative standard deviation) * text in italic refers to W4M functions.*

2.2.4.1. Peak detection and alignment (XCMS)

All the data from the first three analytical batches were processed simultaneously as illustrated in **Figure 3.1**. The XCMS phase includes the following steps. First, the peak

detection and extraction is achieved using the “findChromPeaks” function with the centWave method (Tautenhahn et al., 2008). During this step, the chromatograms are described as a 2D-matrix where each peak is described by a combination of its *m/z* value and retention time (RT), called “feature”. The selected retention time for each peak is the time corresponding to its apex of the intensity value. Then, the “groupChromPeaks” function is used to group the extracted peaks across all the samples. This step is applied to group ions with close RT between the samples. After this step, *m/z* and RT values are averaged in the data matrix. The peaks are next aligned using the “adjustRtime” function to correct the RT across the samples and then grouped again. Finally, the “fillChromPeaks” function is used to identify features where there is no intensity value for some samples and the signal is integrated in the region of the determined feature to avoid missing values. The XCMS parameters for each step were optimized from the QC samples and are presented in **Table 3.A.3** (Supplementary material). A data matrix is then generated, giving the area of each peak for each feature and for each sample. Thus, the features are the variables of the models presented in this study.

2.2.4.2. Data filtration and batch correction

To perform the subsequent data filtration step, the data matrix was split in three distinct data matrices, as shown in **Figure 3.1**, corresponding to the three initial analytical batches, in order to perform filtration steps within each analytical batch. These filtration steps were needed to remove irrelevant information as the number of features detected by XCMS was very high (about 20,000).

First, all peaks corresponding to the dead volume and the column flush were excluded, which means that all the features with a retention time lower than 1.7 min were removed from the data matrix. Then, features that mainly result from blank analyses were removed by calculating the fold change in blanks and samples analyses. For a feature, when the ratio of samples fold change over blanks fold change is lower than 4, this feature is deleted. In this way, between 15% and 20% of the detected features were removed. Finally, features showing a poor stability (relative standard deviation (RSD) higher than 30%) according to QC analyses were also excluded. Similarly, features for which the ratio of RSD pool over RSD sample is higher than 1.25 were deleted. At the end of this step, about 10,000 features remained. Analytical signal drift within the analytical batch was corrected using a LOESS

regression model using the QC sample injections, employing the *Batch Correction* module on the W4M platform.

Then, the three data matrices corresponding to the three sample batches were merged (see **Figure 3.1**) and a second batch correction was applied to correct analytical signal drift between analytical batches by the use of the QC sample injections.

The data matrix was then normalized using the Probabilistic Quotient Normalization method (PQN) (Dieterle et al., 2006) using the QC samples. Its purpose is to limit potential dilution effects that can affect restricted regions of the data. First, the median of each feature in QC samples is calculated, providing a reference vector. Then, the values for each ion in samples are divided by this reference vector. A median of the ratios for each sample is generated. Finally, initial values of each sample are divided by the ratios median.

Prior to chemometrics analysis, the data matrix was Pareto scaled. Then, the data matrix was split to create two distinct authentication studies: the first one contained samples from batches 2 and 3 to evaluate the discrimination between pure and concentrated juice samples (58 samples in the data set); the second study contained samples from batches 1 and 3 to evaluate the discrimination between organic and conventional juice samples (54 samples in the data set).

2.2.4.3. Chemometrics

Multivariate statistical analyses were performed on the W4M platform using unsupervised and supervised techniques. PCA was first performed to have an initial visualization of the data sets and to detect outliers. In order to evaluate the ability of this methodology to discriminate the apple juice samples, PLS-DA and OPLS-DA were used. These models were built using a 7-fold cross validation; by this way, each data set was divided into 7 different parts. Each model was next built using 6 parts (train set) and tested using the 7^{th} part (test set); this step was then iterated until all the parts were used as test set. The cross validation procedure permitted to determine the optimal number of latent variables (LV) to build the PLS-DA and OPLS-DA models (Ballabio and Consonni, 2013; Wold et al., 2001). A new LV was added if the Q2Y obtained with this LV was greater than 0.01. Indeed, the Q2Y was calculated from the ratio of PRESS (predictive residual sum of squares) including the new LV over RSS (residual sum of squares) calculated from the model with the previous

LV (Wold et al., 2001). The quality of the built models was assessed by the goodness of fit (R2X), the proportion of the response matrix variance explained by the model (R2Y) and the predictive performance of the model (Q2Y). These three metrics have values between 0 and 1. The higher they are, the better the performance of the model. The Q2Y metric is particularly important here, as it represents the prediction efficiency of the model. An empirical value of 0.4 for Q2Y has been previously established to judge the quality of the model (Worley and Powers, 2012).

An analysis of variance (ANOVA) was also performed to select significant features between the two studied groups (pure *vs* concentrated juices and organic *vs* conventional juices); a maximum accepted p-value of 0.01 was chosen in order to select significant features. The features identified by the ANOVA were used to build new PLS-DA and OPLS-DA models to compare models quality with lower features.

The biosigner tool (Rinaudo et al., 2016) present on the W4M platform was also used for feature selection. Briefly, this algorithm allows to obtain the smallest number of features which have the most significant contribution in models performance (this module performed PLS-DA, Random Forest (RF) and Support Vector Machine (SVM) models) after performing several iterations. The iterations stop when the number of significant features remains equal to that of the previous iteration. Again, the features selected by biosigner were used to build new PLS-DA and OPLS-DA models.

2.2.4.4. Annotation

Significant features were selected based on their results after the chemometric tools used, particularly OPLS-DA and ANOVA results. After having investigated the MS spectra of those discriminant features, the adduct type of the observed *m/z* was identified which permitted to determine the exact mass of the compound and consequently to suggest molecular formulas. In order to tentatively annotate these features that discriminate the samples, the online databases HMDB (Wishart et al., 2018) and FooDB (FooDB, 2021) (as we are studying apple juice samples) were used. Moreover, the obtained MS/MS spectra were used to confirm the annotation by comparing them to two spectral databases: mzCloud and MassBank. In addition, some commercial compounds known to be present in apple juices

were analyzed thanks to available standards as detailed in section 2.1, enabling to build an in-house database.

2.3. RESULTS AND DISCUSSION

2.3.1. Study 1: Authentication of pure apple juices

2.3.1.1. Principal component analysis

PCA is the most common unsupervised multivariate statistical technique (Medina et al., 2019b; Oliveri and Simonetti, 2016) used for exploratory purposes. It was used here to evaluate the reproducibility of three replicates of the same sample. For this study, 58 samples were considered with three replicates per sample (resulting in a total of 174 samples). PCA was applied on two distinct data matrices, containing either all values (i.e., including separate triplicate values) or only a single value (being the mean of the three replicates) for each sample. In both cases, no outlier was observed on the PCA score plots, so that we considered the three replicates to be reproducible. Consequently, only the average of sample triplicates was considered for the following statistical analyses.

As shown in **Figure 3.A.1** (Supplementary material), the first three principal components explained about 50% of the variance (PC1: 27%; PC2: 14%; PC3: 8%). The replicates of the QC samples were fairly close on the PCA scores plot, showing a good system stability during the analysis. A slight dispersion was noticed in **Figure 3.A.1**a, with two subsequent groups for the QC samples, in line with the two distinct analytical batches; this observation highlights an analytical drift not completely corrected. Interestingly, a trend seemed to appear for the discrimination between the two groups of samples (single strength *vs.* both concentrated juices and juices from concentrate) on the PC3 axis, even though no clear separation could be achieved.

Conversely, group separation of fruit juices based on the type of fruit were already reported using PCA on UHPLC-HRMS data, with a distinct cluster for apple juices (Vaclavik et al., 2012). Guo et al. also reported group separation of fresh squeezed apple juices based on varieties by performing a PCA on their concentrations in 23 polyphenols (Guo et al., 2013). Therefore, it can be assumed that the apple juice production method has fewer differences in the UHPLC-HRMS fingerprint, which explains why no group separation was observed in our PCA.

2.3.1.2. Classification and prediction models: PLS-DA and OPLS-DA

PLS-DA and OPLS-DA models were built using the features left after the between batch correction (9,234 features) and from the 58 samples.

PLS-DA models were already reported for classification purpose of orange juices, with a satisfactory classification rate of samples regarding geographical origin (after cross-validation, the model showed a 100% classification capacity) (Diaz et al., 2014); unfortunately, in our study PLS-DA model remained unsatisfactory (data not shown). The obtained model was built using two latent variables and had a goodness of prediction of 0.37 and a goodness of fit of 0.27. OPLS-DA models were previously found interesting for discrimination of Saffron sample origins (Rubert et al., 2016); results from our data also showed samples discrimination between single strength and both concentrated juices and juices from concentrate, as presented in **Figure 3.2**.

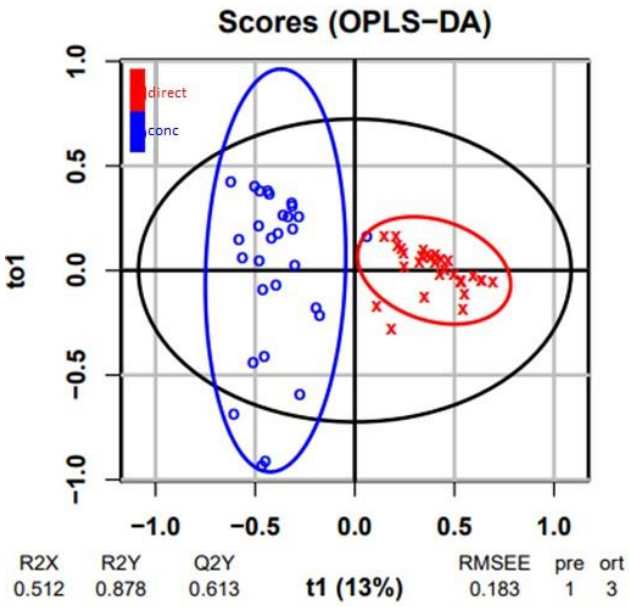

***Figure 3.2**. Scores plot for OPLS-DA obtained with cross-validation (blue circles, both concentrated juices and juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability, the blue and red ellipse are the Mahalanobis ellipse of the sample groups.*

The model metrics indicated that the OPLS-DA model was quite satisfactory (as shown in **Figure 3.2**) with a goodness of fit of about 50% and a prediction capacity of about 60%. This OPLS-DA model was built using 9,236 features, so that these metrics might be improved

with fewer features. In their study on Saffron, Rubert and co-workers reported that the best OPLS-DA model was obtained using 8 features (of about 5,000 features detected) with 85% of prediction capacity and 97% of goodness of fit (Rubert et al., 2016). Nevertheless, on **Figure 3.2** it can be observed that one concentrated juice sample was really close to the direct juice samples group. This observation can lead to incorrect classification or prediction of samples. Consequently, further data processing was tested to improve the modeling. Moreover, the high number of features used for building our OPLS-DA model may have induced overfitting. It was thus important to reduce the number of features used for this model.

2.3.1.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

Selection was needed to reduce as much as possible the number of features to be compare with other analytical batches. This is necessary if we want to implement this methodology as a routine analysis method for apple juice authentication assessment. ANOVA and similar t-test have already been used to identify and select significantly different features between groups of samples (Llano et al., 2018; Bat et al., 2018).

Performing an ANOVA proved to be greatly helpful: from the about 10,000 features obtained at the end of the filtration steps, the ANOVA identified almost 2,000 significantly different features between the two sample groups. Again, the PLS-DA model gave unsatisfactory results (data not shown). The model was built using two latent variables and had a predictive ability of 0.58. In the score plot, the two sample groups were not differentiated. Conversely, the OPLS-DA model obtained with these identified features was again quite satisfactory (based on the metrics values), with more variance being explained by the first latent variable (30% instead of 13% previously). However, the separation of the two groups was not improved, being even quite worse (**Figure 3.A.2**a of Supplementary material).

2.3.1.4. Feature selection using biosigner before PLS-DA and OPLS-DA

The biosigner module present on the W4M platform was also tested for features selection since it allows selecting the fewest number of features to build discrimination models. Accordingly, only 20 features were identified by this tool here. Unfortunately, with these 20 features, the resulting OPLS-DA model showed a worse discrimination of the two groups, even though the direct juice samples stayed close together (**Figure 3.A.2**b of

Supplementary material). The metrics of the model clearly decreased, confirming the low quality of this model. The PLS-DA model obtained was also unsatisfactory (data not shown). One latent variable was used to build this model and a predictive ability of 0.27 was obtained.

Almost all features selected by the biosigner tool were also selected by the ANOVA (90%). The 20 features seemed thus to be discriminant, but it can be emphasized that as the number of samples was quite low (58 samples), the features selected were not sufficiently discriminant to improve the OPLS-DA model quality. A larger set of reference samples would be required to establish a robust routine model.

2.3.2. Study 2: authentication of organic apple juices

2.3.2.1. Principal component analysis

In this study, 54 samples were used to build the PCA. As in the previous study, the reproducibility of the triplicates was evaluated using the PCA scores plots. As the replicates showed to be reproducible, the average of the three replicates per samples was used for the next chemometric analysis. PCA scores plots of the filtered data using the mean of the triplicates are shown in **Figure 3.A.3** of Supplementary material. A good system stability was also observed for this study since the replicates of the QC sample were clustered on the PCA scores plot. As in the previous study, two groups of QC samples could be distinguished, showing that the analytical drift was not completely corrected. However, the correction seemed to be better than in the first study because the QC replicates were less dispersed.

Group separation between organic and conventional juice samples was not achieved by PCA. Using the first three principal components, about 60% of the variance was explained (PC1: 32%; PC2: 15% and PC3: 8%). Cuevas and coworkers also reported previously that PCA did not allow to separate organic from conventional orange juices using UHPLC-HRMS analysis (Cuevas et al., 2017).

2.3.2.2. Classification and prediction models: PLS-DA and OPLS-DA

In order to build models for sample classification, PLS-DA and OPLS-DA analysis were performed (**Figure 3.A.4** of Supplementary material). PLS-DA and OPLS-DA models were both built using a 7-fold cross validation on the 54 samples of the study. OPLS-DA enabled

a clear separation between the two groups. This is in line with another study where organic and conventional juices were discriminated using an OPLS-DA model, with a specificity and sensitivity of nearly 90% for both sample classes after cross-validation (Cuevas et al., 2017). OPLS-DA models satisfactorily discriminated organic and conventional carrot samples analyzed by UHPLC-HRMS, with a classification rate of about 80% using a validation data set (Cubero-Leon et al., 2018).

The predictive ability of the OPLS-DA model was good (Q2Y: 0.746). As this model was obtained with a high number of features (near 8400 features), it could be hypothesized that it could be improved with a reduced number of features. Further works should focus on the external assessment of the models performance, which was not allowed by the number of samples in this study.

2.3.2.3. Feature selection using ANOVA before PLS-DA and OPLS-DA

In this study, the ANOVA found 1,422 significantly different features between the organic and conventional juice samples (from almost 10,000 features detected). PLS-DA and OPLS-DA models obtained from these features are presented in **Figure 3.3**.

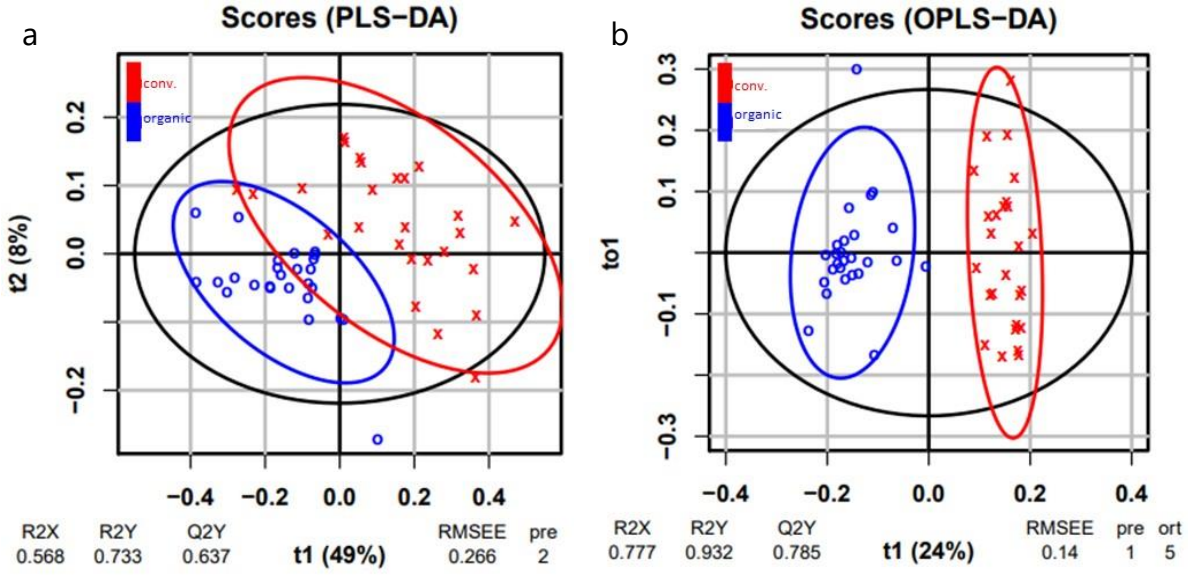

***Figure 3.3**. (a) Scores plot of PLS-DA and (b) scores plot of OPLS-DA obtained after features selection using ANOVA (blue circles: organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.*

The new models obtained with a reduced number of features showed similar metrics compared to the previously obtained models (**Figure 3.A.4** of Supplementary material). The discrimination between the two sample groups was still not observed using PLS-DA model. On the contrary, OPLS-DA model showed a clear separation. The predictive ability of this model indicated that it had a good performance (Q2Y: 0.785) and it was slightly better than the OPLS-DA obtained with all the features. The percentage of variance explained by the first LV had increased to 24% with the feature selection.

2.3.2.4. *Feature selection using biosigner before PLS-DA and OPLS-DA*

Again, the biosigner tool was used to find the smallest number of most significant features. This module found 48 features. To evaluate whether these selected features were the most significant, PLS-DA and OPLS-DA models were built using a 7-fold cross-validation. In contrast to the results obtained from the feature selection using ANOVA, the obtained PLS-DA and OPLS-DA models were not improved. The metrics showed that models were quite worse than with all the features (**Figure 3.A.5** of Supplementary material).

Only 14 of the 48 features selected by the biosigner tool were also selected by the ANOVA. Most of the features selected by this module were chosen based on their performance using SVM models. SVM models can perform very well but they require lots of data (ideally thousands of samples). In this study, there were only 54 samples, so the features selected using SVM models were not discriminant enough to increase the PLS-DA and OPLS-DA model metrics.

2.3.3. Tentative identification of discriminant features in both studies

The number of features remained high, even after selection with ANOVA (about 1,500 features). To reduce this number while keeping the most discriminant features, it was decided to filter them according to their VIP (Variable Importance on Projection) value calculated during the construction of the PLS-DA and OPLS-DA models. The VIP value of a feature indicates its importance on the model building: the higher VIP value, the more discriminating the feature is (Wold et al., 2001). In a previous study, filtration based on the VIP value successfully selected 8 features out of about 5,000 features detected (Rubert et al., 2016). Other authors reported the use of VIP values to select discriminant features by retaining 25 features (out of about 5,000 features detected) that were further tentatively

identified (Cavanna et al., 2020). Cubero-Leon and colleagues also used a similar criterion (VIP greater than 1) applied to remove features contributing to other variability (year of harvest); they were able to build successful OPLS-DA models to discriminate between organic and conventional carrot samples (Cubero-Leon et al., 2018). In the literature, different VIP values between 1 and 2 have been used as a filtration criterion. In this work, a filter was applied to keep features having a VIP value greater than 1, as proposed in different articles (Gorrochategui et al., 2016; Pezzati et al., 2020). After this filtration, about 150 features remained for both authentication applications considered in our work.

To reduce the number of features to be identified, both results from ANOVA and from OPLS-DA were used. We focused on the features with the highest VIP and the lowest p-value, and attempted to identify them using online databases (HMDB and FooDB). Based on this strategy, less than 15 features were selected in each study for further identification, as indicated in **Table 3.1** and **Table 3.A.4** (Supplementary material). Few of these features were also selected by the biosigner tool. Examples of chromatograms for one feature identified as discriminating for each study are shown in **Figure 3.4** and **Figure 3.A.6** (Supplementary material).

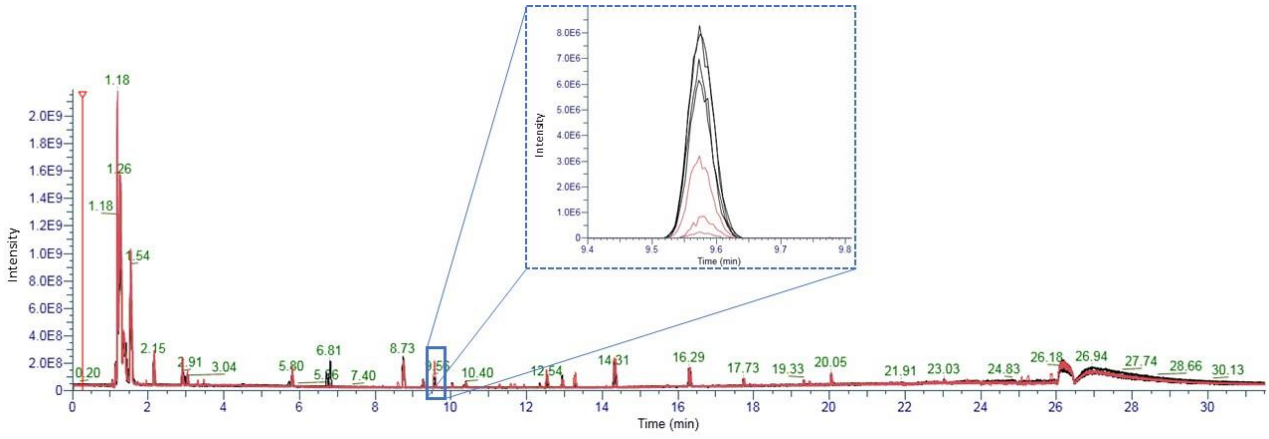

Figure 3.4. *Chromatogram of feature 13 for authentication of organic apple juices (black, organic juice samples; red, conventional juice samples)*

Some features had the same retention time, being either coeluted chromatographic peaks or fragments and/or adducts of a unique compound. The observation of the MS spectra permitted to identify features which correspond to a same molecule (**Table 3.1**); in particular, the presence of certain adducts such as $(M+NH_4)^+$ and $(M+K)^+$ allowed to attribute the adduct type of the observed *m/z*. This was mostly the case for the features identified in the

second study. As presented in **Table 3.1** and **Table 3.A.4**, a majority of features were still unknown as no matches were obtained on the online databases used. For some features, several compounds matched the exact mass defined. It is interesting to observe that results from HMDB and FooDB were really close, being a good starting point to annotate features but still insufficient: the compounds of interest may not be present on these databases.

In order to improve the annotation of these discriminant features, MS/MS acquisitions were performed on a new set of samples (30 samples containing concentrated juice samples, conventional direct juice samples and organic direct juice samples). Only few results were obtained from the databases search, either with the monoisotopic mass or with the proposed molecular formula. Interestingly, two amino acids (methionine and isoleucine or norleucine) could be proposed as discriminant markers between organic and conventional apple juice (**Table 3.1**); this result seems realistic since those compounds were already reported in apple juice samples (Ma et al., 2018). In particular a biosynthesis pathway leading to isoleucine formation in ripening apple fruit has been recently reported (Sugimoto et al., 2021). The same methodology was applied for the features identified in the authentication of pure apple juices with N-(1-deoxy-1-fructosyl)phenylalanine proposed as a marker (**Table 3.A.4** of Supplementary material). Xu et al. also reported an amino-acid (L-glutamine) to discriminate from concentrate and not from concentrate orange juices (Xu et al., 2020). Further investigation is needed to improve the annotation of the identified discriminant features, probably by using other online databases or building in-house database.

Table 3.1. *Discriminant features for authentication of organic apple juices (compounds confirmed based on MS/MS data are indicated in bold characters).*

# Compound	# Feature	Detected *m/z*	Adduct type	RT (min)	p-value	VIP	Characteristic	Monoisotopic mass	Proposed molecular formula	Proposed compounds (FooDB)	Proposed compounds (HMDB)
1	1*	132.1019	$(M+H)^+$	3.32	1.0E-05	10.7	Conv > Org	131.0947	$C_6H_{13}NO_2$	Leucine, **Isoleucine**, 6-Deoxyfagomine, Alloleucine, **Norleucine**, Aminocaproic acid, Alanine betaine	Leucine, **Isoleucine**, 6-Deoxyfagomine, Alloleucine, **Norleucine**, Aminocaproic acid, Methylvaline, N-(2-Hydroxyethyl)-morpholine
	3	133.1052	M+1	3.32	1.0E-05	2.8					
2	2*	133.0317	$(M+H)^+$	1.95	4.4E-08	4.0	Conv > Org	132.0245	$C_3H_2F_2N_4$	n.a.	n.a.
									C_8H_3FN	n.a.	n.a.
									$C_8H_4O_2$	2,4,6-Octatriynoic acid	n.a.
									$C_5H_8O_2S$	n.a.	3-Methyl sulfolene
3	4*	150.0583	$(M+H)^+$	1.95	3.8E-08	10.9	Conv > Org	149.0510	$C_3H_5F_2N_5$	n.a.	n.a.
	5*	151.0616	M+1	1.95	3.2E-08	2.5			$C_8H_6FN_2$	n.a.	n.a.
	6*	152.0541	M+2	1.95	3.4E-08	2.2			$C_5H_{11}NO_2S$	**Methionine**	**Methionine**, Penicillamine
4	7	245.0767	$(M+H)^+$	2.47	5.6E-05	3.6	Conv > Org	244.0693	$C_9H_{12}N_2O_6$	*Pseudouridine, Uridine***	*Pseudouridine, Uridine***
	8	267.0585	$(M+Na)^+$	2.47	6.5E-05	2.8			$C_6H_4N_{12}$	n.a.	n.a.
5	9	271.1149	$(M+H)^+$	6.93	4.5E-06	2.2	Conv > Org	270.1077	$C_9H_{14}N_6O_4$	n.a.	n.a.
									$C_{10}H_{10}N_{10}$	n.a.	n.a.
									$C_{11}H_{16}N_3O_5$	n.a.	n.a.
6	10	331.1724	$(M+H)^+$	12.98	1.3E-05	3.8	Conv > Org	330.1652	$C_{12}H_{22}N_6O_5$	n.a.	n.a.
									$C_{13}H_{18}N_{10}O$	n.a.	n.a.
									$C_{14}H_{24}N_3O_6$	n.a.	n.a.
									$C_{11}H_{26}N_2O_9$	n.a.	n.a.
7	11	433.2040	$(M+H)^+$	13.86	1.8E-05	3.4	Conv > Org	432.1968	$C_{16}H_{28}N_6O_8$	n.a.	n.a.

