UNIVERSITAT DE BARCELONA

Estratègies cromatogràfiques i quimiomètriques per a estudis de metabolòmica no dirigida en arròs (*Oryza sativa* L.)

Meritxell Navarro Reig

UNIVERSITAT DE BARCELONA

Estratègies cromatogràfiques i quimiomètriques per a estudis de metabolòmica no dirigida en arròs (*Oryza sativa* L.)

Meritxell Navarro Reig

ESTRATÈGIES CROMATOGRÀFIQUES I QUIMIOMÈTRIQUES PER A ESTUDIS DE METABOLÒMICA NO DIRIGIDA EN ARRÒS (*ORYZA SATIVA* L.)

Meritxell Navarro Reig

FACULTAT DE QUÍMICA

DEPARTAMENT D'ENGINYERIA QUÍMICA I QUÍMICA ANALÍTICA

Programa de Doctorat:
Química Analítica del Medi Ambient

Estratègies cromatogràfiques i quimiomètriques per a estudis de metabolòmica no dirigida en arròs (*Oryza sativa* L.)

Memòria presentada per

Meritxell Navarro Reig

per optar al grau de Doctora per la Universitat de Barcelona

Directors

Dr. Romà Tauler Ferré

Dr. Joaquim Jaumot Soler

Departament de Química Ambiental Institut de Diagnòstic Ambiental i Estudis de l'Aigua (IDAEA).
Consejo Superior de Investigaciones Científicas (CSIC).

Tutor

Dr. Santiago Hernández Cassou

Departament d'Enginyeria Química i Química Analítica.
Universitat de Barcelona (UB).

El **Dr. Romà Tauler Ferré**, professor d'investigació del Departament de Química Ambiental de l'Institut de Diagnòstic Ambiental i Estudis de l'Aigua del CSIC, i el **Dr. Joaquim Jaumot Soler**, científic titular del mateix Departament,

FAN CONSTAR,

Que la present memòria titulada: "**Estratègies cromatogràfiques i quimiomètriques per a estudis de metabolòmica no dirigida en arròs (*Oryza sativa* L.).**" ha estat realitzada sota la seva direcció per la **Sra. Meritxell Navarro Reig** i que tots els resultats presentats són fruit de la feina realitzada per la doctoranda.

I per tal de que així consti expedeixen el present certificat.

Barcelona, Abril 2018

Dr. Romà Tauler Ferré **Dr. Joaquim Jaumot Soler**

Als meus pares i a l'Alejandro,

"A la vida no hi ha res a témer, només coses a aprendre"
Marie Curie

Agraïments

Arribat aquest moment de la Tesi, toca recordar aquests últims quatre anys i mig i pensar en totes les persones que heu fet possible que aquesta memòria tingui punt i final.

En primer lloc voldria agrair als meus directors, el Dr. Romà Tauler i el Dr. Joaquim Jaumot, per confiar en mi quan vau començar aquest projecte i donar-me la oportunitat de fer la Tesi. Gràcies per guiar-me i aconsellar-me, pels coneixements compartits i pel vostre suport. De la mateixa manera també vull agrair al Dr. Santiago Hernández per haver acceptat ser el meu tutor durant la Tesi.

Gràcies a tots els companys del grup del CSIC passats i presents (Adriana, Alejandro, Aline, Amrita, Carma, Cristian, Guille, Igor, Marc, Marta, Mireia, Miriam, Roger, Stefan, Xin i Yahya), per tots els bons moments compartits al laboratori i fora d'ell: cafès, pastissos, dinars, sortides, congressos, etc. També per ajudar-me sempre que heu pogut contestant les meves preguntes i compartint els vostres coneixements. Vull agrair especialment als altres sis doctorands que vau començar aquesta aventura amb mi: Elba, Elena, Eva, Francesc, Núria i Victor. Vaig començar la Tesi pensant que havia tingut molta sort de trobar uns companys de feina tant bons, i ara l'acabo pensant que tinc molta sort d'emportar-me sis amics dels de veritat. Gràcies a vosaltres aquest camí ha estat més fàcil i més bonic. Sempre recordaré els riures compartits, els descansos de la tarda, els cafès, les cerveses, els sopars, els viatges, les confessions, les teràpies i tot el que hem viscut junts durant aquests quatre anys i mig. No puc acabar aquest paràgraf sense fer una menció especial a l'Elena, simplement gràcies per ser-hi sempre i per tot. Recorda que vals molt nena, no tinc cap dubte de que sempre aconseguiràs el que et proposis.

I would also like to acknowledge Prof. Peter Schoenmakers and Dr. Gabriel Vivió-Truyols for the opportunity to be part of Analytical Chemistry group during my PhD research internship. Thank you for your guidance, support and recommendations and for your kind hospitality.

Pensant en com vaig arribar a començar una Tesi, vull agrair a tots els amics del màster i de la carrera. Anna, Albert, Elena (un altre cop), Gerard, Javi, Laura i Stanis gràcies per fer que l'any del màster fos genial, vosaltres vau ajudar a que tingués ganes de continuar en aquest món. A les meves nenes, Llucia, Elena (ara ja si que és l'últim), Helena, Laura, Núria i Pili, gràcies per seguir juntes encara que passin els anys.

També vull agrair a tots els de fora del món de la Química, però que heu ajudat (i molt) a que jo tiri endavant aquesta Tesi. Als amics de sempre, Abraham, Aida, Cristina, Judith, Maribel i Raquel, per ser la meva segona família. Sé que vosaltres sempre heu cregut en mi i aquesta injecció de confiança m'ha

ajudat a seguir endavant. Gràcies per tots els bons moments, riures i consells. Tot i que ara ens veiem molt menys del que m'agradaria, sé que sou per sempre.

Vull donar les gràcies també a la persona que ha patit aquesta Tesi tant com jo. Ha patit els meus nervis, els meus mal humors, les meves queixes i els meus dubtes. Però també ha celebrat amb mi els bons resultats, els articles acceptats i els objectius assolits. Molts cops hem parlat de si ens penedirem o no de les decisions que hem pres junts, vull dir-te que estic segura que mai em penediré d'haver anat a utilitzar la liofilitzadora del soterrani. Alejandro, ets la millor persona que conec i les paraules se'm queden curtes per agrair-t'ho tot. Gràcies per haver-me convertit en la teva prioritat, sé que sempre has pensat en mi abans que en tu. Gràcies per fer-me treure el millor de mi mateixa en tot moment. Gràcies per confiar en mi i convidar-me a viure la vida junts. Gràcies per fer-me sentir que al teu costat el millor està per arribar. Carmen, Dani i Ana, quan vaig començar aquesta aventura no ens coneixíem i ara sou part de la meva família, gràcies també a vosaltres pel suport.

Per últim, a la meva família i a la colla, gràcies pel vostre suport incondicional, per l'educació que m'heu donat i per fer-me créixer sentint-me la persona més estimada del món. En especial vull agrair als meus pares, gràcies per ajudar-me a sé la persona que sóc ara i per haver-m'ho donat tot. Gràcies també per la vostra mirada amb orgull, espero de tot cor ser digne d'ella. Sou uns pares genials i us admiro per tot el que heu aconseguit junts.

Resum

Les mostres analitzades en els estudis de metabolòmica ambiental presenten una elevada complexitat per la gran quantitat d'informació bioquímica que contenen, especialment, en el cas d'estudis no dirigits. En aquests estudis és necessari l'ús de tècniques analítiques d'alt rendiment com, per exemple, la cromatografia de líquids acoblada a espectrometria de masses (LC-MS), les quals permeten analitzar els canvis en la concentració dels metabòlits en diferents tipus de mostres. D'aquesta forma es generen grans conjunts de dades multivariants (amb desenes de mostres i milers de variables), els quals requereixen l'aplicació d'eines quimiomètriques d'anàlisi de dades per tal d'extreure la informació d'interès. En aquesta Tesi s'han desenvolupat i aplicat nous mètodes d'anàlisi basats en tècniques de cromatografia de líquids uni- i bidimensional acoblada a espectrometria de masses, conjuntament amb noves estratègies d'anàlisi multivariant de dades que permeten extreure la màxima informació bioquímica en estudis de metabolòmica no dirigida d'organismes vegetals.

En un estudi metabolòmic no dirigit, els reptes analítics i quimiomètrics més importants són: l'anàlisi instrumental exhaustiva del màxim nombre de metabòlits presents a les mostres, l'anàlisi quimiomètrica de les dades experimentals generades i la identificació i quantificació dels metabòlits resolts, especialment d'aquells que mostren canvis importants en la seva concentració degut a l'estímul o estrès aplicat sobre les mostres analitzades. En aquesta Tesi s'han avaluat i proposat diverses estratègies i eines que poden resultar útils per solucionar els reptes existents en aquestes etapes. S'ha estudiat per exemple la influència que poden tenir diferents factors experimentals en la separació dels metabòlits mitjançant cromatografia d'interacció hidrofílica (HILIC), com ara el tipus de fase estacionària, el pH i la força iònica de la fase mòbil. Els resultats obtinguts han demostrat que els factors més importants són el tipus de fase estacionària HILIC i el pH de la fase mòbil. Concretament, en l'estudi dels diferents tipus de fase estacionària s'ha determinat que les fases amida i zwitteriònica són les que proporcionen millors resultats per a l'anàlisi de metabòlits polars. D'altra banda, s'han comparat dues estratègies de tractament de dades de metabolòmica no dirigida: l'estratègia de referència en el camp basada en la utilització del programa XCMS, i la que s'ha proposat en aquesta Tesi, que es basa en el mètode quimiomètric de resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS). Els dos procediments han proporcionat resultats similars i, en conseqüència, tots dos resulten apropiats pel tractament de dades de metabolòmica no dirigida. Però cal destacar que el procediment basat en MCR-ALS és més robust ja que s'ha pogut aplicar a una varietat major de dades experimentals. S'ha proposat també l'ús de models de

relació quantitativa estructura-retenció (QSRR) que han permès relacionar l'estructura dels metabòlits amb el seu temps de retenció observat experimentalment. Els resultats obtinguts demostren que l'aplicació de models QSRR pot ser una eina a considerar per a la identificació dels metabòlits, ja que redueix el nombre de possibles candidats i en facilita la seva confirmació.

La complexitat de les mostres analitzades en els estudis metabolòmics sovint excedeix els límits de capacitat de pic aconseguits per la cromatografia de líquids unidimensional. Per aquest motiu, la cromatografia de líquids bidimensional exhaustiva (LC×LC) es presenta com una bona tècnica alternativa que permet ampliar la capacitat de separació de mescles molt complexes, com ara les d'extractes de metabòlits de mostres biològiques. S'ha avaluat l'estructura de les dades obtingudes a partir de l'aplicació de LC×LC-MS, i la seva possible modelització. Els resultats obtinguts han mostrat que, en general, el model de MCR-ALS bilineal és el més indicat per a l'anàlisi d'aquest tipus de dades. S'ha estudiat també el possible ús del procediment ROIMCR (compressió de les dades per cerca de regions d'interès [ROIs] seguida d'una resolució mitjançant MCR-ALS) per a estudis de metabolòmica no dirigida mitjançant LC×LC-MS. Els resultats han mostrat que la combinació de les tècniques de separació multidimensional i dels mètodes quimiomètrics de resolució és adequada en estudis metabolòmics, ja que permet resoldre i identificar un nombre molt elevat de metabòlits i lípids en una sola anàlisi.

Finalment, s'ha investigat la influència de diferents factors estressants ambientals sobre el metaboloma d'un organisme vegetal model com l'arròs (*Oryza sativa* L.). Els factors estressants ambientals estudiats han estat abiòtics, com ara, l'estrès hídric, l'augment de la temperatura i la presència de metalls pesants (Cd, Cu i As) contaminants al sòl. Els resultats obtinguts han demostrat que els tres factors ambientals avaluats afecten de forma significativa el metaboloma de l'arròs durant el seu creixement. Els espectres de masses resolts s'han utilitzat per a la identificació dels metabòlits i lípids de les mostres biològiques analitzades. La interpretació biològica dels canvis de concentració dels metabòlits i lípids identificats ha demostrat que la contaminació per metalls pesants ocasiona una disminució del creixement i de l'activitat fotosintètica de la planta, i que s'indueix un mecanisme de desintoxicació que permet disminuir el dany cel·lular. S'ha demostrat també que els increments de temperatura i d'estrès hídric provoquen un reajustament important del grau d'insaturació dels lípids de les cèl·lules de l'arròs, el qual permet millorar de forma considerable la fluïdesa de les membranes cel·lulars.

Abreviatures i acrònims

1D-LC	cromatografia de líquids unidimensional, *one-dimensional liquid chromatography*
2D-LC	cromatografia de líquids bidimensional, *two-dimensional liquid chromatography*
ALS	mínims quadrats alternats, *alternating least squares*
ANOVA	anàlisi de la variància, *analysis of variance*
ASCA	ANOVA-anàlisi de components simultanis, *ANOVA-simultaneous component analysis*
CE	electroforesi capil·lar, *capillary electrophoresis*
CE-MS	electroforesi capil·lar acoblada a espectrometria de masses, *capillary electrophoresis coupled to mass spectrometry*
CV	validació creuada, *cross validation*
DAD	detecció de díodes en sèrie, *diode array detector*
ESI	electrosprai, *electrospray*
FAO	organització de les Nacions Unides per a l'Agricultura i l'Alimentació, *Food and Agriculture Organization of the United Nations*
fnnls	mínims quadrats ràpids no negatius, *fast non-negative least squares*
FTICR	ressonància ciclotrònica de ions amb transformada de Fourier, *Fourier-transform ion cyclotron resonance*
FWHM	amplada de pic a mitja alçada, *full width at half maximum*
GA	algoritme genètic, *genetic algorithm*
GC	cromatografia de gasos, *gas chromatography*
GC×GC-MS	cromatografia de gasos bidimensional acoblada a espectrometria de masses, *two-dimensional gas chromatography coupled to mass spectrometry*
GC-MS	cromatografia de gasos acoblada a espectrometria de masses, *gas chromatography coupled to mass spectrometry*
GMD	base de dades del metaboloma de Golm, *Golm Metabolome Database*
HILIC	cromatografia d'interacció hidrofílica, *hydrophilic interaction chromatography*
HPLC	cromatografia de líquids d'alta eficàcia, *high performance liquid chromatography*
HRMS	espectrometria de masses d'alta resolució, *high resolution mass spectrometry*
IEX	cromatografia de bescanvi iònic, *ion exchange chromatography*
IPs	punts d'identificació, *identification points*
KEGG	enciclopèdia de Kyoto de gens i genomes, *Kyoto encyclopedia of gens and genomes*
LC	cromatografia de líquids, *liquid chromatography*
LC×LC	cromatografia de líquids bidimensional exhaustiva, *comprehensive two-dimensional liquid chromatography*
LC-LC	cromatografia de líquids bidimensional de talls, *heart-cutting two-dimensional liquid chromatography*
LC-MS	cromatografia de líquids acoblada a espectrometria de masses, *liquid chromatography coupled to mass spectrometry*
LOF	falta d'ajust, *lack of fit*
LOOCV	validació creuada deixant-ne un fora, *leave-one-out cross validation*
LV	variable latent, *latent variable*
m/z	relació de massa i càrrega, *mass-to-charge ratio*
MALDI	desorció/ionització mitjançant làser assistida per matriu, *matrix-assisted laser desorption/ionization*
MANOVA	ANOVA multivariant, *multivariate ANOVA*
MCC	coeficient de correlació de Matthews, *Matthews correlation coefficient*
MCR	resolució multivariant de corbes, *multivariate curve resolution*
MCR-ALS	resolució multivariant de corbes per mínims quadrats alternats, *multivariate curve resolution by alternating least squares*
MDs	descriptors moleculars, *molecular descriptors*

MS	espectrometria de masses, *mass spectrometry*
MS/MS	espectrometria de masses en tàndem, *tandem mass spectrometry*
MSI	espectroscòpia d'imatges de MS, *mass spectrometry imaging*
NIST	institut nacional d'estàndards i tecnologia, *national institute of standards and technology*
nnls	mínims quadrats no negatius, *non-negative least squares*
NP-LC	cromatografia de líquids de fase normal, *normal phase liquid chromatography*
PARAFAC	anàlisi de factors paral·lels, *parallel factor analysis*
PCA	anàlisi de components principals, principal *component analysis*
PLS	mínims quadrats parcials, *partial least squares*
PLS-DA	anàlisi discriminant per mínims quadrats parcials, *partial least squares–discriminant analysis*
QCs	controls de qualitat, *quality controls*
QqQ	triple quadrupol, *triple quadrupole*
QSRR	models de relació quantitativa estructura-retenció, *quantitative structure-activity relationship*
R^2	percentatge de variància explicada, *percentage of explained variance*
RMN	ressonància magnètica nuclear, *nuclear magnetic resonance*
RMSECV	error quadràtic mig de validació creuada, *root mean square error of cross-validation*
RMSEP	error quadràtic mig de predicció, *root mean square error of prediction*
ROI	regions d'interès, *regions of interest*
RP-LC	cromatografia de líquids de fase inversa, *reversed phase liquid chromatography*
SCA	anàlisi de components simultanis, *simultaneous component analysis*
SIMPLISMA	anàlisi iterativa d'auto modelatge senzilla d'utilitzar, *SIMPLe-to-use Iterative Self-Modeling Analysis*
SMILES	especificació d'introducció lineal molecular simplificada, *simplified molecular input line entry specification*
SVD	descomposició en valors singulars, *singular value decomposition*
TIC	corrent de ions totals, *total ion current*
TOF	temps de vol, *time of flight*
UHPLC	cromatografia de líquids de molt alta eficàcia, *ultra high performance liquid chromatography*
UV-Vis	ultraviolat-visible, *ultraviolet-visible*
VIPs	variables importants en projecció, *variable importance on projection*
XCMS	vàries formes de cromatografia (X) acoblada a espectrometria de masses, *various forms (X) of chromatography mass spectrometry*

Notació

En aquesta secció es descriu la notació matemàtica utilitzada en aquesta Tesi. Aquesta és l'acceptada per la comunitat científica.

Les lletres minúscules cursives (per exemple, x) indiquen escalars. Les lletres minúscules en negreta (per exemple, $\mathbf{x}$) indiquen vectors. Les lletres majúscules en negreta (per exemple, $\mathbf{X}$) indiquen matrius. Les lletres majúscules en negreta i subratllades (per exemple, $\underline{\mathbf{X}}$) indiquen cubs. La transposició d'una matrius s'indica amb una "T" com a superíndex (per exemple, $\mathbf{X}^{\mathbf{T}}$).

Índex

CAPÍTOL 1. Objectius i estructura de la Tesi 1

1.1. Context i Objectius 3
1.2. Estructura de la Tesi 4
1.3. Relació dels treballs científics presentats en la memòria 5

CAPÍTOL 2. Introducció 7

2.1. Metabolòmica d'organismes vegetals 9

2.1.1. La metabolòmica en el context de les ciències òmiques 9
2.1.2. Estratègies dirigides i no dirigides 11
2.1.3. Factors ambientals estressants 14
2.1.4. L'arròs (*Oryza sativa* L.) com a organisme vegetal model 20

2.2. Metodologies analítiques 24

2.2.1. Cromatografia de líquids 25
2.2.2. Cromatografia de líquids bidimensional exhaustiva (LC×LC) 31
2.2.3. Espectrometria de masses 35

2.3. Metodologies quimiomètriques 41

2.3.1. Naturalesa de les dades 41
2.3.2. Mètodes de preprocessament de les dades 44
2.3.3. Resolució dels pics cromatogràfics i detecció de variables característiques 51
2.3.4. Mètodes quimiomètrics d'exploració, regressió i classificació de dades metabolòmiques 65

2.4. Referències 79

CAPÍTOL 3. Estudis de metabolòmica no dirigida: estratègies analítiques i de tractament de dades 89

3.1. Introducció 91

3.2. Publicació 1. *Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches* 95

3.3. Publicació 2. *Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies* 119

3.4. Discussió conjunta dels resultats 143

3.4.1. Fases estacionàries HILIC en anàlisis metabolòmiques 143
3.4.2. Resolució de metabòlits per mètodes quimiomètrics 145
3.4.3. Identificació de metabòlits a partir de les seves propietats cromatogràfiques 150

3.5. Referències 155

CAPÍTOL 4. Estudi dels efectes de diversos estressants ambientals sobre l'arròs 159

4.1. Introducció 161

4.2. Publicació 3. *Metabolomic analysis of the effects of cadmium and copper treatment in Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation* 163

4.3. Publicació 4. *Untargeted lipidomic evaluation of hydric and heat stresses on rice growth* 195

4.4. Discussió conjunta dels resultats 237

4.4.1. Resolució dels metabòlits i lípids de l'arròs en diferents condicions ambientals 237
4.4.2. Estudi estadístic dels efectes de diversos factors ambientals sobre el creixement de l'arròs 240
4.4.3. Identificació de metabòlits i lípids 243
4.4.4. Interpretació biològica dels resultats obtinguts 245

4.5. Referències 247

CAPÍTOL 5. Desenvolupament i aplicació de la cromatografia de líquids bidimensional exhaustiva en estudis de metabolòmica no dirigida 251

5.1. Introducció 253

5.2. Publicació 5. *Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil* 257

5.3. Publicació 6. *Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution* 279

5.4. Publicació 7. *An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis* 309

5.5. Discussió conjunta dels resultats 337

5.5.1. Estructura i modelització de les dades de LC×LC-MS 337
5.5.2. Estratègies de compressió de les dades LC×LC-MS 343
5.5.3. Aplicació de les metodologies LC×LC-MS acoblades amb l'anàlisi quimiomètrica de les dades 345

5.6. Referències 354

CAPÍTOL 6. Conclusions 359

Capítol 1

Objectius i estructura de la Tesi

1.1. Objectius i context

L'objectiu principal d'aquesta Tesi és el desenvolupament i l'optimització de nous mètodes d'anàlisi basats en tècniques de cromatografia de líquids uni- i bidimensionals acoblades a l'espectrometria de masses, conjuntament amb l'aplicació de noves estratègies d'anàlisi multivariant de dades que permetin extreure la màxima informació bioquímica en estudis de metabolòmica no dirigida d'organismes vegetals.

De manera més específica, aquesta Tesi presenta els següents objectius secundaris:

- Avaluació del comportament de fases estacionàries de cromatografia d'interacció hidrofílica (HILIC) en diferents condicions experimentals (pH i força iònica de la fase mòbil) pel seu ús en estudis de metabolòmica no dirigida.
- Avaluació i comparació de dues estratègies de tractament de dades de metabolòmica no dirigida: una basada en el mètode XCMS (*various forms (X) of chromatography mass spectrometry*) i l'altra basada en el mètode de resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS) desenvolupada pel nostre grup d'investigació.
- Estimació de la influència de diferents factors ambientals estressants (contaminació per metalls pesants, manca d'aigua i augment de temperatura) sobre el creixement de l'arròs, a partir d'estudis de metabolòmica i lipidòmica no dirigides.
- Estudi de l'estructura de les dades de la cromatografia de líquids bidimensional acoblada a espectrometria de masses (LC×LC-MS), i avaluació de les seves possibilitats en estudis de metabolòmica no dirigida.

Aquesta Tesi s'ha realitzat en el marc del projecte europeu CHEMAGEB (*CHEMometric and High-throughput Omics Analytical Methods for Assessment of Global Change Effects on Environmental and Biological Systems*). L'objectiu principal d'aquest projecte és desenvolupar nous mètodes analítics i quimiomètrics per tal d'avaluar els efectes del canvi climàtic i la pol·lució en diferents organismes model, els quals són representatius dels ecosistemes, mitjançant una aproximació òmica (metabolòmica, transcriptòmica o genòmica). Dins d'aquest projecte, aquesta Tesi s'ha centrat en els estudis de metabolòmica no dirigida d'organismes vegetals, utilitzant com a organisme model l'arròs (*Oryza sativa* L.), i com a metodologia analítica la cromatografia de líquids acoblada a l'espectrometria de masses.

1.2. Estructura de la Tesi

La present memòria està estructurada en sis capítols, que s'introdueixen a continuació.

En el primer capítol es presenten els objectius que han motivat la realització d'aquesta Tesi, la forma en que es troba estructurada i la relació dels treballs científics inclosos en la present memòria.

En el segon capítol es fa una introducció general sobre la metabolòmica d'organismes vegetals, que inclou aspectes com la seva aplicació en estudis ambientals i en els factors ambientals estressants de plantes més estudiats a la bibliografia. S'expliquen amb detall els factors estressants estudiats en aquesta Tesi (manca d'aigua, augment de temperatura i contaminació per metalls pesants) i es descriuen les característiques de l'arròs (*O. sativa* L.) com a organisme model en aquest tipus d'estudis. Es presenten les metodologies analítiques emprades en aquesta Tesi per a l'anàlisi dels metabòlits de l'arròs, basades en la cromatografia de líquids acoblada a l'espectrometria de masses. Finalment, s'introdueixen els mètodes quimiomètrics de processament de les dades experimentals òmiques, entre els que destaquen el mètode de resolució multivariant de corbes (MCR), els mètodes d'exploració (PCA i ASCA), els mètodes de regressió (PLS) i els mètodes de classificació (PLS-DA).

En el tercer capítol es presenten els resultats obtinguts en l'estudi de les tres etapes dels estudis de metabolòmica no dirigida que presenten els reptes analítics i quimiomètrics més importants: anàlisi dels metabòlits, tractament de les dades i identificació dels metabòlits. S'avaluen diverses fases estacionàries de cromatografia d'interacció hidrofílica (HILIC) pel seu ús en metabolòmica no dirigida. Es comparen els resultats obtinguts en processar el mateix conjunt de dades metabolòmiques mitjançant dues estratègies de tractament de dades diferents: el procediment XCMS (*various forms (X) of chromatography mass spectrometry*) i el mètode de resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS). Finalment, es proposa la utilització de models de relació quantitativa estructura-retenció (QSRR) per a millorar la identificació dels metabòlits.

En el quart capítol es presenten els resultats de l'estudi dels efectes de diferents factors ambientals estressants sobre el metaboloma i el lipidoma de l'arròs. D'una banda, s'avaluen les alteracions provocades en el creixement de l'arròs per la contaminació per metalls pesants (cadmi i coure) mitjançant un estudi de metabolòmica no dirigida. D'altra banda, s'estudien els efectes dels canvis de temperatura i de la manca d'aigua sobre el creixement de l'arròs utilitzant una aproximació lipidòmica no dirigida. En els dos treballs d'aquest capítol s'ha utilitzat la cromatografia de líquids unidimensional acoblada a l'espectrometria de masses (LC-MS).

En el cinquè capítol es mostren els resultats de l'aplicació de la cromatografia de líquids bidimensional acoblada a l'espectrometria de masses (LC×LC-MS) en estudis de metabolòmica no dirigida. En primer lloc s'estudia l'estructura de les dades de LC×LC-MS i s'avalua la seva possible multilinealitat. Seguidament, es proposen nous mètodes d'anàlisi mitjançant LC×LC-MS combinats amb estratègies de tractament de dades basades en l'aplicació del procediment MCR-ALS pel seu ús en estudis de metabolòmica no dirigida. La viabilitat de l'estratègia proposada es demostra amb la seva aplicació a l'estudi dels canvis observats sobre el metabolisme de l'arròs en un cicle de 24 h (cicle circadiari). Finalment, es desenvolupa un procediment LC×LC-MS per a l'anàlisi no dirigida dels lípids de l'arròs sotmès a condicions de contaminació ambiental. La viabilitat d'aquest mètode es demostra mitjançant la seva aplicació a un estudi dels efectes de la contaminació per arsènic en el creixement de l'arròs.

El sisè capítol recull les conclusions generals més importants obtingudes en el marc d'aquesta Tesi.

1.3. Relació dels treballs científics presentats en la memòria

1. **Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches.**

   Autors: Meritxell Navarro-Reig, Elena Ortiz-Villanueva, Romà Tauler, Joaquim Jaumot.

   Revista: Metabolites 7 (2017), 54.

2. **Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies.**

   Autors: Meritxell Navarro-Reig, Joaquim Jaumot, Alejandro García-Reiriz, Romà Tauler.

   Revista: Analytical and Bioanalytical Chemistry 407 (2015), 8835-8847.

3. **Metabolomic analysis of the effects of cadmium and copper treatment in *Oryza sativa L.* using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation.**

   Autors: Meritxell Navarro-Reig, Joaquim Jaumot, Benjamín Piña, Encarnación Moyano, Maria Teresa Galceran, Romà Tauler.

   Revista: Metallomics 9 (2017), 660-675.

4. **Untargeted lipidomic evaluation of hydric and heat stresses on rice.**

   Autors: Meritxell Navarro-Reig, Romà Tauler, Guillermo Iriondo-Frias, Joaquim Jaumot.

   Enviat.

5. **Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil.**

   Autors: Meritxell Navarro-Reig, Joaquim Jaumot, Teris A. van Beek, Gabriel Vivó-Truyols, Romà Tauler.

   Revista: Talanta 160 (2016), 624-635.

6. **Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution.**

   Autors: Meritxell Navarro-Reig, Joaquim Jaumot, Anna Baglai, Gabriel Vivó-Truyols, Peter J. Schoenmakers, Romà Tauler.

   Revista: Analytical Chemistry 89 (2017), 7675-7683.

7. **An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis.**

   Autors: Meritxell Navarro-Reig, Joaquim Jaumot, Romà Tauler.

   Enviat.

Capítol 2

Introducció

2.1. Metabolòmica d'organismes vegetals

En aquesta Tesi s'ha estudiat l'efecte que tenen en el creixement de l'arròs (*Oryza sativa* L.) diferents estressos ambientals, com per exemple la manca d'aigua o la contaminació de metalls pesants, mitjançant una aproximació metabolòmica. Per aquest motiu, en aquest apartat s'introdueix la metabolòmica en el context de les ciències òmiques i es presenten les diferents estratègies de treball que existeixen en aquest tipus d'estudis. Finalment, s'exposen les aplicacions de la metabolòmica d'organismes vegetals en estudis ambientals i es justifica l'elecció de l'arròs (*O. sativa* L.) com a organisme model de treball.

2.1.1. La metabolòmica en el context de les ciències òmiques

En la dècada de 1990 els projectes del genoma humà (*The Human Genome Project*) i del genoma d'organismes model (*Model Organism Genome Project*) van marcar un punt d'inflexió en moltes àrees de la ciència [1, 2]. Aquest canvi va consistir en la possibilitat d'analitzar simultàniament un gran nombre d'anàlits (gens, transcrits, proteïnes o metabòlits), enlloc d'utilitzar l'enfocament reduccionista que consistia en estudiar-los individualment i de manera aïllada [1, 2].

En el camp de la biologia molecular, el terme "òmica" es refereix a aquest estudi global dels gens (genoma), transcrits (transcriptoma), proteïnes (proteoma) o metabòlits (metaboloma), així com de les relacions entre ells [3]. Tots aquests estudis han desembocat en l'aparició de les ciències conegudes com a òmiques, que tenen per objectiu l'estudi de l'abundància i/o la caracterització estructural d'un ampli rang de molècules que es troben involucrades en processos biològics [1-5]. Avui en dia, les ciències òmiques són clau per la interpretació i la comprensió dels processos biològics més complexos. Així, les principals ciències òmiques són la genòmica, la transcriptòmica, la proteòmica i la metabolòmica.

El reconeixement de la metabolòmica com a camp científic és molt més recent que en el cas de les altres òmiques (els primers articles daten de finals de la dècada del 1990 [6]). No obstant, la metabolòmica ha adquirit ràpidament una alta rellevància científica, com demostra el creixement exponencial del nombre de publicacions d'aquest camp. Aquesta importància de la metabolòmica rau en el fet de que sigui el punt final de la cascada òmica (Figura 2.1), que relaciona la informació genètica amb el fenotip. El metabolisme d'un organisme expressa directament l'estat biològic en el que aquest es troba. D'aquesta manera, la metabolòmica és la ciència òmica que es relaciona més directament amb el fenotip, el qual depèn tant del genoma com de les circumstàncies en les quals es troba l'organisme [2-4].

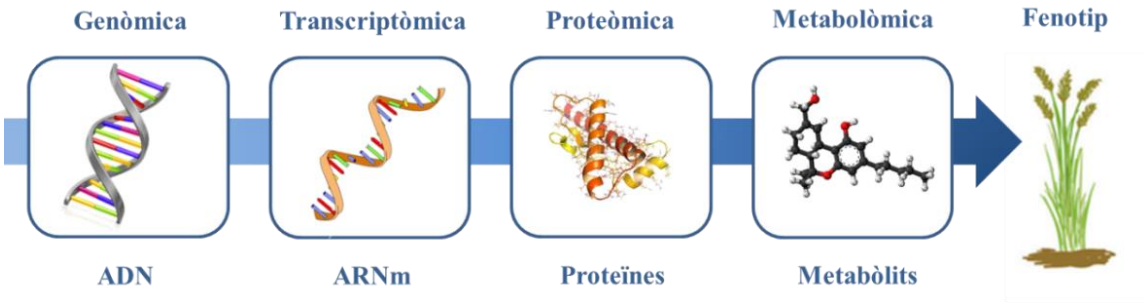


Figura 2.1. Cascada òmica. Les ciències òmiques adopten una perspectiva holística de les molècules que composen un organisme. Tenen com a objectiu la total detecció dels gens (genòmica), transcrits (transcriptòmica), proteïnes (proteòmica) i metabòlits (metabolòmica) d'un sistema biològic específic.

L'any 1999, Nicholson va definir la metabolòmica com el camp de la ciència que té per objectiu caracteritzar la totalitat dels metabòlits que es troben a les cèl·lules, teixits o biofluids d'un organisme [7]. Els metabòlits són compostos de baix pes molecular (inferior a 1500 Da) que es transformen durant el metabolisme i que aporten informació directa sobre l'activitat de les cèl·lules [8, 9]. En aquest context, l'estudi dels metabòlits ajuda a millorar el coneixement sobre els efectes que les malalties, els contaminants o els canvis en les condicions ambientals causen sobre els organismes [3-5]. Cal comentar que hi ha certa controvèrsia en la utilització de les paraules metabolòmica i metabonòmica. D'una banda la metabonòmica té com a objectiu mesurar la resposta metabòlica global de sistemes vius a estímuls biològics o de manipulació genètica i l'atenció es centra en la comprensió biològica dels canvis observats [7]. D'altra banda, la metabolòmica pretén una descripció analítica més global de les mostres biològiques i es centra en caracteritzar i quantificar tots els metabòlits presents en una mostra [7]. A la pràctica els dos termes s'utilitzen indistintament, ja que els procediments de treball són molt semblants [10]. En aquesta tesi s'ha utilitzat el terme metabolòmica en tots els casos.

Existeixen moltes famílies de metabòlits diferents, tals com aminoàcids, nucleòtids, nucleòsids, sucres, coenzims, cofactors, lípids i altres metabòlits secundaris, entre els quals es troben alcaloides, tanins, saponines i flavonoides entre altres. Aquesta varietat resulta en un ampli ventall d'estructures i propietats químiques diferents que fan molt difícil l'anàlisi simultània de tots els metabòlits d'un organisme. Per aquest motiu, s'han desenvolupat subdisciplines de la metabolòmica, que tenen com a objectiu estudiar una família concreta de metabòlits. Exemples d'aquestes disciplines són la glicòmica [11] que estudia els sucres; la volatilòmica [12] que investiga tots els metabòlits volàtils o la lipidòmica [13] que contempla els lípids. Els lípids en particular, són un ampli grup de biomolècules implicades en diverses activitats estructurals i funcionals de les cèl·lules [14-17]. D'una banda són

components estructurals de les membranes cel·lulars, influenciant tant la seva fluïdesa com els intercanvis que s'hi produeixen (transport de nutrients, expulsió de residus, etc.). D'altra banda, els lípids també estan implicats en el transport i l'emmagatzematge de l'energia cel·lular [14, 16, 17]. En els organismes vegetals, els tipus de lípids més abundants són [18, 19]:

- Àcids grassos: cadenes d'hidrocarburs que acaben amb un grup àcid carboxílic. En organismes vegetals els més abundants estan formats per cadenes de 16 o 18 carbonis.
- Glicerolípids: glicerols amb un (-mono), dos (-di) o tres (-tri) substituents, generalment àcids grassos i sucres.
- Fosfolípids: glicerols amb dos àcids grassos i un grup fosfat com a substituents.

Tenint en compte l'àmplia diversitat estructural dels lípids i la importància de les seves implicacions biològiques, no és d'estranyar que la lipidòmica hagi emergit com una branca gairebé independent de la metabolòmica. Així la lipidòmica estudia la totalitat dels lípids presents en un organisme i les molècules amb les quals interactuen i les seves funcions dins la cèl·lula [14-17].

2.1.2. Estratègies dirigides i no dirigides

En els estudis de metabolòmica existeixen dos tipus d'enfocaments principals: anàlisi dirigida i anàlisi no dirigida o global (Figura 2.2) [4, 8, 20].

L'anàlisi dirigida (Figura 2.2A) es centra en la determinació de les concentracions o abundàncies relatives d'un grup específic de metabòlits, generalment de la mateixa família de compostos o relacionats amb una ruta metabòlica concreta d'interès [4, 20]. L'enfocament dirigit habitualment es porta a terme per tal de corroborar una hipòtesi relacionada amb una ruta metabòlica específica. Exemples típics d'aplicacions de l'enfocament dirigit són estudis cinètics del metabolisme de fàrmacs o estudis que mesuren la influència de teràpies i modificacions genètiques sobre un enzim específic [9, 21]. Abans de dur a terme aquest tipus d'experiments, és necessari disposar d'informació sobre els compostos a analitzar.

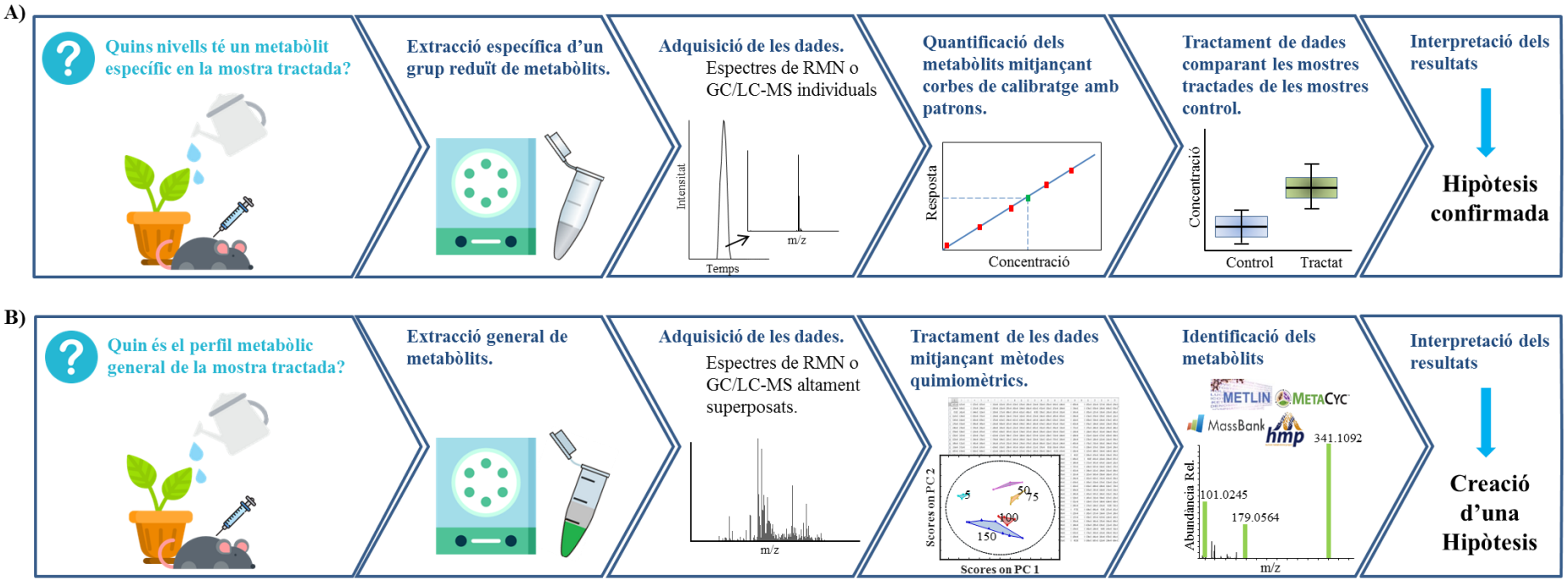


Figura 2.2. Esquema general del procés de treball en estudis de metabolòmica dirigida (A) i no dirigida (B).

Els estudis de metabolòmica dirigida generalment consisteixen en una anàlisi quantitativa, ja que impliquen la comparació dels anàlits amb patrons de referència [20]. Una de les principals característiques de l'enfocament dirigit és que es troben treballs científics relacionats amb l'estudi d'un grup concret de metabòlits des de principis del segle XX [9, 22]. Per tant, en la literatura es disposa d'un ampli ventall d'informació sobre mètodes i protocols per a l'anàlisi de metabòlits en diferents tipus de mostres. A més, amb la creixent millora de les metodologies analítiques disponibles, en la literatura es poden trobar mètodes cada cop més sensibles i robustos per a la mesura d'un nombre gran de metabòlits [4, 23, 24]. No obstant això, el principal inconvenient de l'anàlisi dirigida és la necessitat de seleccionar els metabòlits d'interès *a priori*. A més, els estudis dirigits es troben limitats per la disponibilitat dels patrons de referència necessaris per a dur a terme la quantificació [20].

D'altra banda, l'anàlisi no dirigida (Figura 2.2B) té per objectiu la detecció simultània de tants metabòlits com sigui possible, sense cap mena de filtre i sense disposar de cap coneixement previ sobre les rutes metabòliques d'interès [4, 8, 20, 25]. Aquesta estratègia obre la possibilitat de descobrir metabòlits importants per a un estudi concret que estiguin associats a rutes metabòliques prèviament inexplorades. Al contrari que en el cas de l'anàlisi dirigida, l'enfocament no dirigit normalment dóna lloc a la creació d'hipòtesis relacionades amb rutes metabòliques, el que es coneix com *discovery omics* [8]. Els exemples més típics d'aplicacions de l'anàlisi no dirigida són els estudis que tenen per objectiu trobar nous biomarcadors per un estat biològic específic [2, 4, 20, 26]. En resum, aquesta estratègia es basa en trobar canvis en el metaboloma i, a partir d'aquí, identificar els metabòlits responsables d'aquests canvis. Una altra gran diferència entre l'enfocament dirigit i el no dirigit és el tractament i emmagatzematge de les dades. En el cas de l'anàlisi no dirigida, les dades obtingudes són més complexes ja que es registren milions de senyals per cada mostra analitzada. Això provoca que els arxius de dades generats siguin de l'ordre de gigabytes per mostra. Aquesta major complexitat de les dades requereix la utilització d'eines quimiomètriques adequades pel seu processament [4, 20]. El principal inconvenient de l'enfocament no dirigit és que en la majoria dels casos no és possible realitzar una quantificació absoluta dels metabòlits, si no que és només relativa. A més, sovint no s'aconsegueix la identificació de tots els metabòlits que es detecten com a responsables dels canvis observats en els perfils metabòlics [2, 20, 26]. Malgrat això, aquest enfocament és àmpliament utilitzat perquè ofereix la possibilitat d'explorar noves rutes biològiques i trobar nous marcadors en resposta a diversos factors estressants externs [4, 27]. Finalment, cal destacar que les dues aproximacions poden considerar-se complementàries degut a la relació entre els

dos tipus d'enfocaments. D'aquesta forma a partir dels resultats obtinguts mitjançant la metabolòmica no dirigida es poden plantejar estudis de metabolòmica dirigida específics que permetin corroborar els resultats obtinguts i fer estimacions quantitatives absolutes. Com a conclusió es pot dir que els estudis no dirigits generen hipòtesis relacionades amb possibles rutes metabolòmiques afectades, mentre que l'objectiu dels estudis dirigits és corroborar aquestes hipòtesis i establir estimacions quantitatives absolutes dels efectes estudiats [28].