									$C_{17}H_{24}N_{10}O_4$	n.a.	n.a.
									$C_{18}H_{30}N_3O_9$	n.a.	n.a.
									$C_{24}H_{32}O_5S$	S-Furanopetasitin	S-Furanopetasitin
8	12	495.1744	$(M+H)^+$	13.86	2.5E-06	2.2	Conv > Org	494.1671	$C_{24}H_{24}N_5O_7$	n.a.	n.a.
									$C_{23}H_{28}NO_{11}$	n.a.	n.a.
									$C_{22}H_{22}N_8O_6$	n.a.	n.a.
									$C_{21}H_{26}N_4O_{10}$	n.a.	n.a.
9	13	591.1677	$(M+H)^+$	9.58	9.8E-06	3.4	Org > Conv	590.1606	$C_{24}H_{26}N_6O_{12}$	n.a.	n.a.
									$C_{25}H_{22}N_{10}O_8$	n.a.	n.a.
									$C_{28}H_{30}O_{14}$	Maysin 3'-methyl ether	n.a.

n.a.: not applicable

** These features were detected using biosigner*

*** invalid based on MS/MS data*

During the MS/MS experiments, a full scan analysis was also acquired. It was thus possible to use these independent acquisitions to evaluate the potential of the discriminant features to serve as marker compounds. It is noteworthy that the previously selected features were successfully observed on this fourth analytical batch (only one feature was missing because it has a low intensity); nevertheless, only a few of them still showed a trend in the discrimination of sample groups. In particular, for the authentication of direct apple juices, 5 features still showed a difference in intensity between the two sample groups; these features might be used as markers compounds for the concentrated juice characteristic. On the other hand, for the authentication of organic apple juices, no trend was observed for the discrimination of the two samples groups by observing the intensity of the features between the two sample groups.

It can be emphasized that these independent acquisitions permitted to highlight some marker compounds as they were characteristic of the process type used (juice concentration). For the organic juice characteristic, these new samples came from a different harvest year, which may explain that the discriminant features found previously may fail to discriminate these new acquisitions. Cubero-Leon and co-workers reported that the harvest year was one of the most important variabilities in their studied samples (Cubero-Leon et al., 2018). On the contrary, Diaz et al. identified a biomarker for orange origin which seems to be independent from the harvest year (Diaz et al., 2014). Further investigation is thus needed to find reliable features for the authentication of organic apple juice samples and to confirm the use of the 5 features for the authentication of direct apple juice samples.

Based on the analysis of standards, two detected features might be assigned to phloridzin (p-value: 5.83 E-03) and alpha-terpineol (p-value: 1.08 E-04) according to their *m/z* and retention time; unfortunately, these two features were outside the list of Table 1. It is not surprising that phloridzin was not a discriminant compound as it is a naturally present molecule in apples, with varying concentrations depending on different factors such as variety or processing technology used (Spinelli et al., 2016). The remaining standards were not detected in our samples, possibly because they were not concentrated enough to be observed, while the other one (hydroxymethylfurfural) routinely analyzed by LC-UV may not be present in the juice samples analyzed.

2.4. CONCLUSIONS

This work presents a methodology combining untargeted LC-HRMS analysis and chemometric tools to authenticate apple juice samples. The OPLS-DA models showed good performance in sample classification, especially for the discrimination between organic and conventional sample juices (nearly 80% of predictive ability). To confirm their classification and prediction performance, further validation of these models using an external data set is required.

Coupling the results of ANOVA and OPLS-DA seems to be an interesting methodology to determine the discriminant features as it permitted to reduce the number of detected features (from almost 10,000 features detected to about 150 features) while keeping significant and discriminant features. According to the chemometric tools used (OPLS-DA and ANOVA) about 20 features have been identified as significantly discriminant and tentatively identified for the first time. Some compounds were tentatively annotated as amino-acids and derivatives, and a few markers were confirmed by MS/MS experiments. Interestingly, application of our analytical method to a new set of samples showed that some features retained a tendency to discriminate between the two groups of samples, mainly for authentication of direct apple juices.

The main additional research concerns the annotation workflow, which is the most time-consuming part of this methodology. By building an in-house database, the identification of marker compounds can be faster as the obtained mass spectra will be better compared than by using an online database, the same instrument being used. By identifying the compounds responsible for the discrimination, they could be analyzed in a routine analysis for apple juice authentication. Further investigation is needed to correctly identify the compounds by the analysis of standards.

The proposed analytical methodology enabled, for the very first time, the authentication of apple juice samples in two distinct scenarios using a single analysis (organic *vs*. conventional samples and single strength juice *vs*. both concentrated juice and juice from concentrate samples). Other chemometric models could be developed to implement juice discrimination based on variety and/or geographical origin, in addition to the scenarios presented here.

Acknowledgement

The authors warmly thank Dr. Peter Rinke from SGF for kindly providing them with several samples of apple juices used in this study. They are thankful for the financial support provided by the Association Nationale Recherche et Technologie (ANRT) through the CIFRE program (CIFRE n°2018/0937).

Conflict of interest

The authors declare that they have no commercial or financial relationships that could have influence the research conducted in this paper.

2.5. REFERENCES

AIJN Code of Practice (2020), AIJN European Fruit Juice Association.

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods 5, 3790. https://doi.org/10.1039/c3ay40582f

Bat, K.B., Vodopivec, B.M., Eler, K., Ogrinc, N., Mulič, I., Masuero, D., Vrhovšek, U., (2018). Primary and secondary metabolites as a tool for differentiation of apple juice according to cultivar and geographical origin. *LWT – Food Science and Technology,* 90, 238–245. https://doi.org/10.1016/j.lwt.2017.12.026

Brooks, S., Elliott, C.T., Spence, M., Walsh, C., Dean, M., (2017). Four years post-horsegate: an update of measures and actions put in place following the horsemeat incident of 2013. *npj Science of Food*, 1. https://doi.org/10.1038/s41538-017-0007-z

Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends in Food Science and Technology*, 80, 223-241. https://doi.org/10.1016/j.tifs.2018.08.007

Cavanna, D., Loffi, C., Dall'Asta, C., Suman, M., (2020). A non-targeted high-resolution mass spectrometry approach for the assessment of the geographical origin of durum wheat. *Food Chemistry,* 317, 126366. https://doi.org/10.1016/j.foodchem.2020.126366

Chaleckis, R., Meister, I., Zhang, P., Wheelock, C.E., (2019). Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology*, 55, 44-50. https://doi.org/10.1016/j.copbio.2018.07.010

Chambers, M.C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., ... Mallick, P., (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology, 30*(10), 918–920. https://doi.org/10.1038/nbt.2377

Codex Alimentarius: Codex General Standard for Fruit Juices and Nectars (2005) CODEX STAN 247-2005

Cuberon-Leon, E., Peñalver, R., Maquet, A., (2014). Review on metabolomics for food authentication. *Food Research International*, 60, 95-107. https://doi.org/10.1016/j.foodres.2013.11.041

Cubero-Leon, E., De Rudder, O., Maquet, A., (2018). Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry,* 239, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

Cuevas, F.J., Pereira-Caro, G., Moreno-Rojas, J.M., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., (2017). Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics. *Food Control*, 82, 203–211. https://doi.org/10.1016/j.foodcont.2017.06.031

Cuevas, F.J., Pereira-Caro, G., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., Montenegro, J.C., Moreno-Rojas, J.M., (2019). A holistic approach to authenticate organic sweet oranges (Citrus Sinensis L. cv Osbeck) using different techniques and data fusion. *Food Control,* 104, 63–73. https://doi.org/10.1016/j.foodcont.2019.04.012

Danezis, G.P., Tsagkaris, A. S., Camin, F., Brusic, V., Georgiou, C.A., (2016). Food authentication: Techniques, trends & emerging approaches. *Trends in Analytical Chemistry,* 85, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Dasenaki, M.E. & Thomaidis, N.S. (2019). Quality and authenticity control of fruit juices - A review. *Molecules,* 24, 1014. https://doi.org/10.3390/molecules24061014

Delaporte, G., Cladiere, M., Jouan-Rimbaud Bouveresse, D., Camel, V., (2019). Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study on tea. *Food Chemistry*, 277, 54–62. https://doi.org/10.1016/j.foodchem.2018.10.089

Diaz, R., Pozo, O.J., Sancho, J.V., Hernandez, F., (2014). Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry*, 157, 84-93. https://doi.org/10.1016/j.foodchem.2014.02.009

Dieterle, F., Ross, A., Schlotterbeck, G., Senn, H., (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13), 4281–4290. https://doi.org/10.1021/ac051632c

Directive 2012/12/EC, (2012). Council Directive Relating to Fruit Juices and Certain Similar Products Intended for Human Consumption of 19 April 2012

Dubin, E., Dumas, A.-S., Rebours, A., Jamin, E., Ginet, J., Lees, M., Rutledge, D.N., (2017). Detection of Blackcurrant Adulteration by Aronia Berry Using High Resolution Mass Spectrometry, Variable Selection and Combined PLS Regression Models. *Food Analytical Methods,* 10, 683–693. https://doi.org/10.1007/s12161-016-0638-8

Esteki, M., Simal-Gandarab, J., Shahsavaria, Z., Zandbaafa, S., Dashtakia, E., Heydenc, Y.V., (2018). A review on the application of chromatographic methods, coupled to chemometrics, for food authentication. *Food Control*, 93, 195-182. https://doi.org/10.1016/j.foodcont.2018.06.015

FooDB, (2021). Food Database, https://foodb.ca/ accessed on November 24th 2021

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... Caron, C., (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31*(9), 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R, (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. *Trends in Analytical Chemistry*, 82, 425−442. https://doi.org/10.1016/j.trac.2016.07.004

Guo, J., Yue, T., Yuan, Y., Wang, Y., (2013). Chemometric classification of apple juices according to variety and geographical origin based on polyphenolic profiles. *Journal of Agricultural and Food Chemistry*, 61, 6949−6963. https://doi.org/10.1021/jf4011774

IFU, (2021). International Fruit and Vegetable Juice Association, https://ifu-fruitjuice.com/ accessed on May 28th 2021

Jandric, Z., Roberts, D., Rathor, M.N., Abrahim, A., Islam, M., Cannavan, A., (2014). Assessment of fruit juice authenticity using UPLC-QToF MS: A metabolomics approach. *Food Chemistry*, 148, 7–17. https://doi.org/10.1016/j.foodcont.2018.06.015

Jandric, Z., Islam, M., Singh, D.K., Cannavan, A., (2017). Authentication of Indian citrus fruit/fruit juices by untargeted and targeted metabolomics. *Food Control*, 71, 181–188. https://doi.org/10.1016/j.foodcont.2015.10.044

Knolhoff, A. M., Croley, T. R., (2016). Non-targeted screening approaches for contaminants and adulterants in food using liquid chromatography hyphenated to high resolution mass spectrometry. *Journal of Chromatography A*, 1428, 86–96. https://doi.org/10.1016/j.chroma.2015.08.059

Llano, S.M., Muñoz-Jiménez, A.M., Jiménez-Cartagena, C., Londoño-Londoño, J., Medina, S., (2018). Untargeted metabolomics reveals specific withanolides and fatty acyl glycoside as tentative metabolites to differentiate organic and conventional *Physalis peruviana* fruits. *Food Chemistry*, 244, 120–127. https://doi.org/10.1016/j.foodchem.2017.10.026

Ma, S., Neilson, A.P., Lahne, J., Peck, G.M., O'Keefe, S.F., Stewart, A.C., (2018). Free amino acid composition of apple juices with potential for cider making as determined by UPLC-PDA. *Journal of the Institute of Brewing*, 124, 467–476. https://doi.org/10.1002/jib.519

Medina, S., Pereira, J.A., Silva P., Perestrelo, R., Câmara, J.S., (2019a). Food fingerprints – a valuable tool to monitor food authenticity and safety. *Food Chemistry*, 278, 144-162. https://doi.org/10.1016/j.foodchem.2018.11.046

Medina, S., Perestrelo, R., Silva, P., Pereira, J., Câmara, J.S., (2019b). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology*, 85, 163-176. https://doi.org/10.1016/j.tifs.2019.01.017

Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T. (2021). High-resolution mass spectrometry-based metabolomics for the discrimination between organic and conventional crops: A review. *Trends in Food Science & Technology*, 110, 142-154. https://doi.org/10.1016/j.tifs.2021.01.071

Moore, J., Spink, J., Lipp, M., (2012). Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *Journal of Food Science*, 77, R118-R126. https://doi.org/10.1111/j.1750-3841.2012.02657.x.

Oliveri, P., & Simonetti, R., (2016). Chemometrics for Food Authenticity Applications. In G. Downey (Eds.), *Advances in Food Authenticity Testing* (pp.701-728). Woodhead Publishing is an imprint of Elsevier

Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., Gonzalez-Ruiz, V., Rudaz, S., (2020). Implementation of liquid chromatography - high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: a tutorial. *Analytical Chimica Acta,* 1105, 28e44. https://doi.org/10.1016/j.aca.2019.12.062.

Rinke, P., (2016). Tradition Meets High Tech for Authenticity Testing of Fruit Juices. In G. Downey (Ed.), *Advances in Food Authenticity Testing* (pp.625-665). Woodhead Publishing is an imprint of Elsevier

Rinke, P., & Jamin, E., (2018). Fruit juices. In Morin, J.-F., Lees, M. (Eds.), *FoodIntegrity Handbook: A guide to food authenticity issues and analytical solutions*, 1st ed. Eurofins Analytics France. https://doi.org/10.32741/fihb

Rinaudo, P., Boudah, S., Junot, C., Thévenot, E.A., (2016). biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences* 3. https://doi.org/10.3389/fmolb.2016.00026

Rubert, J., Lacina, O., Zachariasova, M., Hajslova, J., (2016). Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chemistry,* 204, 201–209. https://doi.org/10.1016/j.foodchem.2016.01.003

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78, 779-787. https://doi.org/10.1021/ac051437y

Sobolev, A.P., Thomas, F., Donarski, J., Ingallina, C., Circi, S., Marincola, F.C., Capitani, D., Mannina, L., (2019). Use of NMR applications to tackle future food fraud issues. *Trends in Food Science & Technology*, 91, 347-353. https://doi.org/10.1016/j.tifs.2019.07.035

Spinelli, F.R., Dutra, S.V., Carnieli, G., Leonardelli, S., Drehmer, A.P., Vanderlinde, R., (2016). Detection of addition of apple juice in purple grape juice. *Food Control,* 69, 1–4. https://doi.org/10.1016/j.foodcont.2016.04.005

Sugimoto, N., Engelgau, P., Jones, A.D., Song, J., Beaudry, R., (2021). Citramalate synthase yields a biosynthetic pathway for isoleucine and straight- and branched-chain ester formation on ripening apple fruit. *PNAS,* 118(3), e2009988118. https://doi.org/10.1073/pnas.2009988118

Tautenhahn, R., Bottcher, C., Neumann, S., (2008). Highly sensitive feature detection for high resolution LC/MS*. BMC Bioinformatics*, 9. https://doi.org/10.1186/1471-2105-9-504

Vaclavik, L., Schreiber, A., Lacina, O., Cajka, T., (2012). Liquid chromatography-mass spectrometry-based metabolomics for authenticity assessment of fruit juices. *Metabolomics*, 8, 793-803. https://doi.org/10.1007/s11306-011-0371-7

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., … Scalbert, A., (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research,* 46, D608–D617. https://doi.org/10.1093/nar/gkx1089

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems,* 58, 109–130.

Wolter, C., Gessler, A., Winterhalter, P., (2008). Aspects when evaluating apple-juice aroma. Fruit processing, 64-80

Worley, B., Powers, R., 2012. Multivariate Analysis in Metabolomics. Current Metabolomics 1, 92–107. https://doi.org/10.2174/2213235X11301010092

Xu, L., Xu, Z., Kelly, S., Liao, X., (2020). Integrating untargeted metabolomics and targeted analysis for not from concentrate and from concentrate orange juices discrimination and authentication. *Food Chemistry*, 329, 127130. https://doi.org/10.1016/j.foodchem.2020.127130

2.6. APPENDIX A. SUPPLEMENTARY DATA

Table 3.A.1. *Information about the analyzed samples.*

N° batch	Sample code	Type	Farming type	Geographical origin	° Brix
1	B1_JC_C_01	Juice from concentrate	Conventionnal	n.a.	n.a.
	B1_DJ_C_02	Direct juice	Conventionnal	n.a.	n.a.
	B1_DJ_O_03	Direct juice	Organic	n.a.	n.a.
	B1_JC_C_04	Juice from concentrate	Conventionnal	n.a.	n.a.
	B1_DJ_O_05	Direct juice	Organic	n.a.	n.a.
	B1_DJ_C_06	Direct juice	Conventionnal	n.a.	11.5
	B1_DJ_C_07	Direct juice	Conventionnal	n.a.	11.3
	B1_JC_C_08	Juice from concentrate	Conventionnal	n.a.	11.4
	B1_JC_C_09	Juice from concentrate	Conventionnal	n.a.	11.5
	B1_DJ_C_10	Direct juice	Conventionnal	n.a.	12.4
	B1_DJ_O_11	Direct juice	Organic	n.a.	12.5
	B1_DJ_O_12	Direct juice	Organic	n.a.	11.6
	B1_DJ_O_13	Direct juice	Organic	n.a.	12.4
	B1_DJ_C_14	Direct juice	Conventionnal	Germany	12.7
	B1_DJ_C_15	Direct juice	Conventionnal	Germany	12.6
	B1_DJ_C_16	Direct juice	Conventionnal	Germany	13.4
	B1_DJ_C_17	Direct juice	Conventionnal	Germany	14.8
	B1_DJ_O_18	Direct juice	Organic	Germany	13.0
	B1_DJ_O_19	Direct juice	Organic	n.a.	14.5
	B1_DJ_O_20	Direct juice	Organic	n.a.	13.9
	B1_DJ_O_21	Direct juice	Organic	n.a.	14.0
	B1_DJ_O_22	Direct juice	Organic	n.a.	15.3
	B1_DJ_O_23	Direct juice	Organic	n.a.	14.7
	B1_DJ_O_24	Direct juice	Organic	n.a.	11.5
2	B2_DJ_C_01	Direct juice	Conventionnal	Poland	n.a.
	B2_JC_C_02	Concentrated juice	Conventionnal	Poland	n.a.
	B2_DJ_C_03	Direct juice	Conventionnal	Turkey	n.a.
	B2_JC_C_04	Concentrated juice	Conventionnal	Turkey	n.a.
	B2_DJ_C_05	Direct juice	Conventionnal	China	n.a.
	B2_JC_C_06	Concentrated juice	Conventionnal	China	n.a.
	B2_DJ_C_07	Direct juice	Conventionnal	Brazil	n.a.
	B2_JC_C_08	Concentrated juice	Conventionnal	Brazil	n.a.
	B2_DJ_C_09	Direct juice	Conventionnal	Spain	n.a.

	B2_JC_C_10	Concentrated juice	Conventionnal	Spain	n.a.
	B2_DJ_C_11	Direct juice	Conventionnal	Hungary	n.a.
	B2_JC_C_12	Concentrated juice	Conventionnal	Hungary	n.a.
	B2_DJ_C_13	Direct juice	Conventionnal	Poland	n.a.
	B2_JC_C_14	Concentrated juice	Conventionnal	Poland	n.a.
	B2_DJ_C_15	Direct juice	Conventionnal	Poland	n.a.
	B2_JC_C_16	Concentrated juice	Conventionnal	Poland	n.a.
	B2_DJ_C_17	Direct juice	Conventionnal	Poland	n.a.
	B2_JC_C_18	Concentrated juice	Conventionnal	Poland	n.a.
	B2_DJ_C_19	Direct juice	Conventionnal	Turkey	n.a.
	B2_JC_C_20	Concentrated juice	Conventionnal	Turkey	n.a.
	B2_DJ_C_21	Direct juice	Conventionnal	Ukraine	n.a.
	B2_JC_C_22	Concentrated juice	Conventionnal	Ukraine	n.a.
	B2_JC_C_23	Juice from concentrate	Conventionnal	n.a.	11.4
	B2_JC_C_24	Concentrated juice	Conventionnal	China	70.0
	B2_JC_C_25	Concentrated juice	Conventionnal	n.a.	70.0
	B2_DJ_C_26	Direct juice	Conventionnal	n.a.	12.8
	B2_DJ_C_27	Direct juice	Conventionnal	n.a.	12.8
	B2_JC_C_28	Juice from concentrate	Conventionnal	Germany	n.a.
	B2_DJ_C_29	Direct juice	Conventionnal	Germany	11.7
	B2_DJ_C_30	Direct juice	Conventionnal	Germany	12.8
3	B3_JC_C_01	Concentrated juice	Conventionnal	n.a.	69.8
	B3_JC_C_02	Concentrated juice	Conventionnal	n.a.	69.0
	B3_JC_C_03	Concentrated juice	Conventionnal	n.a.	69.4
	B3_JC_O_04	Concentrated juice	Organic	n.a.	69.2
	B3_JC_O_05	Concentrated juice	Organic	n.a.	69.4
	B3_JC_O_06	Concentrated juice	Organic	n.a.	53.5
	B3_JC_C_07	Juice from concentrate	Conventionnal	n.a.	11.2
	B3_JC_C_08	Juice from concentrate	Conventionnal	n.a.	11.2
	B3_JC_C_09	Juice from concentrate	Conventionnal	n.a.	11.3
	B3_JC_C_10	Juice from concentrate	Conventionnal	n.a.	11.3
	B3_JC_O_11	Juice from concentrate	Organic	n.a.	11.3
	B3_JC_O_12	Juice from concentrate	Organic	n.a.	11.3
	B3_JC_O_13	Juice from concentrate	Organic	n.a.	11.3
	B3_JC_O_14	Juice from concentrate	Organic	n.a.	11.2
	B3_DJ_C_15	Direct juice	Conventionnal	France	12.2
	B3_DJ_C_16	Direct juice	Conventionnal	n.a.	11.9
	B3_DJ_O_17	Direct juice	Organic	n.a.	11.2

	B3_DJ_C_18	Direct juice	Conventionnal	n.a.	11.6
	B3_DJ_O_19	Direct juice	Organic	n.a.	11.5
	B3_DJ_O_20	Direct juice	Organic	France	12.4
	B3_DJ_C_21	Direct juice	Conventionnal	France	12.4
	B3_DJ_C_22	Direct juice	Conventionnal	n.a.	13.4
	B3_DJ_C_23	Direct juice	Conventionnal	n.a.	13.5
	B3_DJ_C_24	Direct juice	Conventionnal	n.a.	11.7
	B3_DJ_O_25	Direct juice	Organic	n.a.	11.6
	B3_DJ_O_26	Direct juice	Organic	n.a.	11.7
	B3_DJ_O_27	Direct juice	Organic	n.a.	11.8
	B3_DJ_O_28	Direct juice	Organic	n.a.	11.4
4	B4_JC_C_01	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_02	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_03	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_04	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_05	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_06	Juice from concentrate	Conventionnal	n.a.	n.a.
	B4_JC_C_07	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_08	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_09	Concentrated juice	Conventionnal	n.a.	n.a.
	B4_JC_C_10	Juice from concentrate	Conventionnal	n.a.	n.a.
	B4_DJ_O_11	Direct juice	Organic	France	n.a.
	B4_DJ_O_12	Direct juice	Organic	France	n.a.
	B4_DJ_O_13	Direct juice	Organic	France	n.a.
	B4_DJ_O_14	Direct juice	Organic	France	n.a.
	B4_DJ_O_15	Direct juice	Organic	France	n.a.
	B4_DJ_O_16	Direct juice	Organic	France	n.a.
	B4_DJ_O_17	Direct juice	Organic	France	n.a.
	B4_DJ_O_18	Direct juice	Organic	France	n.a.
	B4_DJ_O_19	Direct juice	Organic	France	n.a.
	B4_DJ_O_20	Direct juice	Organic	France	n.a.
	B4_DJ_C_21	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_22	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_23	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_24	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_25	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_26	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_27	Direct juice	Conventionnal	France	n.a.