La metabolòmica no dirigida s'ha aplicat en diversos àmbits de recerca, com el clínic [26, 29], l'alimentari [30, 31] i l'ambiental [29, 32]. En particular, l'interès d'aquesta Tesi es centra en l'àmbit ambiental, en el qual la metabolòmica avalua les interaccions entre els organismes i el seu medi [33, 34]. Això s'aconsegueix mesurant els metabòlits endògens d'un teixit o d'un fluid biològic de l'organisme estudiat, els quals proporcionen una descripció del seu fenotip metabòlic funcional [34]. En els estudis de metabolòmica ambiental s'avaluen els efectes que els canvis ocasionats en el medi ambient causen en els organismes. Aquests canvis en el medi ambient inclouen tant factors abiòtics, com ara la sequera, la salinitat, alteracions en la temperatura o presència de contaminants [30, 33, 35, 36], com factors biòtics, com per exemple la competició entre espècies o la simbiosis [37, 38].

2.1.3. Factors ambientals estressants

En la metabolòmica ambiental destaquen els esforços que s'han fet en estudiar els efectes de diferents factors estressants sobre els organismes vegetals [29]. D'una banda, això és degut a l'avantatge que presenta el fet que la metabolòmica no dirigida permeti l'anàlisi simultània dels metabòlits primaris i els secundaris. Els primaris són importants pel creixement i desenvolupament dels organismes vegetals, mentre que els secundaris tenen un paper molt rellevant en les funcions de protecció i defensa [32, 39]. En la Taula 2.1 es mostren alguns exemples de metabòlits secundaris i les seves funcions en els organismes vegetals.

Taula 2.1. Exemples de les funcions dels metabòlits secundaris en organismes vegetals [40, 41].

Tipus de Metabòlit Secundari	Funció
Alcaloides	Defensa contra herbívors (sobretot mamífers) per intoxicació.
Glucosinolats	Defensa contra plagues per olor.
Glucòsids cianogènics	Defensa contra herbívors per alliberament del verí àcid cianhídric (HCN).
Monoterpens	Defensa contra herbívors (sobretot insectes) per intoxicació.
Sesquiterpens	Protecció contra estrès d'aigua per modificació de les propietats de les membranes.
Cumarines	Activitat antimicrobiana contra fongs i bactèries.
Flavonoides	Protecció contra radiació UV-B.

D'altra banda, els estudis de metabolòmica també tenen l'avantatge de no necessitar informació genètica prèvia, al contrari que els de transcriptòmica i proteòmica [42]. Aquest fet va fer possible l'estudi metabolòmic d'organismes vegetals comestibles abans de que se'n completés la seqüenciació del genoma, com va ser el cas del blat (*Triticum aestivum* L.), l'ordi (*Hordeum vulgare* L.), la patata (*Solanum tuberosum* L.*)* o la soja (*Glycine max* L.) [42]. Tot i així, cal destacar que els estudis metabolòmics d'organismes vegetals són un gran repte, degut a la diversitat química dels metabòlits presents en les plantes i a que la seva dependència de les condicions ambientals és molt gran. S'ha estimat que en el regne vegetal hi ha entre 20.000 i 100.000 metabòlits diferents, i que una sola espècie de planta en pot contenir milers (per exemple, en l'*Arabidopsis thaliana* L. s'estima que pot presentar-ne uns 5.000) [43].

La metabolòmica dels organismes vegetals ha esdevingut una eina potent per a l'exploració de diferents aspectes de la fisiologia i la biologia de les plantes, augmentant considerablement el coneixement sobre els mecanismes metabòlics de regulació del creixement, el desenvolupament i la resposta a factors estressants. En conseqüència, hi ha nombroses aplicacions de la metabolòmica no dirigida en organismes vegetals, des d'estudis sobre els efectes d'un canvi ambiental fins a investigacions sobre la millora de la productivitat i qualitat de cultius agrícoles [39-45].

Els organismes vegetals s'han d'enfrontar a diversos tipus d'estressos ambientals durant el seu desenvolupament. A més, com a organismes estàtics les plantes són particularment vulnerables a alguns d'aquests estressos ambientals. Per aquest motiu, el desenvolupament de respostes metabòliques a les condicions adverses és clau per a la seva supervivència.

En general, els estressos ambientals es poden classificar en dos tipus: biòtics i abiòtics. Els biòtics provenen de patògens i plagues, mentre que els abiòtics són el resultat d'alteracions en els factors

ambientals per fenòmens naturals, com per exemple la sequera, les inundacions, les temperatures extremes, la radiació severa, la presència de contaminants o la limitació dels nutrients. Tant bon punt els receptors són estimulats pels senyals d'estrès, l'expressió dels gens responsables de la resposta a aquest estrès s'activa ràpidament. A conseqüència d'això, els metabòlits especialitzats es sintetitzen, es transformen o s'eliminen per tal poder adaptar-se i sobreviure a la situació adversa. En aquest punt, la metabolòmica no dirigida permet investigar els mecanismes bioquímics de protecció i resposta dels organismes vegetals, així com els efectes que causen les alteracions ambientals [39, 42, 44, 45].

La distribució geogràfica dels organismes vegetals es troba limitada pels estressos ambientals presents en unes determinades condicions. L'any 1982, Boyer va indicar que aquesta limitació arribava al 70% [46]. Segons un estudi de 2007 de l'organització de les Nacions Unides per a l'Agricultura i l'Alimentació (FAO), només un 3,5% de l'àrea global del sòl no es troba afectada per alguna restricció ambiental [47]. Per exemple, el dèficit d'aigua (sequera) afecta al 64% de l'àrea global del sòl, les inundacions (anòxia) al 13%, la salinitat al 6%, el dèficit de minerals al 9%, l'acidificació dels sòls al 15% i el fred al 57% [35]. És difícil determinar els efectes actuals de l'estrès abiòtic sobre la distribució geogràfica de les plantes. Tot i així, tenint en compte el percentatge de sòl afectat, la continua reducció de terra cultivable i de recursos hídrics, així com la tendència d'escalfament global i canvi climàtic, sembla evident que l'impacte ambiental és encara més significatiu avui en dia. Aquesta creixent preocupació es reflecteix en l'increment del nombre de publicacions centrades en l'estrès abiòtic [48]. Per exemple, des de l'any 2002, el nombre de publicacions relacionades amb l'estrès abiòtic en organismes vegetals i que utilitzen metabolòmica ha crescut de forma gradual fins a l'any 2017 (Figura 2.3).

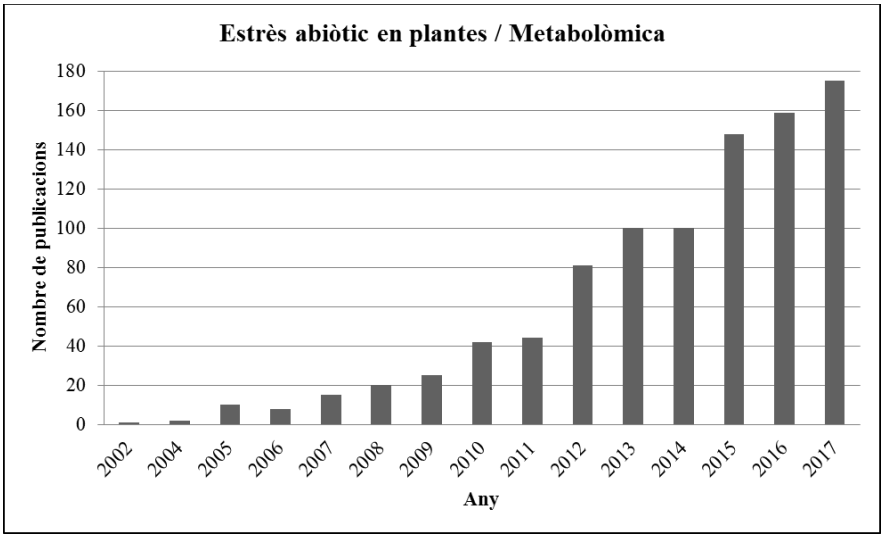


Figura 2.3. Nombre de publicacions per any relacionades amb metabolòmica i estrès abiòtic en plantes. Les paraules utilitzades en la cerca a Scopus van ser: *metabolomics, plants, abiotic stress.*

Fonamentalment, els organismes vegetals necessiten energia (llum), aigua, carboni i minerals per a créixer, desenvolupar-se i reproduir-se. Els estressos abiòtics es poden definir com les condicions ambientals que limiten aquestes necessitats i redueixen el creixement per sota dels nivells òptims. La resposta dels organismes vegetals a aquests estressos abiòtics és altament dinàmica i complexa, pot ser tant elàstica (reversible) com plàstica (irreversible) i canvia en funció de l'òrgan o teixit al qual afecti. A més, el nivell i la duració de l'estrès també influeix en la complexitat de la resposta [35, 48, 49]. En general, aquesta reacció davant l'estrès implica la interacció i l'encreuament de diferents rutes metabòliques. Per exemple, sota els efectes d'un determinat estrès abiòtic el metabolisme dels organismes vegetals es pot alterar degut a la inhibició d'enzims, l'escassetat de substrat, l'excés de demanda de compostos específics o una combinació d'aquests factors. L'adaptació de la xarxa metabòlica permet mantenir el metabolisme essencial i alhora adoptar un nou estat estacionari que permeti fer front a la situació d'estrès [35, 49-51].

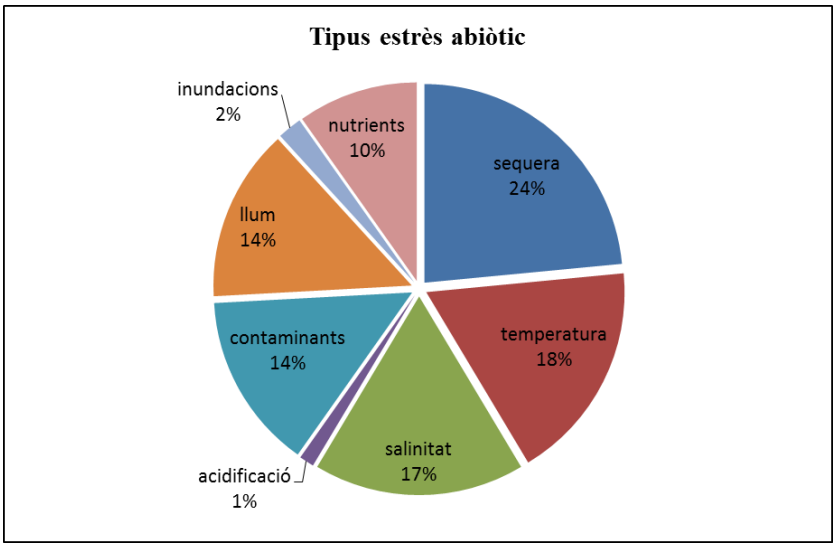


Figura2.4. Percentatge de publicacions per cada tipus d'estrès abiòtic. La cerca a Scopus es va realtizar pels articles publicats des de l'any 2002 fins al 2017 amb les següents paraules: *metabolomics, plants, abiotic stress, drought, temperature, salinity, acidic, pollutants, light, flood, nutrient starvation*.

Els principals estressos abiòtics són: la sequera, les temperatures extremes, la salinitat i l'acidificació dels sòls, la presència de contaminants, la intensitat lumínica, les inundacions i la falta de nutrients [35, 49, 52]. En la Figura 2.4 es poden veure el nombre de publicacions de metabolòmica d'organismes vegetals relacionades amb cada tipus d'estrès abiòtic des de l'any 2002 fins el 2017. Els tres factors ambientals que han estat més estudiats en plantes són la sequera, les temperatures extremes i la salinitat. En aquesta Tesi s'han estudiat la sequera, les variacions de temperatura i la presència de contaminants en el medi, en concret de metalls pesants.

Estrès hídric

L'estrès hídric per manca d'aigua o sequera és l'estrès abiòtic que més limita el creixement, el desenvolupament i la reproducció de les plantes. A més, els models del canvi climàtic prediuen que els episodis de sequera seran més freqüents i severs en un futur proper. Així, es creu que la freqüència de la sequera augmentarà al voltant d'un 20% cap al final d'aquest segle, especialment a les regions d'Amèrica del Sud i del centre i oest d'Europa [49-51, 53]. Per aquest motiu, és necessari entendre millor la resposta de les plantes a aquest estrès abiòtic. La sequera es relaciona amb una restricció en l'abastament d'aigua, el qual pot ser tant un cessament total com un dèficit continuat durant les diferents etapes de la vida dels organismes vegetals. Un dels paràmetres fisiològics més afectats per la sequera és la fotosíntesi, ja que causa una reducció en la capacitat d'assimilar el diòxid de carboni (CO_2) [50, 51, 54]. A nivell metabòlic destaca l'augment dels metabòlits involucrats en el metabolisme del carboni i la fixació del nitrogen, la qual cosa es relaciona amb un ajust en el nivell osmòtic i amb la reducció de la fotosíntesi. També s'observa un increment en la concentració dels sucres i els aminoàcids. A més, la composició dels lípids es veu alterada per ajustar la fluïdesa de la membrana cel·lular. El grau de insaturació dels àcids grassos es modifica per tal d'incremental la fluïdesa de la membrana [55].

Variacions de temperatura

Les variacions de temperatura són un altre dels factors abiòtics més estudiats. Els estudis sobre el canvi climàtic preveuen que la mitjana global de la temperatura de l'aire augmenti entre 1,3 i 3,1ºC cap al final del segle XXI. Per tant, el risc de que les plantes pateixin estrès degut a temperatures elevades augmentarà en un futur [56]. Aquest increment de temperatura afecta a les plantes en diferents nivells del desenvolupament, causant una reducció de la germinació de les llavors, una disminució en la permeabilitat de la membrana o pèrdues en el rendiment fotosintètic.

L'estrès per alta temperatura es detecta a la membrana plasmàtica de les cèl·lules, la qual es modifica físicament i actua com a termòmetre. Un augment en la temperatura causa un desequilibri metabòlic sever, altera principalment l'estabilitat de les proteïnes i dels àcids nucleics, el nivell de fitohormones, els metabòlits primaris i secundaris, l'estructura del citoesquelet, l'eficiència de les reaccions enzimàtiques i el sistema fotosintètic [35, 51, 57, 58].

D'altra banda, l'efecte de les temperatures baixes per sobre el punt de congelació (0-15ºC) també és un factor limitant important en la productivitat dels cultius. Tal com passa en el cas de l'augment de temperatura, la fotosíntesi es veu greument afectada pel fred. En condicions de temperatura baixes el

creixement de les plantes s'interromp. En conseqüència, la capacitat d'aprofitament energètic disminueix i s'inhibeix la fotosíntesi. Igual que en el cas d'altes temperatures, la percepció d'una disminució de la temperatura és deguda principalment als canvis ocasionats en la composició i fluïdesa de les membranes cel·lulars. A baixes temperatures el plasma cel·lular de les membranes redueix la seva fluïdesa fins a formar un gel sòlid, que és utilitzat per les cèl·lules vegetals per a combatre l'estrès del fred [51]. Moltes espècies de plantes augmenten la resistència al fred durant la seva exposició a temperatures fredes per sobre del punt de congelació (0-15ºC) mitjançant un procés conegut com "aclimatació al fred". Molts dels estudis de metabolòmica de plantes sotmeses a baixes temperatures estan orientats a entendre les bases moleculars d'aquest procés [50, 51, 59].

Contaminació per metalls pesants

La contaminació dels sòls amb diferents productes químics, com els pesticides o els fertilitzants, és un dels estressos ambientals més estudiats en el camp de la metabolòmica ambiental. Una de les manifestacions més importants de la contaminació química és la contaminació per metalls pesants. Els metalls pesants són constituents importants de l'escorça terrestre i dels processos geològics. Alguns processos naturals, com l'erosió del material geològic subterrani, les emissions dels volcans o els incendis forestals, contribueixen a la seva presència al medi ambient. A més, alguns d'aquests metalls pesants, com el coure (Cu), el ferro (Fe) o el manganès (Mn), a nivell traça són essencials per diversos processos metabòlics dels organismes. Malgrat això, la intervenció humana ha fet que el nivell de metalls pesants en el medi ambient augmenti contínuament, convertint-los en els contaminants presents en el sòl més perjudicials per a les plantes. Degut a l'alta reactivitat dels metalls pesants, una concentració alta d'aquests contaminants en el medi ambient afecta al creixement, la senescència i la producció d'energia de les plantes. A més, redueix el percentatge de germinació i la grandària de les llavors. Les reaccions més comunes a la presència de metalls pesants són la generació d'espècies reactives oxidants (ROS), el bloqueig de grups funcionals essencials de les biomolècules i la substitució de ions metàl·lics essencials en les biomolècules [49, 50, 60].

2.1.4. L'arròs (*Oryza sativa* L.) com a organisme vegetal model

En aquesta Tesi s'ha utilitzat l'arròs (*O. sativa* L.) com a organisme vegetal model. L'arròs compleix les principals característiques que ha de tenir un organisme per poder ser utilitzat com a model del seu regne [61, 62]:

- Fàcil manipulació.
- Manteniment en el laboratori relativament econòmic.
- Fàcilment accessible per a la majoria dels investigadors.
- Creixement ràpid.
- Es disposa d'informació sobre els seus gens, transcrits, proteïnes i metabòlits (dades òmiques multinivell).
- S'ha estudiat àmpliament, arribant a nivells molt alts de comprensió del seu funcionament.

Els organismes model són un factor clau en els estudis de metabolòmica ambiental, ja que el coneixement obtingut per ells pot extrapolar-se a altres organismes relacionats. La recerca amb bacteris, llevat, insectes, larves, peixos, rosegadors i plantes ha demostrat que els principis funcionals bàsics són molt semblants per tots els éssers vius. Per aquest motiu, els organismes models també són útils per a obtenir informació sobre altres especies que són més difícils d'estudiar directament. Gran part del coneixement que s'ha obtingut sobre processos biològics, ja siguin implicats en malalties humanes, farmacologia, toxicologia, fisiologia de plantes, ecologia o evolució, s'han obtingut a partir de la investigació amb un conjunt reduït d'organismes model. A més, al llarg dels anys s'han acumulat moltes dades sobre aquests organismes, la qual cosa encara els fa més atractius d'estudiar [61, 62]. En la Taula 2.2 es resumeixen els organismes models més habituals per cadascun dels grans regnes.

Taula 2.2. Principals organismes model de cada regne [62].

Regne	Nom Llatí	Nom Comú
Bacteris	*Escherichia coli*	-
Fongs	*Saccharomyces cerevisiae*	llevat
Animals (invertebrats)	*Caenorhabditis elegans*	cuc nematode
	Daphnia magna	puça d'aigua
	Drosophila melanogaster	mosca de fruita
Animals (vertebrats)	*Danio rerio*	peix zebra
	Mus musculus	ratolí
Plantes	*Arabidopsis thaliana*	arabidopsis
	Medicago truncatula	melgó truncat
	Oryza sativa	arròs
	Solanum lycopersicum	tomàquet

L'*A. thaliana* ha estat l'organisme model en plantes més utilitzat pels botànics durant molts anys [63]. La seqüenciació completa del seu genoma es va aconseguir l'any 2000, sent el primer genoma vegetal en ser seqüenciat completament. L'*A. thaliana* presenta molts avantatges, com un ràpid creixement, un temps de generació curt (de 1 a 2 mesos) i una mida petita de genoma (125 Mbp i 26422 gens) [64, 65]. Tot i així, l'*A. thaliana* és un planta dicotiledònia (caracteritzades per tenir dues fulles embrionàries a la llavor) i es va trobar que aquestes tenen diferències importants en el seu desenvolupament amb les plantes monocotiledònies (caracteritzades per tenir una sola fulla embrionària a la llavor) [63, 66]. Per exemple, els cereals són plantes monocotiledònies i tenen una gran importància social i econòmica perquè abasteixen una gran part de l'alimentació humana. Aquests motius van fer evident la necessitat d'introduir una planta monocotiledònia com a organisme model [63]. L'arròs (*O. sativa* L.) es va proposar com a aquest model ja que és un cereal amb una mida de genoma petita (466 Mbp i 55615 gens), en comparació amb la mida molt més gran del genoma del blat de moro, l'ordi o el blat (aproximadament 3000, 5000 i 16000 Mbp respectivament) [64, 66, 67]. A més, també és fàcil de manipular genèticament, el seu creixement és ràpid i econòmic i té un temps de generació bastant curt (de 3 a 6 mesos) [63, 67, 68]. La seqüenciació completa del genoma de l'arròs es va aconseguir l'any 2002, sent el segon genoma vegetal seqüenciat completament [66, 69].

L'arròs és un organisme vegetal de gran importància alimentaria i econòmica, sent un dels cereals més consumits per la població mundial. Juntament amb el blat (*Triticum aestivum* L.) i el blat de moro (*Zea mays* L.), supleixen més del 50% de les calories consumides pels humans. Cada any es cultiven un total de 154 milions d'hectàrees d'arròs, i el consum humà representa un 85% d'aquesta producció [70]. Existeixen principalment dues subespècies importants d'arròs cultivat, *Oryza sativa ssp. japonica* i *Oryza sativa ssp. indica*. La major part de les poblacions segregades d'arròs es generen a partir de creuaments entre aquestes dues subespècies [69]. L'any 2002 es van seqüenciar completament els genomes de les dues subespècies. La subespècie *indica* està més adaptada a un clima tropical, per això és la més cultivada a les regions del sud d'Àsia i alguns països d'Àfrica i els seus conreus es caracteritzen per créixer majoritàriament submergits. En canvi, la subespècie *japonica* està adaptada a climes més temperats i creix en camps més secs, es cultiva àmpliament a l'est d'Àsia i a les zones altes del sud-est asiàtic [71]. D'una banda, les plantes de la subespècie *japonica* es caracteritzen per ser baixes i amb les fulles de color verd fosc. Els grans d'aquesta subespècie són fins, curts i difícils de trencar. D'altra banda, les plantes de la subespècie *indica* són altes i amb les fulles de color verd clar. Els grans d'aquesta subespècie són

llargs, gruixuts, una mica plans i fàcils de trencar. En aquesta Tesi s'ha utilitzat la subespècie *japonica* com a organisme model en els estudis de metabolòmica d'organismes vegetals. En la Figura 2.5 es mostra una representació esquemàtica de plantes d'arròs d'aquesta subespècie.

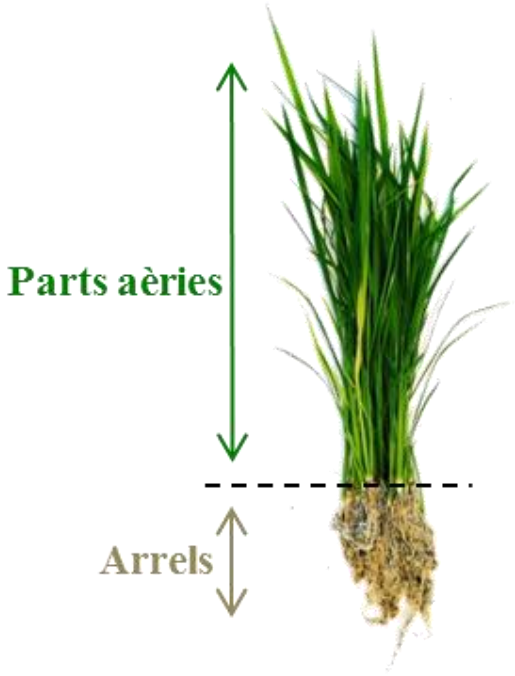


Figura 2.5. Imatge de plantes *Oryza sativa ssp. japonica* utilitzades en aquesta Tesi com a organisme model de treball. En la imatge es mostra la diferenciació de les dues parts de les planta analitzades: parts aèries i arrels.

En la Taula 2.3 es resumeixen alguns estudis dels darrers cinc anys sobre els efectes de l'estrès abiòtic en els lípids i/o metabòlits de l'arròs. En aquesta taula es pot veure que els tipus d'estressos abiòtics estudiats en l'arròs són diversos i que en la majoria de treballs el teixit utilitzat són les fulles. També que la majoria d'estudis utilitzen una anàlisi no dirigida, que és l'aproximació emprada en aquesta Tesi. En el cas dels estudis que utilitzen una aproximació dirigida s'ha especificat la família de metabòlits analitzada.

Taula 2.3. Resum d'estudis sobre els efectes de l'estrès abiòtic sobre els metabòlits i/o lípids de l'arròs publicats entre el 2013 i el 2017.

Tipus d'estrès	Teixit	Anàlisi	Anàlits	Any	Referència
Salinitat					
	Fulles	No dirigida	Metabòlits	2017	[69]
	Fulles	No dirigida	Metabòlits	2017	[70]
	Fulles	No dirigida	Metabòlits	2015	[72]
	Cèl·lules	No dirigida	Metabòlits	2013	[73]
	Llavors	Dirigida	Sucres	2016	[71]
Temperatura					
Alta	Fulles	No dirigida	Metabòlits	2017	[74]
Alta	Fulles	No dirigida	Metabòlits primaris	2015	[75]
Baixa	Llavors	No dirigida	Metabòlits	2016	[76]
Baixa	Part Aèria	No dirigida	Metabòlits	2014	[78]
Baixa	Fulles	No dirigida	Metabòlits	2013	[79]
Baixa	Arrels	Dirigida	Àcids tricarboxílics	2015	[77]
Metalls pesants					
Arsènic	Fulles	Dirigida	Sucres	2016	[71]
Arsènic	Fulles i arrels	Dirigida	Aminoàcids	2013	[80]
Cadmi	Fulles, arrels i tiges	Dirigida	Fitoquelatines	2016	[81]
Pesticides					
	Fulles	No dirigida	Metabòlits i lípids	2014	[84]
	Fulles	Dirigida	Sucres, aminoàcids, àcids tricarboxílics, fenilpropanoids	2016	[82]
	Fulles	Dirigida	Sucres, aminoàcids, àcids tricarboxílics, fenilpropanoids	2015	[83]
Deficiència de nutrients					
Fòsfor	Grans	No dirigida	Metabòlits	2017	[86]
Nitrogen	Fulles	No dirigida	Metabòlits	2014	[87]
Ferro	Arrels	Dirigida	Metabòlits amb grup amina	2017	[85]
Aigua					
Inundació	Part Aèria	No dirigida	Metabòlits	2013	[92]
Sequera	Grans	No dirigida	Metabòlits	2016	[88]
Sequera	Grans	Dirigida	Aminoàcids, àcids grassos, vitamines B1, B2 i E.	2014	[89]
Sequera	Part Aèria	No dirigida	Metabòlits	2014	[78]
Sequera	Fulles	No dirigida	Metabòlits	2013	[90]
Sequera	Fulles	No dirigida	Metabòlits primaris	2013	[91]

2.2. Metodologies analítiques

La gran diversitat en l'estructura i les propietats químiques dels metabòlits presents en els organismes vegetals fa que l'anàlisi metabolòmica de les mostres de plantes sigui molt complexa. En conseqüència, les metodologies analítiques que s'utilitzen en aquests estudis han de tenir una capacitat de separació molt gran i, a més, han de ser altament sensibles i selectives. Les tècniques que més s'empren en metabolòmica d'organismes vegetals són la ressonància magnètica nuclear (RMN) i l'espectrometria de masses (MS) [44, 50, 72].

La RMN és una tècnica apropiada pels estudis de metabolòmica no dirigits, especialment en els casos enfocats a mostres que contenen molts metabòlits semblants (per exemple sucres) [73]. La RMN presenta l'avantatge de produir un senyal que es pot relacionar directament i de forma lineal amb l'abundància dels metabòlits. Tot i així, és menys sensible que les tècniques acoblades a MS i la seva capacitat de detectar metabòlits poc abundants és limitada. A més, a mesura que el pes molecular dels anàlits augmenta, com per exemple en el cas dels lípids amb àcids grassos de cadena llarga, la capacitat d'identificar-los empitjora, degut a que la complexitat i solapament dels senyals augmenta [44, 50, 72].

La MS és una tècnica sensible i versàtil, avantatges que la fan molt útil pels estudis de metabolòmica no dirigida. Aquesta tècnica es pot utilitzar mitjançant injecció directa, però en metabolòmica és més habitual utilitzar-la acoblada a tècniques de separació, com la cromatografia de gasos (GC), la cromatografia de líquids (LC) i l'electroforesi capil·lar (CE) [44, 50, 72].

La CE ofereix una alta eficiència de separació. Una de les propietats úniques de la CE-MS és que requereix molt poca quantitat de mostra (nanolitres de mostra). Malauradament, això implica que la CE té una sensibilitat més baixa que les tècniques cromatogràfiques. D'altra banda, també presenta un altre inconvenient important, com és la baixa reproductibilitat dels temps de migració. En conseqüència, la CE-MS s'utilitza molt menys en estudis de metabolòmica d'organismes vegetals que la GC-MS i la LC-MS [44, 50].

La MS acoblada a GC és la tècnica més emprada tradicionalment en metabolòmica d'organismes vegetals. La GC-MS ofereix una gran capacitat de separació i una alta resolució. És una tècnica robusta i reproduïble per a l'anàlisi d'un ampli rang de metabòlits derivatitzables i tèrmicament estables. A més, el principal avantatge que presenta la GC-MS és que existeixen moltes llibreries de compostos de referència per diferents models d'espectròmetres de masses, les quals resulten molt útils a l'hora d'identificar els metabòlits. Alguns exemples d'aquestes llibreries són la de l'Institut Nacional d'Estàndards i Tecnologia

(NIST), la base de dades del metaboloma de Golm (GMD) [74] i la FiehnLib [75]. Malgrat això, el principal inconvenient que presenta la GC-MS és que no és capaç de detectar els metabòlits termolàbils, com per exemple els di- i tri-fosfats o les lisofosfatidilcolines (LPC), ni els compostos que no són volàtils ni després d'una derivatització, com les fosfatidiletanolamines (PE). Això limita el seu us en els estudis de metabolòmica no dirigits [44, 72]. L'altre gran desavantatge de la GC-MS rau en la necessitat de dur a terme una etapa de derivatització. Aquest pas afegeix més temps de treball i, a més, limita el rang de metabòlits que es poden analitzar, pot causar confusions (per exemple la sililació pot convertir l'arginina en ornitina) i genera espectres de masses més complexos. També augmenta la dificultat a l'hora d'identificar els metabòlits en base a un patró de fragmentació, ja que l'estructura d'aquests pot alterar-se. A més, la derivatització pot donar lloc a la formació de més d'un producte a partir d'un sol metabòlit degut a la presència de més d'un hidrogen actiu a la molècula [2].

En canvi, la LC-MS no requereix la derivatització prèvia de les mostres per a poder analitzar els metabòlits. A més, s'ha demostrat que és una tècnica apropiada per a la detecció d'un ampli rang de classes de metabòlits diferents. Per aquests motius, en aquesta Tesi s'ha utilitzat la LC-MS com a metodologia analítica per a l'anàlisi de metabòlits i lípids. Degut a la gran diversitat de metabòlits del regne vegetal, s'espera que la LC-MS esdevingui la metodologia analítica més utilitzada en els estudis de metabolòmica d'organismes vegetals. El desenvolupament i millora de les tècniques de ionització que permeten l'acoblament de la LC amb els espectròmetres de masses també ha contribuït a que augmenti l'ús d'aquesta tècnica en metabolòmica. A més, els espectròmetres de masses es troben en un continu desenvolupament, oferint cada cop més resolució de massa i precisió, la qual cosa ajuda en la detecció i identificació dels metabòlits [27, 72, 76]. A continuació s'expliquen amb més detall les estratègies que s'han fet servir en aquesta Tesi basades en LC i MS.

2.2.1. Cromatografia de líquids

La LC en columna es va introduir a principis del segle XX, quan el botànic Mikhail S. Tsvet va aconseguir separar pigments de plantes utilitzant com a fase estacionària carbonat de calci i alúmina dins una columna de vidre i com a fase mòbil una mescla d'etanol i èter de petroli [77]. Des d'aleshores la LC ha esdevingut una de les tècniques de separació més rellevants i amb nombroses aplicacions, entre les quals es troba la metabolòmica.

La LC es pot definir com la tècnica analítica que separa els anàlits d'una mescla en funció de les velocitats a les que aquests elueixen d'una fase estacionària típicament durant un gradient de fase mòbil.

La separació dels compostos presents en la mescla es deu a les diferents afinitats que presenten aquests compostos amb les fases mòbil i estacionària. De tal manera que els compostos que tinguin més afinitat amb la fase mòbil eluiran abans, mentre que els compostos que tinguin més afinitat amb la fase estacionària quedaran retinguts durant més temps i eluiran més tard. El temps en el qual un compost elueix (conegut com a temps de retenció) depèn de diversos factors: el tipus de fase estacionària, la fase mòbil utilitzada i les propietats fisicoquímiques del compost (polaritat, mida, càrrega, etc.) [78].

Un dels avantatges de la LC és la seva versatilitat envers el rang de compostos que pot analitzar. Aquesta versatilitat rau en els diversos tipus de fases estacionàries disponibles (fase inversa, fase normal, exclusió iònica, exclusió de mida, interacció hidrofílica, etc.). La cromatografia de líquids de fase inversa (RP-LC) és el mode més àmpliament utilitzat en diferents aplicacions, ja que permet l'anàlisi d'un ampli rang de compostos, des de molècules de baix pes molecular fins a molècules grans, com les proteïnes [2, 72, 76]. En el camp de la lipidòmica la RP-LC és el mode de separació més habitual, degut a les propietats apolars dels lípids. En aquesta Tesi s'ha utilitzat la RP-LC per dur a terme els estudis de lipidòmica de l'arròs. En RP-LC es fan servir fases mòbils més polars que les fases estacionàries. Les fases estacionàries típiques en fase inversa consisteixen en un suport de partícules de sílice químicament lligat a grups funcionals hidrofòbics. Els alquils d'alta densitat (per exemple C_4, C_8 i C_{18}) són els grups funcionals utilitzats més freqüentment. Particularment, en el cas dels estudis de lipidòmica les columnes més habituals són les d'octadecilsilà (C_{18}) i d'octasilà (C_8). Aquestes columnes tenen diversos avantatges com, per exemple, la disponibilitat comercial en diferents mides, un cost relativament baix en comparació amb altres tipus de fases estacionàries i el fet que són compatibles amb les fases mòbils més típiques, que consisteixen en mescles acidificades d'aigua i solvents orgànics (per exemple acetonitril/aigua o metanol/aigua amb 0,1% d'àcid fòrmic), les quals afavoreixen la ionització dels lípids en MS [79, 80]. La separació en el mode RP-LC es basa en la partició dels anàlits entre la fase estacionària no-polar i la fase mòbil polar, de manera que els anàlits no-polars es retenen en la fase estacionària i s'elueixen més lentament que els polars [79, 80].

Molts dels metabòlits presents en els organismes vegetals són altament polars (aminoàcids, amines, àcids orgànics, sucres, etc.) i la capacitat de les columnes de fase inversa (C_8 i C_{18}) per retenir aquest tipus de molècules d'elevada polaritat és limitada i la majoria es perden en el volum mort [2, 72, 76]. Aquests compostos polars es poden analitzar mitjançant altres modes de separació, com la cromatografia de líquids en fase normal (NP-LC) o la cromatografia de bescanvi iònic (IEX). Malauradament aquets modes

presenten diversos inconvenients. Per exemple, en el cas de la NP-LC les fases mòbils utilitzades són poc polars i, per tant, els anàlits molt polars no hi són solubles. En el cas de la IEX, el principal problema és que els solvents utilitzats tenen una alta concentració de sals, la qual cosa dificulta l'acoblament amb MS. Per aquests motius, la cromatografia d'interacció hidrofílica (HILIC) s'ha presentat en els darrers anys com una bona alternativa per a l'anàlisi dels compostos polars [81, 82]. Així, en aquesta Tesi s'ha utilitzat la HILIC en els estudis de metabolòmica d'arròs i, a més, s'han comparat diferents tipus de fases estacionàries HILIC pel seu ús en estudis de metabolòmica no dirigida. Per aquest motiu, aquest mode cromatogràfic s'explica detalladament a continuació.

Cromatografia d'interacció hidrofílica (HILIC)

La HILIC va ser introduïda l'any 1990 per A.J. Alpert [83] i la primera publicació en la que s'utilitza HILIC-MS en un estudi de metabolòmica de plantes és de l'any 2002 [84]. Des d'aleshores fins a l'actualitat el nombre d'estudis d'aquest camp en els que s'utilitza HILIC ha crescut de forma gradual [85-87]. Això es deu a l'avantatge que presenta el fet que la HILIC utilitza fases estacionàries polars semblants a les de NP i fases mòbils similars a les de RP. Aquestes fases mòbils solen estar formades per mescles d'aigua (una quantitat mínima del 2,5%), en la qual els metabòlits polars són solubles, amb solvents orgànics, els quals permeten un augment de la sensibilitat del MS ja que milloren l'eficiència de la ionització. No obstant això, el mode HILIC també presenta alguns inconvenients. Per exemple, les columnes cromatogràfiques requereixen més temps per equilibrar-se i assegurar una bona repetibilitat que els altres modes de separació [81, 82].

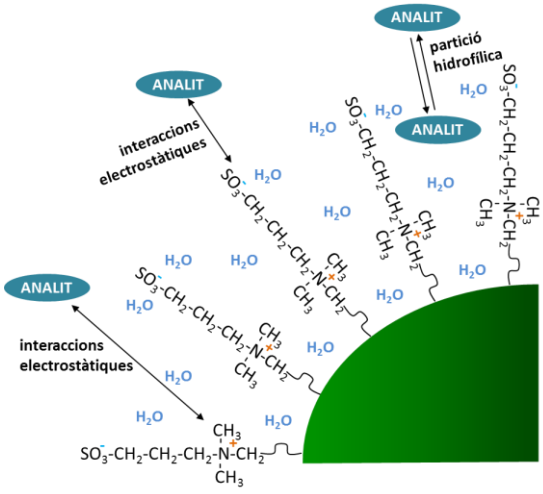


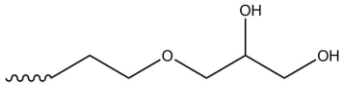
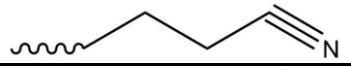
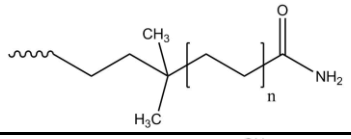
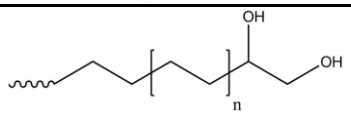
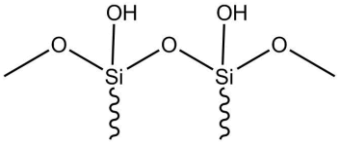
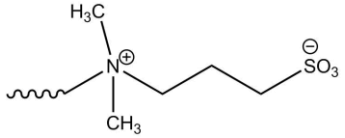
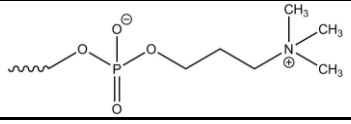
Figura 2.6. Representació esquemàtica del mecanisme de retenció en una fase estacionària HILIC zwitteriònica. La retenció es basa en la partició hidrofílica dels anàlits entre la fase mòbil i la capa hidrofílica de la fase estacionària i en les interaccions electrostàtiques amb les càrregues positives i negatives dels grups funcionals.

En el mode de separació HILIC es crea una capa superficial d'aigua adsorbida sobre la fase estacionària polar. La separació s'aconsegueix per la partició dels anàlits entre la fase mòbil i aquesta

capa hidrofílica. Així, les interaccions per pont d'hidrogen i dipol-dipol són factors importants en el mecanisme de retenció en HILIC. D'altra banda, en les fases estacionàries carregades, les interaccions electrostàtiques per bescanvi iònic també tenen influència en la retenció. En la Figura 2.6 es mostra una representació esquemàtica d'aquest procés de retenció, pel cas concret d'una fase estacionària HILIC zwitteriònica, que presenta càrregues positives i negatives [81, 82, 85].

En els darrers anys, s'han millorat les fases estacionaries HILIC tradicionals i també s'han desenvolupat nous materials per a la separació en HILIC. En conseqüència, avui en dia existeix un ampli ventall de fases estacionàries disponibles per aquest mode de LC. Les fases estacionàries HILIC típiques consisteixen en una columna de sílice o gel de sílice modificat amb grups funcionals polars. Aquestes fases estacionàries es poden classificar en tres grups: neutres, carregades i zwitteriòniques. En la Taula 2.4 es resumeixen les estructures dels grups funcionals de les fases estacionàries més habituals en HILIC i les seves aplicacions [81, 82, 88].

Taula 2.4. Fases estacionàries més habituals en HILIC.

	Fase Estacionària	**Grup Funcional**	**Aplicacions**
Neutra	Diol		Metabòlits polars, proteïnes, vitamines, compostos fenòlics de baix pes molecular
	Ciano		Fàrmacs, metabòlits polars.
	Amida		Oligosacàrids, pèptids, glicoproteïnes, glicòsids.
	Mode Mixt (diol)		Compostos polar i no-polars.
Carregada	Sílice pura		Metabòlits, toxines, contaminants, fàrmacs.
	Amino		Metabòlits polars, aminoàcids, nucleòsids.
Zwitteriònica	Sulfobetaïna		Metabòlits polars, pèptids, fàrmacs (més afinitat amb cations).
	Fosforilcolina		Metabòlits polars, pèptids, fàrmacs (més afinitat amb anions).

Les fases estacionàries neutres contenen grups funcionals polars que no tenen càrrega en el rang de pH habitual de les fases mòbils en HILIC, entre 3 i 8. Això implica que no presenten interaccions electrostàtiques de bescanvi iònic amb cap tipus d'anàlit, de manera que la retenció es basa principalment en la partició hidrofílica. Tot i així, cal tenir en compte que a valors de pH més bàsics, les fases estacionàries de base sílice poden contenir grups silanols residuals desprotonats que aportin una càrrega negativa donant lloc a interaccions electrostàtiques febles.

Les fases estacionàries carregades contenen grups polars amb una càrrega positiva o negativa. En aquest cas, les interaccions electrostàtiques de bescanvi iònic són fortes i tenen un paper important en el mecanisme de retenció dels anàlits. Els anàlits neutres quedaran retinguts per les interaccions hidrofíliques, mentre que els anàlits carregats ho faran principalment gràcies a les interaccions electrostàtiques.

Finalment, les fases estacionàries zwitteriòniques contenen la mateixa proporció de grups amb càrrega negativa i grups amb càrrega positiva. Generalment aquests lligands zwitteriònics, tenen un fort comportament àcid i bàsic independentment del pH de la fase mòbil. Els lligands zwitteriònics afavoreixen l'adsorció de l'aigua a la superfície de la fase estacionària i, per aquest motiu, la partició hidrofílica dels anàlits és el principal mecanisme de retenció. Les interaccions electrostàtiques són més febles que en el cas de les fases carregades però, tot i així, influeixen en l'ordre d'elució [81, 82, 88].

Com ja s'ha esmentat, les fases mòbils en mode HILIC estan formades per mescles d'aigua (mínim el 2,5%) amb solvents orgànics en una proporció alta. Per tal d'afavorir la formació de la capa d'aigua a la superfície de la fase estacionària, el tipus de solvent orgànic ideal ha de ser miscible en aigua. A més, no ha de tenir activitat donadora ni acceptadora de protons, ja que si no podria compatir amb l'aigua per adherir-se a la superfície de la fase estacionària. L'acetonitril compleix aquests requisits i és el solvent orgànic més utilitzat en mode HILIC [81, 82, 88, 89].