	B4_DJ_C_28	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_29	Direct juice	Conventionnal	France	n.a.
	B4_DJ_C_30	Direct juice	Conventionnal	France	n.a.

n.a.: not available

Table 3.A.2. *Parameters and their corresponding values for the MS source.*

Parameter	**Value**
Sheath gas flow (A.U.)	40
Auxiliary gas flow (A.U.)	12
Sweep gas flow (A.U.)	0
Spray voltage (|kV|)	3.2
Capillary temperature (°C)	275
S-lens RF level	50
Auxiliary gas heater temperature (°C)	300

A.U.: Arbitrary Unit

Table 3.A.3. *Parameters and their corresponding values for the different steps using XCMS.*

Step	Parameter	Value
findChromPeaks	method	centWave
	ppm	5
	peakwidth	5-30
	snthresh	10
	prefilter	3 / 100000
	mzCenterFun	wMean
	integrate	1
	mzdiff	0.01
	noise	50000
groupChromPeaks - 1	method	density
	bw	6
	minFraction	0.25
	minSample	6
	binSize	0.02
adjustRtime	method	peakgroups
	smooth	loess
	extra	1
	minFraction	0.85
	span	0.2
	family	gaussian
	plottype	mdevden
groupChromPeaks - 2	method	density
	bw	4
	minFraction	0.25
	minSample	6
	binSize	0.02
fillChromPeaks	method	chrom
	convertRTMinute	TRUE
	numDigitsMZ	4
	numDigitsRT	2
	intval	into

Table 3.A.4. *Discriminant features for authentication of pure apple juices (compounds confirmed based on MS/MS data are indicated in bold characters).*

# Feature	Detected *m/z*	Adduct type	RT (min)	p-value	VIP	Characteristic	Monoisotopic mass	Tentative formula	Proposed compounds (FooDB)	Proposed compounds (HMDB)
1	249.0441	$[M+2H]^{2+}$	2.62	2.5E-07	3.91	Conc > Direct	496.0731	$C_{18}H_{16}N_4O_{13}$	n.a.	n.a.
								$C_{23}H_{16}N_2O_{11}$	n.a.	n.a.
2	249.0441	$[M+2H]^{2+}$	2.27	8.6E-07	3.14	Conc > Direct	496.0731	$C_{18}H_{16}N_4O_{13}$	n.a.	n.a.
								$C_{23}H_{16}N_2O_{11}$	n.a.	n.a.
3	261.0603	$[M+H]^+$	2.61	4.6E-06	2.96	Conc > Direct	260.0530	$C_{10}H_{12}O_8$	n.a.	n.a.
								$C_8H_{10}N_3O_7$	n.a.	n.a.
4	294.1179	$[M+H]^+$	1.85	1.6E-07	2.68	Conc > Direct	293.1108	$C_{10}H_{13}N_8O_3$		
								$C_{11}H_{19}NO_8$	Galactosyl 4-hydroxyproline, 4-Hydroxyproline galactoside, (3R)-3,4-Dihydroxy-3-(hydroxymethyl)butanenitrile 4-glucoside, N-Acetyl-muramic acid, Sophorose	Galactosyl 4-hydroxyproline, 4-Hydroxyproline galactoside, (3R)-3,4-Dihydroxy-3-(hydroxymethyl)butanenitrile 4-glucoside, N-Acetyl-muramic acid
5	303.0684	$[M+H]^+$	1.89	1.6E-06	2.41	Conc > Direct	302.0611	$C_8H_{10}N_6O_7$	n.a.	n.a.
								$C_{10}H_{12}N_3O_8$	n.a.	n.a.
6	308.1337	$[M+H]^+$	3.64	4.8E-07	3.09	Conc > Direct	307.1263	$C_{11}H_{15}N_8O_3$	n.a.	n.a.
								$C_{12}H_{21}NO_8$	n.a.	n.a.
								$C_{13}H_{17}N_5O_4$	n.a.	n.a.
								$C_9H_{13}N_{11}O_2$	n.a.	n.a.
7*	312.1108	$[M+H]^+$	2.17	2.4E-06	5.95	Conc > Direct	311.1034	$C_{17}H_{15}N_2O_4$	n.a.	n.a.
								$C_{11}H_{21}NO_7S$	N-(1-Deoxy-1-fructosyl)methionine	N-(1-Deoxy-1-fructosyl)methionine
								$C_{18}H_{11}N_6$	n.a.	n.a.

								$C_{14}H_{15}N_8O_2$	n.a.	n.a.
								$C_{15}H_{21}NO_7$	**N-(1-Deoxy-1-fructosyl)phenylalanine**	**N-(1-Deoxy-1-fructosyl)phenylalanine**
8	328.1387	$[M+H]^+$	6.16	1.2E-06	3.48	Conc > Direct	327.1316	$C_{13}H_{19}N_4O_6$	n.a.	n.a.
								$C_{16}H_{17}N_5O_3$	n.a.	n.a.
								$C_{12}H_{13}N_{11}O$	n.a.	n.a.
								$C_{14}H_{15}N_8O_3$	n.a.	n.a.
								$C_{15}H_{21}NO_8$	N-(1-Deoxy-1-fructosyl)tyrosine	N-(1-Deoxy-1-fructosyl)tyrosine
9	344.1336	[M+H]+	2.6	1.1E-06	3.09	Conc > Direct	343.1264	$C_{13}H_{19}N_4O_7$	n.a.	n.a.
								$C_{16}H_{17}N_5O_4$	n.a.	n.a.
								$C_{12}H_{13}N_{11}O_2$	n.a.	n.a.
10	481.1160	$[M+H]^+$	2.61	3E-06	3.87	Conc > Direct	480.1087	$C_{14}H_{20}N_6O_{13}$	n.a.	n.a.
								$C_{15}H_{16}N_{10}O_9$	n.a.	n.a.
								$C_{33}H_{58}N_3O_5$	n.a.	n.a.
								$C_{32}H_{52}N_{10}$	n.a.	n.a.
11*	599.4275	$[M+Na]^+$	24.05	1.9E-09	2.47	Direct > Conc	576.4383	$C_{35}H_{60}O_6$	Coriandrinol, Schottenol 3-glucoside beta-Sitosterol 3-O-beta-D-galactopyranoside	Schottenol 3-glucoside, beta-Sitosterol 3-O-beta-D-galactopyranoside
								$C_{31}H_{56}N_6O_4$	n.a.	n.a.
								$C_{36}H_{56}N_4O_2$	n.a.	n.a.

n.a.: not applicable

* These features were detected using biosigner

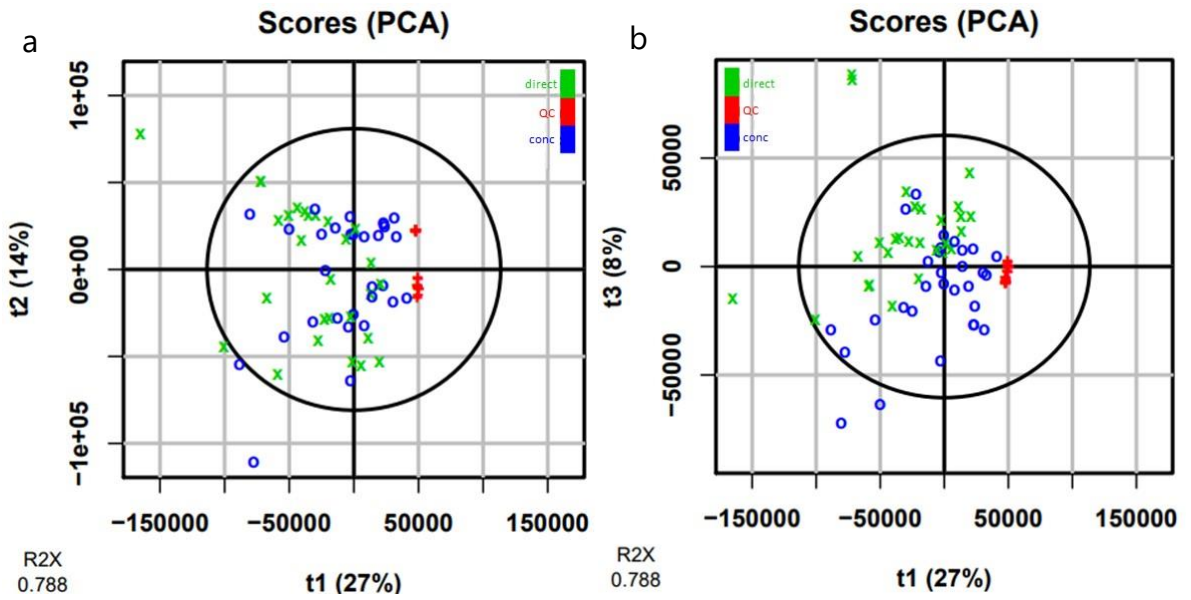

Figure 3.A.1. *PCA scores plot (blue circles, both concentrated juices and juices from concentrate; green crosses, direct juices; red plus signs, QC samples) obtained with the mean of the three replicates for each sample (a) using the first and second principal components and (b) using the first and third principal components. The black ellipse represents 95% of the variability.*

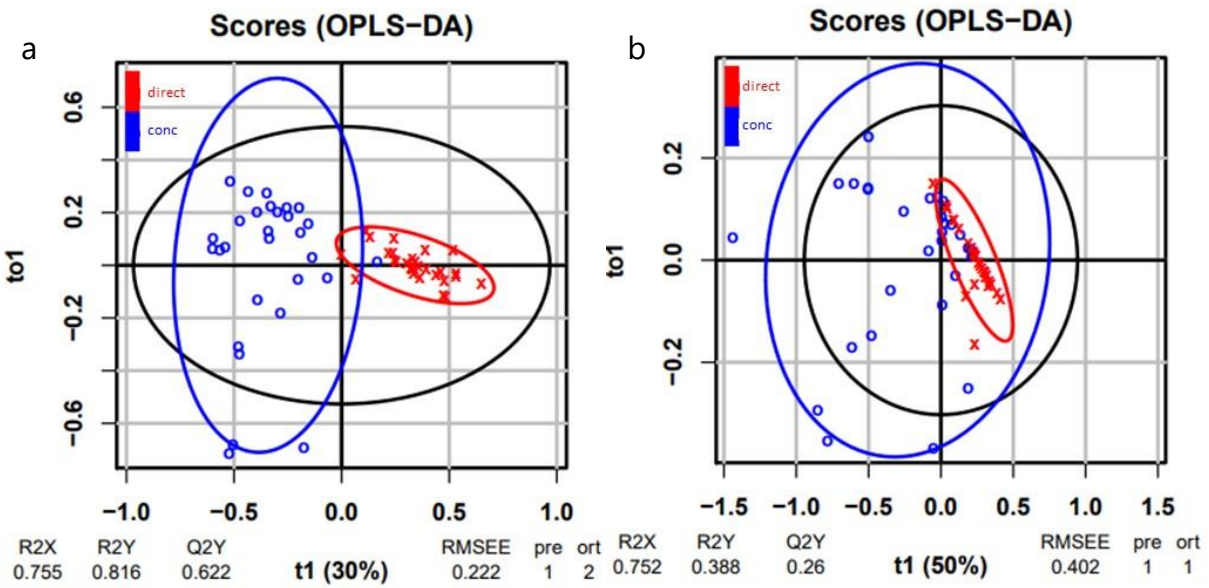

Figure 3.A.2. *(a) OPLS-DA obtained after features selection using ANOVA and (b) OPLS-DA obtained after features selection using biosigner (blue circles, both concentrated juices and juices from concentrate; red crosses, direct juices). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.*

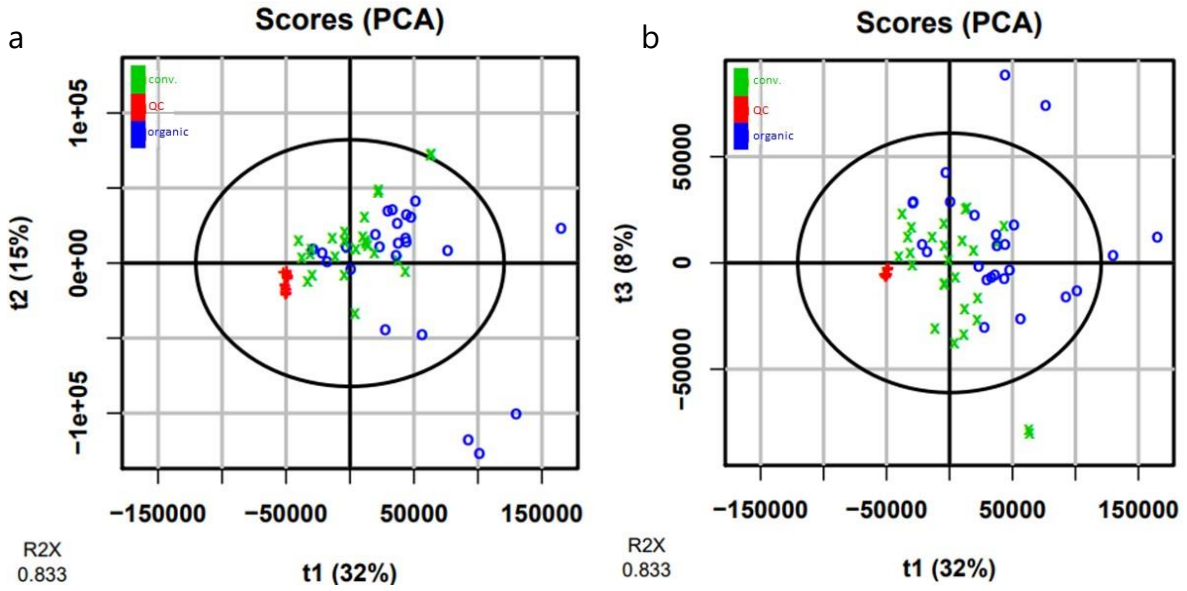

Figure 3.A.3. *PCA scores plot (blue circles, organic juice samples; green crosses, conventional juice samples; red plus signs, QC samples) (a) using the first and second principal components and (b) using the first and third principal components. The black ellipse represents 95% of the variability.*

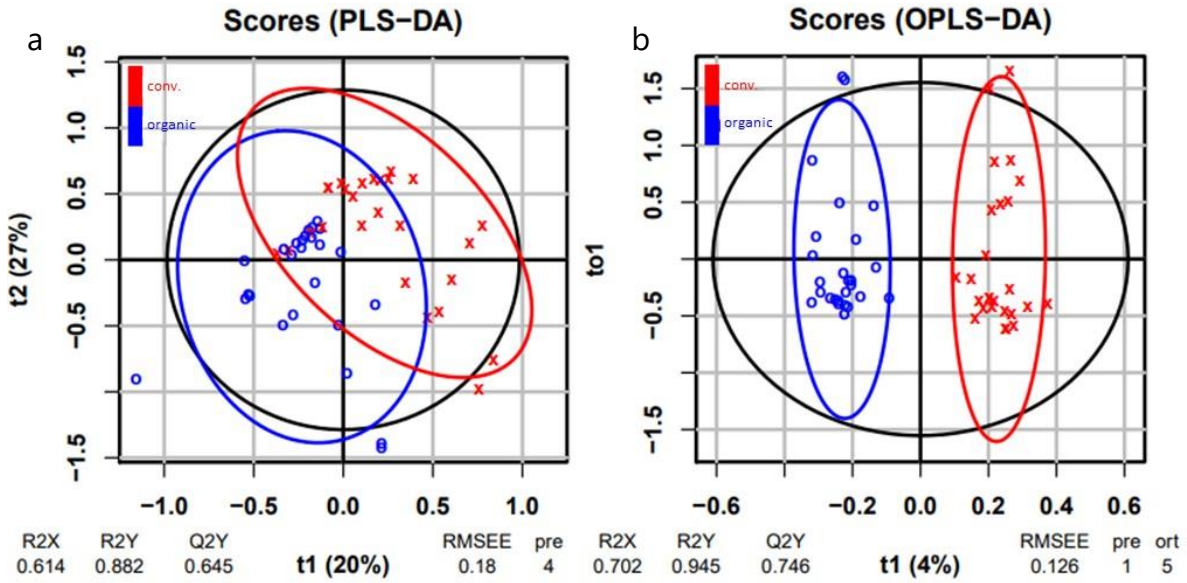

Figure 3.A.4. *(a) Score plot of PLS-DA and (b) scores plot of OPLS-DA obtained with cross-validation (blue circles, organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses correspond to 95% of the multivariate distributions for each sample groups.*

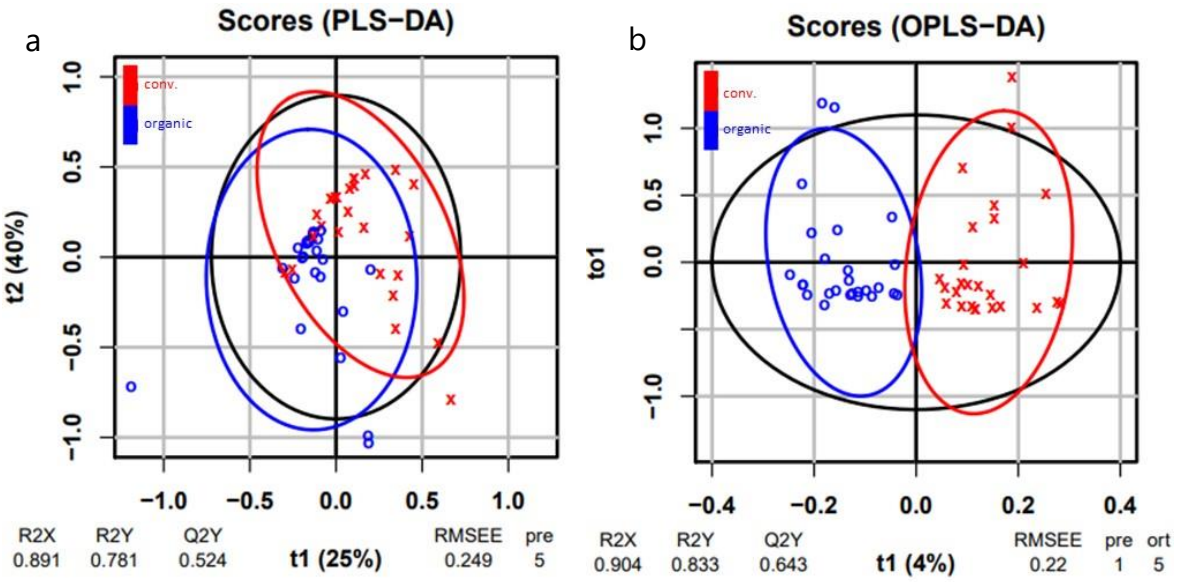

Figure 3.A.5. *(a) PLS-DA and (b) OPLS-DA obtained after features selection using biosigner (blue circles: organic juice samples; red crosses, conventional juice samples). The black ellipse represents 95% of the variability, the blue and red ellipses represent 95% of the multivariate distributions for each sample groups.*

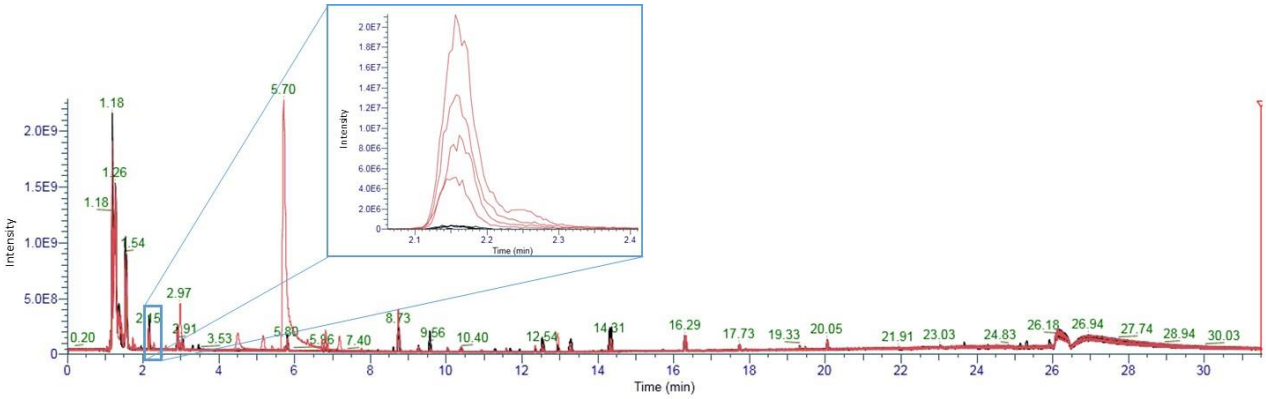

Figure 3.A.6. *Chromatogram of feature 7 for the authentication of pure apple juices (black, direct juice samples; red, concentrated juice samples)*

3. CONCLUSION

Cet article présente une méthodologie combinant une analyse non ciblée par LC-HRMS couplée à des outils chimiométriques pour l'authentification de jus de pommes.

Les modèles de classification des échantillons obtenus par OPLS-DA ont montré des performances intéressantes, notamment dans le cas de l'authentification des jus issus de l'agriculture biologique (avec près de 80 % de capacité prédictive obtenue). Ces performances ont été estimées par validation croisée. Afin d'avoir une estimation plus

réaliste des performances de ces modèles, une validation utilisant un jeu externe de données reste nécessaire. En effet cela permettra, d'une part d'évaluer les performances des modèles dans des conditions d'analyse de routine, et d'autre part d'inclure de la variabilité aux échantillons testés (par exemple l'année de production) et ainsi de s'assurer que les modèles sont toujours performants pour identifier les échantillons authentiques. Dans le but de se focaliser sur les features les plus discriminants, l'approche proposée ici (combinant les résultats de l'OPLS-DA et de l'ANOVA) semble être prometteuse. En effet, celle-ci a permis de passer d'environ 10 000 features détectés à environ 150 features. L'annotation des features a été effectuée sur 24 features sélectionnés selon leurs valeurs de VIP obtenues par OPLS-DA et de p-value obtenues par ANOVA. L'étape d'annotation a permis d'identifier pour la première fois des composés appartenant à la famille des acides aminés et dérivés. Certaines annotations ont été confirmées par les acquisitions MS/MS : la méthionine et la N-(1-deoxy-1-fructosyl)phenylalanine. Les analyses effectuées sur de nouveaux échantillons ont permis de mettre en évidence une tendance pour certains features dans la discrimination des groupes d'échantillons ; ces features pourraient donc servir comme composés marqueurs.

La méthodologie proposée dans cet article requiert des optimisations, notamment pour l'étape d'annotation. Cette étape est particulièrement longue à réaliser. L'annotation des composés discriminants est effectuée par comparaison des spectres MS/MS expérimentaux avec ceux présents dans des bases de données spectrales. Leur identification doit ensuite être confirmée par l'analyse de standard permettant ainsi de vérifier le RT et le spectre MS/MS obtenus. Cette étape d'annotation des composés est particulièrement importante car, dès lors que des composés ont été identifiés, ceux-ci pourraient être analysés en routine par des analyses ciblées pour contrôler l'authenticité des jus de pommes.

De plus, afin de permettre l'implémentation en routine de ce type de méthodologie, il est nécessaire d'être en capacité de traiter les données à partir d'outils internes. Les différentes étapes du traitement des données doivent donc être mises en place sur des outils libres présents au laboratoire tels que le logiciel R.

CHAPITRE 4 ANALYSE NON CIBLEE POUR L'AUTHENTIFICATION DE CAROTTES

1. INTRODUCTION ET RESUME DE L'ARTICLE

Les travaux menés dans cet article ont été réalisés dans le cadre du projet TOFoo (*True Organic Food*), dont l'objectif est de développer des méthodologies d'analyse pour le contrôle d'authenticité des produits biologiques. Les échantillons collectés dans le cadre de ce projet sont analysés *via* différentes techniques d'analyse, et les données acquises par chacune de ces techniques sont ensuite traitées par des analyses chimiométriques.

Différents échantillons de carottes issus de l'agriculture biologique et de l'agriculture conventionnelle ont été collectés dans diverses régions de France, sur une période de temps limitée (25 janvier à 24 février 2021).

Une méthode d'analyse non ciblée par LC-HRMS a ensuite été mise en place. Cette méthode est proche de celle utilisée dans le chapitre précédent mais elle a subi quelques ajustements : (i) dans la phase mobile le méthanol est remplacé par l'acétonitrile (permettant ainsi de réduire la pression du système chromatographique et d'obtenir une élution plus rapide des composés) ; (ii) un flux d'acétonitrile est délivré avant l'entrée dans la source d'ionisation, et ce tout au long du gradient, afin de favoriser l'ionisation des composés ; (iii) deux colonnes analytiques sont utilisées (silice greffée C18 et amide-HILIC) en parallèle dans l'objectif de détecter une plus grande gamme de composés. En effet, la colonne de silice greffée C18 est la colonne majoritairement utilisée pour ce type d'application dans la littérature car elle permet la rétention de composés moyennement polaires à apolaires. A l'inverse, la colonne amide-HILIC permet une rétention des composés polaires, et parait donc tout à fait complémentaire.

Pour implémenter en routine ce type de méthodologie, le traitement des données proposé dans le chapitre précédent a été mis en place en interne sur le logiciel R. Les étapes du traitement des données acquises restent les mêmes comme présenté en **Figure 4.A.1**. Le traitement des données mis en place a pour but ici d'évaluer la capacité de la méthode LC-

MS non ciblée à discriminer les échantillons selon leur origine géographique et leur mode de production.

Dans cette étude, une séquence comprenant 30 échantillons de carottes (provenant de différentes origines géographiques, et issus soit de l'agriculture biologique soit de l'agriculture conventionnelle) a été analysée *via* la méthode d'analyse non ciblée par LC-HRMS puis retraitée par le workflow mis en place sous R. Les données obtenues pour chaque colonne analytique (silice C18 et amide-HILIC) ont été traitées séparément *via* le même workflow (certains paramètres ont été adaptés à la colonne utilisée comme présenté dans le **Table 4.A.4**). Cela permettra ainsi de comparer les performances de discrimination selon la colonne utilisée, et de ce fait de savoir si une famille de composés (polaires ou apolaires) semble être liée à la discrimination des groupes d'échantillons. De plus, les outils chimiométriques ont également été utilisés sur un jeu de données compilant à la fois les données brutes acquises sur la colonne de silice C18 et sur la colonne amide-HILIC ; cette fusion des données a été effectuée dans le but d'améliorer les performances des modèles de classification.

Dans un deuxième temps, un second jeu de 30 échantillons de carottes (dont 16 déjà analysés précédemment) a été analysé avec des expériences MS/MS afin d'aider à l'identification des features discriminants.

2. CORPS DE L'ARTICLE

Untargeted metabolomics-based approach using UHPLC-HRMS to authenticate carrots (*Daucus carota* L.) based on geographical origin and production mode

Katy Dinis [a,b], Lucie Tsamba [a]*, Eric Jamin [a], Valérie Camel [b]

[a] *Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323, Nantes Cedex 3, France*

[b] *UMR SayFood, Université Paris-Saclay, INRAE, AgroParisTech, 91300 Massy, France*

***** *Corresponding author**:** Eurofins Analytics France, 9 rue Pierre Adolphe Bobierre, B.P. 42301, F-44323 NANTES Cedex 3, France, tel.: +33 2 51 82 55 39, fax: +33 2 51 83 21 11. E-mail address: LucieTsamba@eurofins.com*

Soumis dans Food Chemistry

Abstract

Carrot samples produced in different agricultural production regions with either organic or conventional mode were analyzed by untargeted UHPLC-HRMS using two chromatographic modes: reversed-phase (on a C18-silica column) and HILIC (on a dedicated polar column). Data obtained using both types of chromatographic column were first treated separately to assess the most suitable mode, and further combined to possibly improve the results. An in-house data processing workflow was applied to identify relevant features after peak detection. Then, based on these features, discrimination models were built using chemometrics. A tentative annotation of chemical markers for discrimination was then performed using online databases and UHPLC-HRMS/MS analyses. An independent set of samples was also analyzed to assess the discrimination potential of these marker compounds. Carrots produced in the New Aquitaine region could be successfully discriminated from carrots originating from the Normandy region by an OLPS-DA model. Arginine and 6-methoxymellein could be identified as potential markers with the C18-silica column. Interestingly, additional markers (N-acetylputrescin and L-carnitin) could be identified thanks to the polar column. Discrimination of carrots based on production mode was more challenging: some trend was observed but the metrics of the models built remained unsatisfactory.

Keywords

Food authenticity, High resolution mass spectrometry, Liquid chromatography, Metabolomics, PCA, PLS-DA, OPLS-DA

Highlights

- Discrimination of carrot samples based on their geographical region of production
- Relevant markers selected by OPLS-DA or PLS-DA and ANOVA tentatively identified
- Arginine and 6-methoxymellein as potential biomarkers of geographical origin
- N-Acetylputrescin and L-carnitin as additional biomarkers

2.1. INTRODUCTION

Carrot (*Daucus carota* L., a member of the *Apiaceae* family) is a root vegetable produced and consumed worldwide. Its production underwent a major increase in the 1980s (30% increase between 1980 and 1990), and is still increasing (Arscott and Tanumihardjo, 2010). Worldwide 13.7 million tons were produced in 1990 (Arscott and Tanumihardjo, 2010) and 27 million tons in 2008 (Stolarczyk and Janick, 2011). Carrot offers interesting nutritional benefits, mainly due to the presence of carotenoids especially provitamin A (i.e. beta-carotene converted to vitamin A in the body) and numerous phenolic compounds (Arscott and Tanumihardjo, 2010; Stolarczyk and Janick, 2011; Ahmad et al., 2019). Several positive impacts on consumers health are assumed for this vegetable, such as anticarcinogenic and antioxidant effects (Akhtar et al., 2017). Another striking feature of carrots is their possibility to be consumed under various forms: fresh, processed (i.e. juice), dried, boiled or fried. This vegetable has interesting technological properties for the food industry and the cosmetic industry as well (Stolarczyk and Janick, 2011), and a growing demand is foreseen in the near future since a recent study has suggested carrot could be a valuable ingredient for several processed foods (Rocchetti et al., 2020).

China is the major producing country, with other top producers Uzbekistan, Russia and the USA, contributing together to nearly 50% of the world carrot production; among other leading producers are Ukraine, Poland, the United Kingdom, Germany, Indonesia, Turkey and France (Arscott and Tanumihardjo, 2010). Besides physical quality attributes (such as size, shape, uniformity, color and texture), sensory and nutritional quality attributes are also important for carrots. As with other vegetables, these attributes are dependent on different factors: the cultivar, the geographical origin, the production method and post-harvest conditions (Pereira et al., 2016; Koudela et al., 2021; Ahmad et al., 2019). Since carrots are among the most popular agricultural commodities in the world and among the world's most economically important vegetables, food control is an important issue to guaranty the quality of carrots put on the market, as well as the fair labelling of their geographical origin and production mode.

Plant metabolites deserve extensive research in the food control field, especially secondary metabolites that are responsible for several bioactive properties of plants and may serve as biomarkers for plant characterization as recently reviewed (Pedrosa et al., 2021). For

vegetables, several interesting applications have been reported in food authentication, mainly based on LC methods targeted on several molecular markers and combined with chemometrics (Campmajo et al., 2019). As an illustration, UHPLC-HRMS or UHPLC-HRMS/MS were applied for the determination of phenolic compounds: phenolic profiles and concentration levels were good chemical descriptors when using chemometric tools (principal component analysis (PCA) and partial least squares - discriminant analysis (PLS-DA)) to separate paprika samples based on the production region and flavor varieties (Barbosa et al., 2020). For carrots, phenolic acids (such as 5-O-caffeoylquinic acid) were investigated as possible biomarkers of the production mode, but no statistical differences between organic and conventional growth systems were found (Soltoft et al., 2010).

In recent years, metabolomics-based untargeted analytical approaches have raised a growing interest in the food control field. They seem to be promising in this field as they allow to obtain a global view of the sample, opening the way to the classification of food samples based on several descriptors referring to different quality attributes. Several applications to plant-based foods (fruits, vegetables, spices) have been reported. The phenolic compounds fingerprinting of paprika samples acquired by LC and fluorescence detection enabled their classification according to the production region after data treatment by PLS-DA (Campmajo et al., 2021). Mie et al. (2014) discriminated conventional and organic white cabbage during a long-term study (2 years) thanks to the untargeted analysis of 1,600 compounds of the plant metabolome by UHPLC-HRMS combined with chemometrics (PCA separated samples by production year, and orthogonal partial least squares - discriminant analysis (OPLS-DA) models discriminated 83% samples based on the production mode); in their work, white cabbage samples were produced in only three farms in Denmark, covering a narrow geographical zone. Using a similar methodology (UHPLC-HRMS, PCA and linear discriminant analysis (LDA)), discrimination between organic and conventional tomato crops was achieved by other authors with 73% of samples being correctly classified (Martinez Bueno et al., 2018). Other authors successfully authenticated organic oranges using a similar approach (LC-HRMS and OPLS-DA model), with approximately 90% of samples correctly classified (Cuevas et al., 2017). Another work on goldenberry samples applying UHPLC-HRMS combined to PCA permitted the separation of sample groups based on organic and conventional productions (Llano et al., 2018).

Yet, applications of LC-MS-based untargeted analytical approaches to carrot samples are scarce. To our knowledge, only one recent study in the area of organic food authenticity has been reported (Cubero-Leon et al., 2018). In this work, carrot samples of two varieties (Nerac and Namur) were collected in two Belgian Walloon regions over four consecutive years – each time both organic and conventional production modes were considered. Combining LC-HRMS with OPLS-DA models correctly classified 100% of unknown carrot samples if models were refined to exclude variables contributing to the production year (Cubero-Leon et al., 2018). These promising results need to be confirmed by other studies that take into account a greater diversity of geographical origins. Interestingly, Cubero-Leon et al. (2018) identified several markers related to carbohydrate metabolism and plant defense mechanism as responsible for the differences between organic and conventional growth systems.

For carrots, sample discrimination based on both geographical origin and production mode would be of great value in practice for the laboratories and control authorities due to their worldwide production. In this study, we have investigated the capability of a metabolomics-based untargeted analytical method combined with chemometrics tools for carrots discrimination based on these two factors. Carrot samples produced in several French agricultural production regions with either organic or conventional methods were analyzed by untargeted UHPLC-HRMS. The novelty of our work lies in the realization of a representative sampling of the carrot production in France. In addition, all carrot samples were systematically analyzed using two chromatographic modes: reversed-phase (on a non-polar C18-silica column, classically used in applications reported in literature) and HILIC (on a dedicated polar column, to preferentially retain polar to highly polar compounds). Indeed, as amino acids were identified as possible marker compounds for authentication of organic foods (Dinis et al., 2022; Cuevas et al., 2017; Mihailova et al., 2021), polar chromatographic columns might be interesting. One previous study reported the targeted analysis of polar phospholipids, small peptides and amino acids using HILIC mode for the assessment of garlic authenticity (Hrbek et al. 2018). In our work, results obtained from both chromatographic columns were compared to assess which family of compounds (polar or non-polar compounds) provides better discrimination between carrot samples.

2.2. Materials and Method

2.2.1. Reagents and chemicals

Acetonitrile (ACN), methanol, water and formic acid (FA), all LC-MS grade, were purchased from Fisher Scientific. Ammonium formate (LC-MS grade) was purchased from Sigma-Aldrich.

Different internal standards were used to assess the analytical performance and stability by monitoring their *m/z* and retention time. Carbaryl-d7 (purity: 98%) and a mix containing 9 pesticides unexpected on carrot samples (namely: boscalid (purity: 98%), captan (purity: 99%), chlorantraniliprole (purity: 98%), daminozide (purity: 99%), dimethomorph (purity: 99%), fenhexamid (purity: 99%), flonicamid (purity: 99%), hexythiazox (purity: 99%) and pyridaben (purity: 99%)) were both purchased from Restek. The carbaryl-d7 standard (20 µg/mL) as well as the mix standard (each pesticide at 100 µg/mL) were in acetonitrile. The pesticide mix was chosen to cover a wide range of retention times and m/z values.

The MS detector was weekly calibrated using the Pierce™ positive and negative ion calibration solution purchased from Thermo Fisher Scientific.

2.2.2. Samples description and preparation

Forty-four carrot samples were collected in the frame of the TOFoo (True Organic Food) collaborative project (see https://www.tofoo-project.com/en/). Carrots were sent by their producers from several regions of France to the laboratory. Thus, the samples collection depended on the producers' possibilities and on the harvest time. Most of the carrots were collected between January, 25th and February, 24th of year 2021.

After collection, the samples were crushed, homogenized, and placed in a 50 mL flask in the laboratory and stored at -20 °C before analysis. After defrosting, aliquots (5 g) of samples were extracted in 5 ml of water and 15 ml of methanol. Next, 10 µl of the pesticide mix standard was added to control the sample preparation. Samples were then homogenized using an Ultra-Turrax for 30 sec. Samples were next agitated for 10 min using a mechanical stirring and further centrifuged for 5 min at 4,000 rpm. The supernatant was collected and introduced directly in a vial before analysis. Then, 20 µl of the carbaryl-d7 standard was also

added in the vial as an internal standard for injection. Two replicates per sample were prepared for each column analysis.

Sample vials were randomized in the analytical sequence. Quality Control (QC) samples (pool of the samples) and diluted QC samples (2-fold and 5-fold in water) were also prepared and analyzed every 10 injections. The repeated injections of QC samples were used to assess the analytical performance. Also, analytical blanks were analyzed regularly (every 20 samples) to check for carry over. Moreover, these blanks were useful for detecting the residual peaks corresponding to the mobile phases used.

Samples were analyzed in two different analytical batches. The first one was composed of 30 samples (19 organic and 11 conventional carrot samples). The second one was composed of 30 samples (14 organic and 16 conventional carrot samples) in which MS/MS acquisitions were performed. A detailed list of the samples is presented in **Table 4.A.1** and **Table 4.A.2** (Supplementary material). It can be observed that our study samples came mainly from two French regions, namely Normandy and New Aquitaine. Two samples in the first batch were from two additional geographical origins (Brittany and Hauts-de-France).

2.2.3. Analytical conditions

Analyses were performed on a ThermoFisher® Vanquish Flex UHPLC system, composed of a binary pump, a refrigerated sampler and a column oven, connected to a ThermoFisher® high resolution Orbitrap® mass spectrometer QExactive Plus with a heated electrospray ion source (HESI). The UHPLC separation was achieved using either a hydrophobic C18-silica (reversed-phase mode) or a hydrophilic one (HILIC mode). In both cases, the column temperature was set at 30°C, and the injection volume was 1 µL.

2.2.3.1. Reversed-phase mode

The UHPLC separation was achieved using a C18 Hypersil Gold column (150 x 2.1 mm, 1.9 µm) at a 0.3 mL/min flow-rate. The mobile phases were water acidified with 0.05% FA and 5 mM ammonium formate (A), and ACN acidified with 0.05% FA (B), with the following linear gradient elution: 0-2 min, B: 3%; 2-20 min, B: 3-98%; 20-24 min: B: 98%; 24-24.1 min, B: 98-3%; 24.1-32 min, B: 3%. Also, a 0.1 ml/min of ACN was added before the HRMS acquisition to improve compounds ionization.

2.2.3.2. HILIC mode

The UHPLC separation was achieved using an amide-HILIC column (150 x 2.1 mm, 2.6 µm) at a 0.6 ml/min flow-rate. The mobile phases were water acidified with 0.05% FA and 5 mM ammonium formate (A), and ACN (B), with the following linear gradient: 0-2 min, B: 98%; 2-8 min, B: 98-75%; 8-8.1 min: B: 75-40%; 8.1-10 min, B: 40%; 10-10.1 min, B: 40-98%; 10.1-15 min, B: 98%.

2.2.3.3. MS and MS/MS data acquisition

Raw data were acquired using TraceFinder software (version 3.1, ThermoFisher®). MS data were acquired using positive ion mode (ESI+) with a mass range set at *m/z* 100-1000 in full scan mode, and with a resolution of 140,000. The parameters applied on the electrospray ion source are presented in **Table 4.A.3** (Supplementary material). The MS data was acquired in centroid mode.

For the MS/MS acquisition, full scan data-dependent analyses were carried out using an inclusion list dedicated for each chromatographic mode. This inclusion list was built after the data processing of the first batch where several features were identified as discriminant; for the reversed-phase mode, the inclusion list contained 25 features and for the HILIC mode, it contained 12 features. The resolution was set at 17,500. An isolation window of ± 1 uma was used to select the *m/z* of interest at the expected retention time of the features (± 1 min). Three normalized collision energies were applied (10; 30 and 60 eV) for the MS/MS spectrum acquisition.

2.2.4. Data processing

Raw data obtained from the two different analytical columns were processed separately using the same workflow, which was built in the R software, after data files conversion to mzXML format using ProteoWizard. The main steps of data treatment were: (1) peak detection; (2) retention time alignment; (3) peaks grouping; (4) data filtration and normalization; (5) chemometric analysis and (6) discriminant peak annotation. The first three steps were performed using functions of the XCMS (an acronym for various forms (X) of chromatography mass spectrometry) R-package. This workflow, illustrated in Erreur !

Source du renvoi introuvable. (Supplementary material), was adapted from a previous study and implemented in-house (Dinis et al., 2022).

The “features”, defined by their *m/z* and retention time (RT), and their intensities in different samples were used for the statistical analysis as commonly reported (Cavanna et al., 2018). The chemometric tools used were PCA for exploratory purposes and possible sample group separation, as well as PLS-DA and OPLS-DA in order to build models for discrimination and classification of sample groups as detailed below. Also, analysis of variance (ANOVA) was used to reduce the number of features selected for model building which may improve model quality. Finally, features that were found to be the most discriminant between the sample groups were tentatively identified using online databases, as described below.

2.2.4.1. Peak detection and alignment (XCMS)

The XCMS part of the workflow consists of the following steps. The peak detection and extraction were achieved using the “centWave” method of the “findChromPeaks” function (Tautenhahn et al., 2008). This first step allowed to transform the chromatograms into a 2D-matrix where each peak was described by its feature. The selected RT for each peak corresponded to its apex of the intensity value. Then, the detected peaks were grouped across the samples according to their RT using the “groupChromPeaks” function. Values of *m/z* and RT were then averaged in the data matrix. The RT deviation was corrected using the “adjustRtime” function, and peaks were next grouped again. Finally, missing intensity values were completed using the “fillChromPeaks” function. The XCMS parameters used are presented in **Table 4.A.4** (Supplementary material). A data matrix was then generated, giving the integration of each peak made by XCMS for each feature and for each sample. Thus, the features are the variables of the models presented in this study.

2.2.4.2. Data filtration and analytical drift correction

These filtration steps were used to remove irrelevant information present on the previously obtained data matrix, as the number of features detected by XCMS was very high (about 20,000 for the C18-silica analysis and about 10,000 for the HILIC analysis). Firstly, all features detected in the dead volume and during the column equilibration were removed (for C18-silica analysis: before 1.7 min and after 26 min; for HILIC analysis: before 0.75 min and after 12 min). Secondly, features that were primarily detected in blank analyses were

removed based on their fold change (FC(samples)/ FC(blanks) < 4). Thirdly, features with low stability in the QC analyses were removed based on their relative standard deviation (RSD > 30%). Also, features for which the ratio of RSD pool over RSD sample was higher than 1.25 were deleted. Next, the analytical signal drift was corrected using a LOESS regression method from the replicate injections of the QC sample. Finally, dilution effects were removed using the probabilistic quotient normalization (PQN) (Dieterle et al., 2006). Before chemometrics, the data matrix was pareto scaled.

2.2.4.3. Chemometrics

Multivariate statistical analyses were performed using unsupervised and supervised techniques. PCA was first considered for exploratory purpose of the data sets and to detect outliers. In order to evaluate the ability of this methodology to discriminate the samples, PLS-DA and OPLS-DA models were built using a 7-fold cross validation; by this way, each data set was divided into 7 different parts. Each model was next built using 6 parts (train set) and tested using the 7^{th} part (test set); this step was then iterated until all the parts were used as test set. This cross validation permitted to determine the optimal number of latent variables (LV) to build the models (Ballabio and Consonni, 2013). A new LV was added based on the predictive performance of the model (Q2Y) value: if the Q2Y obtained with the new LV was greater than 0.01, this LV was added to the model. The quality of the built models was assessed by the goodness of fit (R2X), the proportion of the response matrix variance explained by the model (R2Y), and the Q2Y. These three metrics have values between 0 and 1: the higher they are, the better the performance of the model. The Q2Y metric and the root mean square error of prediction (RMSEP) were particularly considered here, as they represent the prediction efficiency of the model.

To evaluate the classification performance of the PLS-DA and OPLS-DA models, the data set was divided into four different subsets between the calibration set and validation set. Four PLS-DA models and four OPLS-DA models were thus obtained using the four subsets. The confusion matrix of each model was then extracted. The confusion matrix contained the number of true positive samples (TP, samples correctly classified in class A), the number of true negative samples (TN, samples correctly classified in class B), the number of false positive samples (FP, samples incorrectly classified in class A) and false negative samples (FN, samples incorrectly classified in class B). The resulting confusion matrices for these

four subsets were used to calculate the class sensitivity, the class specificity, the model accuracy, and the number of misclassifications (NMC) as detailed below (Ballabio and Consonni, 2013; Riedl et al., 2015):

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$NMC = FP + FN$$

As the number of detected features was quite high, feature selection was necessary to identify the most discriminant features that could be considered as discriminant markers. An ANOVA was then performed to select significant features between the two studied groups (the two regions of interest or organic *vs.* conventional carrots respectively); a maximal accepted p-value of 0.01 was chosen in order to select significant features. The Bonferroni correction method was used to limit the number of false positive results, as recommended when dealing with multiple hypothesis testing (Pezzatti et al., 2020). New PLS-DA and OPLS-DA models were then built based on the features selected by ANOVA to improve the performance of the models.

The chemometrics were applied on the data matrix obtained for each chromatographic column separately, as well as after fusion of both data matrices. We aimed to investigate if the discrimination models could be improved by combining the information from both chromatographic modes. Some factors such as geographical origin and sample variety are described as the most important sources of sample variability (Cubero-Leon et al., 2018; Mihailova et al., 2021). Thus, PLS-DA and OPLS-DA models were first built based on the geographical origin of the samples. The features with a high VIP (variable importance on projection) were then removed from the data matrix to remove variability related to this factor; a VIP value filtration criterion of 1 was applied as proposed in different papers (Gorrochategui et al., 2016; Pezzati et al., 2020). Then, new PLS-DA and OPLS-DA models were built for the discrimination between organic and conventional samples.

2.2.4.4. Annotation

The number of features was quite high as detailed above. It was reduced by filtering the features according to their VIP value calculated during the construction of the PLS-DA and OPLS-DA models. Filtration based on this criterion has been successfully used in previous studies, selecting 8 features (out of about 5,000 features) in a saffron study (Rubert et al., 2016) or 25 features (out of about 5,000 features) in a durum wheat study (Cavanna et al., 2020). Since feature annotation is a time-consuming part of the workflow, it was decided to reduce further the number of features by using the ANOVA results, focusing on features with the lowest p-value.

The MS spectra of the remaining features were first studied, which allowed the determination of the adduct types of the observed *m/z* and thus the exact mass of the compounds and, consequently, molecular formulas were suggested for each feature. An online database (FooDB) was then used to tentatively annotate these discriminant features. Moreover, the MS/MS experiments were used to confirm the tentative annotation by the help of spectral databases (MassBank and mzCloud).

2.3. Results and Discussion

The results obtained with the two analytical columns were quite similar. It was therefore decided to describe mainly the results obtained with the reversed-phase column because it is a widely used column. Some results obtained with the HILIC column are presented in the Supplementary material. Combining the data provided by both chromatographic modes was tested but this did not improve our results.

2.3.1. Authentication of the geographical origin of carrots

The dataset used for the reversed-phase mode contained 30 samples (in columns) and 4,652 features (in rows). The dataset used for the HILIC mode contained 30 samples (in columns) and 1,426 features (in rows). After fusion of both data matrices, the dataset used contained 30 samples (in columns) and 6,078 features (in rows).

2.3.1.1. Principal component analysis

PCA was first used as an exploratory purpose, and results from the C18-silica column are presented in **Figure 4.1**. The PCA score plot permitted to observe a cluster of the QC

samples, highlighting a good system stability during the analysis. Near 50% of variance was explained by the first three principal components (PC1: 22%, PC2: 19%; PC3: 8%). Similar results were observed for data obtained with the HILIC column (**Figure 4.A.2** of supplementary material).

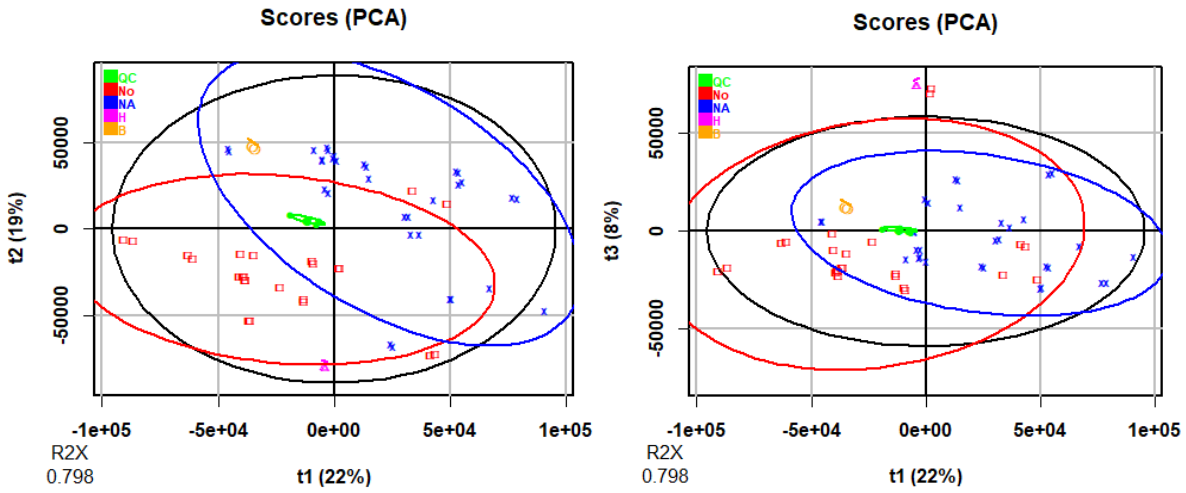

Figure 4.1. *PCA score plots of PC1 vs. PC2 and PC1 vs. PC3, obtained using the reversed-phase mode analysis for the first batch of samples (red squares: Normandy (No) samples, blue crosses: New Aquitaine (NA) samples, orange circles: Brittany (B) sample, magenta triangles: Hauts-de-France (H) sample, and green filled circles: quality control (QC) samples)*

Despite the fact that no clear separation could be obtained between the sample groups based on their geographical origin, a trend seemed to emerge on the PC1 axis regarding the separation of the New Aquitaine and the Normandy sample groups.

The two replicates per samples were grouped, showing good reproducibility. Thus, mean of the two replicates was used for the following chemometric tools.

2.3.1.2. Classification and prediction models: PLS-DA and OPLS-DA

Since two French regions were represented by only one sample each (namely Brittany and Hauts-de-France), it was decided to discard these two samples as these two geographical origins were not sufficiently represented to be modeled. The discriminant and classification models built were therefore only based on the Normandy and New Aquitaine samples.

PLS-DA and OPLS-DA models for sample discrimination and classification were built using a 7-fold cross validation. The PLS-DA model was built using three latent variables and showed a good predictive ability (Q2Y: 0.632) as shown in **Figure 4.2**. The OPLS-DA model

was built using three orthogonal latent variables and showed a similar predictive ability (Q2Y: 0.697).

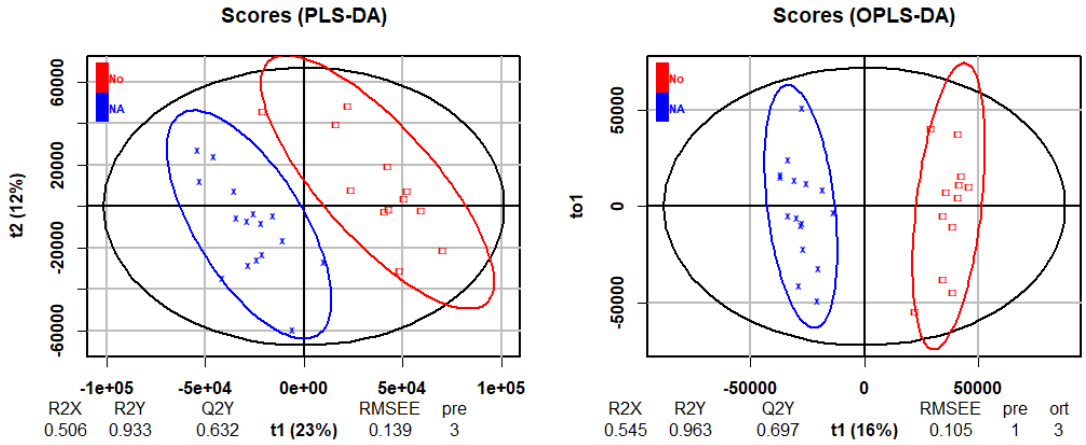

Figure 4.2. *PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the reversed-phase mode analysis for the geographical origin discrimination (blue: New Aquitaine (NA) samples, red: Normandy (No) samples)*

The PLS-DA and OPLS-DA models obtained with the four different subsets showed very similar results, with acceptable sensitivity and specificity as indicated in **Table 4.A.5**, showing good capacity of the models to correctly classify the samples in their belonging class. Overall, about 85% of model accuracy was obtained. In all the models, at least one sample was misclassified. The observed Q2Y values (about 0.7) showed good predictive ability for the obtained models, but the quite high RMSEP values (around 0.4), indicated that an error on the prediction of new samples may occur. The HILIC mode analysis showed very close results (**Figure 4.A.3** and **Table 4.A.6** of supplementary materials), with slightly better Q2Y and RMSEP values (about 0.75 and 0.35 respectively).

Geographical origin of carrots has already been assessed by other methods in previous studies. Jandric et al. used isotope ratios of strontium combined with the content of 12 elements to classify carrots from different Austrian regions with OPLS-DA (Jandric et al., 2021). Targeted LC-MS also showed good classification rates on the same dataset. Magdas et al. reported the potential of rare earth elements (REEs) measurement by ICP-MS to discriminate geographical origin of several food matrices, among which carrots (Magdas et al., 2019). Models built on REEs values showed a good stability with harvest year, which

could not be tested in this study. However, isotopic ratios and REEs are more likely to fail to discriminate geographical origins on processed or mixed food due to fractionation.