El pH i la força iònica de la fase mòbil també són factors que influeixen en la contribució de les possibles interaccions que es poden establir en mode HILIC. És habitual utilitzar additius iònics (solucions amortidores o tampons), com l'acetat o el formiat d'amoni, per a controlar aquests factors. Per exemple, el pH pot afectar a la polaritat dels anàlits i alterar-ne la retenció. A més, un augment en la força iònica (increment en la concentració de la solució amortidora), pot disminuir la retenció quan les interaccions electrostàtiques de bescanvi iònic controlen el mecanisme de separació. Contràriament, quan la partició hidrofílica té la major contribució en el mecanisme de separació, un augment de la

concentració dels additius iònics afavoreix l'adsorció de l'aigua a la superfície de la fase estacionaria i, per tant, augmenta la retenció. També es poden utilitzar altres tipus de sals, com el perclorat de sodi, que augmenten la polaritat de la fase mòbil i afavoreixen l'elució. El principal inconvenient d'utilitzar tampons és que en HILIC-MS poden crear supressió iònica i disminuir la sensibilitat de l'espectròmetre de masses. Finalment, cal comentar que la temperatura de la columna influeix molt poc en la separació en mode HILIC, al contrari del que passa quan es treballa amb RP-LC [81, 82, 88, 89]. En la Taula 2.5 es resumeixen les principals característiques de la HILIC i es compara amb la RP-LC.

Taula 2.5. Resum de les característiques principals de la RP-LC i la HILIC en estudis de metabolòmica.

	RP-LC	**HILIC**
Aplicació	Lipidòmica	Metabolòmica
Fase estacionària	Apolar	Polar
Fase mòbil	Polar	Polar
Mecanisme de retenció	Partició dels anàlits entre les fases mòbil i estacionària.	Partició dels anàlits entre les fases mòbil i estacionària. Interaccions electrostàtiques de bescanvi iònic.
Factors influents	pH de la fase mòbil. Temperatura.	pH de la fase mòbil. Força iònica de la fase mòbil.

A banda del desenvolupament de nous tipus de fases estacionàries, una altra de les tendències en el camp de la LC ha estat la reducció progressiva de la mida de partícula del rebliment de les columnes per aconseguir una millora en l'eficàcia de la separació cromatogràfica. Una reducció de la mida de partícula provoca un augment de l'àrea superficial, la qual cosa millora la retenció dels compostos i, per això, incrementa l'eficàcia de la separació cromatogràfica. En aquest sentit, s'ha passat de treballar amb columnes de partícules d'entre 100 i 200 µm de diàmetre durant la dècada del 1950, a utilitzar partícules esfèriques de 5 µm en la dècada del 1980, d'entre 3 i 3,5 µm durant la dècada del 1990 i, finalment, partícules inferiors a 2 µm (sub-2 µm) durant els primers anys del segle XXI. Malauradament, la disminució de la mida de partícula implica un augment en la pressió de fons del sistema. Per aquest motiu, els instruments de LC també han evolucionat i, actualment, disposen de bombes que poden suportar pressions superiors a 1000 bar. La combinació de les columnes sub-2 µm i instruments capaços de suportar pressions de més de 1000 bar va donar lloc a la cromatografia de líquids de molt alta eficàcia (UHPLC). A part de l'alta eficàcia de la separació cromatogràfica, un altre avantatge molt important de la UHPLC és la rapidesa de les anàlisis. La capacitat de suportar pressions altres permet augmentar la velocitat del flux de la fase mòbil i, per tant, disminuir el temps d'anàlisi [72, 90]. En aquesta Tesi s'ha utilitzat la UHPLC en els estudis de lipidòmica, però no en els de metabolòmica. Això es deu a que quan

es va començar la Tesi, la columna HILIC que es va trobar més apropiada pels treballs de metabolòmica (TSK-gel amida 80) no es trobava disponible comercialment en UHPLC. Actualment, aquesta columna si que està disponible amb una mida de partícula inferior a 2 µm, la qual cosa és un exemple de la continua millora de les fases estacionàries en HILIC.

2.2.2. Cromatografia de líquids bidimensional exhaustiva (LC×LC)

Les mostres analitzades en els estudis de metabolòmica no dirigida són altament complexes. Això és degut a que contenen centenars de compostos i, a més, molts d'aquests compostos tenen propietats i estructures molt similars, el que dificulta la seva resolució cromatogràfica. Per aquest motiu, malgrat els avanços que s'han fet en la LC, no és possible aconseguir una resolució total d'aquestes mostres biològiques complexes utilitzant cromatografia de líquids unidimensional (1D-LC). Una solució a aquest problema és la utilització de sistemes de separació multidimensionals, en els quals cada una de les dimensions es basa en un sistema de separació cromatogràfica diferent [91-93].

La cromatografia de líquids bidimensional (2D-LC) es pot dur a terme tant *off-line* com *on-line*. La primera opció té l'avantatge que és més fàcil d'operar, ja que per exemple no cal preocupar-se per la miscibilitat de les fases mòbils de les dues dimensions. Malauradament, presenta alguns inconvenients importants: requereix molt temps, és difícil de reproduir i és susceptible a la pèrdua i contaminació de la mostra. En canvi, la 2D-LC *on-line*, és més difícil d'operar i necessita interfícies específiques per a la connexió de les dues dimensions. Malgrat això, presenta avantatges destacats, com que és més ràpida, reproduïble i fàcil d'automatitzar. Aquests avantatges són el motiu de que en aquesta Tesi s'hagi utilitzat només l'opció *on-line* [94]. En la Taula 2.6 es resumeixen les principals característiques de les 2D-LC *off-line* i *on-line* [95].

Taula 2.6. Principals característiques de les 2D-LC *off-line* i *on-line.*

	Off-line	***On-line***
Interfície	Col·lector de fraccions	Vàlvula
Temps d'anàlisi total	Molt elevat (d'hores a dies)	Elevat (de minuts a hores)
Temps d'anàlisi a la segona columna	No té cap limitació	Ha de ser curt
Tractament de mostra abans de la segona dimensió	Si	No
Capacitat de pic	Molt elevada si les columnes de les dues dimensions són llargues.	Elevada
Automatització	No, té risc de contaminació o pèrdua de la mostra.	Sí, sense risc de contaminació o pèrdua de la mostra.

La 2D-LC també es pot classificar en termes generals en 2D-LC de talls (*heart-cutting*) o exhaustiva (*comprehensive*), tal com es pot veure en la Figura 2.7. La notació utilitzada en aquesta Tesi en relació a la 2D-LC es basa en l'establert l'any 2012 en el treball de P. J. Marriott, Z. Wu i P. Schoenmakers [96].

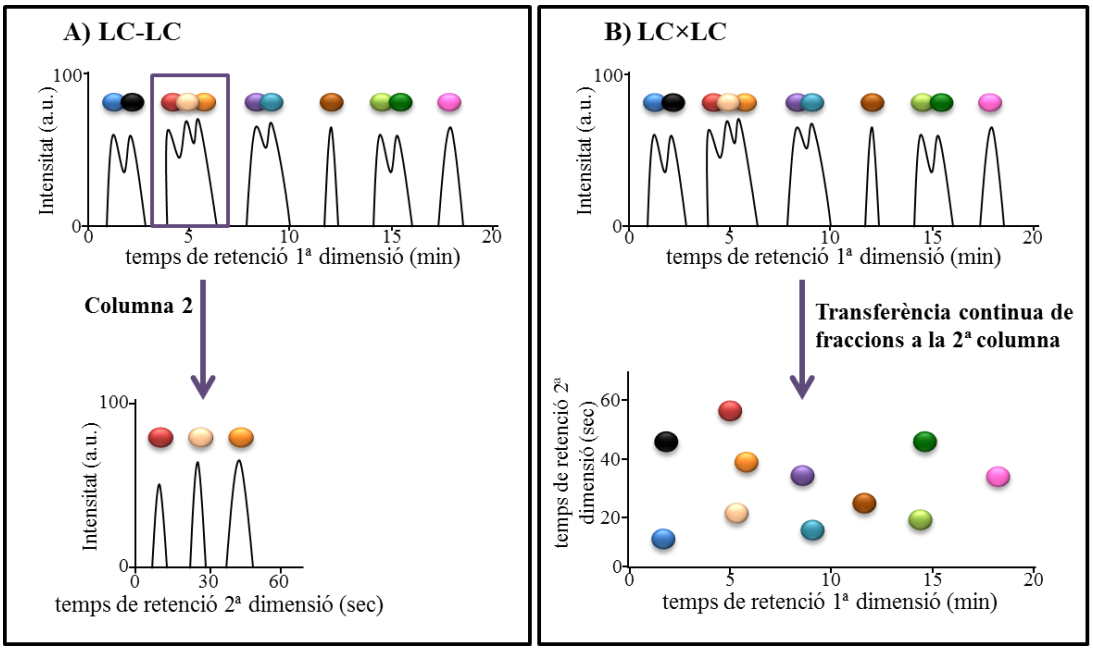


Figura 2.7. Representació de les diferents implementacions generals de la 2D-LC. El panell A mostra el cas de la LC-LC (2D-LC *heart-cutting*), en la qual només les fraccions seleccionades es re-injecten en la segona columna. El panell B mostra el cas de la LC×LC (2D-LC exhaustiva), en la qual la mostra sencera es separa en les dues dimensions cromatogràfiques.

En la 2D-LC *heart-cutting* (LC-LC, Figura 2.7A) es selecciona un nombre concret de fraccions de la primera dimensió de separació que continguin compostos solapats i només aquestes fraccions es re-injecten en la segona dimensió, en la qual s'aconsegueix la seva separació. La LC-LC és una implementació simple i fàcil de la 2D-LC, però és limitada en termes de l'àmbit d'aplicació i del nombre de compostos que pot resoldre. L'aplicació més típica de la LC-LC és la quantificació d'un nombre reduït d'anàlits en una matriu complexa [91].

En la 2D-LC exhaustiva (LC×LC, Figura 2.7B) la mostra sencera passa per les dues dimensions de separació [91, 93]. L'objectiu principal de la LC×LC és aconseguir la màxima separació possible de les mostres fent servir les dues dimensions. Les aplicacions més típiques son les anàlisis de mostres complexes, com és el cas dels estudis de metabolòmica, en les quals es busca separar el màxim possible els compostos presents en la mostra per tal de facilitar-ne la detecció. A més, en aquests estudis no dirigits no es disposa d'informació prèvia sobre els compostos presents en la mostra, la qual cosa fa que no sigui recomanable utilitzar la LC-LC ja que no és possible la selecció *a priori* de les fraccions més

interesants. Per aquests motius, en aquesta Tesi només s'ha utilitzat la LC×LC, la qual a continuació s'explica més detalladament [91].

La LC×LC presenta dos avantatges importants. El primer està relacionat amb un augment en la capacitat de resolució en comparació als sistemes unidimensionals. En aquest punt és important tenir en compte la capacitat de pic, que és el paràmetre més utilitzat per a mesurar la limitació de la 1D-LC per a la resolució de mostres altament complexes. La capacitat de pic és un paràmetre teòric que estima el nombre màxim de pics que es poden ajustar a la mateixa resolució (normalment 1,0) en un cromatograma entre el volum mort i l'últim pic. Recentment s'ha estimat que la màxima capacitat de pic que es pot aconseguir en 1D-LC per a molècules petites varia des de 100 en anàlisis de 5 a 10 minuts fins a alguns centenars en anàlisis de poques hores [97]. En canvi, una capacitat de pic al voltant de 3000 es pot assolir en LC×LC en un temps d'anàlisi d'entre 1 i 2 hores [91, 97]. Aquesta superioritat en la capacitat de pic de la LC×LC es deu a que en circumstàncies ideals, quan els mecanismes de separació de les dues dimensions són totalment ortogonals (independents i complementaris), la capacitat de pic total (n_{2D}) és igual al producte de les capacitats de pic de la primera (n_1) i la segona (n_2) dimensions. A la pràctica, aquesta ortogonalitat total és molt difícil d'aconseguir, però es pot arribar a una capacitat de pic molt elevada si s'utilitzen dos modes de separació poc correlacionats, com per exemple RP i HILIC. L'altre avantatge important de la LC×LC és que té una gran capacitat a l'hora d'identificar els metabòlits, ja que aporta informació de dos mecanismes de retenció diferents. A més, en els cromatogrames bidimensionals s'ha observat que els pics es poden ordenar al llarg de vectors o arcs en relació als grups funcionals dels anàlits (alcans, aldehids, grau de insaturació, etc.), la qual cosa és molt útil a l'hora d'identificar els compostos desconeguts [91].

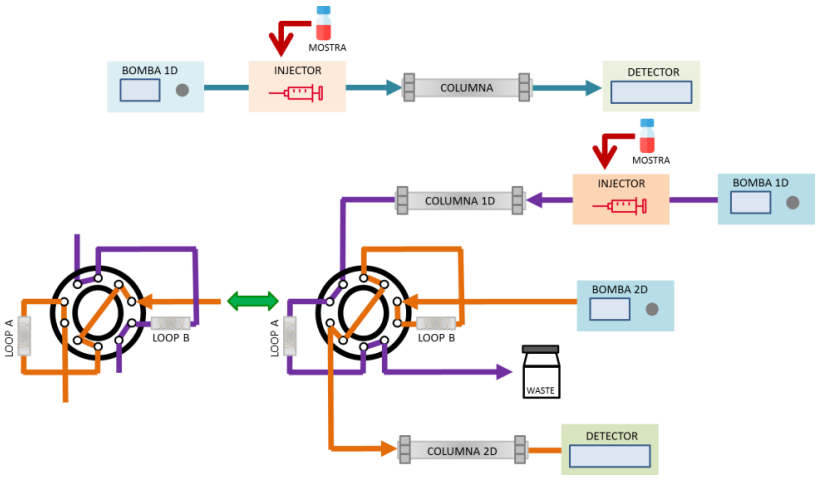


Figura 2.8. Diagrama d'un sistema de LC×LC comú. El modulador representat és una vàlvula de 10 ports i dues posicions. La doble fletxa verda indica el canvi de posició del modulador.

A la pràctica, la LC×LC es porta a terme mitjançant dos sistemes de HPLC equipats amb dues columnes cromatogràfiques connectades mitjançant una interfície. En la Figura 2.8 es representa el diagrama d'un sistema comú de LC×LC.

La interfície que connecta les dues columnes s'anomena modulador i habitualment és una vàlvula de 8 o 10 ports i dues posicions. Aquest modulador talla les fraccions que elueixen de la primera columna i les condueix cap a la segona columna. Aquests talls fets pel modulador es coneixen amb el nom de modulacions. La doble fletxa verda de la Figura 2.8 indica el canvi de posició del modulador, que permet el canvi de funció de cadascun dels bucles (*loops*) entre recollir la mostra que elueix de la primera columna i injectar cada modulació en la segona columna. Un dels requeriments de la LC×LC és que tota la mostra ha de passar pel mecanisme de separació de les dues dimensions i, per tant, és important conservar la separació aconseguida en la primera dimensió. Per a poder mantenir aquesta separació, es recomana que cada pic de la primera dimensió es re-injecti a la segona dimensió en un mínim de 3 modulacions. Un altre requeriment és que la fracció injectada a la segona columna s'ha de separar complement abans que entri la següent modulació. Com a conseqüència d'això, la separació a la segona columna acostuma a ser molt ràpida, ja que el temps d'anàlisi de la segona dimensió ha de ser menor o igual al temps de modulació [91-93].

En l'àmbit de les ciències òmiques, la LC×LC s'ha utilitzat sobretot en estudis de proteòmica. En la bibliografia es troben aplicacions de la LC×LC per a proteòmica on s'acostuma a determinar el perfil peptídic de proteïnes utilitzades com a fàrmacs [98, 99]. Al contrari que en el cas de la proteòmica, l'aplicació de LC×LC en estudis de metabolòmica és molt recent. A més, la majoria d'aquests treballs estan orientats a la innovació i la millora de l'aplicació pràctica de la LC×LC. Per exemple, l'any 2013 P. Jandera i els seus col·laboradors van publicar l'anàlisi d'una mescla de flavonoides per LC×LC [100]. En aquest treball es va destacar l'ús d'un nou tipus de columnes monolítiques. Un altre exemple d'aquestes publicacions és el treball de l'any 2015 de M. Holčapek i els seus col·laboradors, en el que van publicar un nou mètode de LC×LC per analitzar els lípids de mostres complexes (plasma humà i cervell porcí) [101]. En aquest treball, d'una banda es va comparar la utilització de la LC×LC *on-line* amb mètodes anteriors que utilitzaven una configuració *off-line*. D'altra banda es va proposar l'ús de columnes HILIC en la segona dimensió, fet que no és habitual.

Tot i que la utilització de LC×LC en metabolòmica no dirigida té un gran potencial, aquesta àrea encara està en desenvolupament. La majoria de publicacions de la bibliografia en les que s'utilitza 2D-LC

en l'àmbit de la metabolòmica no dirigida utilitzen configuracions alternatives a la LC×LC *on-line*, com la LC-LC o la 2D-LC *off-line* [102-104].

2.2.3. Espectrometria de masses

Durant la dècada passada es van desenvolupar i millorar diverses tecnologies de MS. L'ús d'aquestes tecnologies es va generalitzar en diferents àmbits de la ciència. En metabolòmica, s'ha demostrat la seva utilitat en la detecció i caracterització de compostos biològics a nivells de concentració molt baixos [105]. En comparació amb altres detectors, com l'absorció molecular, l'ultraviolat-visible (UV-Vis) o el de fluorescència, els espectròmetres de masses són més universals i versàtils [2, 27, 72, 105].

Breument, la MS detecta els valors de la relació de massa i càrrega (*m/z*) i l'abundància dels diversos anàlits generats durant la ionització d'una mostra o fracció cromatogràfica. La informació que aporten els patrons de fragmentació obtinguts per MS en tàndem (MS/MS) i l'elevada exactitud de les lectures de *m/z* de la MS d'alta resolució (HRMS) permeten l'elucidació estructural i en alguns casos la completa identificació dels metabòlits [72, 105].

Un espectròmetre de masses és un instrument analític que consta principalment de tres parts: la font de ionització, l'analitzador i el detector. A la font de ionització els anàlits de la fase mòbil es transfereixen directament a ions en fase gasosa. La ionització és un pas clau, ja que els ions són molt més fàcils de manipular que les molècules neutres. Un cop ionitzats, l'analitzador separa els ions en funció de la seva *m/z* i, finalment, el detector enregistra la càrrega induïda o el corrent produït quan un ió passa o colpeja una superfície [72, 105]. A continuació s'expliquen amb més detall la font de ionització i els analitzadors que s'han emprat en aquesta Tesi.

Font de Ionització. Electrosprai

L'electrosprai (ESI) es considera una tècnica de ionització tova, ja que aplica molt poca energia sobre l'anàlit i, per tant, pràcticament no es produeix cap fragmentació durant el procés de ionització. Avui en dia, l'ESI és la font de ionització més emprada en l'anàlisi de mostres líquides. És una molt bona opció quan la MS s'acobla a LC, ja que ionitza directament les molècules en la fase líquida. També cal destacar que és una font de ionització universal i amb molt poca especificitat química [106, 107].

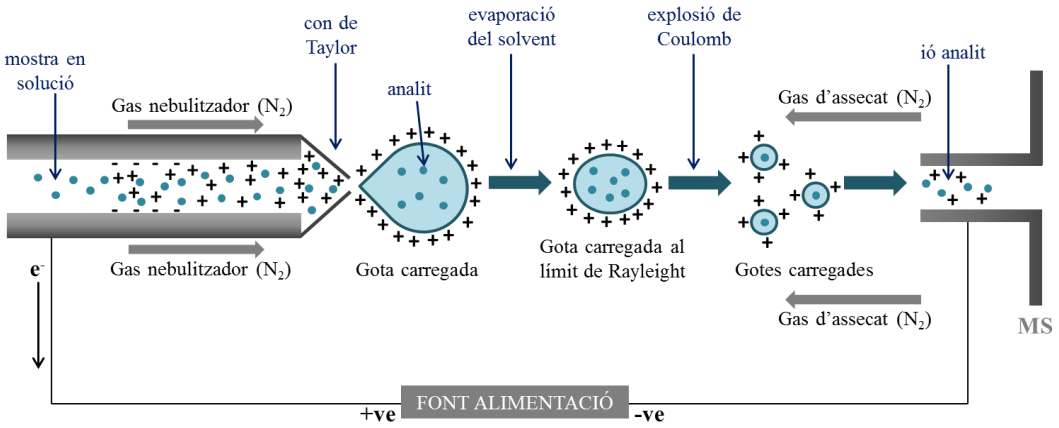


Figura 2.9. Descripció del procés d'obtenció dels ions en fase gasosa a partir de molècules en solució en una interfície d'ESI per a LC-MS.

En la Figura 2.9 es mostra esquemàticament el procés de formació dels ions en fase gasosa a partir de les molècules en solució, el qual s'explica en tres passos principals. En un primer pas, la solució que conté els anàlits passa per un capil·lar metàl·lic al qual se li aplica un potencial elèctric (habitualment entre 2 i 4.5 kV). Aquest potencial promou una migració de càrregues entre la superfície del capil·lar i la solució, que dóna lloc a una doble capa elèctrica. Com a resultat d'això, es formen gotes carregades al final del capil·lar. El següent pas és l'evaporació del solvent, que s'aconsegueix fent passar gas (habitualment N_2) en la direcció oposada. La reducció de la mida de les gotes carregades es produeix fins al moment en que les gotes arriben al límit de Rayleight, que és l'estat en el que la tensió superficial que manté unida una gota carregada és igual a la repulsió de Coulomb entre les càrregues de la superfície. El tercer pas correspon a aquesta explosió de Coulomb. Amb la reducció de la mida de les gotes, les càrregues s'aproximen entre sí i les forces de repulsió electrostàtiques augmenten. Conseqüentment, la tensió superficial de les gotes disminueix produint que aquestes es trenquin. El trencament ocasiona la formació de gotes més petites i estables, les quals es sotmeten successivament a diverses rondes de dessolvatació fins a aconseguir l'alliberament dels ions a la fase gasosa [106-108].