To our knowledge, our study reports for the first time the successful discrimination of carrot samples based on geographical origin using untargeted LC-HRMS analysis. It would be an interesting perspective to test the validity of our models presented here on processed or mixed carrots, and compare their performances with other reported methods such as NMR. Indeed, NMR could discriminate carrot juices from three locations in Italy (Tomassini et al., 2016). The authors showed that the different profiles were linked to the pedoclimatic conditions, which could be the case in our study as well: there were indeed around 1,500 hours of sunlight in Normandy in 2020, whereas in New Aquitaine, this number rises to 2,100 hours.

2.3.2. Authentication of organic carrots

For the sample discrimination based on their production mode, all 30 samples of the first batch were considered: 19 organic and 11 conventional carrot samples (see **Table 4.A.1**). Similar to the methodology applied by Cubero-Leon et al. (2018), features having a VIP higher than 1 for the geographical origin discrimination were removed from the data matrix to discard variability related to this factor. As previously stated, the mean value of the two replicates was used to build the models. After this filtration step, 4,268 features were left in the dataset for the reversed-phase mode and 1,314 features for the HILIC mode dataset.

PCA was unsuccessful to distinguish between organic and conventional samples as illustrated in **Figure 4.A.4** (Supplementary material). Similar results were reported for carrot samples by Cubero-Leon et al. (2018). They suggested that this may be due to the fact that PCA is an unsupervised technique and also to other factors (such as variety) that may further influence differences in metabolites.

PLS-DA and OPLS-DA models for sample discrimination and classification were next built using a 7-fold cross validation. The PLS-DA model was built using three latent variables and the OPLS-DA model was built using three orthogonal latent variables. The Q2Y obtained with these models were not as good as those obtained for the discrimination based on the geographical origin as presented in **Figure 4.3** (Q2Y: 0.304 for PLS-DA and 0.154 for OPLS-DA). Globally, for all the distributions tested, about 65% of model accuracy was obtained

(see **Table 4.A.7**). In particular, conventional samples were rarely recognized as being conventional samples by the different models as described by the specificity values obtained. On the contrary, organic samples were mostly correctly recognized, as described by the sensitivity values. The NMC obtained were higher than when dealing with the discrimination of the geographical origin (between 3 and 5 samples). The RMSEP values were also higher (around 0.5) and the Q2Y values lower (about 0.35), showing that the obtained models had a worse classification performance. The results from the HILIC mode gave very similar results (**Figure 4.A.5** and **Table 4.A.8** of supplementary material), with even higher NMC values (between 5 and 8 samples). We may attribute this unsatisfactory result to the unbalanced composition of our sample batch, and subsequently the low number of conventional samples present in the data set (11 in total). As already reported, classification models describe better the majority class, leading to poor prediction accuracy for the minority class (Nikulin et al., 2009). Unfortunately, the dataset of this study was too small to allow the use of most data balancing techniques.

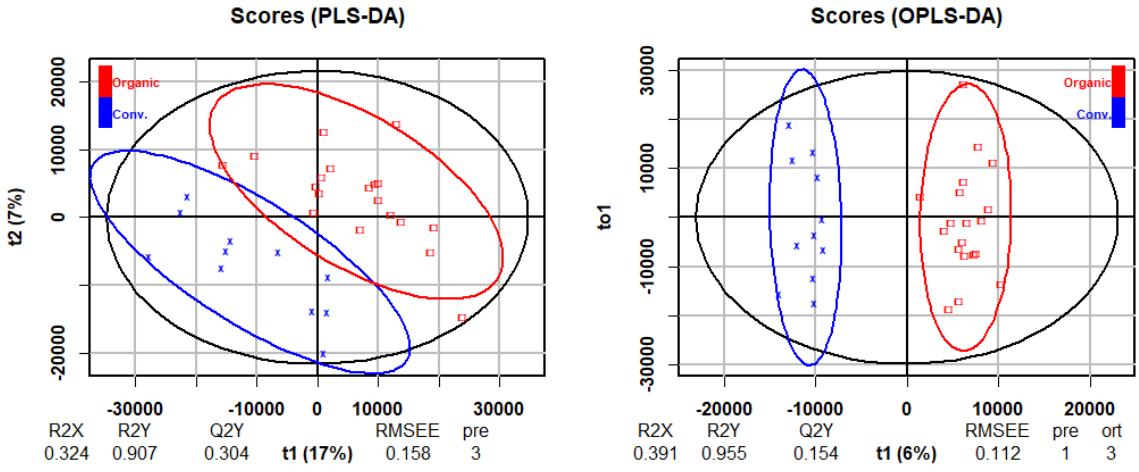

Figure 4.3. *PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the reverse phase mode analysis for the production mode discrimination (blue crosses: conventional samples, red squares: organic samples)*

The poor performances obtained for these discrimination models may be linked to the variety of the studied samples. Ten different varieties were present in our sample set, most of them having only one sample per variety (see **Table 4.A.1** in Supplementary material). Variety has already been identified as a strong source of variability for this type of study as the composition of carrots may vary with the variety (Arscott and Tanumihardjo, 2010). In their long-term study of carrots, Cubero-Leon et al. (2018) considered only two varieties which

could explain the good metrics of their model. It also has to be noticed that their carrot samples came from the same geographical area (all fields were within 20 km). Another explanation for the poor performance of the model presented here is that removing the features which contribute the most to the geographical discrimination is not sufficient to reduce the variability caused by the geographical origin. However, the number of samples from the same area was too low to test this hypothesis here.

Globally, few studies could successfully differentiate conventional from organic carrots. In particular, the nitrogen isotope ratios failed to discriminate organic and conventional carrot crops, whereas the same method was successfully applied on tomatoes and lettuce (Bateman et al., 2007). An explanation suggested by the authors is the need for nitrogen during growth which is lower for carrots than for tomatoes and lettuces. Similarly, no discrimination model using NMR has been published for carrots, although this technique has proven its ability to discriminate organic and conventional plants for many other different fruits and vegetables (Hohmann et al., 2014; Pacifico et al., 2013). To our knowledge, only one study reported successful discrimination between organic and conventional carrots by analyzing oxygen isotope ratio of sulphate (Novak et al., 2019); however, the built model was not tested on commercial samples.

2.3.3. Annotation and tentative identification of discriminant features for the geographical origin

A VIP value filtration criterion of 1 remained approximately 100 features left after filtration. Using further the ANOVA results and focusing on features with the lowest p-value, 18 features were selected for the reverse-phase mode dataset as indicated in **Table 4.1**, and 10 features for the HILIC mode dataset (see **Table 4.A.9** in Supplementary material). No additional discriminant feature could be obtained by combining the information from both chromatographic modes.

Arginine was identified and confirmed by its MS/MS spectrum (see Erreur ! Source du renvoi introuvable.), in line with previous work reporting this amino acid as a potential marker of the agricultural practice for carrots (Cubero-Leon et al., 2018). In particular, these authors linked the difference observed in the carrot metabolome to the impact of fertilization on the soil metabolome. If this hypothesis is valid, the difference observed in our study could result

from a difference in soil metabolome between Normandy and New Aquitaine. Arginine was also reported in garlic as a possible biomarker of the geographical origin (Hrbek et al., 2018).

Other amino acid derivatives were suspected as markers, although the MS/MS spectra were not confirmed (see **Table 4.1**). Our results are in line with previous ones showing that amino acids are strong markers of the geographical origin for carrots because their concentrations are linked to acclimation processes and thus depend on pedoclimatic conditions (Tomassini et al., 2016). Sciubba et al. observed a correlation between some amino acid concentrations in carrots and the harvest time (Sciubba et al., 2020). However, this hypothesis could not be confirmed in our study, as carrots from geographical areas were regularly collected during the harvest period of time.

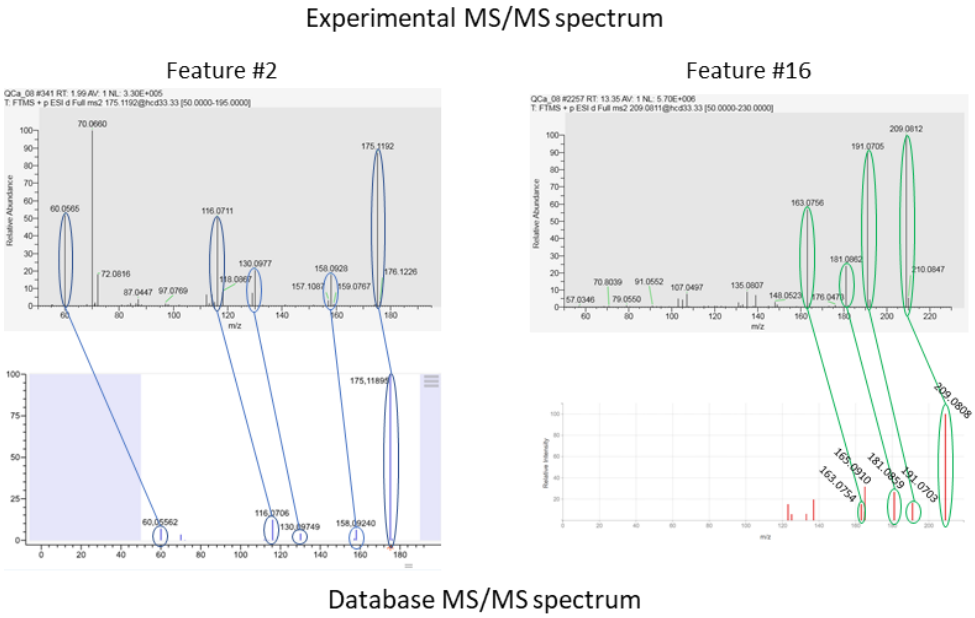

Figure 4.4. *Experimental MS/MS spectrum of features 2 and 16 for authentication of the geographical origin using the reversed-phase mode analysis, and their corresponding database MS/MS spectrum (arginine and 6-methoxymellein respectively)*

Another discriminant marker (feature 16) could also be identified and confirmed by MS/MS data. Several suspected compounds were discarded based on their MS/MS spectra and also since they had never been reported in carrots (neither from FoodDB nor from the literature). Sinapaldehyde was a suspected compound for this feature, already reported in foods according to FoodDB. However we discarded this molecule due to some differences observed between its experimental MS/MS spectrum with the spectra reported in FoodDB as shown in Erreur ! Source du renvoi introuvable. (Supplementary material).

Table 4.1. *Discriminant features for the authentication of geographical origin (compounds confirmed based on MS/MS data are indicated in bold characters) with the C18-silica column.*

# Features	Detected *m/z*	RT (min)	p-value	VIP	Characteristic	Adduct type	Monoisotopic mass	Proposed molecular formulas	Tentative identification
1	256.1291	1.78	1,67E-06	6,5	Normandy	$[M+H]^+$	255.1219	$C_{11}H_{17}N_3O_4$	*N-alpha-(tert-Butoxycarbonyl)-L-histidine**
2	175.1190	2.02	7,07E-03	4,4	New Aquitaine	$[M+H]^+$	174.1117	$C_6H_{14}N_4O_2$	**Arginine**
3	229.1548	2.64	5,11E-03	3,1	Normandy	$[M+H]^+$	228.1474	$C_{11}H_{20}N_2O_3$	*L-Isoleucyl-L-proline*, L-Leucyl-L-proline**
4	146.0812	2.85	3,32E-04	2,2	New Aquitaine	$[M+H]^+$	145.0739	$C_6H_{11}NO_3$	*Allysine*, L-cis-4-(Hydroxymethyl)-2-pyrrolidinecarboxylic acid*, 4-Acetamidobutanoic acid**
5	188.0706 205.0971	6.5	1,22E-04	7,6	New Aquitaine	$[M+H]^+$ $[M+NH_4]^+$	187.0633	$C_{11}H_9NO_2$	*3-(1H-Indol-3-yl)-2-propenoic acid**
6	383.1312	6.64	9,66E-04	1,4	Normandy	$[M+H]^+$	382.1237 382.1232 382.1250 382.1250 382.1224	$C_{14}H_{18}N_6O_7$ $C_{28}H_{16}NO$ $C_{15}H_{14}N_{10}O_3$ $C_{16}H_{20}N_3O_8$ $C_{13}H_{22}N_2O_{11}$	n.a.
7	378.1757	6.65	5,70E-04	1,8	Normandy	$[M+H]^+$	377.1686 377.1672 377.1672 377.1699	$C_{16}H_{27}NO9$ $C_{14}H_{25}N_4O_8$ $C_{13}H_{19}N_{11}O_3$ $C_{17}H_{23}N_5O_5$	n.a.
8	342.1758	7.15	1,76E-08	1,1	Normandy	$[M+H]^+$	341.1686 341.1672 341.1699	$C_{13}H_{27}NO_9$ $C_{10}H_{19}N_{11}O_3$ $C_{14}H_{23}N_5O_5$	n.a.
9	390.1393	8.31	3,83E-05	1,8	Normandy	$[M+H]^+$	389.1322 389.1309 389.1308 389.1335	$C_{16}H_{23}NO_{10}$ $C_{14}H_{21}N_4O_9$ $C_{13}H_{15}N_{11}O_4$ $C_{17}H_{19}N_5O_6$	n.a.
10	490.2282	8.45	3,37E-07	1,2	Normandy	$[M+H]^+$	489.2210 489.2205 489.2218	$C_{22}H_{35}NO_{11}$ $C_{35}H_{27}N_3$ $C_{37}H_{29}O$	n.a.

							489.2197	$C_{20}H_{33}N_4O_{10}$	
							489.2223	$C_{23}H_{31}N_5O_7$	
11	432.1864	9.08	1,60E-04	2,0	Normandy	$[M+H]^+$	431.1791	$C_{17}H_{17}N_{15}$	n.a.
							431.1791	$C_{18}H_{23}N_8O_5$	
							431.1791	$C_{19}H_{29}NO_{10}$	
							431.1800	$C_{34}H_{23}$	
							431.1778	$C_{17}H_{27}N_4O_9$	
							431.1778	$C_{16}H_{21}N_{11}O_4$	
							431.1805	$C_{20}H_{25}N_5O_6$	
12	354.1757	9.16	8,72E-03	1,2	New Aquitaine	$[M+H]^+$	353.1699	$C_{15}H_{23}N_5O_5$	*Dihydrozeatin riboside**, *2-[[1-(2-amino-3-hydroxybutanoyl)pyrrolidine-2-carbonyl]amino]-3-(1H-imidazol-5-yl)propanoic acid**
13	348.2744	12.27	2,36E-03	16,5	Normandy	$[M+H]^+$	347.2672	$C_{17}H_{31}N_8$	n.a.
	313.2371					Fragment	347.2672	$C_{18}H_{37}NO_5$	
							347.2658	$C_{16}H_{35}N_4O_4$	
							347.2685	$C_{19}H_{33}N_5O$	
14	683.4699	12.27	7,09E-03	4,6	Normandy	$[M+H]^+$	682.4629	$C_{34}H_{62}N_6O_8$	n.a.
	353.2296					Fragment	682.4616	$C_{33}H_{66}N_2O_{12}$	
							682.4611	$C_{46}H_{58}N_4O$	
							682.4597	$C_{45}H_{62}O_5$	
							682.4656	$C_{38}H_{66}O_{10}$	
15	699.4347	12.27	2,94E-03	2,7	Normandy	$[M+H]^+$	698.4241	$C_{37}H_{62}O_{12}$	*Cyclopassifloside I**, *Cyclopassifloside IV**, *Cyclopassifloside X**
16	209.0809	13.46	6,06E-05	14,5	New Aquitaine	$[M+H]^+$	208.0736	$C_{11}H_{12}O_4$	**6-Methoxymellein**, Sinapaldehyde, *Caffeic acid ethyl ester**, *Methyl ferulate**, *3-(3,4-Dimethoxyphenyl)-2-propenoic acid**, *Furapiole**, *1-(2-Methoxy-3,4-methylenedioxyphenyl)-1-propanone**, *3-(3-Methoxy-4,5-methylenedioxyphenyl)-2-propen-1-ol**, *Anthriscinol**, *(R)-2-Benzylsuccinate**, *5-(3',4'-Dihydroxyphenyl)-gamma-valerolactone**, *5-(3',5'-Dihydroxyphenyl)-gamma-valerolactone**
	191.0703					Fragment			
17	374.2901	14.42	4,69E-03	2,3	New Aquitaine	$[M+H]^+$	373.2842	$C_{21}H_{35}N_5O$	*1-[[(2R,4S,5R)-5-(5-cyclohexyl-2-methylpyrazol-3-yl)-1-azabicyclo[2.2.2]octan-2-yl]methyl]-3-ethylurea**

18	305.2474	19.31	8,86E-04	1,7	New Aquitaine	$[M+H]^+$	304.2402	$C_{20}H_{32}O_2$	*Oryzalexin D*, Oryzalexin F*, Arachidonic acid*, Sideridiol*, ent-17-Hydroxy-16b-kauran-19-al*, Yucalexin P21*, Copalic acid*, Junicedral*, 7,13-Eperudien-15-oic acid*, Oryzalexin S*, Oryzalexin E*, Mesterolone*, Abietadiene-diol*, Sodium oleate**

*n.a.: not applicable; * invalid based on MS/MS data*

We finally identified feature 16 as being 6-methoxymellein (or phytoalexin 8-hydroxy-3-methyl-6-methoxy-3,4-dihydroisocoumarin) as indicated in **Table 4.1** and shown in Erreur ! Source du renvoi introuvable.. This compound has already been reported in fresh carrots, with varying levels (0.02 to 76 µg/g) depending on carrot genotype, maturity as well as environmental and growing conditions (De Girolamo et al., 2004). It is generally considered of being associated with the bitterness in strained carrots, and partly responsible for the sensory quality of carrots. A threshold concentration of 94 µg/g has been reported to bitter flavor in carrots (De Girolamo et al. 2004). Levels of 6-methoxymellein in carrots were found stable at low temperature for quite long periods.

Examples of chromatograms for the discriminant features identified according to the MS/MS acquisitions for each analytical column are presented in Erreur ! Source du renvoi introuvable. and Erreur ! Source du renvoi introuvable. (Supplementary material).

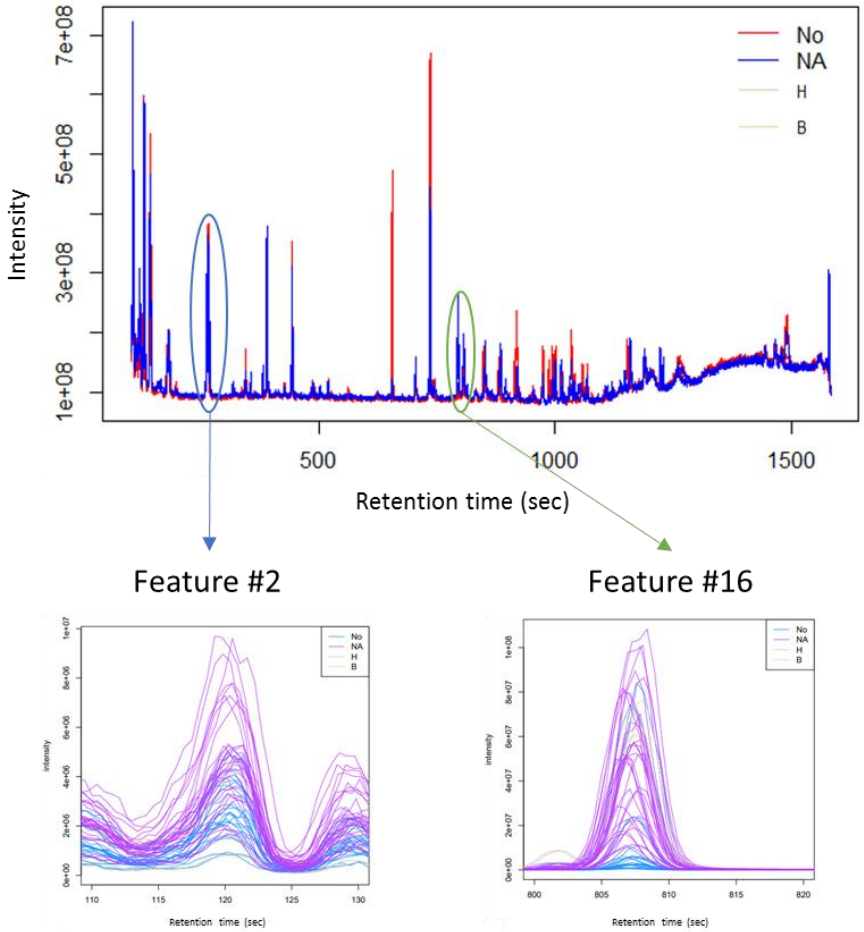

Figure 4.5. *Chromatogram of features 2 and 16 for authentication of the geographical origin using the reversed phase mode analysis*

Two additional biomarkers could be identified based on their MS/MS spectra thanks to the results from the HILIC polar column as indicated in **Table 4.A.9**, Erreur ! Source du renvoi introuvable. and Erreur ! Source du renvoi introuvable. (Supplementary material): N-acetylputrescine and L-carnitine.

N-Acetylputrescine was reported in plants, especially in *Daucus carota*, being linked to plant defense mechanism (Lou et al., 2016). This molecule comes from the acetylation of putrescine, a polyamine biosynthetised from arginine that may accumulate in plants in response to several biotic and abiotic stresses. Interestingly, in our work both arginine avec N-acetylputrescine are found as biomarkers of the New Aquitaine geographical origin.

L-Carnitine (or 2-hydroxy-4-trimethylammonium butyrate) is an amino acid already reported in plants, playing a role in their lipid metabolism (Bourdin et al., 2007). Concentrations of 37 µg/g have been recently reported in carrots (Kepka et al., 2021), and nutritional sources of carnitine are considered as important for the human body. It is suspected to be biosynthetised in the plant from gamma-butyrobetaine (or 4-trimethylammonium butyrate) as a precursor (Rippa et al., 2012). In our work, we found both L-carnitine as a marker of the Normandy geographical origin (see **Table 4.A.9** and Erreur ! Source du renvoi introuvable.), and gamma-butyrobetaine was considered as a possible compound for feature 7. However, 3-dehydroxycarnitine was also a possible compound for this feature, even though to our knowledge it has never been reported in plants.

During the MS/MS experiments, a full scan analysis was also acquired. It was thus possible to use these independent acquisitions to evaluate the potential of the discriminant features to serve as marker compounds. It is noteworthy that the majority of the selected features were successfully observed on this second sample batch (only three features for the reversed-phase mode and one feature for the HILIC mode showed a very low intensity). No significant difference in intensity was observed between the two groups of samples, but a trend was noted for some features (12 features for the reversed-phase mode and 6 features for the HILIC mode).

2.4. CONCLUSION

Forty-four samples of carrots were analysed by UHPLC-HRMS using two different chromatographic phases (C18-silica and HILIC). The samples were collected from various

production sites at the beginning of 2021 and originated from different regions of France. PLS-DA and OPLS-DA models were constructed to discriminate production modes (conventional *versus* organic) and geographical origin (New Aquitaine *versus* Normandy). Despite some trend in sample discrimination based on the production mode, the models metrics were not satisfying (Q2Y below 0.5). On the contrary, good results were obtained for the geographical discrimination. An additional batch was analysed to validate the discrimination performance and tentatively identify discriminant markers. Arginine and 6-methoxymellein were proposed as biomarkers of the geographical origin based on the C18-silica analyses and their MS/MS spectra. The HILIC analyses provided the identification of two additional biomarkers confirmed by their MS/MS spectra: N-acetylputrescin and L-carnitin.

Acknowledgements

The authors warmly thank GRAB, Carottes de France, Larrère, Legum'land, GIE de l'Ombrière, SICA Altus, Valorex, Agrial and SILEBAN for kindly providing them with several samples of carrots from their fields or from suppliers of their network. These samples were collected through the TOFoo (True Organic Food) project, supported by the French Government in the framework of the "Investissements d'avenir" program. The authors are also thankful for the financial support provided through this program, as well as for the financial support provided by the Association Nationale Recherche et Technologie (ANRT) through the CIFRE program (CIFRE n°2018/0937).

Conflict of interest

The authors declare that they have no commercial or financial relationships that could have influence the research conducted in this paper.

2.5. REFERENCES

Ahmad, T., Cawwod, M., Iqbal, Q., Arino, A., Batool, A., Tariq, R.M.S., Azam, M., & Akhtar, S. (2019). Phytochemicals in Daucus carota and their health benefits – review article. *Foods*, *8*(9), 424. https://doi.org/10.3390/foods8090424

Akhtar, S., Rauf, A., Imran, M., Qamar, M., Riaz, M., & Mubarak, M. S. (2017). Black carrot (*Daucus carota* L.), dietary and health promoting perspectives of its polyphenols: A

review. *Trends in Food Science & Technology*, *66*, 36– 47. https://doi.org/10.1016/j.tifs.2017.05.004

Arscott, S. A., & Tanumihardjo, S. A. (2010). Carrots of many colors provide basic nutrition and bioavailable phytochemicals acting as a functional food. *Comprehensive Reviews in Food Science and Food Safety, 9*(2), 223–239. https://doi.org/10.1111/j.1541-4337.2009.00103.x

Ballabio, D., & Consonni, V., (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods, 5*(16), 3790. https://doi.org/10.1039/c3ay40582f

Barbosa, S., Campmajo, G., Saurina, J., Puignou, L., & Nunez, O. (2020). Classification and authentication of paprika by UHPLC-HRMS fingerprinting and multivariate calibration methods (PCA and PLS-DA). *Foods, 9*(4), 486. https://dx.doi.org/10.3390/foods9040486

Bateman, A. S., Kelly, S. D., & Woolfe, M. (2007). Nitrogen Isotope Composition of Organically and Conventionally Grown Crops. *Journal of Agricultural and Food Chemistry*, *55*(7), 2664–2670. https://doi.org/10.1021/jf0627726

Bourdin, B., Adenier, H., & Perrin, Y. (2007). Carnitine is associated with fatty acid metabolism in plants. *Plant Physiology and Biochemistry, 45*(12), 926-931. http://dx.doi.org/10.1016/j.plaphy.2007.09.009

Campmajó, G., Núñez, N., & Núñez, O. (2019). The Role of Liquid Chromatography-Mass Spectrometry in Food Integrity and Authenticity. In G. Shamrao Kamble (Ed.), *Mass Spectrometry—Future Perceptions and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.85087

Campmajó, G., Rodríguez-Javier, L. R., Saurina, J., & Núñez, O. (2021). Assessment of paprika geographical origin fraud by high-performance liquid chromatography with fluorescence detection (HPLC-FLD) fingerprinting. *Food Chemistry*, *352*, 129397. https://doi.org/10.1016/j.foodchem.2021.129397

Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends in Food Science and Technology*, *80*, 223-241. https://doi.org/10.1016/j.tifs.2018.08.007

Cavanna, D., Loffi, C., Dall'Asta, C., & Suman, M. (2020). A non-targeted high-resolution mass spectrometry approach for the assessment of the geographical origin of durum wheat. *Food Chemistry, 317*, 126366. https://doi.org/10.1016/j.foodchem.2020.126366

Cubero-Leon, E., De Rudder, O., & Maquet, A. (2018). Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry*, *239*, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

Cuevas, F. J., Pereira-Caro, G., Moreno-Rojas, J. M., Muñoz-Redondo, J. M., & Ruiz-Moreno, M. J. (2017). Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics. *Food Control*, *82*, 203–211. https://doi.org/10.1016/j.foodcont.2017.06.031

De Girolamo A., Solfrizzo M., Vitti C., & Visconti, A. (2004). Occurrence of 6-methoxymellein in fresh and processed carrots and relevant effect of storage and processing. *Journal of Agricultural and Food Chemistry, 52*(21), 6478-6484, https://doi.org/10.1021/jf0491660

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, *78*(13), 4281–4290. https://doi.org/10.1021/ac051632c

Dinis, K., Tsamba, L., Thomas, F., Jamin, E., & Camel, V. (2022). Preliminary authentication of apple juices using untargeted UHPLC-HRMS analysis combined to chemometrics, Food Control, in press. https://doi.org/10.1016/j.foodcont.2022.109098

Gorrochategui, E., Jaumot, J., Lacorte, S., & Tauler, R. (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. *Trends in Analytical Chemistry*, *82*, 425−442. https://doi.org/10.1016/j.trac.2016.07.004

Hohmann, M., Christoph, N., Wachter, H., & Holzgrabe, U. (2014). ^{1}H NMR Profiling as an Approach To Differentiate Conventionally and Organically Grown Tomatoes. *Journal of Agricultural and Food Chemistry*, *62*(33), 8530–8540. https://doi.org/10.1021/jf502113r

Hrbek V., Rektorisova M., Chmelarova H., Ovesna J., & Hajslova J. (2018). Authenticity assessment of garlic using a metabolomic approach based on high resolution mass spectrometry, *Journal of Food Composition and Analysis, 67*, 19-28. https://doi.org/10.1016/j.jfca.2017.12.020

Jandric Z., Tchaikovsky A., Zitek A, Causon, T, Stursa V, Prohaska T, & Hann, S. (2021). Multivariate modelling techniques applied to metabolomic, elemental and isotopic fingerprints for the verification of regional geographical origin of Austrian carrots. *Food Chemistry, 338*, 127924. https://doi.org/10.1016/j.foodchem.2020.127924.

Kepka, A., Ochocinska, A., Chojnowska, S., Borzym-Kluczyk, M., Skorupa, E., Knas, M., & Waszkiewicz, N. (2021). Potential Role of L-Carnitine in Autism Spectrum Disorder. *Journal of Clinical Medicine, 10*(6), 1202. https://doi.org/10.3390/jcm10061202

Koudela, M., Schulzova, V., Krmela, A., Chmelarova, H., Hajslova, J., & Novotny, C. (2021). Effect of agroecological conditions on biologically active compounds and metabolome in carrot. *Cells, 10*(4), 784. https://doi.org/10.3390/cells10040784

Llano, S.M., Muñoz-Jiménez, A.M., Jiménez-Cartagena, C., Londoño-Londoño, J., & Medina, S. (2018). Untargeted metabolomics reveals specific withanolides and fatty acyl glycoside as tentative metabolites to differentiate organic and conventional Physalis peruviana fruits. *Food Chemistry, 244*, 120–127. https://doi.org/10.1016/j.foodchem.2017.10.026

Lou, Y-R., Bor, M., Yan, J., Preuss, A.S., & Lander, G. (2016). Arabidopsis NATA1 acetylates putrescine and decreases defense-related hydrogen peroxide accumulation. *Plant Physiology, 171*, 1443–1455. https://doi.org/10.1104/pp.16.00446

Magdas, D. A., Marincas, O., Cristea, G., Feher, I., & Vedeanu, N. (2020). REEs – a possible tool for geographical origin assessment? *Environmental Chemistry*, *17*(2), 148. https://doi.org/10.1071/EN19163

Martínez Bueno, M.J., Díaz-Galiano, F.J., Rajski, Ł., Cutillas, V., & Fernández-Alba, A.R. (2018). A non-targeted metabolomic approach to identify food markers to support discrimination between organic and conventional tomato crops. *Journal of Chromatography A, 1546*, 66–76. https://doi.org/10.1016/j.chroma.2018.03.002

Mie, A., Laursen K.H., Aberg, K.M., Forshed, J., Lindahl, A., Thorup-Kristensen, K., Olsson, M., Knuthsen, P., Larsen, E.H., & Husted, S. (2014). Discrimination of conventional and organic white cabbage from a long-term field trial study using untargeted LC-MS-based metabolomics. *Analytical and Bioanalytical Chemistry, 406*(12), 2885–2897. https://doi.org/10.1007/s00216-014-7704-0

Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T., Maestroni, B.M., & Cannavan, A. (2021). High-resolution mass spectrometry-based metabolomics for the discrimination between organic and conventional crops: A review. *Trends in Food Science & Technology, 110*, 142–154. https://doi.org/10.1016/j.tifs.2021.01.071

Nikulin, V., McLachlan, G. J., & Ng, S. K. (2009). Ensemble Approach for the Classification of Imbalanced Data. In A. Nicholson & X. Li (Eds.), *AI 2009: Advances in Artificial Intelligence* (Vol. 5866, pp. 291–300). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10439-8_30

Novak, V., Adler, J., Husted, S., Fromberg, A., & Laursen, K. H. (2019). Authenticity testing of organically grown vegetables by stable isotope ratio analysis of oxygen in plant-derived sulphate. *Food Chemistry*, *291*, 59–67. https://doi.org/10.1016/j.foodchem.2019.03.125

Pacifico, D., Casciani, L., Ritota, M., Mandolino, G., Onofri, C., Moschella, A., Parisi, B., Cafiero, C., & Valentini, M. (2013). NMR-Based Metabolomics for Organic Farming Traceability of Early Potatoes. *Journal of Agricultural and Food Chemistry*, *61*(46), 11201–11211. https://doi.org/10.1021/jf402961m

Pedrosa, M.C., Lima, L., Heleno, S., Carocho, M., Ferreira, I.C.F.R., & Barros, L. (2021). Food metabolites as tools for authentication, processing, and nutritive value assessment. *Foods, 10*(9), 2213. https://doi.org/10.3390/foods10092213

Pereira, F.d.O., Pereira, R.d.S., Rosa, L.d.S., & Teodoro, A.J. (2016). Organic and conventional vegetables: comparison of the physical and chemical characteristics and antioxidant activity. *African Journal of Biotechnology, 15*(33), 1746-1755. https://doi.org/10.5897/1JB2016.15386

Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., Gonzalez-Ruiz, V., & Rudaz, S. (2020). Implementation of liquid chromatography - high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: a tutorial. *Analytical Chimica Acta, 1105*, 28-44. https://doi.org/10.1016/j.aca.2019.12.062.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885*, 17–32. https://doi.org/10.1016/j.aca.2015.06.003

Rippa, S., Zhao, Y., Merlier, F., Charrier, A., & Perrin, Y. (2012). The carnitine biosynthetic pathway in *Arabidopsis thaliana* shares similar features with the pathway of mammals and fungi. *Plant Physiology and Biochemistry, 60*, 109–114. https://dx.doi.org/10.1016/j.plaphy.2012.08.001

Rocchetti, G., Pateiro, M., Campagnol, P.C.B., Barba, F.J., Tomasevic, I., Montesano, D., Lucini, L., & Lorenzo, J.M. (2020). Effect of partial replacement of meat by carrot on physicochemical properties and fatty acid profile of fresh turkey sausages: a chemometric approach. *Journal of the Science of Food and Agriculture, 100*(13), 4968-4977. https://doi.org/10.1002/jsfa.10560

Rubert, J., Lacina, O., Zachariasova, M., & Hajslova, J. (2016). Saffron authentication based on liquid
chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food*
Chemistry, 204, 201–209. https://doi.org/10.1016/j.foodchem.2016.01.003

Sciubba, F., Tomassini, A., Giorgi, G., Brasili, E., Pasqua, G., Capuani, G., Aureli, W., & Miccheli, A. (2020). NMR-Based Metabolomic Study of Purple Carrot Optimal Harvest Time for Utilization as a Source of Bioactive Compounds. *Applied Sciences*, *10*(23), 8493. https://doi.org/10.3390/app10238493

Søltoft, M., Nielsen, J., Holst Laursen, K., Husted, S., Halekoh, U., & Knuthsen, P. (2010). Effects of Organic and Conventional Growth Systems on the Content of Flavonoids in Onions and Phenolic Acids in Carrots and Potatoes. *Journal of Agricultural and Food Chemistry*, *58*(19), 10323–10329. https://doi.org/10.1021/jf101091c

Stolarczyk, J., & Janick, J. (2011). Carrot: history and iconography. *Chronia Horticulturae*, *51*(2).

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, *9*(1). https://doi.org/10.1186/1471-2105-9-504

Tomassini, A., Sciubba, F., Di Cocco, M. E., Capuani, G., Delfini, M., Aureli, W., & Miccheli, A. (2016). ^{1}H NMR-Based Metabolomics Reveals a Pedoclimatic Metabolic Imprinting in Recady-to-Drink Carrot Juices. *Journal of Agricultural and Food Chemistry*, *64*(25), 5284–5291. https://doi.org/10.1021/acs.jafc.6b01555

2.6. APPENDIX A. SUPPLEMENTARY MATERIAL

Table 4.A.1. *Information about the carrot samples of the first batch.*

Sample Code	Farming production	Region	Variety
car_b_00569	Organic	New Aquitaine	Brillyance
car_b_00570	Organic	New Aquitaine	Boléro
car_b_00571	Organic	New Aquitaine	Brillyance
car_b_00572	Organic	New Aquitaine	Maestro
car_c_00587	Conventional	New Aquitaine	Natuna
car_c_00588	Conventional	New Aquitaine	Maestro
car_b_00589	Organic	Normandy	Norway
car_b_00590	Organic	Normandy	Nipomo
car_b_00591	Organic	Normandy	Nerac
car_b_00592	Organic	Normandy	Norway
car_b_00593	Organic	Normandy	Nerac
car_c_00594	Conventional	Normandy	Nazareth
car_b_00595	Organic	Normandy	Negovia
car_b_00603	Organic	New Aquitaine	Maestro
car_c_00604	Conventional	New Aquitaine	Maestro
car_b_00605	Organic	New Aquitaine	Nipomo
car_c_00606	Conventional	New Aquitaine	Natuna
car_b_00607	Organic	New Aquitaine	Brillyance
car_b_00608	Organic	New Aquitaine	Boléro
car_b_00609	Organic	New Aquitaine	Maestro
car_c_00610	Conventional	New Aquitaine	Boléro
car_b_00612	Organic	New Aquitaine	Maestro
car_b_00613	Organic	New Aquitaine	Brillyance
car_c_00619	Conventional	Brittany	Maestro
car_c_00620	Conventional	Normandy	Nerac
car_c_00621	Conventional	Normandy	Maestro
car_c_00622	Conventional	Normandy	Subito
car_b_00623	Organic	Normandy	Norway
car_b_00624	Organic	Hauts-de-France	Norway
car_c_00625	Conventional	Normandy	Nerac

Table 4.A.2. *Information about the carrot samples of the second batch used for MS/MS acquisitions.*

Sample code	Farming production	Region	Variety
car_c_00182	Conventional	Normandy	Rodelika
car_c_00185	Conventional	Normandy	Touchon
car_c_00186	Conventional	Normandy	Berlikum
car_b_00188	Organic	Normandy	Napoli
car_c_00230	Conventional	Normandy	Touchon
car_c_00231	Conventional	Normandy	Napoli
car_b_00569	Organic	New Aquitaine	Brillyance
car_b_00572	Organic	New Aquitaine	Maestro
car_c_00587	Conventional	New Aquitaine	Natuna
car_b_00589	Organic	Normandy	Norway
car_b_00590	Organic	Normandy	Nipomo
car_b_00591	Organic	Normandy	Nerac
car_b_00593	Organic	Normandy	Nerac
car_c_00594	Conventional	Normandy	Nazareth
car_c_00604	Conventional	New Aquitaine	Maestro
car_b_00607	Organic	New Aquitaine	Brillyance
car_c_00610	Conventional	New Aquitaine	Boléro
car_b_00612	Organic	New Aquitaine	Maestro
car_c_00620	Conventional	Normandy	Nerac
car_c_00621	Conventional	Normandy	Maestro
car_b_00623	Organic	Normandy	Norway
car_c_00625	Conventional	Normandy	Nerac
car_c_01216	Conventional	New Aquitaine	Speedo
car_b_01217	Organic	New Aquitaine	Napoli
car_b_01289	Organic	New Aquitaine	Laguna
car_b_01290	Organic	New Aquitaine	Speedo
car_b_01291	Organic	New Aquitaine	Napoli
car_c_01292	Conventional	New Aquitaine	Speedo
car_c_01387	Conventional	New Aquitaine	Romance
car_c_01388	Conventional	New Aquitaine	Laguna

Table 4.A.3. *Parameters and their corresponding values for the MS source.*

Parameters	Value
Sheath gas flow rate (A.U.)	50
Aux gas flow rate (A.U.)	13
Sweep gas flow rate (A.U.)	1
Spay voltage (|kV|)	3.50
Capillary temperature (°C)	275
S-lens RF level	50
Aux. gas heater temperature (°C)	400

A.U.: Arbitrary units

Table 4.A.4. *Parameters and their corresponding values for the different steps using XCMS and for the two analytical columns used.*

Step	Parameter	Value (C18)	Value (HILIC)
findChromPeaks	method	centWave	centWave
	ppm	5	5
	peakwidth	5 - 30	10 - 60
	snthresh	5	5
	prefilter	3 / 1000	3 / 100
	mzCenterFun	apex	apex
	integrate	1	1
	mzdiff	0.01	0.01
	noise	100	100
groupChromPeaks - 1	method	density	density
	bw	10	10
	minFraction	0.25	0.25
	minSample	5	5
	binSize	0.02	0.02
adjustRtime	method	peakgroups	peakgroups
	smooth	loess	loess
	extraPeaks	1	1
	minFraction	0.8	0.8
	span	0.2	0.5
	family	gaussian	gaussian
groupChromPeaks - 2	method	density	density
	bw	5	5
	minFraction	0.25	0.25
	minSample	5	5
	binSize	0.02	0.02
fillChromPeaks	method	FillChromPeaksParam	FillChromPeaksParam

Table 4.A.5. *Metrics obtained from the confusion matrices on four different subsets to assess the performance of PLS-DA and OPLS-DA models for carrot samples discrimination based on their geographical origin using the reversed-phase mode analysis.*

# Subset	Number of samples (calibration set)	Number of samples (test set)	Sensitivity	Specificity	Accuracy	Q2Y	RMSEP	NMC
PLS-DA models								
1	20	8	100.0%	75.0%	87.5%	0.707	0.349	1
2	20	8	100.0%	75.0%	87.5%	0.699	0.396	1
3	19	9	60.0%	100.0%	77.8%	0.733	0.441	2
4	19	9	100.0%	80.0%	88.9%	0.546	0.402	1
OPLS-DA models								
1	20	8	100.0%	75.0%	87.5%	0.854	0.348	1
2	20	8	100.0%	75.0%	87.5%	0.784	0.368	1
3	19	9	60.0%	100.0%	77.8%	0.739	0.450	2
4	19	9	100.0%	80.0%	88.9%	0.659	0.402	1

Table 4.A.6. *Metrics obtained from the confusion matrices on four different subsets to assess the performance of PLS-DA and OPLS-DA models for carrot samples discrimination based on their geographical origin using the HILIC mode analysis.*

# Subset	Number of samples (calibration set)	Number of samples (test set)	Sensitivity	Specificity	Accuracy	Q2Y	RMSEP	NMC
PLS-DA models								
1	20	8	100.0%	75.0%	87.5%	0.751	0.364	1
2	20	8	100.0%	75.0%	87.5%	0.743	0.310	1
3	19	9	80.0%	100.0%	88.9%	0.727	0.360	1
4	19	9	100.0%	60.0%	77.8%	0.730	0.388	2
OPLS-DA models								
1	20	8	100.0%	75.0%	87.5%	0.808	0.337	1
2	20	8	100.0%	75.0%	87.5%	0.833	0.303	1
3	19	9	80.0%	100.0%	88.9%	0.762	0.360	1
4	19	9	100.0%	60.0%	77.8%	0.749	0.388	2

Table 4.A.7. *Metrics obtained from the confusion matrix on four different subsets to assess the performance of PLS-DA and OPLS-DA models for carrot samples discrimination based on their production mode using the reversed-phase mode analysis.*

# Subset	Number of samples (calibration set)	Number of samples (test set)	Sensitivity	Specificity	Accuracy	Q2Y	RMSEP	NMC
PLS-DA models								
1	19	11	71.4%	75.0%	72.7%	0.205	0.478	3
2	19	11	80.0%	50.0%	63.6%	0.613	0.591	4
3	18	12	100.0%	20.0%	66.7%	0.708	0.578	4
4	18	12	100.0%	40.0%	58.3%	0.443	0.569	5
OPLS-DA models								
1	19	11	71.4%	75.0%	72.7%	0.080	0.478	3
2	19	11	80.0%	50.0%	63.6%	0.060	0.602	4
3	18	12	100.0%	20.0%	66.7%	0.099	0.583	4
4	18	12	100.0%	28.6%	58.3%	0.224	0.569	5

Table 4.A.8. *Metrics obtained from the confusion matrix on four different subsets to assess the performance of PLS-DA and OPLS-DA models for carrot samples discrimination based on their production mode using the HILIC mode analysis.*

# Subset	Number of samples (calibration set)	Number of samples (test set)	Sensitivity	Specificity	Accuracy	Q2Y	RMSEP	NMC
PLS-DA models								
1	19	11	57.1%	50.0%	54.5%	0.012	0.521	5
2	19	11	40.0%	50.0%	45.5%	0.537	0.507	6
3	18	12	71.4%	40.0%	58.3%	0.638	0.532	5
4	18	12	80.0%	0.0%	33.3%	0.550	0.620	8
OPLS-DA models								
1	19	11	57.1%	50.0%	54.5%	0.233	0.486	5
2	19	11	40.0%	50.0%	45.5%	0.626	0.502	6
3	18	12	71.4%	40.0%	58.3%	0.362	0.528	5
4	18	12	80.0%	0.0%	33.3%	0.414	0.612	8

Table 4.A.9. *Discriminant features for the authentication of geographical origin (compounds confirmed based on MS/MS data are indicated in bold characters) with the HILIC mode*

# Features	Detected m/z	RT (min)	p-value	VIP	Characteristic	Adduct type	Monoisotopic mass	Proposed molecular formulas	Tentative identification
1	285.1021	6.95	9,13E-03	2,0	Normandy	$[M+H]^+$	284.0950 284.0955	$C_{19}H_{12}N_2O$ $C_4H_8N_{14}O_2$	n.a.
2	188.0705 205.0970	7.00	4,22E-04	6,3	New Aquitaine	$[M+H]^+$ $[M+NH_4]^+$	187.0633	$C_{11}H_9NO_2$	*3-(1H-Indol-3-yl)-2-propenoic acid**
3	221.0920	7.05	1,59E-03	1,8	New Aquitaine	$[M+H]^+$	220.0848	$C_{11}H_{12}N_2O_3$	*5-Hydroxy-L-tryptophan*, Oxitriptan 1**
4	131.1180	7.69	3,33E-04	2,4	New Aquitaine	$[M+H]^+$	130.1106	$C_6H_{14}N_2O$	**N-Acetylputrescine** *N-Nitrosodipropylamine**
5	279.0475	8.05	1,73E-05	1,1	Normandy	$[M+H]^+$	278.0400 278.0413	$C_9H_6N_6O_5$ $C_{11}H_8N_3O_6$	n.a.
6	265.1116	8.18	4,48E-08	2,7	Normandy	$[M+H]^+$	264.1043 264.1056 264.1056	$C_4H_{12}N_{10}O_4$ $C_5H_8N_{14}$ $C_6H_{14}N_7O_5$	n.a.
7	146.1175	8.44	2,57E-05	3,1	Normandy	$[M+H]^+$	145.1103	$C_7H_{15}NO_2$	3-Dehydroxycarnitine Gamma-butyrobetaine *(3R,4S)-3-Amino-4-methylhexanoic acid*, (S)-3-Amino-5-methylhexanoic acid*, L-beta-Homoisoleucine hydrochloride**
8	162.1124	8.74	3,77E-03	3,7	Normandy	$[M+H]^+$	161.1052	$C_7H_{15}NO_3$	**L-Carnitine**
9	256.1290	9.32	8,95E-08	1,4	Normandy	$[M+H]^+$	255.12	$C_{11}H_{17}N_3O_4$	*N-alpha-(tert-Butoxycarbonyl)-L-histidine**
10	309.1290	9.34	1,46E-06	2,2	New Aquitaine	$[M+H]^+$	308.1220 308.1206	$C_{11}H_{20}N_2O_8$ $C_8H_{12}N_{12}O_2$	n.a.

*n.a.: not applicable; * invalid based on MS/MS data*

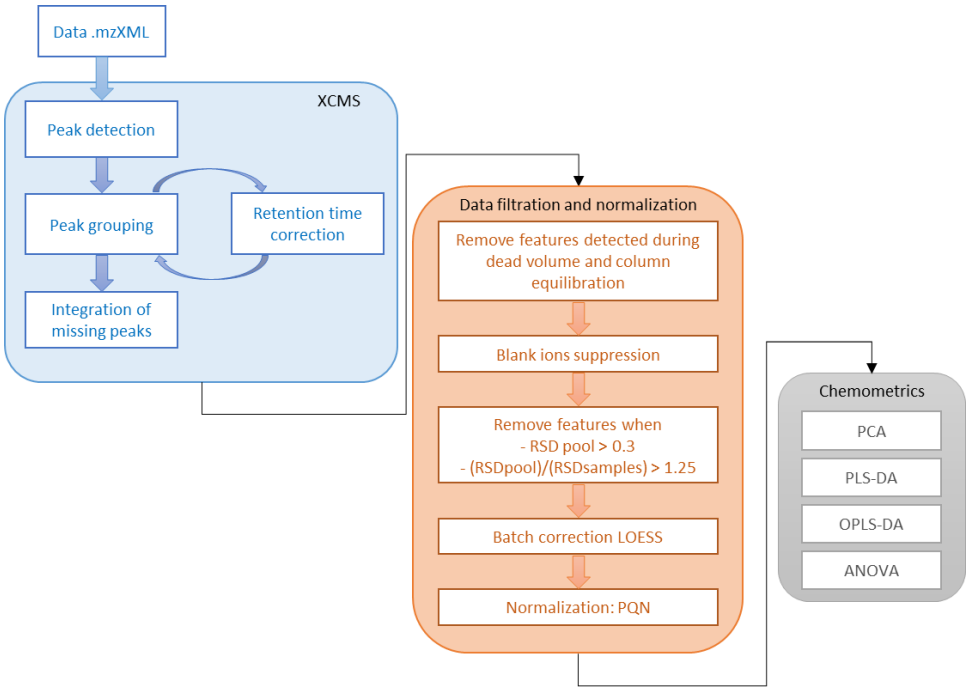

Figure 4.A.1. *Workflow of the data treatment (RSD: relative standard deviation)*

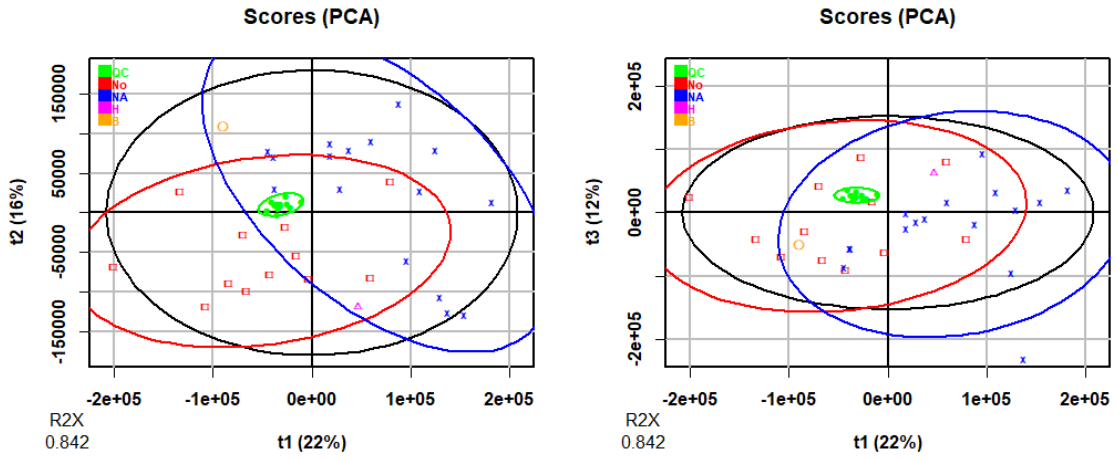

Figure 4.A.2. *PCA score plots of PC1 vs. PC2 and PC1 vs. PC3, obtained using the HILIC mode analysis for the first batch of samples (red squares: Normandy (No) samples, blue crosses: New Aquitaine (NA) samples, orange circles: Brittany (B) sample, magenta triangles: Hauts-de-France (H) sample, and green filled circles: quality control (QC) samples)*

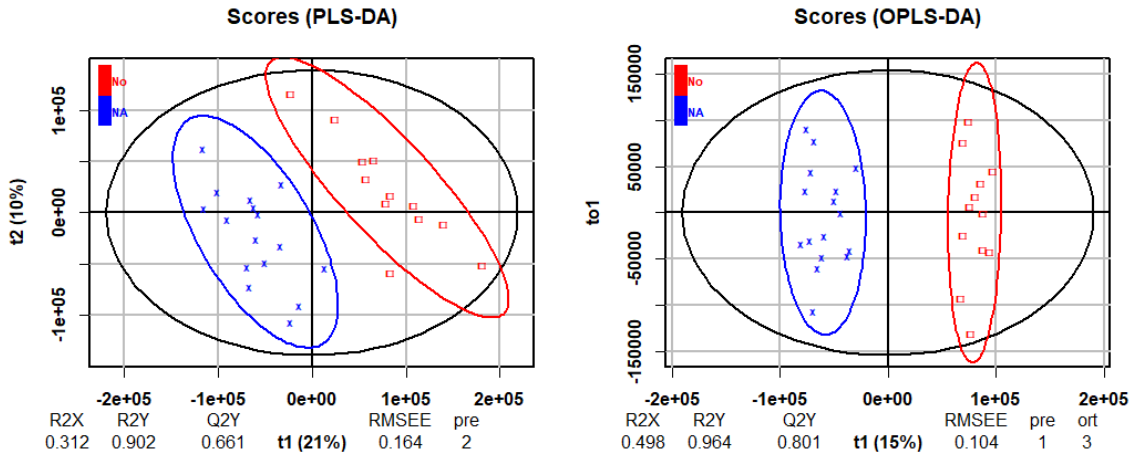

Figure 4.A.3. *PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the HILIC mode analysis for the geographical origin discrimination (blue: New Aquitaine (NA) samples, red: Normandy (No) samples)*