Diversos paràmetres influeixen en l'eficiència de la formació de l'esprai i, per tant, en la sensibilitat del mètode. El més important d'aquests paràmetres és el voltatge de con. Malgrat que l'ESI es considera una font de ionització tova, cal vigilar amb el potencial que s'aplica en la formació de l'esprai quan s'analitzen molècules làbils. Cal trobar un terme mig entre aplicar un voltatge massa elevat, que pugui ocasionar fragmentació, o un voltatge tan baix que disminueixi considerablement la sensibilitat. Altres paràmetres important són la composició i el pH de la fase mòbil, així com el nombre d'espècies presents

en la mostra. Una concentració elevada de sals o espècies diferents pot portar a la supressió iònica i, per tant, disminuir la sensibilitat de la mesura [72, 106-108].

Analitzadors

L'analitzador és la part de l'espectròmetre de masses en la qual els ions es separen en funció del valor de la seva relació *m/z*. És una de les parts més importants de l'instrument, ja que d'ell depenen en gran mesura característiques com la selectivitat i la sensibilitat, així com les possibilitats de mode de treball. Avui en dia existeixen diversos tipus d'analitzadors, els quals varien en termes de resolució, rang de massa i possibilitat de realitzar MS/MS. La majoria d'aquests espectròmetres es van desenvolupar cap a la meitat del segle XX, però en l'última dècada s'han desenvolupat alguns analitzadors com l'Orbitrap, i d'altres com el temps de vol (TOF) han evolucionat. Això ha facilitat la utilització de l'espectrometria de masses d'alta resolució (HRMS) de manera robusta i rutinària. En el camp de la metabolòmica, els més utilitzats són el triple quadrupol (QqQ), el TOF, l'Orbitrap i el de ressonància ciclotrònica de ions amb transformada de Fourier (FTICR). A continuació es descriuen amb més detall els analitzador utilitzats en aquesta Tesi.

Triple Quadrupol (QqQ).

Un QqQ és un espectròmetre de masses en tàndem format per dos analitzadors quadrupol en sèrie (Q1 i Q3) amb una cel·la de fragmentació entre ells, la qual correspon a un quadrupol de radiofreqüència (q2). Essencialment, un triple quadrupol opera sota els mateixos principis que un quadrupol simple. D'una banda, cadascun dels dos filtres de masses (Q1 i Q3) contenen quatre barres de metall paral·leles. Cada parell de barres oposades es connecten entre si elèctricament. S'aplica un potencial de corrent continu (*U*) a un dels parells de les barres mentre que l'altre parell està lligat a un potencial altern (*V*) de radiofreqüència (*w*). D'aquesta manera els ions oscil·len entre les barres i arriben fins al detector. Per uns determinats valors de *U*, *V* i *w*, només els ions amb un valor de *m/z* dins d'un rang concret podran arribar fins al detector, la resta de ions tenen trajectòries inestables i xoquen amb les barres metàl·liques. Això permet tant la selecció d'un ió amb un valor de *m/z* concret, com l'exploració d'una gamma de valors de *m/z* a partir d'una rampa dels valors de *U*, *V* i *w*. En la Figura 2.10A es mostra un esquema del funcionament d'aquest quadrupol. D'altra banda, la cel·la de col·lisió (q2) és un quadrupol només sotmès a radiofreqüència [72, 109, 110].

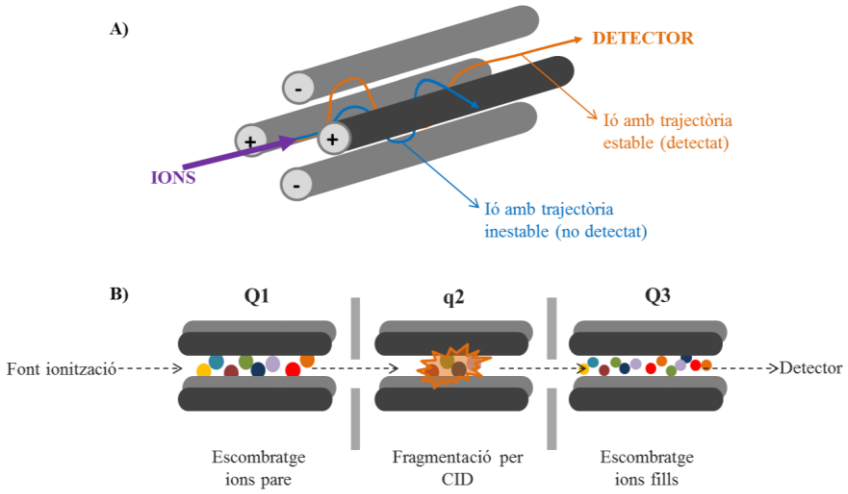


Figura 2.10. A) Esquema del funcionament del quadrupol. B) Esquema de l'arranjament d'un QqQ.

En el QqQ els ions que surten de la font de ionització entren al primer quadrupol (Q1), en el qual es seleccionen els ions pares. Aquests ions pares entren a la cel·la de col·lisió (q2) on es fragmenten. Finalment, aquests fragments s'analitzen en el tercer quadrupol (Q3). La Figura 2.10B mostra un esquema del funcionament del QqQ [72, 110].

L'arranjament del QqQ permet treballar en quatre modes d'anàlisi diferents. En aquesta Tesi s'ha utilitzat l'exploració dels ions producte, en el qual es selecciona un ió pare de massa coneguda en el Q1 que es fragmenta en la cel·la de col·lisió (q2) i, finalment, en el Q3 s'enregistren tots els fragments generats. Aquest mètode generalment s'utilitza per la identificació de compostos, ja que l'estructura de l'ió pare pot deduir-se a partir dels fragments trobats [72, 110].

Temps de Vol (TOF).

L'analitzador de TOF és un dels analitzadors de masses d'alta resolució més antics i més emprats avui en dia. La metodologia de separació de ions que utilitza el TOF és una de les més simples i es basa en el vol lliure de les molècules ionitzades en un tub de 1-2 m de llargada abans d'arribar al detector. L'analitzador TOF permet establir una relació directa entre el valor de *m/z* i aquest temps de vol. Així, els ions amb una massa petita i una càrrega elevada travessaran el tub més ràpid que els ions amb una massa més gran i una càrrega més baixa [109, 111].

En els seus inicis, la resolució que oferia el TOF podia ser insuficient en alguns casos. Aquest problema es va solucionar amb l'ús de reflectors, que permeten aconseguir resolucions de fins 60000 FWHM (amplada de pic a mitja alçada) a *m/z* 1222. Aquesta utilització de reflectors va solucionar la influència de la linearitat del tub d'un TOF convencional en el poder de resolució. Els ions que entren al

TOF tenen diferents energies cinètiques i això pot afectar la resolució i la mesura dels ions moleculars. Els reflectors són instruments òptics que modifiquen el feix de ions dins el TOF, tal com es mostra en la Figura 2.11. Els ions amb una energia cinètica més gran penetren més profundament en el reflector que els ions amb una energia cinètica menor, d'aquesta manera els ions es repelen gradualment, millorant així la resolució del TOF. A més, l'ús de reflectors també augmenta la distància al detector, la qual cosa també contribueix en una millora en la resolució i en la precisió de les mesures [109, 111].

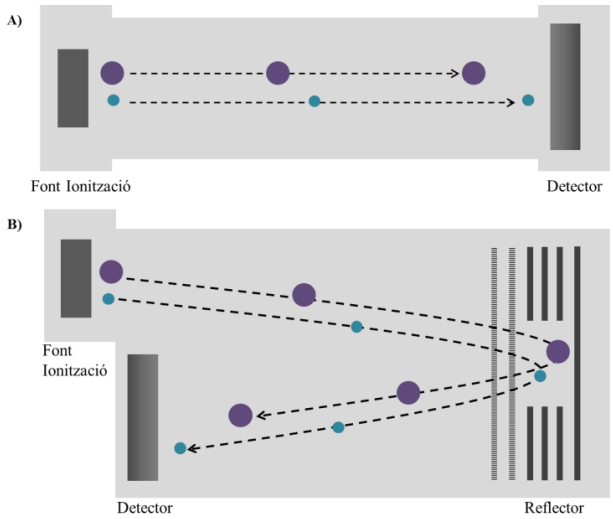


Figura 2.11. Representació esquemàtica d'un TOF convencional (A) i d'un TOF amb reflector (B).

Un dels principals avantatges del TOF és que tots els ions que es formen a la font arriben al detector, al contrari que en el cas del quadrupol. A més, és capaç d'analitzar un ampli rang de masses. Tot i així, el principal inconvenient del TOF és que per tal d'obtenir una bona exactitud en la mesura dels valors de *m/z* (error inferior a 5 ppm) cal utilitzar un calibratge intern en l'eix de masses o aplicar correccions post-anàlisi [109, 111].

Orbitrap.

L'Orbitrap és un analitzador de masses d'alta resolució presentat per Makarov l'any 2000 [112]. Aquest analitzador consisteix en un elèctrode exterior en forma de barril i un elèctrode interior coaxial en forma de fus que forma un camp electrostàtic amb potencial quadro-logarítmic [111, 113].

En un Orbitrap, els ions són injectats dins el camp elèctric generat entre els dos elèctrodes i queden atrapats perquè la seva atracció electrostàtica cap a l'elèctrode interior es contraresta per forces centrífugues. D'aquesta manera, els ions donen voltes al voltant de l'elèctrode central en òrbites, alhora que es mouen endavant i endarrere al llarg de l'eix d'aquest elèctrode. Així, els ions amb una *m/z* determinada es mouen en anells que oscil·len al voltant del fus central, on la freqüència d'aquestes

oscil·lacions harmòniques és independent de la velocitat de l'ió i és inversament proporcional a l'arrel quadrada de *m/z* [111, 113]. En la Figura 2.12 es mostra una representació esquemàtica d'aquests elèctrodes.

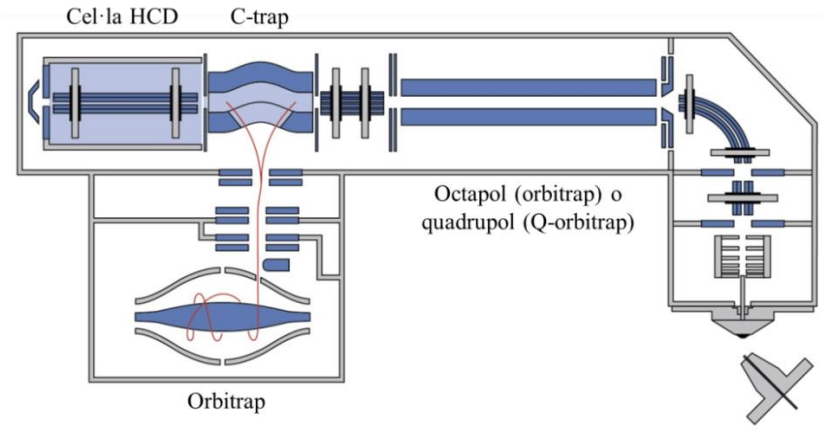


Figura 2.12. Representació esquemàtica de l'espectròmetre de masses amb analitzador Q-orbitrap (imatge adaptada de ThermoFisher Scientific, http://planetorbitrap.com).

L'Orbitrap pot assolir resolucions de 100000 a 240000 FWHM a *m/z* 200 i proporciona una molt bona exactitud de la mesura de la massa, habitualment amb un error inferior a 3 ppm. A més, presenta una gran estabilitat en l'eix de masses (és possible el calibratge extern). En comparació amb els altres analitzadors, actualment només el de FTICR és capaç d'aconseguir una resolució espectral de la mateixa magnitud que l'Orbitrap [111].

Finalment, és important destacar que l'Orbitrap treballa en el mode d'escombratge de ions (*full scan)* i no permet dur a terme la fragmentació dels ions dins el mateix analitzador. Per tal de superar aquesta limitació, els sistemes que incorporen aquest analitzador disposen d'una cel·la de dissociació HCD (*Higher-energy Collision Dissociaton*, Figura 2.12) que permet dur a terme la fragmentació dels ions i obtenir-ne informació estructural addicional.

El model d'Orbitrap que s'ha utilitzat en aquesta Tesi és l'híbrid quadrupol-orbitrap (Q-Orbitrap). Aquest analitzador combina la velocitat d'escombratge d'un quadrupol amb l'alta resolució de l'Orbitrap. La particularitat del Q-Orbitrap és que la fragmentació es pot realitzar seleccionant l'ió precursor al quadrupol per fragmentar-lo a la cel·la HCD (anàlisi MS/MS dirigida), o bé es pot utilitzar el quadrupol en mode de transmissió i fragmentar tots els ions simultàniament a la cel·la HCD (mode *all-ion fragmentation*, AIF). La utilitat d'aquest analitzador en estudis de metabolòmica no dirigida resulta evident, ja que facilitat l'elucidació estructural i la identificació de compostos desconeguts a partir de la seva fragmentació i de les lectures de massa en alta resolució.

2.3. Metodologies quimiomètriques

Els estudis de metabolòmica no dirigida generen conjunts de dades complexos. Aquesta complexitat de les dades requereix l'ús d'eines quimiomètriques per a poder fer la seva anàlisi. En les següents seccions es descriuran els mètodes emprats en aquesta Tesi per a realitzar aquest tractament de les dades. En primer lloc s'explica la naturalesa de les dades metabolòmiques obtingudes. Seguidament, es descriuen els mètodes emprats pel preprocessament de les dades i per a la resolució de pics cromatogràfics. Finalment, s'expliquen els mètodes quimiomètrics d'exploració, regressió i classificació de dades experimentals de metabolòmica

2.3.1. Naturalesa de les dades

Les dades obtingudes en els estudis de metabolòmica es classifiquen segons la seva complexitat en:

- unidireccionals (*one-way data*)
- bidireccionals (*two-way data*)
- tridireccionals (*three-way data*)
- multidireccionals (*multi-way data*)

En aquesta Tesi s'ha preferit utilitzar el terme direcció al de dimensió per definir l'estructura de les dades estudiades. El terme dimensió s'ha deixat per expressar la mida de les diferents estructures de dades: dimensions d'un vector de dades com el seu nombre d'elements, d'una matriu de dades com el seu nombre de files i de columnes i així successivament. En el cas cromatogràfic el terme dimensió s'utilitza per descriure les diferents columnes cromatogràfiques utilitzades de forma acoblada en una mateixa anàlisi (veure més endavant).

El cas més senzill és el de les dades unidireccionals, que correspon a la descripció d'una sola mostra per una sèrie de mesures ordenades en un vector. Per exemple, l'anàlisi d'una mostra mitjançant MS en infusió directa proporciona dades ordenades en una direcció. Quan s'analitzen vàries mostres, els vectors obtinguts es poden ordenar en forma de matriu de dades generant dades bidireccionals. Aquesta matriu té tantes files com mostres analitzades i tantes columnes com variables mesurades. L'anàlisi de dades ordenades en més d'una direcció requereix l'ús de mètodes multivariants.

Un exemple de dades bidireccionals són les tècniques de cromatografia de líquids acoblades a espectrometria de masses (LC-MS). Cada mostra es descriu mitjançant dues direccions, una que conté la informació dels temps de retenció i l'altra que representa els espectres de masses. Aquestes dades es representen en forma de matriu (**D**), en la qual els valors de *m/z* analitzats es troben en la direcció *x* de la

matriu i els temps de retenció s'ubiquen a la direcció *y* de la mateixa. En la figura 2.13A es mostra una representació d'aquesta matriu de dades de LC-MS, en la qual es pot veure que per cada valor de *m/z* (columna) s'obté el cromatograma dels seus compostos corresponents. Mentre que per cada temps de retenció (fila) es té l'espectre de masses individual. Quan es considera més d'una mostra mesurada per LC-MS, el conjunt de dades obtingut és tridireccional (usualment representat per un cub, Figura 2.13B).

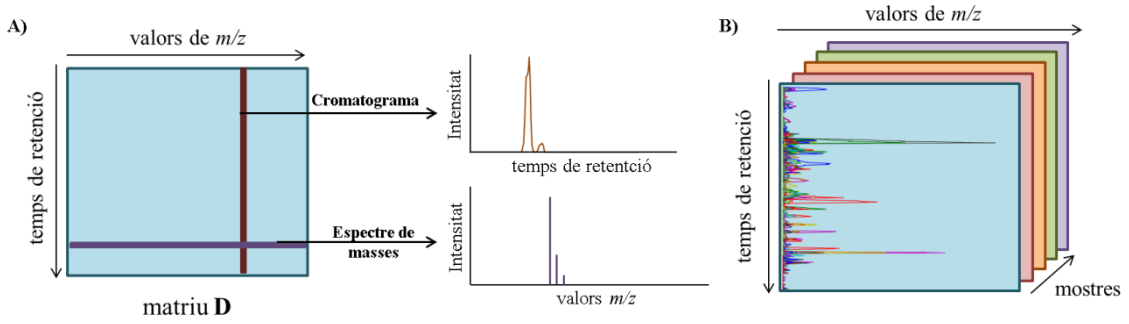


Figura 2.13. A) Representació d'una matriu de dades LC-MS (matriu **D**). B) Representació d'un cub de dades per vàries mostres analitzades mitjançant LC-MS.

Finalment, les tècniques de separació multidimensional, com la cromatografia de líquids bidimensional acoblada a espectrometria de masses (LC×LC-MS), proporcionen dades tridireccionals. En aquest cas, cada mostra es descriu mitjançant tres direccions, dues que contenen la informació de les dues dimensions o modes cromatogràfics i l'altra que representa els espectres de masses. Aquestes dades s'ordenen en un cub (**B**), en el qual l'eix *x* correspon als espectres de masses, els temps de retenció de la segona dimensió cromatogràfica s'ubiquen en l'eix *y* i els temps de retenció de la primer dimensió cromatogràfica es representen en la dimensió *z*. En la Figura 2.14A es mostra una representació d'aquest cub de dades de LC×LC-MS, en la qual es pot veure que un tall d'aquest cub a un valor determinat de *m/z* correspon al cromatograma bidimensional dels compostos corresponents a aquest valor de *m/z*. A més, el cromatograma bidimensional TIC (*total ion corrent*) es pot obtenir en sumar les intensitats de tots els valors de *m/z* en les dues dimensions cromatogràfiques. També és possible estudiar l'espectre de masses individual per cada combinació de les dues dimensions cromatogràfiques. Finalment, els talls del cub a un temps concret de la primera dimensió donen les modulacions. Cada modulació és una separació cromatogràfica sencera en la segona columna, i que permet construir una matriu de dades de LC-MS (matriu $\mathbf{B_m}$, Figura 2.14A). Aquesta matriu conté els valors de *m/z* a les columnes i els temps de retenció de la segona dimensió a les files.

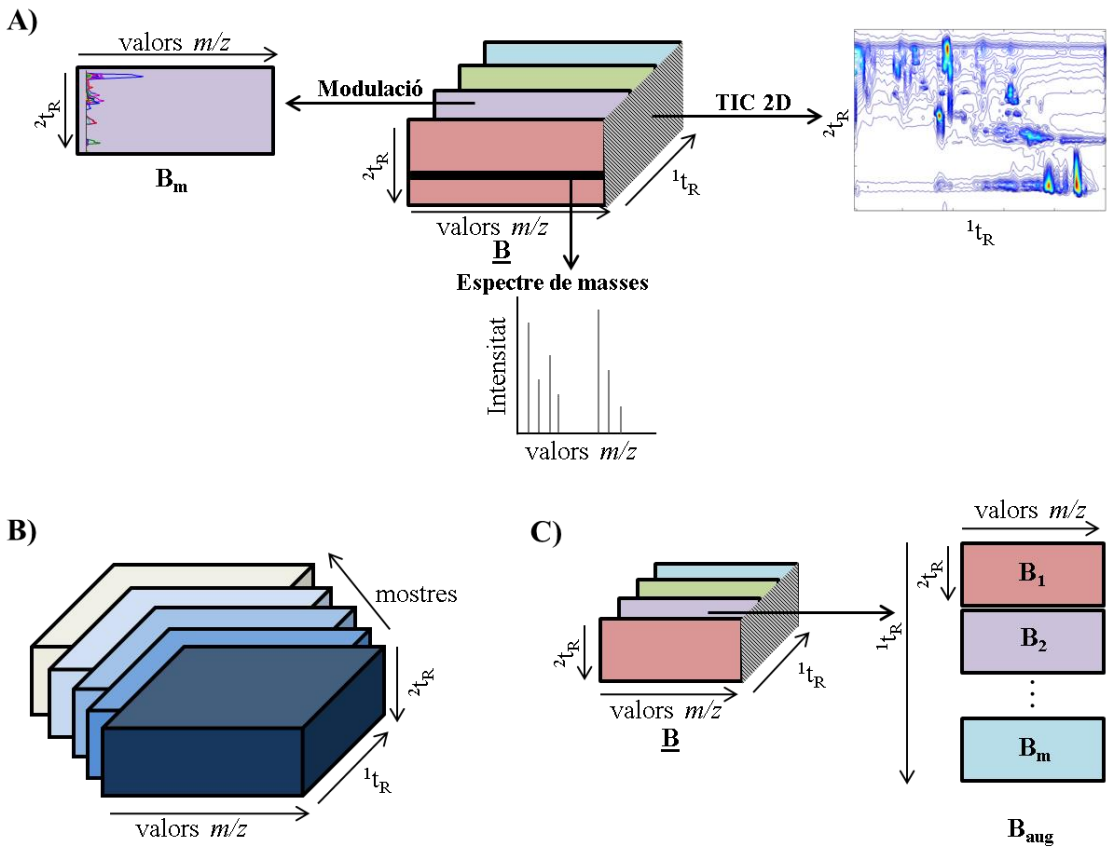


Figura 2.14. A) Representació d'un cub de dades LC×LC-MS (cub **B**), mostrant una matriu de dades LC-MS per cada modulació, un espectre de masses per cada combinació de les dues dimensions cromatogràfiques i un cromatograma bidimensional considerant el costat del cub. B) Representació d'un hipercub de dades per vàries mostres analitzades mitjançant LC×LC-MS. C) Dades LC×LC-MS ordenades en forma de matriu augmentada en la direcció de les columnes (*column-wise*).