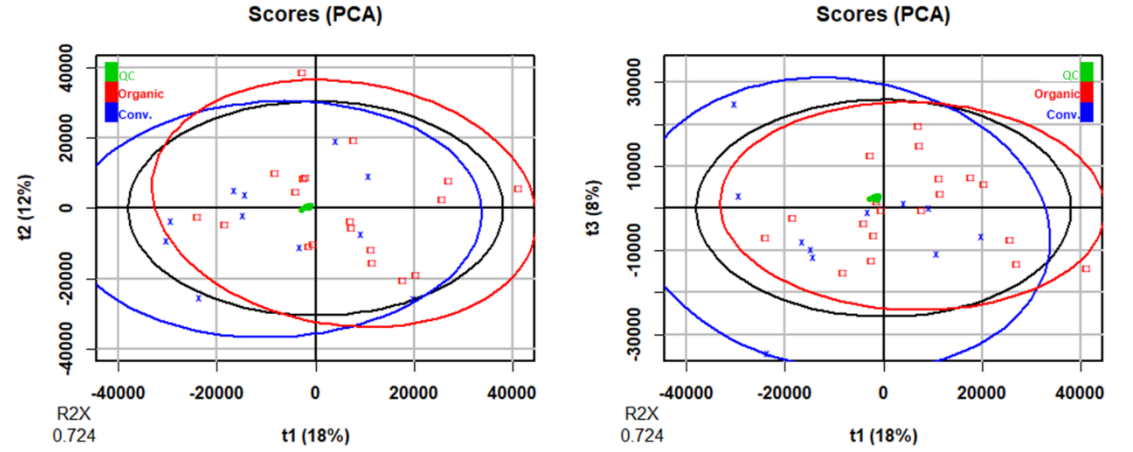

Figure 4.A.4. *PCA scores obtained using the reversed-phase mode analysis for tentative discrimination of samples based in their production mode (blue crosses: conventional samples, red squares: organic samples, green filled circles: QC samples)*

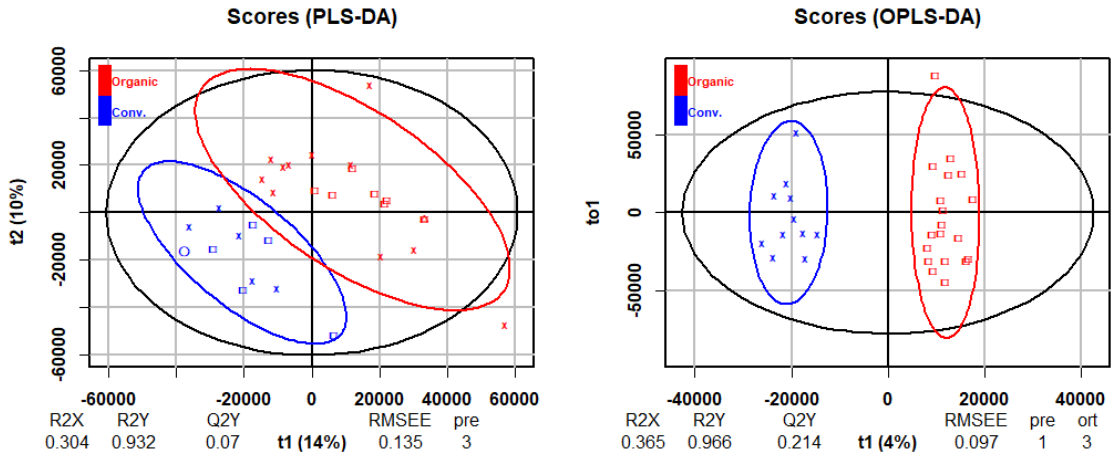

Figure 4.A.5. *PLS-DA and OPLS-DA score plots of LV1 vs. LV2 obtained using the HILIC mode analysis for the production mode discrimination (blue crosses: conventional samples, red squares: organic samples)*

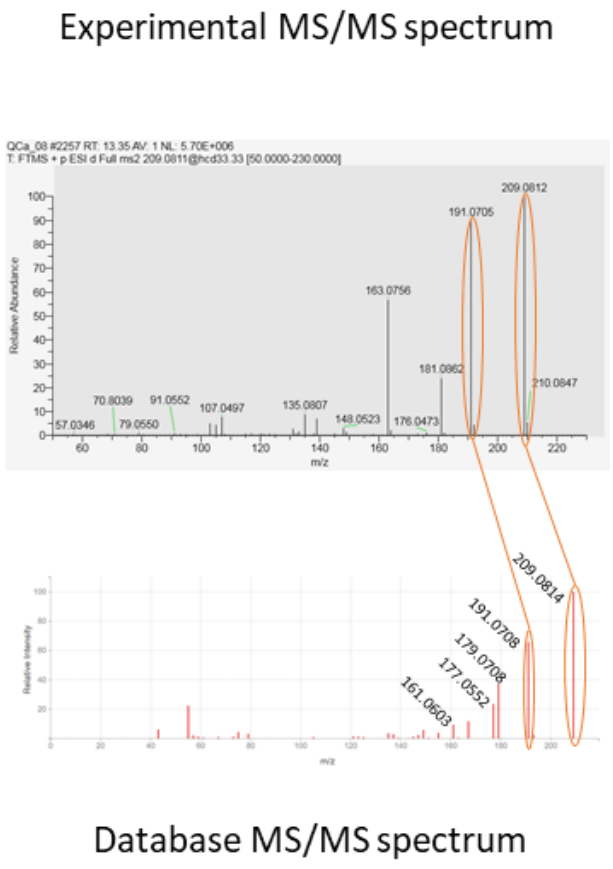

Figure 4.A.6. *Experimental MS/MS spectrum of feature 16 for authentication of the geographical origin using the HILIC mode analysis and the sinapaldehyde database MS/MS spectrum as suspected compound*

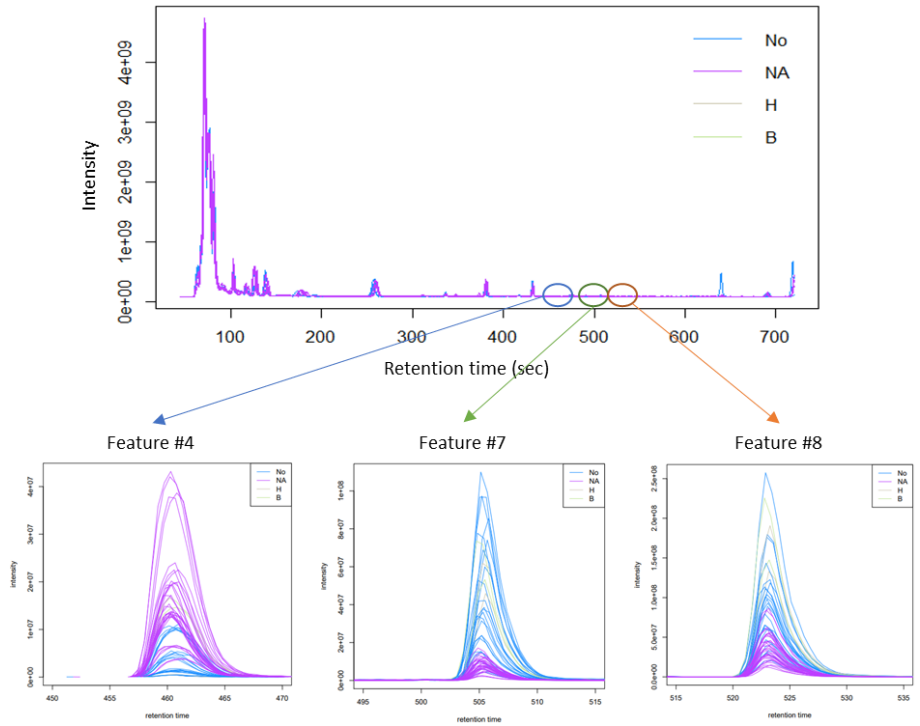

Figure 4.A.7. *Chromatogram of features 4, 7 and 8 for authentication of the geographical origin using the HILIC mode analysis*

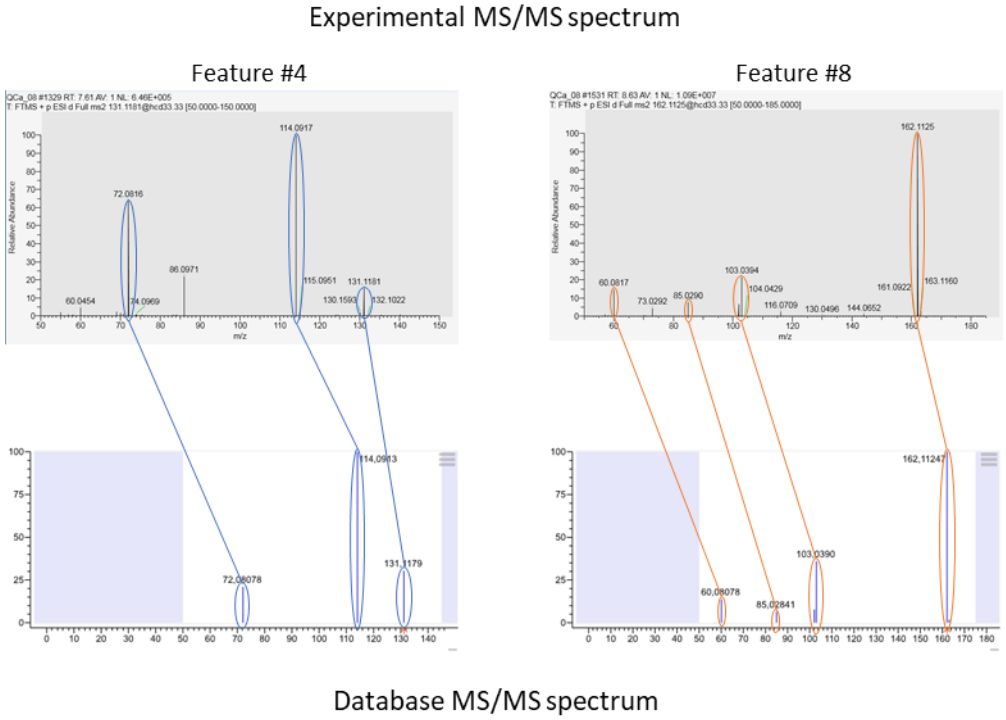

Figure 4.A.8. *Experimental MS/MS spectrum of features 4 and 8 for authentication of the geographical origin using the HILIC mode analysis, and their corresponding database MS/MS spectrum (N-acetylputrescine and L-carnitine, respectively)*

3. CONCLUSION

Dans cette étude, plusieurs échantillons de carottes provenant de différentes origines géographiques et issus de l'agriculture biologique ou conventionnelle ont été analysés par une méthode LC-HRMS non ciblée en utilisant deux colonnes différentes : une silice greffée C18 et une amide-HILIC.

Les résultats obtenus pour chacune de ces colonnes sont très similaires : les modèles PLS-DA et OPLS-DA donnent des performances satisfaisantes pour la discrimination de l'origine géographique (région Normandie *versus* Nouvelle Aquitaine) avec près de 80 % de capacité prédictive obtenue avec les données acquises sur chaque colonne chromatographique. En ce qui concerne la discrimination selon le mode de production (biologique *versus* conventionnel), les modèles sont nettement moins performants pour chaque colonne, avec moins de 50 % de capacité prédictive. Dans les deux cas les résultats obtenus par la fusion des données C18 et HILIC ont été similaires : les performances des modèles obtenus sont restées du même ordre de grandeur.

Les analyses MS/MS ont permis de confirmer l'identification de plusieurs features discriminants : l'arginine et la 6-méthoxymelleine (détectées sur silice greffée C18), ainsi que la N-acétylputrescine et la L-carnitine (détectées en HILIC). L'annotation de ces composés marqueurs reste toutefois à confirmer par l'analyse des standards de ces molécules.

Les performances des modèles générés dans cette étude ont été estimées par validation croisée. Afin de confirmer ces performances dans un contexte d'utilisation en routine de cette méthodologie, il est nécessaire de conforter ces modèles par une validation externe sur un jeu de données indépendants. Cela permettra ainsi de s'assurer de la fiabilité de ces modèles pour contrôler l'authenticité de nouveaux échantillons. Ceci n'a pas pu être réalisé dans le cadre de ce travail de thèse par manque de temps.

Afin d'implémenter ce type de méthodologie en analyse de routine, un traitement des données comprenant une étape de correction intersessions est nécessaire. En effet, il sera ainsi possible de prendre en compte un nombre plus important d'échantillons, et par suite

une plus grande variabilité d'échantillons. Ceci permettra d'asseoir la robustesse des features discriminants identifiés.

CHAPITRE 5 DISCUSSION GENERALE

Cette thèse a pour but le développement de nouvelles méthodologies pour le contrôle d'authenticité des aliments par LC-HRMS. La stratégie expérimentale adoptée peut se décomposer selon les trois axes suivants : (1) développement et implémentation en routine de méthodes LC-HRMS ciblées sur certains marqueurs connus d'authenticité ; (2) développement et validation d'une méthode d'analyse non ciblée (incluant un traitement poussé des données non ciblées acquises) ; (3) ajustement de la méthodologie non ciblée en vue de son automatisation pour une implémentation en routine (avec deux volets à considérer : automatisation du traitement des données brutes et gestion de plusieurs sessions d'analyse).

En effet, à terme l'objectif d'Eurofins Analytics France est de disposer d'une méthodologie LC-HRMS non ciblée pour le contrôle en routine de l'authenticité des aliments. Ce laboratoire effectue déjà en routine des analyses d'authenticité non ciblées par RMN. Ainsi, l'authentification des purs jus de pomme s'effectue par la méthode de routine JuiceScreener™ (Spraul et al., 2009) ; les échantillons étudiés sont comparés à une base de données en constante augmentation, répertoriant en 2009 plus de 6 000 échantillons, et donnant des résultats avec 95 % de précision.

1. DEVELOPPEMENT DE NOUVELLES METHODES CIBLEES POUR DES ANALYSES DE ROUTINE

Dans un premier temps, les travaux de cette thèse ont été focalisés sur le développement de méthodes d'analyse ciblées par LC-HRMS. Ils ont débouché sur l'implémentation en routine d'une méthode d'analyse permettant de contrôler l'authenticité du citron jaune *via* la quantification de différents composés de la famille des polyméthoxyflavones (PMF) et composés apparentés. Cette méthode a été développée en interne à Eurofins au cours d'un projet interlaboratoires coordonné par l'association internationale SGF. Les marqueurs de différentes espèces de citrus ont été confirmés, et les résultats d'analyse obtenus par les trois laboratoires participant à ce projet (chacun utilisant sa méthode développée en interne) ont permis la proposition de valeurs limites pour ces composés. Ces résultats constituent donc une avancée importante pour le contrôle d'authenticité du citron jaune. La

quantification des PMF permet de distinguer les différents procédés subis par le citron jaune (jus ou huile essentielle), et de discriminer les échantillons de citron jaune et ceux de citron vert.

Ce projet montre notamment l'intérêt de continuer à développer de nouvelles méthodes ciblées pour garantir l'authenticité des aliments, les fraudes étant de plus en plus sophistiquées. De plus, il est important de travailler avec différents laboratoires ou organismes afin de définir des valeurs seuils pour les marqueurs d'authenticité. Toutefois, à ce jour, peu de composés sont connus et validés comme marqueurs d'authenticité. Il est également important d'identifier et de valider de nouveaux composés marqueurs d'authenticité pour continuer de lutter contre les fraudes. Mais l'identification de ces potentiels marqueurs d'authenticité est difficile car la composition chimique d'un aliment est complexe et influencée par de nombreux facteurs (variété, origine géographique et/ou botanique, mode de production, procédés de transformation, présence d'adultération, etc.) comme présenté dans le paragraphe 1.4.1. du chapitre bibliographique. La **Figure 5.1** présente des exemples de marqueurs connus pour le contrôle d'authenticité des jus de fruits par des méthodes ciblées. Cette figure illustre notamment les limites de ces analyses ciblées puisque pour certains facteurs de variabilité aucun marqueur n'est à ce jour connu. La **Figure 5.1** présente également quelques études utilisant une méthodologie non ciblée par LC-HRMS pour évaluer l'authenticité, ce qui montre le potentiel de ces approches non ciblées pour contrôler l'authenticité des aliments.

L'étude de Medina et al. présente un recensement de marqueurs potentiels d'authenticité (Medina et al., 2019). Cette revue compile différents composés identifiés comme discriminants par le biais de méthodes statistiques dans des articles scientifiques parus entre 2014 et 2019 sur différents aspects d'authenticité (variété, origine géographique et/ou botanique, mode de production, adultération), et inclue également des indicateurs de qualité sanitaire (l'altération et la fraicheur). Comme indiqué par les auteurs, ces potentiels marqueurs d'authenticité nécessitent d'être validés (Medina et al., 2019).

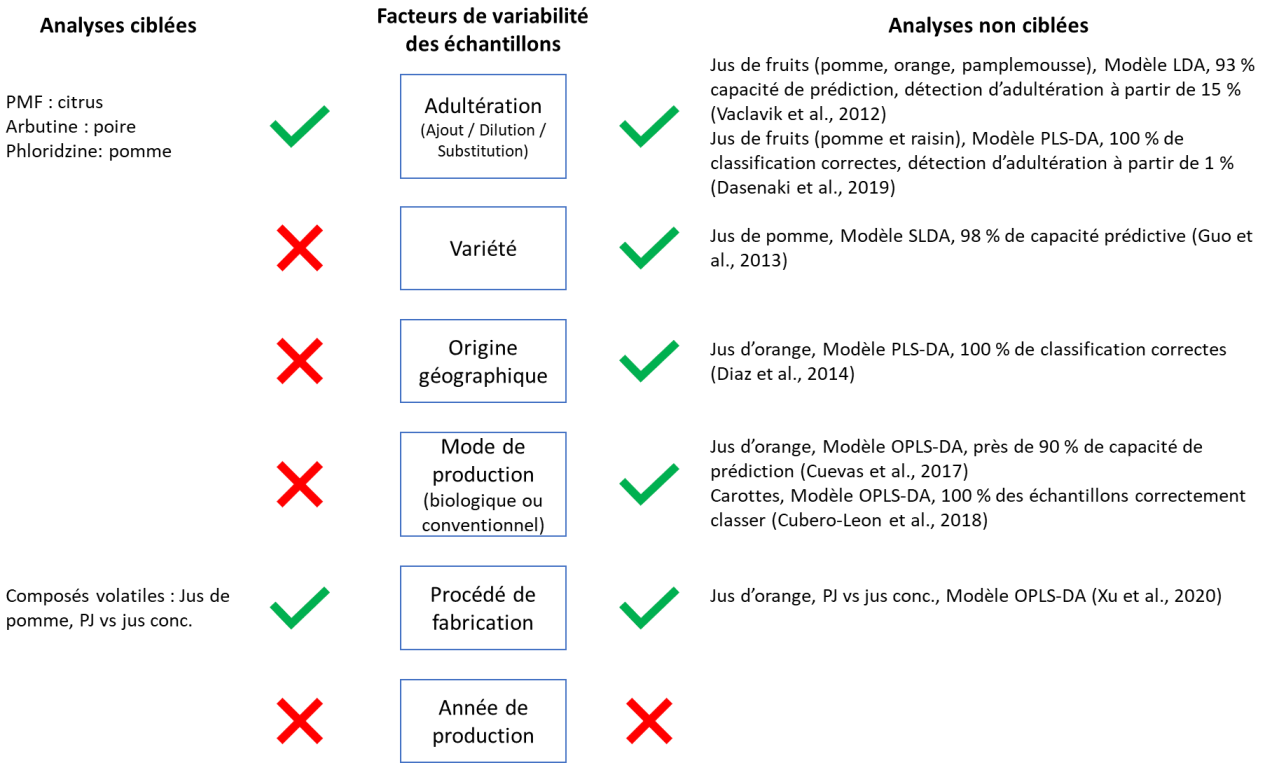

***Figure 5.1** : Présentation des marqueurs connus analysés par des méthodes ciblées et différentes études appliquant des méthodologies non ciblées par LC-HRMS pour le contrôle d'authenticité des jus de fruits et des carottes.*

La vérification de la robustesse des marqueurs d'authenticité est une étape importante à réaliser. Elle nécessite cependant d'être effectuée sur de nouveaux échantillons pour intégrer la variabilité liée à certains facteurs (par exemple avec d'autres variétés ou origines). En effet, cela permet ainsi de s'assurer que le marqueur est indépendant des facteurs influant sur la composition chimique des échantillons.

2. DEVELOPPEMENT DE METHODES NON CIBLEES

Dans un second temps, les travaux de cette thèse ont été focalisés sur le développement de méthodes d'analyse non ciblées par LC-HRMS, susceptibles de pallier aux difficultés d'identification de nouveaux marqueurs par les méthodes ciblées décrites précédemment, du fait de l'obtention d'une empreinte globale.

Une méthodologie s'inspirant des études métabolomiques a été mise en place : analyse non ciblée couplée à des outils chimiométriques. Ces premiers développements ont conduit à une étude préliminaire sur les jus de pommes. Cette étude a permis de confirmer l'intérêt de cette méthodologie pour le contrôle d'authenticité des jus de pommes. Des résultats

intéressants ont été obtenus sur deux scénarios : l'authentification des purs jus et l'authentification des jus issus de l'agriculture biologique. Des composés discriminants ont été identifiés pour la première fois par le biais d'acquisitions MS/MS. Cette étude préliminaire a également permis de prendre en main les différentes étapes du traitement des données à l'aide d'outils en ligne.

Dans les deux scénarios étudiés sur les jus de pommes, l'ACP s'est révélée insatisfaisante pour discriminer les deux groupes d'échantillons. Cette observation est en accord avec les résultats de différentes études sur les jus d'orange, aussi bien pour la discrimination entre jus concentré et pur jus que pour l'authentification des jus issus de l'agriculture biologique (Xu et al., 2020 ; Cuevas et al., 2017).

Les modèles OPLS-DA construits pour le scénario 1 (authentification des purs jus de pommes) ont des performances intéressantes avec environ 60 % de capacité prédictive. Ces performances ne sont qu'une estimation, les modèles étant construits par validation croisée. L'utilisation des résultats de l'ANOVA pour réduire le nombre de features en amont de la construction des modèles (O)PLS-DA a donné des résultats pertinents : l'ANOVA a identifié environ 2 000 features (sur près de 10 000 features détectés) et le modèle OPLS-DA construit à partir de ces features a des performances très similaires. Ces résultats montrent donc que des features non pertinents face à la problématique ont bien été supprimés. Cette méthodologie est proche de celle utilisée par Xu et al., où leurs modèles OPLS-DA pour discriminer les purs jus des jus concentrés d'orange ont été construits après sélection des features par un test de Student (Xu et al., 2020).

En ce qui concerne le scénario 2 (authentification des jus issus de l'agriculture biologique), les modèles OPLS-DA construits ont également montré de bonnes performances avec 75 % de capacité prédictive. Les modèles ont été encore plus performants après sélection des features par l'ANOVA, avec près de 80 % de capacité prédictive obtenue. D'autres auteurs s'intéressant à l'authentification des aliments issus de l'agriculture biologique ont également obtenu de bonnes performances de prédiction et de classification avec des modèles OPLS-DA. Ainsi, Cuevas et al. ont obtenu près de 90 % de sensibilité et spécificité sur des jus d'oranges issus de l'agriculture biologique. Dans leur étude sur les carottes, Cubero-Leon et al. ont construit un modèle OPLS-DA capable de classer correctement 100 % des échantillons testés dans leur étude.

Dans notre étude sur les jus de pommes, les modèles ont été construits par validation croisée. Les performances obtenues sont donc estimatives. Il faudra, pour poursuivre cette étude, prévoir des jeux externes de validation pour mieux asseoir les performances des modèles de prédiction et de classification OPLS-DA construits.

Dans les deux scénarios étudiés, l'utilisation de l'outil *biosigner* s'est révélée être insatisfaisante. En effet, bien que peu de features ont été identifiés comme discriminants (20 et 48 pour les scénarios 1 et 2 respectivement), les modèles construits sur la base de ces features ont montré une diminution des performances.

Comme détaillé dans le paragraphe 3.3 du chapitre bibliographique, la sélection des variables est une étape importante à considérer dans les approches non ciblées car elle permet de réduire le nombre de variables, et ainsi de se focaliser sur les variables les plus pertinentes. Les résultats d'analyses univariées comme l'ANOVA ou les valeurs de VIP obtenues par des modèles (O)PLS-DA sont les méthodes les plus répandues dans la littérature pour sélectionner les variables. Comme appliqué dans ce travail sur les jus de pommes, il semble être intéressant de combiner les résultats de l'ANOVA et de l'OPLS-DA pour la sélection de features discriminants.

Les features discriminants identifiés dans cette étude (identification de niveau 2) ont été annotés comme appartenant à la famille des acides aminés et dérivés. Ces résultats sont cohérents avec le fait que les acides aminés ont été répertoriés comme marqueurs d'authenticité dans différents articles (Xu et al., 2020 ; Cubero-Leon et al., 2018 ; Medina et al., 2019 ; Mihailova et al., 2021). Dans cette thèse, la méthionine et l'isoleucine (ou la norleucine) ont été trouvées comme marqueurs du mode de production (conventionnel *vs* biologique), et la N-(1-deoxy-1-fructosyl)phenylalanine comme marqueur du procédé de fabrication (jus concentré *vs* pur jus). Ces résultats restent à confirmer par l'analyse des standards correspondants (ce qui n'a pas été fait faute de temps).