En el cas de les tècniques de separació multidimensionals, quan s'analitzen vàries mostres s'obtenen conjunts de dades multidireccionals. Aquests conjunts de dades tenen quatre direccions i geomètricament es representen per hipercubs. En la figura 2.14B es mostra una possible representació gràfica d'aquests hipercubs.

La majoria de les eines d'anàlisi multivariant i d'àlgebra lineal que s'han utilitzat en aquesta Tesi requereixen que les dades es trobin ordenades únicament en dues direccions. Per aquest motiu, les dades de LC×LC-MS també s'han arranjat en forma de matriu augmentada en la direcció de les columnes (matriu $\mathbf{B}_{aug}$, Figura 2.14C). Aquesta matriu es construeix desplegant el cub **B** en la direcció de l'eix *x* (temps de retenció de la segona columna). Això s'aconsegueix col·locant les matrius individuals de cada modulació (matrius $\mathbf{B}_m$) una sota de l'altra, mantenint els valors de *m/z* en comú. En la Figura 2.14C es representa esquemàticament aquesta matriu.

Cal diferenciar l'estructura de les dades multidireccionals i els models matemàtics que s'utilitzen pel seu estudi. En general, s'empraran models matemàtics bilineals, ja que són els que s'adapten més fàcilment a les dades de tipus cromatogràfic estudiades en aquesta Tesi. De tota manera, quan es tracten dades multidireccionals cal tenir en compte la possibilitat de la seva modelització mitjançant mètodes multilineals. Per exemple, en el cas de dades generades per LC×LC-MS, la condició de trilinealitat només es compleix si els perfils d'elució d'un mateix component en les diferents modulacions tenen la mateixa forma i apareixen en el mateix temps de retenció, de manera que només varien entre modulacions en un factor d'escala. El compliment o no d'aquesta condició influeix en la selecció del mètode d'anàlisi més adient per cada conjunt de dades. La trilinealitat o multilinealitat de les dades LC×LC-MS és un tema de controvèrsia en l'àmbit de la quimiometria i es discutirà en detall més endavant en aquesta Tesi.

2.3.2. Mètodes de preprocessament de les dades

Sovint és necessari dur a terme tractaments previs (preprocessament) que permetin millorar la qualitat de les dades abans de procedir a la seva anàlisi. En les dades de LC-MS, el nivell de soroll i les irregularitats en la forma de la línia base són problemes típics que deterioren els senyals originals dels anàlits. A banda d'aquesta millora en la qualitat, el preprocessament de les dades també és rellevant per a reduir-ne la mida i així facilitar el seu tractament. Aquesta darrera propietat és evident sobretot en els grans conjunts de dades multivariants, com és el cas dels generats en els estudis metabolòmics [114]. A continuació es descriuen els pretractaments que s'han emprat en aquesta Tesi.

Compressió i construcció de les matrius de dades

Les estructures de dades descrites en la secció anterior es construeixen a partir de les dades cromatogràfiques experimentals originals (dades *raw*) mitjançant diverses estratègies. L'anàlisi de conjunts de dades multivariants de gran mida requereix ordinadors amb una alta capacitat d'emmagatzematge i processament per a poder tractar-les. Per tal de facilitar aquest processament, les estratègies utilitzades en aquesta Tesi per a la construcció de les matrius de dades de LC-MS també permeten la reducció de la mida dels conjunts de dades originals sense perdre informació química o biològica important. La manipulació de les dades *raw* no és difícil només per la seva gran mida, sinó que a més els conjunts inicials de dades LC-MS contenen per a cada temps de retenció un nombre de valors *m/z* separats de manera no equidistant. En organitzar aquestes dades en forma de matriu, a les files es representen cadascun dels temps de retenció i a les columnes els valors de *m/z*, que han de ser comuns per

tots els temps de retenció. Les estratègies de compressió de les dades es poden diferenciar entre les que comprimeixen la direcció espectral (columnes) i les que comprimeixen la direcció cromatogràfica (files). Les estratègies de compressió utilitzades en aquesta Tesi s'expliquen a continuació.

Binning

El *binning* és un dels procediments més senzills i utilitzats per a la compressió de les dades en la direcció espectral. El procediment de *binning* suma les lectures d'un interval de canals espectrals veïns (per exemple valors de *m/z* d'alta resolució) per a formar una sola mesura (valors de *m/z* de baixa resolució, compressió espectral) [115].

L'aplicació del procediment de compressió *binning* implica la transformació de les dades cromatogràfiques originals (dades r*aw*) en una matriu de dades (*x*,*y*) amb els valors de *m/z* en les columnes de la matriu i els temps de retenció en les files de la matriu. La conversió dels espectres de masses originals d'alta resolució (amb valors de *m/z* no equidistants) a aquesta matriu de dades, requereix l'establiment d'un nou eix de *m/z* amb una mida prèviament especificada (*bin*). D'aquesta manera, la compressió de les dades i la seva representació en forma de matriu es realitza simultàniament [115, 116]. En la Figura 2.15 es mostra un exemple gràfic del procediment del *binning* aplicat a l'escombratge en un determinat temps de retenció.

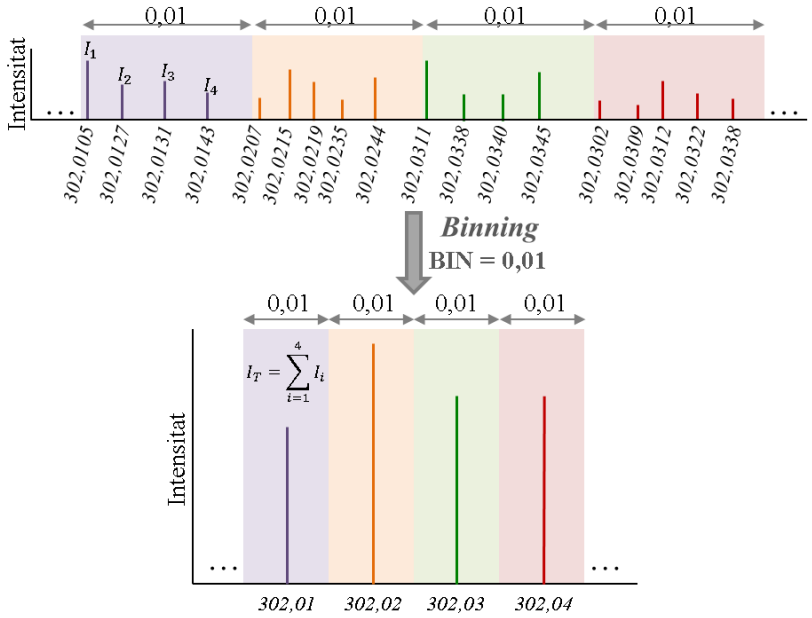


Figura 2.15. Representació gràfica del procediment de compressió del *binning*. Exemple mostrat per un escombrat en un temps de retenció concret, utilitzant una mida de *bin* de 0,01 Da.

En els estudis que analitzen més d'una mostra cal construir una matriu augmentada per poder analitzar la informació de totes les mostres simultàniament. En aplicar *binning* es generen nous valors de *m/z* (columnes de la matriu), els quals són comuns per a totes les matrius de dades de LC-MS construïdes per les diferents mostres. D'aquesta manera la matriu augmentada en la direcció de les columnes es pot construir fàcilment.

Malgrat tot, el procediment de compressió *binning* presenta inconvenients importants, com és la dificultat d'escollir la mida del *bin* apropiada per a cada conjunt de dades i la pèrdua de resolució espectral. Si es selecciona una mida massa petita de *bin*, els pics cromatogràfics poden dividir-se entre *bins* consecutius i no detectar-se per la pèrdua de la forma del pic cromatogràfic. En canvi, si s'escull una mida de *bin* massa gran, els pics cromatogràfics corresponents a un valor *m/z* es poden solapar amb els altres valors *m/z* i, a més, els pics més petits poden desaparèixer per l'augment del nivell de soroll. Finalment, la mida del *bin* també determina la resolució espectral de les dades i la velocitat del processament. Per exemple, si s'escull una mida de *bin* igual a la resolució instrumental original (0,0001 Da per alta resolució) es manté l'exactitud original, però el processament és molt lent. En canvi, si s'escull una mida de *bin* més gran (0,01 Da per alta resolució) la compressió realitzada en la direcció dels valors de *m/z* comporta una pèrdua de resolució espectral en comparació a la resolució instrumental original de les dades, però permet un processament de les dades més ràpid [115, 116].

Regions d'interès (ROI)

La compressió de les dades mitjançant la cerca de regions d'interès (ROI) és una tècnica alternativa al *binning*, que permet una alta compressió de les dades sense perdre la resolució espectral de les dades originals. Aquest mètode es basa en considerar els anàlits com a regions de punts de dades d'alta densitat [117]. Aquestes regions d'anàlits són les anomenades regions d'interès (ROI) i contenen les traces corresponents als valors de *m/z* més rellevants, és a dir aquelles que tenen una intensitat significativament superior al llindar prèviament fixat de la relació senyal/soroll (*signal to noise ratio threshold*, SNR_{Thr}). Els valors de *m/z* continguts en una mateixa ROI han d'estar dins d'un rang d'error (*m/z* error, μ), similar a l'exactitud instrumental de l'analitzador de MS utilitzat. A més, les ROIs han de contenir un nombre mínim de valors consecutius en la direcció dels temps (p_{min}) que permetin definir correctament la forma dels pics cromatogràfics. Aquestes condicions eviten que els valors associats només a soroll de fons es puguin considerar en les ROIs [116, 117]. En la figura 2.16 es mostra una representació gràfica de la compressió de les dades mitjançant la cerca de ROIs.

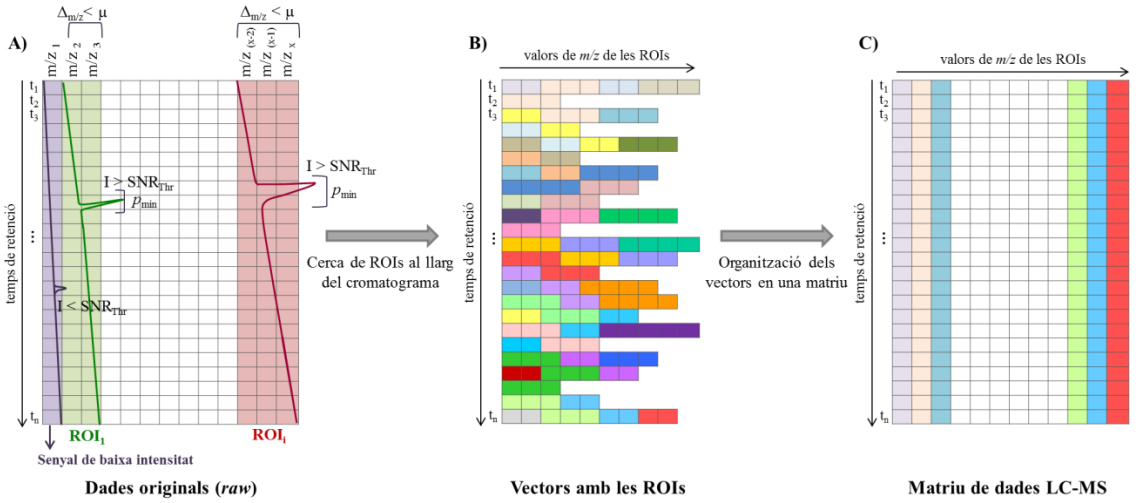


Figura 2.16. Representació gràfica del procediment de compressió per cerca de ROIs. A) Dades originals en les quals es representen dos exemples de ROIs (verd i vermell) i es distingeixen dels senyals de baixa intensitat (lila). B) Vectors que contenen les ROIs trobades a cada temps de retenció. Cada vector té una mida diferent. C) Matriu de dades LC-MS construïda reorganitzant els vectors que contenen la informació de les ROIs.

Tal com es mostra en la Figura 2.16, les ROIs es busquen per cada temps de retenció al llarg de tot el cromatograma. Aquesta cerca genera un parell de vectors per a cada temps de retenció, un que conté els valors de *m/z* trobats i l'altre que conté les seves corresponents intensitats. Com que a cada temps de retenció es detecten uns anàlits concrets, els valors de *m/z* trobats són diferents entre els diversos temps de retenció i, conseqüentment, els vectors obtinguts en cada temps de retenció tenen longituds diferents (en funció del nombre de ROIs trobades a cadascun). Finalment, aquests vectors s'han d'organitzar en una matriu de dades. Per poder fer això, d'una banda les ROIs comunes entre tots els temps de retenció s'agrupen i el valor de *m/z* final de cada ROI es calcula a partir de la mitjana de tots els valors de *m/z* de la sèrie de dades contingudes dins d'aquesta ROI específica. D'altra banda, en el cas de que una ROI no es trobi present en un temps de retenció determinat, la intensitat associada a aquesta ROI a aquest temps s'estableix com un valor baix i aleatori pròxim al nivell del soroll. D'aquesta manera es manté la informació detectada per les ROIs comunes i no comunes al llarg de tot el cromatograma [116].

En els estudis de metabolòmica cal considerar tot el conjunt de mostres simultàniament, i per això cal construir una matriu augmentada a partir de les matrius individuals comprimides per cada mostra. Cada una de les mostres de l'estudi conté uns anàlits concrets, per tant, cada una de matrius individuals té un nombre de ROIs diferent (columnes). Per poder construir la matriu augmentada en la direcció de les columnes es realitza una nova cerca de ROIs comunes i no comunes per a totes les matrius (mostres) individuals. D'una banda, s'agrupen les ROIs amb una diferència en el valor de *m/z* per sota de l'error de

massa tolerat (ROIs comunes). D'altra banda, quan una ROI no té una intensitat significant en una determinada mostra, aquesta s'estableix com un valor baix i aleatori pròxim al nivell del soroll. Seguint aquest procediment es manté la informació obtinguda per les ROIs comunes i no comunes en totes les mostres d'un estudi.

Compressió en la direcció dels temps de retenció

A causa de les grans dimensions dels conjunts de dades originats, en alguns casos la reducció de la mida en la direcció espectral (columnes de la matriu de dades) no és suficient per a poder tractar totes les mostres simultàniament. En aquests casos, es realitza una reducció addicional en la direcció temporal (files de la matriu de dades). Existeixen diverses estratègies per a la compressió de les dades en la direcció temporal i en aquesta Tesi s'han utilitzat les finestres de temps (*time windowing*) i les transformades d'ondetes (*wavelet transform*).

L'estratègia de les finestres de temps es basa en dividir els cromatogrames de LC-MS en diferents regions de temps i analitzar-les de forma independent. La matriu augmentada de dades LC-MS que conté la informació de totes les mostres d'un mateix estudi es divideix en submatrius que contenen diferents finestres cromatogràfiques. Cadascuna d'aquestes submatrius és una matriu augmentada que té un nombre de files igual al nombre de temps de retenció considerat per a la determinada finestra cromatogràfica multiplicat pel nombre de mostres de l'estudi (no es requereix que els temps de retenció siguin exactament els mateixos en les diferents mostres simultàniament analitzades). Les diverses submatrius augmentades s'analitzen separadament.

Les transformades d'ondetes (*wavelets*) s'han utilitzat per a la compressió de les dades en diferents àrees de la química, ja que són estratègies de processament ràpides i simples. La transformació d'ondetes és una transformació de base, és a dir, expressa el senyal original en termes de funcions bàsiques diferents a les coordenades originals. En el cas de la transformació d'ondetes, el conjunt de bases d'aquesta transformació consisteix en petites ones. En aquesta Tesi, la transformació d'ondetes s'ha utilitzat per comprimir les matrius de dades de LC×LC-MS en la direcció dels temps de retenció. Amb aquesta finalitat, la transformació d'ondetes s'aplica individualment a cada columna (valor de *m/z*) de la matriu de dades LC×LC-MS. En la Figura 2.17A es mostra un esquema de la transformació del senyal original en senyal en base d'ona. Tal com es mostra en la figura 2.17A, aquesta transformació es pot realitzar a diferents nivells. Cada nivell divideix el senyal original en dos conjunts de coeficients, un que conté una descripció aproximada del senyal i l'altre que en conté els seus detalls. El coeficient que conté la

descripció aproximada és el que s'utilitza per a representar el senyal de forma comprimida. Seguint aquest procediment, la transformada d'ondetes redueix la mida del senyal 2^n vegades, sent n el nivell de compressió [118-120].

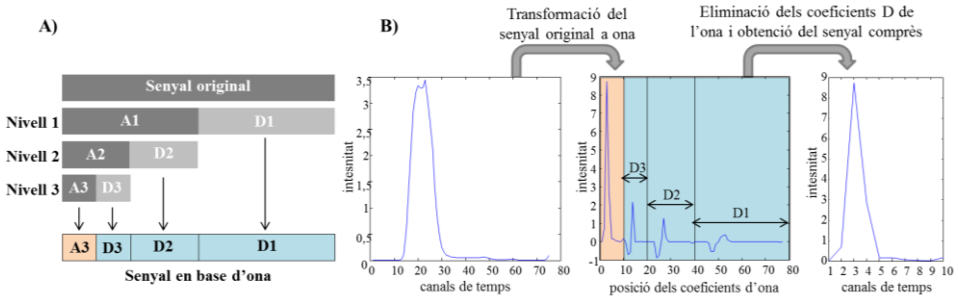


Figura 2.17. A) Esquema del procediment de transformació d'ondetes de nivell 3. Els coeficients que contenen la informació aproximada del senyal es representen amb una A i els que en conten els detalls amb una D. B) Exemple de compressió d'un pic cromatogràfic mitjançant una transformació d'ondetes de nivell 3.

En la Figura 2.17B es mostra un exemple de compressió d'un pic cromatogràfic (columna de la matriu de dades de LC-MS) mitjançant una transformació d'ondetes de nivell 3. Tal com es representa en aquesta figura, el primer pas és transformar el senyal original en un nou vector en forma d'ona que descriu el senyal (coeficient A). Seguidament, s'eliminen els elements d'aquest vector que corresponen als coeficients que contenen els detalls del senyal (coeficients D). El vector filtrat es torna a transformar a les coordenades de les dades originals per obtenir el senyal comprimit [118-120].

Normalització i escalat de les dades

La normalització i l'escalat són etapes del preprocessament de les dades que minimitzen la variació sistemàtica no desitjada. La normalització s'utilitza per eliminar les diferències entre mostres, i per tant, s'aplica en la direcció de les files. Els mètodes de normalització més utilitzats en metabolòmica són estratègies químiques: patrons interns i controls de qualitat (QCs). D'altra banda, l'escalat permet comparar entre els diferents metabòlits (eliminar la variació entre les variables) i s'aplica en la direcció de les columnes. Els mètodes d'escalat utilitzats en aquesta Tesi per disminuir les variacions entre variables han estat el centrat, l'autoescalat, la transformació MinMax i la transformació logarítmica [114, 121].

La variació sistemàtica no desitjada entre diferents mostres apareix a causa de petites diferències que es produeixen durant el procés experimental, com per exemple diferents recuperacions durant l'extracció dels metabòlits o diferents respostes a la ionització en l'espectròmetre de masses. Per poder minimitzar aquestes diferències, cal tenir en compte diferents aspectes. En primer lloc, cal analitzar tot el conjunt de mostres preferiblement agrupades en un mateix grup o lot (el mateix dia) i en un ordre aleatori per tal de

minimitzar la variació interna entre les mostres de diferents classes (per exemple controls i exposades). En segon lloc, cal utilitzar mostres de control de qualitat (QCs), que són mostres de composició similar a les que s'analitzen dins d'un mateix grup/lot, així com diferents patrons interns (afegits després de l'extracció) i patrons d'extracció (*surrogates*, afegits abans de l'extracció). D'una banda, els patrons interns i d'extracció permeten controlar les desviacions degudes a diferències experimentals i a derives instrumentals. D'altra banda, l'ús de QCs permet avaluar l'estabilitat de tot el procés analític i corregir les desviacions d'intensitat [116, 122].

L'escalat consisteix en dividir cada valor de les variables originals per un factor d'escala, el qual és diferent per a cada variable. Freqüentment s'utilitza la desviació estàndard de cadascuna de les variables com a factor d'escala, d'aquesta manera s'aconsegueix que cada variable de la nova matriu de dades quedi amb una nova desviació estàndard igual a 1. L'escalat és especialment útil quan els metabòlits mesurats tenen magnituds molt diferents, ja que els metabòlits amb valors més grans tendiran a dominar la variància observada, mentre que els que tinguin valors més petits quedaran més emmascarats [114, 123].

El pretractament de centrat consisteix en restar a cada valor de les variables originals el valor de la seva mitjana. D'aquesta manera, s'aconsegueix que cada variable de la nova matriu de dades (matriu centrada) quedi amb una mitjana igual a 0. El centrat ajusta les diferències en el desplaçament respecte l'origen entre els diferents metabòlits, facilitant així la visualització de les variacions respecte el valor de la seva mitjana. Per tant, s'omet la informació relativa als metabòlits que no varien entre les diferents mostres (*offsets*) [114].

L'autoescalat consisteix en l'aplicació simultània del centrat i l'escalat. La distribució dels valors dels metabòlits obtinguts en aplicar l'autoescalat és similar al cas de l'escalat, però al mateix temps les dades experimenten una translació del seu origen de variació degut al centrat amb la mitjana [114, 123]. Així, totes les variables tenen la mitjana igual a 0 i la desviació estàndard a 1.