Des recherches plus approfondies restent nécessaires pour pouvoir implémenter en routine ce type de méthodologie. Dans un premier temps, le traitement des données nécessite d'être internalisé. L'étape d'annotation des features discriminants s'avère être très longue : il serait donc judicieux d'implémenter une base de données interne de spectres MS/MS.

3. IMPLEMENTATION EN ROUTINE DES METHODES NON CIBLEES

Les différentes étapes du traitement des données non ciblées LC-HRMS étant mieux maitrisées, celles-ci ont ensuite été mises en place en interne sur le logiciel R à partir du « *workflow* » développé sur la plateforme W4M. Cette nouvelle version du traitement des données a été utilisée pour traiter des données issues du projet TOFoo. Le jeu de données contenait des échantillons de carottes provenant de plusieurs régions de France et issus de différents modes de production (biologique ou conventionnel). Le traitement des données a été effectué sur les résultats acquis sur deux colonnes chromatographiques de polarité très différente (silice greffée C18 apolaire et amide-HILIC polaire). Il a été montré dans cette étude que les deux colonnes conduisaient à des résultats très similaires en termes de discrimination des échantillons. Les modèles de prédiction et de classification PLS-DA et OPLS-DA se sont avérés performants pour la discrimination de l'origine géographique des échantillons. Plusieurs marqueurs discriminants ont pu être identifiés, et l'utilisation des deux types de colonnes chromatographiques s'est avérée complémentaire. Néanmoins, les performances de ces modèles pour la discrimination du mode de production ont été moins satisfaisantes, même après avoir éliminé des jeux de données les features marqueurs de l'origine géographique pour enlever la variabilité liée à ce facteur.

Lors de leur étude sur les carottes, Cubero-Leon et al. ont construit un modèle OPLS-DA permettant de classer correctement 100 % de leurs échantillons. Ce modèle a été construit après avoir retiré les variables liées à l'année de production, principal facteur influant sur les empreintes LC-HRMS (Cubero-Leon et al., 2018). Bien que dans notre étude une méthodologie similaire ait été appliquée (en supprimant les features liés à l'origine géographique), les modèles obtenus pour l'authentification des carottes issues de l'agriculture biologique avaient au mieux 65 % d'exactitude. Ce résultat un peu décevant pourrait s'expliquer par la présence de différentes variétés dans les échantillons que nous avons étudiés, alors que dans l'étude de Cubero-Leon et al. seules deux variétés étaient considérées (Cubero-Leon et al., 2018). Or la variété est connue pour être l'un des facteurs contribuant le plus à la variabilité des échantillons. Ceci pourrait donc justifier la moins bonne performance des modèles que nous avons obtenus pour la discrimination du mode de production.

Afin que les modèles de classification et de prédiction utilisés dans les analyses de contrôle d'authenticité soient les plus fiables possibles face à de nouveaux échantillons, il est essentiel que ces modèles soient construits (et validés) à partir d'échantillons représentatifs des différents facteurs influant sur la variabilité. En effet, ces facteurs impactant la composition chimique des échantillons, ils influent également sur l'empreinte obtenue (Donarski et al., 2019). Il faut donc s'assurer que la variabilité interne liée à la matrice étudiée n'impacte pas les résultats de classification ou de prédiction des modèles. De même, il est pertinent d'analyser des échantillons frauduleux pour s'assurer de la robustesse des modèles, ce type d'échantillons pouvant être analysé lors des contrôles d'authenticité.

Il est donc nécessaire d'analyser un grand nombre d'échantillons pour pouvoir modéliser un maximum de variabilité dans les échantillons. Pour ce faire, ceci implique de réaliser l'analyse des échantillons sur plusieurs sessions (en raison du temps requis pour réaliser un nombre important d'analyses), et par conséquent d'être en capacité de pouvoir traiter les données issues de différentes sessions d'analyse.

3.1. IMPLEMENTER DES ETAPES DE CORRECTION INTERSESSIONS

Comme introduit dans le paragraphe 3.3.2.3. du chapitre bibliographique, il est nécessaire de mettre en place des étapes de correction de l'effet intersession pour comparer des échantillons provenant de différentes sessions d'analyse. En effet, pour comparer des jeux de données issus de différentes sessions d'analyse, il est important de prendre en compte les déviations en *m/z* et en RT entre les sessions, la LC-HRMS souffrant de problèmes de répétabilité et reproductibilité. Différentes méthodologies pour corriger cet effet intersession ont été proposées dans la littérature : l'utilisation d'une table de référence (Dunn et al., 2011), ou en se basant sur les features détectés dans les échantillons (Wehrens et al., 2016).

Dans notre étude sur les jus de pommes, trois sessions d'analyse étaient considérées. Afin de s'affranchir des effets intersessions, ces trois sessions ont été prétraitées simultanément avec les fonctions de XCMS. De par cette méthodologie, il a été aisé de comparer les échantillons des différentes sessions d'analyse. Néanmoins, celle-ci n'est pas adaptée à

une implémentation en routine. Il faut donc se tourner vers d'autres méthodologies pour mettre en place les corrections intersessions.

L'approche proposée par Dunn et al. a montré son potentiel sur leurs données LC-HRMS (Dunn et al., 2011). Néanmoins, celle-ci ne semble pas correspondre aux problématiques d'authenticité. En effet, la limite de l'utilisation d'une table de référence est que seuls les features présents dans cette table seront comparés entre les différentes sessions d'analyse. Il est difficile d'avoir une telle table de référence dans le cadre des analyses d'authenticité, du fait de la complexité des échantillons et de la diversité des marqueurs.

L'approche se basant sur les échantillons pour corriger les effets intersessions, proposée par Wehrens et al., a également montré son potentiel sur des données LC-HRMS et GC-MS (Wehrens et al., 2016). Cette méthodologie semble plus adaptée aux problématiques de contrôle d'authenticité puisqu'elle permet de prendre en compte tous les features détectés, et vise même en particulier les features peu intenses qui sont souvent mal corrigés avec les méthodologies utilisant uniquement les QC (Wehrens et al., 2016).

Avant de mettre en place les corrections des effets intersessions, il faut fixer des seuils de tolérance pour les valeurs de *m/z* et de RT qui seront comparées entre les différents jeux de données. Ces seuils ne doivent être ni trop petits (car ils conduiraient à ne pas grouper ensemble des features identiques) ni trop grands (car des features différents pourraient alors être groupés ensemble). Dunn et al. proposent dans leur méthodologie d'appliquer une tolérance de 10 secondes en RT et de 5 ppm en *m/z* (Dunn et al., 2011) ; ces seuils de tolérance semblent adaptés à la LC-HRMS.

Au cours de ces travaux de thèse, des premiers essais ont été réalisés pour tester ces valeurs de seuils. Deux jeux de données (correspondant à un jeu d'échantillons analysé dans deux sessions différentes) contenant chacun près de 20 000 features (18 074 et 16 816 respectivement) ont ainsi été comparés en utilisant ces tolérances. La comparaison des deux jeux de données avec ce seuil de tolérance permet d'obtenir une correspondance pour un très grand nombre de features (plus de 5 400). Il a été constaté que pour quelques features (deux pour chaque jeu de données), l'application du seuil de tolérance à ces features d'un jeu de données donnait deux features possibles dans l'autre jeu de données. Ces résultats sont encourageants puisque le nombre de features problématiques est très

faible vis à vis du nombre de features concordants. En outre, dans ces essais, les jeux de données utilisés n'avaient pas subi les étapes de filtration des données et de correction intra-session. Il est donc vraisemblable de supposer qu'aucun feature problématique n'aurait été observé si les corrections intra-sessions avaient été effectuées. De plus, à la vue du faible nombre de features problématiques, il est possible de simplement supprimer ces features. Ces premiers essais ont conduit à la mise en place sous R d'un script pour la comparaison de deux jeux de données.

Une fois les seuils de tolérance en *m/z* et en RT fixés, les différents jeux de données pourront être combinés pour ne donner plus qu'un seul jeu de données contenant les informations des différentes sessions. Afin de rendre les intensités de chaque feature comparables entre tous les échantillons, il faudra ensuite lancer une correction du signal avec la méthode LOESS comme ce qui est déjà fait dans les étapes de corrections intra-session. La normalisation des intensités devra ensuite être effectuée par la méthode PQN (cette étape ne sera donc pas effectuée lors des corrections intra-session). Les différences entre les traitements de données en intra- et en intersession sont illustrées en **Figure 5.2**.

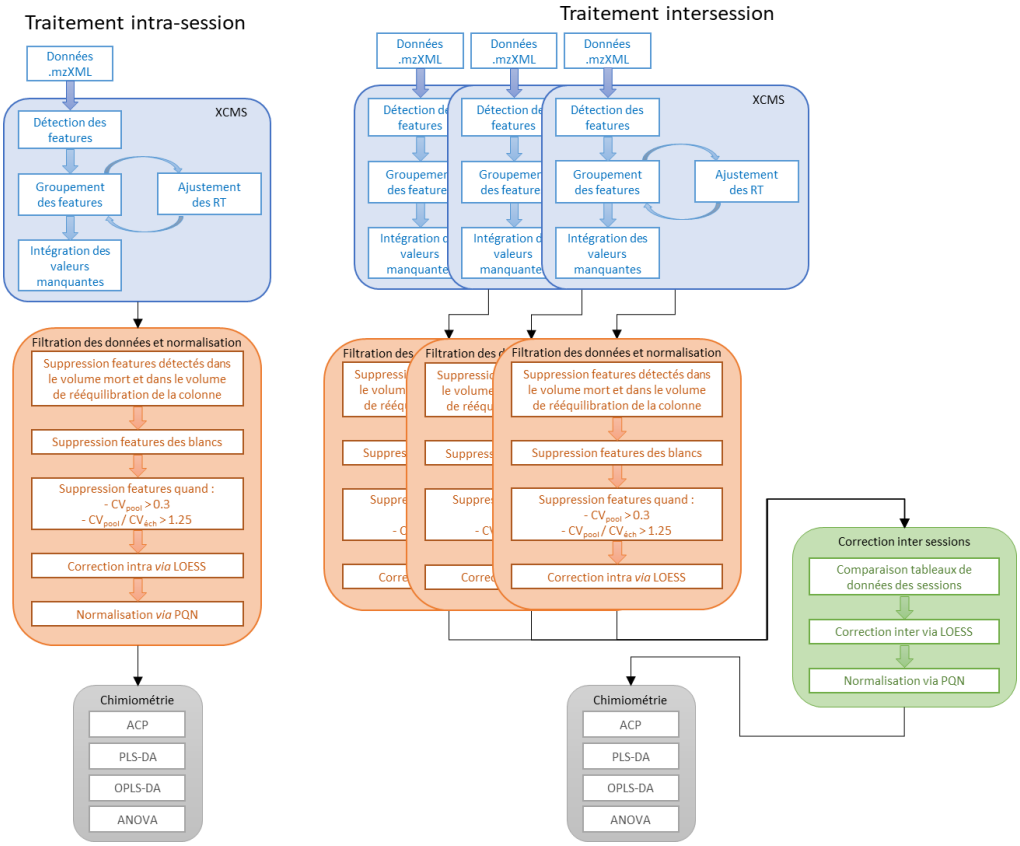

Figure 5.2 *: Schéma des étapes du traitement des données en intra- et en intersession.*

Une fois la mise en place des corrections intersession, il sera plus aisé d'utiliser un jeu externe de validation pour valider les performances des modèles de classification et de prédiction PLS-DA et OPLS-DA.

3.2. AMELIORER L'ETAPE DE CARACTERISATION DES FEATURES

Pour poursuivre ce travail, il pourrait également être intéressant d'améliorer l'étape d'annotation des features. Cette étape est longue car elle nécessite d'une part d'acquérir les spectres MS/MS des features identifiés comme discriminants, et d'autre part de comparer les spectres MS/MS expérimentaux à des spectres présents dans des bases de données spectrales. En pratique, des améliorations pourraient être apportées sur chacun de ces deux volets dans la méthodologie développée.

Dans les travaux présentés dans cette thèse, l'acquisition des spectres MS/MS a été réalisée dans une session d'analyse dédiée et différée dans le temps, dès lors qu'une liste restreinte de features significativement discriminants était établie par les traitements chimiométriques suite à l'analyse des échantillons par LC-HRMS. Il serait donc judicieux de modifier la méthodologie afin de lancer, dès la séquence d'analyse des échantillons, des acquisitions en MS/MS. Une méthode a été décrite pour les instruments de la marque ThermoFisher, permettant d'acquérir des spectres MS/MS pour près de70 % de features en plus qu'avec une simple expérience MS/MS (Kloelmel et al., 2017). Cette méthode consiste à lancer des acquisitions en mode « *full scan data dependent analysis* » (DDA) durant lesquelles, à chaque scan, les 5 ions les plus intenses sont fragmentés en vue d'obtenir leurs spectres MS/MS. Ces acquisitions sont ensuite lancées en mode itératif, avec à chaque itération une liste d'exclusion créée et mise à jour pour permettre la recherche d'ions de moins en moins intenses et l'acquisition de leurs spectres MS/MS. Cette méthode peut s'appliquer en alternant dans la même séquence d'analyse les deux modes d'ionisation (positif et négatif) afin de couvrir un maximum de features, comme illustré sur la **Figure 5.3**. Selon le logiciel utilisé, la création et la mise à jour de la liste d'exclusion peut se faire de façon automatique. Il est également possible de l'effectuer en utilisant R et le package IE-OMICS (Kloelmel et al., 2017).

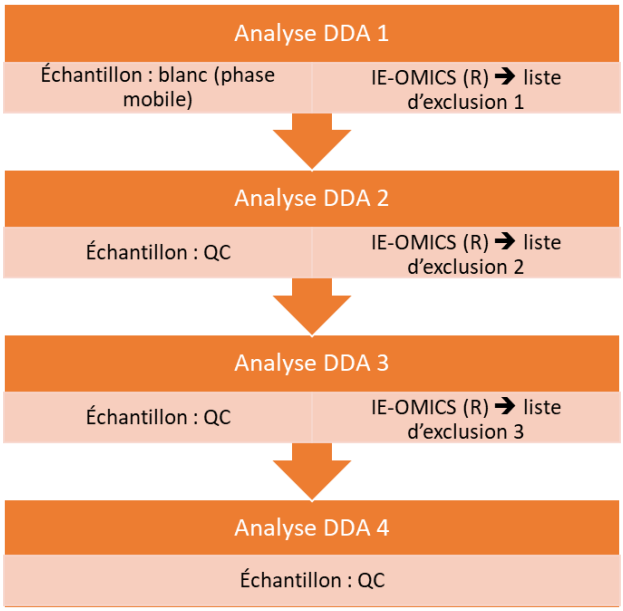

***Figure 5.3** : Illustration de la méthodologie DDA itératif.*

Les différentes acquisitions en DDA ne doivent pas nécessairement se suivre, surtout si la création et la mise à jour de la liste d'exclusion doivent être effectuées en utilisant R. Elles peuvent donc être lancées à n'importe quel moment dans la séquence d'analyse. Comme illustré sur la **Figure 5.3**, cette méthodologie débute par l'analyse d'un blanc pour identifier et noter dans la liste d'exclusion les features liés à la phase mobile. Les acquisitions suivantes sont réalisées sur les échantillons QC qui sont un mélange des échantillons de la session d'analyse. Cette méthode nécessite d'ajouter entre 4 et 6 analyses à la séquence d'acquisition (Kloelmel et al., 2017), ce qui est faible par rapport aux nombres d'analyses que peut contenir une séquence d'analyses non ciblées (environ une centaine).

Cette méthodologie permet alors un gain de temps puisqu'il n'est plus nécessaire de lancer de nouvelles acquisitions de données pour obtenir les spectres MS/MS. De plus, cela permet d'avoir les spectres MS/MS d'une grande partie des features détectés, et non uniquement ceux des features discriminants. Il est ainsi possible de se constituer une base de données spectrales en interne, ce qui pourrait permettre de faciliter ensuite les comparaisons des spectres MS/MS. En effet la comparaison de spectres MS/MS avec les bases spectrales en ligne achoppe souvent sur des conditions d'analyse ou instrumentales différentes – dans le cas présent les spectres MS/MS de la base de données spectrales pourraient être acquis dans les mêmes conditions opératoires que l'analyse des échantillons. Ceci faciliterait ensuite le travail d'annotation des features discriminants.

4. REFERENCES

Cubero-Leon, E., De Rudder, O., Maquet, A., 2018. Metabolomics for organic food authentication: Results from a long-term field study in carrots. *Food Chemistry 239*, 760–770. https://doi.org/10.1016/j.foodchem.2017.06.161

Cuevas, F.J., Pereira-Caro, G., Moreno-Rojas, J.M., Muñoz-Redondo, J.M., Ruiz-Moreno, M.J., 2017. Assessment of premium organic orange juices authenticity using HPLC-HR-MS and HS-SPME-GC-MS combining data fusion and chemometrics. *Food Control 82*, 203–211. https://doi.org/10.1016/j.foodcont.2017.06.031

Dasenaki, M.E., Drakopoulou, S.K., Aalizadeh, R., Thomaidis, N.S., 2019. Targeted and Untargeted Metabolomics as an Enhanced Tool for the Detection of Pomegranate Juice Adulteration. *Foods 8*, 212. https://doi.org/10.3390/foods8060212

Diaz, R., Pozo, O.J., Sancho, J.V., Hernández, F., 2014. Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry 157*, 84–93. https://doi.org/10.1016/j.foodchem.2014.02.009

Donarski, J., Camin, F., Fauhl-Hassek, C., Posey, R., Sudnik, M., 2019. Sampling guidelines for building and curating food authenticity databases. *Trends in Food Science & Technology 90*, 187–193. https://doi.org/10.1016/j.tifs.2019.02.019

Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N., Nicholls, A.W., Wilson, I.D., Kell, D.B., Goodacre, R., 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols 6*, 1060–1083. https://doi.org/10.1038/nprot.2011.335

Gattuso, G., Barreca, D., Gargiulli, C., Leuzzi, U., Caristi, C., 2007. Flavonoid composition of citrus juices. *Molecules 12*, 1641–1673. https://doi.org/10.3390/12081641

Guo, J., Yue, T., Yuan, Y., Wang, Y., 2013. Chemometric Classification of Apple Juices According to Variety and Geographical Origin Based on Polyphenolic Profiles. J. Agric. Food Chem. 61, 6949–6963. https://doi.org/10.1021/jf4011774

Koelmel, J. P., Kroeger, N. M., Gill, E. L., Ulmer, C. Z., Bowden, J. A., Patterson, R. E., Yost, R. A., & Garrett, T. J. (2017). Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation. Journal of the American Society for Mass Spectrometry, 28(5), 908–917. https://doi.org/10.1007/s13361-017-1608-0

Lehnert N., & Ara V. 2014. Authenticity analysis of lemon juices concerning the adulteration lime. *Fruit Processing*, 242–248.

Lehnert N., Schmidt M., & Ara V. 2017. Authenticity proof of lemon juices by means of fingerprint methods. *Fruit Processing*, 314–318.

Mihailova, A., Kelly, S.D., Chevallier, O.P., Elliott, C.T., Maestroni, B.M., Cannavan, A., 2021. High-resolution mass spectrometry-based metabolomics for the discrimination between organic and conventional crops: A review. *Trends in Food Science & Technology 110*, 142–154. https://doi.org/10.1016/j.tifs.2021.01.071

Medina, S., Pereira, J.A., Silva, P., Perestrelo, R., Câmara, J.S., 2019. Food fingerprints – A valuable tool to monitor food authenticity and safety. *Food Chemistry 278*, 144–162. https://doi.org/10.1016/j.foodchem.2018.11.046

Spraul, M., Schütz, B., Rinke, P., Koswig, S., Humpfer, E., Schäfer, H., Mörtter, M., Fang, F., Marx, U., Minoja, A., 2009. NMR-Based Multi Parametric Quality Control of Fruit Juices: SGF Profiling. *Nutrients 1*, 148–155. https://doi.org/10.3390/nu1020148

Vaclavik, L., Schreiber, A., Lacina, O., Cajka, T., Hajslova, J., 2012. Liquid chromatography–mass spectrometry-based metabolomics for authenticity assessment of fruit juices. *Metabolomics 8*, 793–803. https://doi.org/10.1007/s11306-011-0371-7

Wehrens, R., Hageman, Jos.A., van Eeuwijk, F., Kooke, R., Flood, P.J., Wijnker, E., Keurentjes, J.J.B., Lommen, A., van Eekelen, H.D.L.M., Hall, R.D., Mumm, R., de Vos, R.C.H., 2016. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics 12*. https://doi.org/10.1007/s11306-016-1015-8

Xu, L., Xu, Z., Kelly, S., Liao, X., 2020. Integrating untargeted metabolomics and targeted analysis for not from concentrate and from concentrate orange juices discrimination and authentication. *Food Chemistry 329*, 127130. https://doi.org/10.1016/j.foodchem.2020.127130

CONCLUSION GENERALE ET PERSPECTIVES

Le but de ces travaux de thèse était de développer une méthodologie d'analyse non ciblée par LC-HRMS combinée à des outils chimiométriques pour le contrôle d'authenticité des jus de fruits. Cette méthodologie a également pour vocation d'être implémentée en routine au sein du laboratoire d'Eurofins Analytics France. Bien que différentes publications présentaient l'intérêt de cette méthodologie pour répondre à la problématique du contrôle d'authenticité et montraient des résultats prometteurs, ces études se focalisaient souvent sur un seul scénario d'authenticité. De plus, les outils disponibles pour traiter ces données non ciblées ne permettent pas à ce jour d'avoir une telle méthodologie implémentée en routine dans un laboratoire de contrôle.

Les travaux de cette thèse ont ainsi été axés sur le développement méthodologique d'une approche non ciblée par LC-HRMS et sur l'automatisation des différentes étapes du traitement des données afin de rendre cette méthodologie implémentable en routine au sein du laboratoire d'Eurofins Analytics France. A terme, cette approche non ciblée sera utilisée en complément d'approches ciblées, qui restent incontournables pour des analyses de routine et qui nécessitent encore quelques développements dans le champ de l'analyse d'authenticité des aliments.

Dans un premier temps, une méthode d'analyse ciblée a été développée pour l'authentification du jus de citron jaune *via* la quantification des polyméthoxyflavones et de composés apparentés. Ce développement s'est inscrit dans le cadre d'un projet interlaboratoires visant d'une part à valider certains composés comme marqueurs d'authenticité, et d'autre part à définir des valeurs seuils pour les composés étudiés pour garantir l'authenticité des échantillons. La méthode développée a été implémentée en routine dans le laboratoire d'Eurofins Analytics France. Cette méthode permet notamment de détecter l'ajout d'autres espèces de citrus (en particulier le citron vert), l'une des principales fraudes du jus de citron jaune.

Dans un second temps, une méthodologie d'analyse non ciblée par LC-HRMS a été développée pour caractériser l'authenticité du jus de pommes. La séparation des composés s'est effectuée sur une colonne de type silice greffée C18, en accord avec la majorité des

articles traitant de méthodologie non ciblée pour l'authenticité. La préparation des échantillons est restée minime avec une approche « *dilute and shoot* » particulièrement adaptée aux jus de fruits. Cette méthodologie a été testée sur deux scénarios : l'authenticité des purs jus, et l'authenticité des jus issus de l'agriculture biologique. Le traitement des données a été effectué à l'aide d'outils en ligne, sur la plateforme *Workflow4Metabolomic*. Les modèles de prédiction et de classification des échantillons obtenus par (O)PLS-DA ont montré des performances intéressantes, en particulier pour authentifier des jus issus de l'agriculture biologique, avec près de 80 % de capacité prédictive.

Pour pouvoir implémenter cette méthodologie en routine, le traitement des données a été mis en place en interne. Ceci a nécessité une adaptation de la méthodologie développée précédemment. Celle-ci a ensuite été appliquée à des échantillons de carottes et testée sur deux scénarios : l'authentification de l'origine géographique, et la discrimination du mode de production (biologique ou conventionnel). Dans ce cas, deux colonnes chromatographiques ont été comparées : une silice greffée C18 et une amide-HILIC. Les résultats obtenus sur chaque colonne ont été très similaires. Les modèles construits pour authentifier l'origine géographique sont là aussi intéressants, avec à nouveau près de 80 % de capacité prédictive. Les modèles obtenus pour la discrimination du mode de production ont conduit à des résultats moins satisfaisants.

Dans les deux types d'étude (jus de fruits, carottes), des analyses MS/MS ont été réalisées sur de nouveaux échantillons afin d'identifier les composés marqueurs d'authenticité responsables de la discrimination observée entre les échantillons. Pour les jus de pommes, la méthionine et la N-(1-deoxy-1-fructosyl)phenylalanine ont été identifiées. Pour les carottes, l'utilisation combinée de deux types de colonne chromatographique a permis d'identifier quatre marqueurs : l'arginine et la 6-méthoxymelleine (sur silice greffée C18), ainsi que la N-acétylputrescine et la L-carnitine (en HILIC). L'annotation de ces différents composés marqueurs reste toutefois à confirmer par l'analyse des standards de ces molécules.

Les résultats obtenus au cours de ces travaux de thèse confirment l'intérêt de l'utilisation d'une approche d'analyse non ciblée par LC-HRMS combinée à des outils chimiométriques pour contrôler l'authenticité des aliments. Les premiers modèles de classification et de prédiction construits montrent des performances satisfaisantes, bien que celles-ci ne soient

qu’estimatives. La méthodologie mise en place dans ces travaux ne permet pas encore une implémentation en routine dans le laboratoire d’Eurofins Analytics France. Le traitement des données doit encore être automatisé, en particulier pour intégrer des corrections intersessions. L’acquisition de spectres MS/MS et l’annotation des features discriminants sont également des étapes qui nécessitent d’être optimisées.

La poursuite de ces travaux de thèse peut s’orienter autour de la mise en place de nouveaux modèles de classification et de prédiction dans le but d’être en capacité de caractériser au maximum l’authenticité d’un échantillon. Au cours des travaux réalisés dans le cadre de cette thèse, deux scénarios d’authenticité ont été traités, mais cela reste faible comparé au nombre de facteurs influant sur la variabilité des échantillons.

Ces travaux peuvent également se poursuivre vers des méthodologies combinant les résultats d’analyses obtenus par LC-HRMS et par RMN non ciblée *via* des approches de fusion des données. En effet, ces deux techniques d’analyse étant complémentaires, cela pourrait permettre d’accroitre les performances des modèles pour répondre aux problématiques d’authenticité. Ce type de méthodologie a déjà été appliqué pour fusionner des données LC-HRMS et GC-MS et a montré de meilleurs résultats qu’en appliquant les techniques seules.