La transformació per MinMax és un tipus d'escalat que consisteix en restar a cada valor de la variable el valor mínim d'aquesta i posteriorment dividir per la diferència entre el valor mínim i el valor màxim (rang de la variable). D'aquesta manera, s'aconsegueix que el valor mínim de totes les variables sigui igual a 0 i el valor màxim, a 1 [114].

Per últim, la transformació logarítmica de les dades actua de manera similar a l'escalat, augmentant el pes de les variables que presenten uns valors més baixos respecte aquelles que presenten els valors més

alts. Aquest procediment és especialment útil per conjunts de dades esbiaixats cap a l'extrem baix de l'escala, quan la major part dels metabòlits presenten valors baixos i només una petita part d'aquests tenen valors alts. Amb l'aplicació de la transformació logarítmica s'obté una distribució més simètrica de les dades. Malgrat això, aquesta transformació pot ocasionar efectes no desitjats i s'ha d'utilitzar amb cura. Per exemple, pot alterar l'estructura interna de les dades i augmentar la seva no linealitat, i per tant, el nombre de components necessaris per explicar la seva variància [114].

2.3.3. Resolució dels pics cromatogràfics i detecció de variables característiques

Les dades obtingudes en estudis de metabolòmica contenen molta informació. En aquests estudis, per poder distingir la informació útil relacionada amb els metabòlits d'interès de la resta d'informació continguda en les matrius de dades de LC-MS (per exemple, el soroll de fons o les contribucions del solvent) s'utilitzen tant mètodes de resolució de pics com mètodes de detecció de variables característiques (*features*). La detecció de variables característiques i la resolució de perfils d'elució cromatogràfica són dos conceptes molt propers. El primer té com a objectiu detectar les variables característiques presents en les dades de LC-MS, entenent com a variable característica un senyal bidimensional de LC-MS, és a dir un senyal amb un determinat valor de *m/z* que elueix en un temps de retenció concret [115]. En canvi, la resolució de pics té per objectiu trobar els perfils d'elució i espectres de masses purs dels components responsables d'aquestes variables característiques. És a dir, resol tots els valors *m/z* que formen l'espectre de masses d'un component determinat, enlloc de detectar només la intensitat corresponent a un valor de *m/z*.

En aquesta Tesi s'han comparat els resultats obtinguts tractant les dades amb el mètode de detecció de variables característiques XCMS (vàries formes de cromatografia (X) acoblada a espectrometria de masses), basat en la cerca de ROIs, amb els resultats que s'obtenen quan s'utilitza la resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS) com a mètode de resolució de pics després d'utilitzar un dels mètodes de compressió de dades explicats en l'apartat anterior (*binning* o cerca de ROIs).

Mètodes de detecció de variables característiques (features), XCMS.

El XCMS [124] és un mètode de tractament de dades de LC-MS i GC-MS molt popular i àmpliament utilitzat a la bibliografia metabolòmica [125, 126]. Aquest mètode permet la detecció de les variables característiques de forma automàtica i el càlcul de les àrees cromatogràfiques corresponents.

En l'actualitat, el XCMS es basa en l'aplicació de l'algoritme *centWave*, el qual inclou dues etapes principals. En primer lloc, els senyals de *m/z* més importants es seleccionen utilitzat l'estratègia de la cerca de ROIs, explicada detalladament a l'apartat 2.3.2. La segona etapa consisteix en la identificació i modelització dels perfils cromatogràfics mitjançant la transformació d'ondetes (*wavelet transform*) i l'aplicació d'un mètode d'ajust de forma Gaussiana. La transformació d'ondetes continuada modela els perfils cromatogràfics de diferents amplades [127]. La modelització dels perfils es completa utilitzant un model d'ajust que els forci a tenir una forma gaussiana (*matched filtration*). Aquest ajust es basa en aplicar un filtre que tingui els mateixos coeficients que la forma esperada en el senyal. En el cas de dades cromatogràfiques, una funció Gaussiana és un bon filtre per a definir la forma dels perfils cromatogràfics. Aquest ajust s'aplica a cada un dels cromatogrames extrets per cada valor de *m/z* de les ROIs trobades [124]. En la Figura 2.18 es representa gràficament aquest procés d'identificació de variables característiques i modelització dels perfils cromatogràfics. Finalment, les variables característiques considerades com a no rellevants s'eliminen (filtren), descartant totes aquelles que no es trobin presents en un percentatge prèviament determinat de mostres.

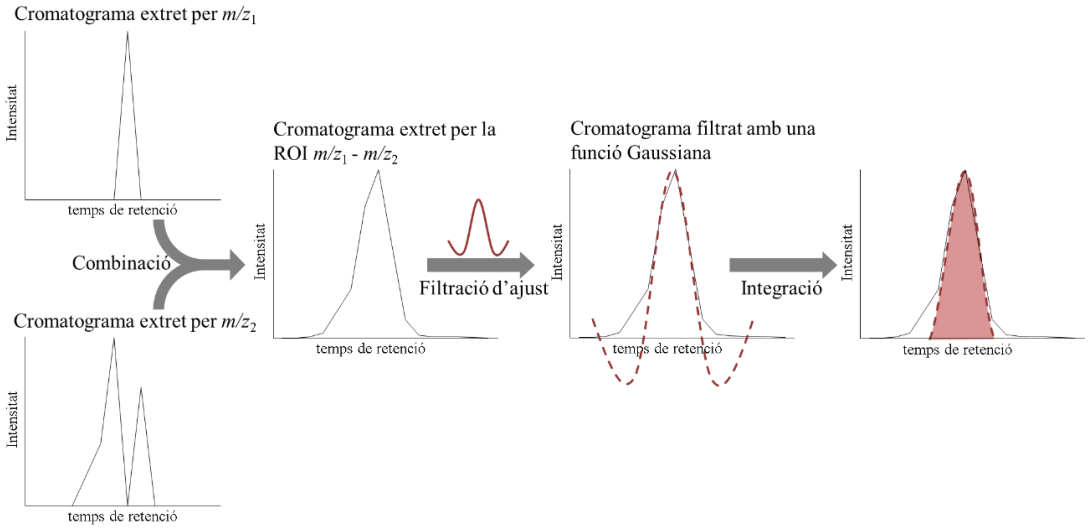


Figura 2.18. Representació gràfica de la detecció i modelització dels pics per XCMS. Primer l'algoritme crea el cromatograma combinat per tots els senyals de *m/z* compresos entre m/z_1i m/z_2. Seguidament es modela el pic obtingut aplicant una funció Gaussiana com a filtre d'ajust.

A continuació, després de la modelització descrita anteriorment de cada perfil cromatogràfic a cada valor de *m/z* seleccionat, el procediment XCMS aplica un procés d'alineament dels perfils cromatogràfics d'un mateix element en les diferents mostres. Aquest pas comença amb l'agrupació de les variables característiques detectades al llarg de totes les mostres. A continuació s'identifiquen quins dels grups es poden classificar com a "grups amb bon comportament". Aquests "grups amb bon comportament" (*"well-*

behaved" peak groups) es caracteritzen per tenir per a una determinada ROI molt poques mostres sense perfil o pic cromatogràfic, i molt poques mostres amb més d'un pic. Per cadascun d'aquests grups, l'algoritme calcula la mitjana dels temps de retenció i la desviació d'aquesta mitjana per cada mostra. Com que normalment els "grups amb bon comportament" es troben ben distribuïts al llarg del cromatograma, el gràfic de la desviació en el temps de retenció es pot construir per cada mostra. Els gràfics de desviació obtinguts s'utilitzen per corregir els temps de retenció originals i, seguidament, els grups de variables característiques es tornen a calcular. Aquest procés d'agrupament i alineament es repeteix de manera iterativa [124]. En la Figura 2.19 s'observa un exemple d'aquest procés d'alineament dels pics.

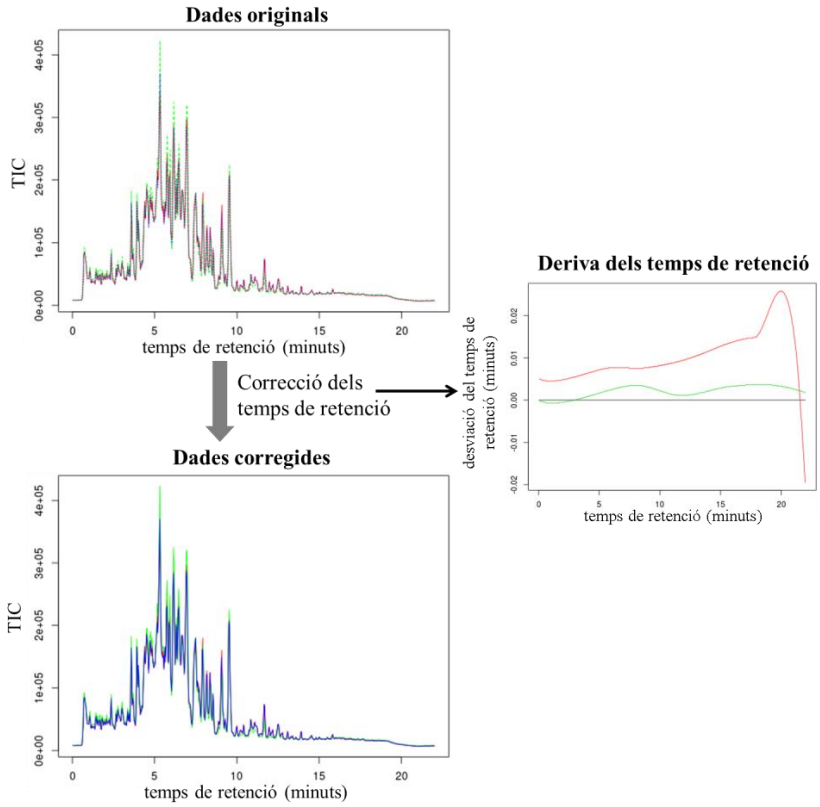


Figura 2.19. Exemple del procés d'alineament de pics del XCMS mitjançant el gràfic de deriva dels temps de retenció. En l'exemple es mostra l'alineament dels pics cromatogràfic de tres mostres de lipidoma d'arròs.

El mètode de XCMS es tracta doncs d'un algoritme complex que requereix l'optimització de diversos paràmetres. No obstant això, la seva implementació en format web (https://xcmsonline.scripps.edu) ha automatitzat molt aquest procés, fent-lo fàcilment accessible als usuaris no experts. A més, els resultats obtinguts es poden processar en la mateixa web per la base de dades pública METLIN[128], la qual cosa resulta molt útil a l'hora de fer una primera identificació dels metabòlits. Aquests motius han fet que actualment el XCMS sigui el mètode estàndard de tractament de

dades en estudis de metabolòmica no dirigida basats en tècniques cromatogràfiques acoblades a espectrometria de masses [125].

Resolució de pics mitjançant el mètode MCR-ALS

La resolució multivariant de corbes (MCR) [129-133] és un mètode quimiomètric utilitzat per a identificar i resoldre les contribucions existents en una mescla. Aquest procediment s'utilitza per a la resolució dels múltiples components químics presents en mescles complexes [134-136]. En la bibliografia es troben molts treballs de l'aplicació del mètode MCR en conjunts de dades d'àmbits diversos, però en aquesta secció s'explicarà només pel cas particular de dades de LC-MS i LC×LC-MS en estudis de metabolòmica.

El mètode MCR s'aplica als conjunts de dades de LC-MS agrupades en forma de matriu de dades (**D**) amb la finalitat de descriure la màxima quantitat de variància observada a partir d'un model bilineal amb un nombre reduït de components. Aquests components es defineixen mitjançant un perfil d'elució i un espectre de masses. El mètode MCR es basa en la descomposició de la informació continguda en la matriu de dades **D**, segons el model bilineal següent:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equació 2.1}$$

On **D** representa la matriu de dades de LC-MS per una mostra. Aquesta matriu **D** es descompon en el producte de dues matrius: **C**, que conté els perfils d'elució cromatogràfics dels components (metabòlits) presents a la mostra analitzada i $\mathbf{S}^{\mathbf{T}}$, que conté els espectres de masses associats als components (metabòlits) resolts. Cal destacar que mentre el nombre de files en les matrius **D** i **C** és el mateix, el nombre de columnes en les matrius **D** i $\mathbf{S}^{\mathbf{T}}$ coincideix. El nombre de components escollit (*i.e.* nombre de columnes en **C** i files en $\mathbf{S}^{\mathbf{T}}$) és el mínim necessari per descriure la major part de la variància de les dades de LC-MS originals. Finalment, la matriu **E** conté l'error de les dades, la part de la variància que no explica el model [126, 131].

En la Figura 2.20 es mostra un exemple de dades de LC-MS contingudes en la matriu **D** i de la seva descomposició en **C** i $\mathbf{S}^{\mathbf{T}}$. En aquesta figura també es descriu el procediment general utilitzat en MCR, que inclou els passos que s'expliquen a continuació:

1) Determinació del nombre de components presents en la matriu de dades original **D**.
2) Obtenció d'una estimació inicial de la matriu **C** o $\mathbf{S}^{\mathbf{T}}$.
3) Optimització de les matrius **C** i $\mathbf{S}^{\mathbf{T}}$ mitjançant mínims quadrats alternats (ALS) sota l'aplicació de determinades restriccions fins que s'assoleix el criteri de convergència.

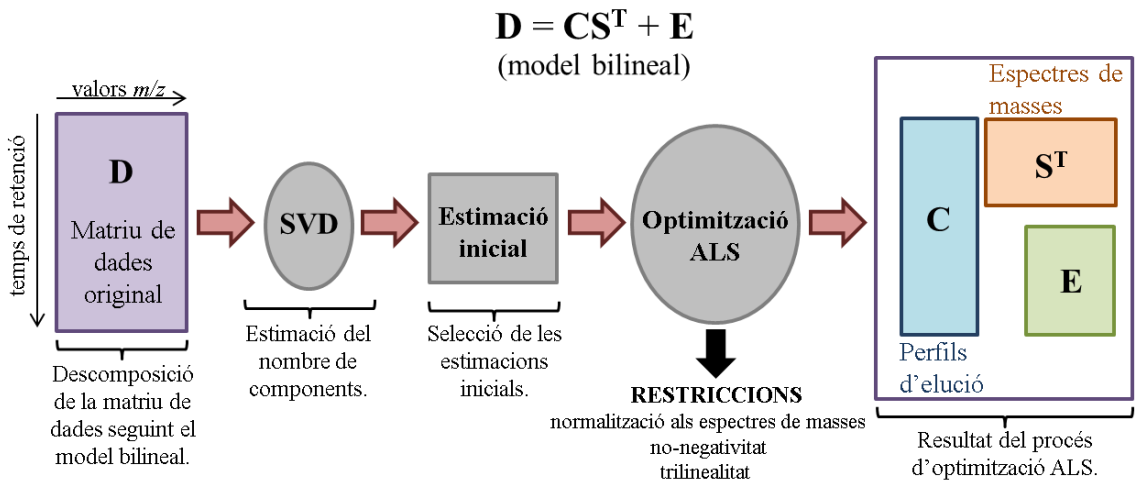


Figura 2.20. Diagrama d'implementació del mètode MCR-ALS per la resolució d'una matriu de dades de LC-MS.

En els treballs de metabolòmica no dirigida realitzats en aquesta Tesi el nombre de components (*N*) es refereix al nombre de metabòlits presents en la mostra analitzada i, per tant, no es coneix prèviament. En aquests casos, el nombre de components importants (*N*) presents en la matriu **D** s'estima de forma aproximada utilitzant l'algoritme de descomposició en valors singulars (SVD) [137], de manera que s'expliqui la major part de la variància en la matriu de dades original **D**.

Un cop determinat el nombre de components a fer servir en l'anàlisi, la descomposició bilineal comença a partir d'una estimació inicial d'una de les matrius, **C** o $\mathbf{S}^{\mathbf{T}}$. Aquesta estimació inicial es pot obtenir a partir de la selecció d'un conjunt de columnes o d'un conjunt de files de la matriu original **D** igual en nombre als components estimats en la mostra (*N*). Les files o columnes seleccionades seran les més pures, és a dir les que siguin més diferents entre sí i que no continguin només soroll experimental. En aquesta Tesi l'estimació inicial de **C** i $\mathbf{S}^{\mathbf{T}}$ s'ha realitzat mitjançant una adaptació del mètode SIMPLISMA (*SIMPLe-to-use Iterative Self-Modeling Analysis*) [138-140]. Aquest mètode selecciona iterativament les variables més diferents de la mitjana en la direcció considerada (files o columnes).

Un cop es disposa d'aquesta estimació inicial, les matrius **C** i $\mathbf{S}^{\mathbf{T}}$ s'optimitzaran mitjançant cicles iteratius d'un procediment de mínims quadrats alternats (ALS) [134-136]. Per tal de que els perfils d'elució i els espectres de masses resolts tinguin sentit químic, cal aplicar restriccions durant el procés d'optimització. Les restriccions són propietats químiques o matemàtiques generals que donen un sentit físic a les matrius **C** i $\mathbf{S}^{\mathbf{T}}$, que es compleixen en els perfils reals i que actuen dirigint el procediment d'optimització iteratiu cap a una solució interpretable en termes químics. Per aplicar les restriccions, cal aprofitar el coneixement físic que es té sobre el sistema i traduir-lo a condicions matemàtiques. En el cas

de les dades de LC-MS estudiades en aquesta Tesi, les restriccions que s'utilitzen són la no-negativitat (**C** i $\mathbf{S}^T$) i la normalització dels espectres de masses ($\mathbf{S}^T$) [131, 136]. Aquetes restriccions s'expliquen amb més detall a continuació.

La restricció de no-negativitat, tal com el seu nom indica, evita la presència de valors negatius en les matrius **C** i $\mathbf{S}^T$. S'aplica sempre als perfils de concentració, ja que les concentracions dels compostos químics han de ser valors positius o zero. D'altra banda, la no-negativitat també s'aplica en els espectres, ja que l'espectrometria de masses sempre proporciona lectures d'intensitat positives. Hi ha diferents maneres d'aplicar la restricció de no-negativitat, d'una banda es poden substituir directament els valors negatius per zero durant els diferents passos iteratius. D'altra banda, hi ha algoritmes més rigorosos basats en procediments de mínims quadrats no negatius (*non-negative least squares*, nnls) [141] i les seves variants més ràpides, com per exemple, la de mínims quadrats ràpids no negatius (*fast non-negative least squares*, fnnls) [142].

La restricció de normalització als espectres de masses s'aplica amb l'objectiu d'eliminar l'ambigüitat d'escala en els perfils resolts per MCR-ALS. Els espectres de masses es normalitzen per tal d'evitar que surtin fora d'escala durant el procés d'optimització per ALS. Aquesta normalització es pot fer de diverses maneres, per exemple dividint tots els valors de l'espectre d'un component determinat pel seu valor màxim d'intensitat. D'aquesta manera, es forcen els perfils de tots els components a tenir el màxim d'intensitat igual a 1[131].

En el cas de les dades de LC×LC-MS, a part de les restriccions de no-negativitat i la normalització dels espectres de masses, també es pot utilitzar la restricció de trilinealitat quan els perfils d'elució obtinguts en les diferents modulacions són altament reproduïbles (el que es coneix com a MCR trilineal). Aquesta restricció força als perfils d'un mateix component en les diferents modulacions a estar sincronitzats (mateix temps de retenció) i presentar la mateixa forma, però permet que variïn en un factor d'escala. A més, també és possible aplicar la restricció de trilinealitat d'una manera més flexible. Així, es poden permetre petites diferències en la sincronització dels perfils cromatogràfics, és a dir els perfils d'un mateix component en les diferents modulacions es forcen a ser iguals en forma, però poden variar en el temps de retenció i en un factor d'escala (MCR-ALS trilineal amb llibertat de desplaçament dels perfils d'elució en la direcció temporal). L'aplicació de la restricció de trilinealitat permet obtenir solucions úniques (sense ambigüitats), però cal ser molt estricte en els casos on es pot aplicar, ja que és una restricció molt forta i només es compleix en el cas de dades totalment multilineals.

En aquesta Tesi, l'avaluació de la qualitat del model MCR-ALS s'ha fet mitjançant dos paràmetres: el percentatge de falta d'ajust (*lack of fit*, LOF %, equació 2.2) i el percentatge de variància explicada (R^2 %, equació 2.3).

$$\text{LOF}\,\% = 100 \times \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \qquad e_{i,j} = d_{i,j} - \hat{d}_{i,j} \qquad \text{Equació 2.2}$$

$$R^2\,\% = \left(\frac{\sum d_{i,j}^2 - \sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}\right) \times 100 \qquad \text{Equació 2.3}$$

On $d_{i,j}$ és l'element de la matriu de dades originals a la fila *i* i a la columna *j*, $e_{i,j}$ és el residual relacionat, que resulta de la diferència entre l'element $d_{i,j}$ de les dades originals i l'anàleg reproduït amb el model MCR-ALS ($\hat{d}_{i,j}$).

L'objectiu principal és que el model expliqui el màxim de variància possible (R^2). Si el valor de R^2 és petit, una opció és provar models amb un nombre creixent de components i observar com el nombre de components afecta a la variància explicada. Si el valor de R^2 millora, això significa que el primer model no tenia suficient informació per descriure completament el sistema. En canvi, si el valor de R^2 empitjora o no varia de forma significativa, això indica que l'addició de components al model és innecessària. A més, també cal observar la influència del nombre de components en el valor de la manca d'ajust (*LOF*). Per defecte no existeix un valor òptim d'aquest paràmetre, ja que depèn de la qualitat de les dades originals i de la relació senyal/soroll de les mesures. Tot i així, si el valor del *LOF* és gran, es construeixen models amb un nombre creixent de components i s'observa com aquest increment en el nombre de components afecta a l'ajust del model. Per últim, per acabar d'assegurar la qualitat d'un model MCR-ALS és necessari avaluar la versemblança dels perfils de **C** i $\mathbf{S}^\mathbf{T}$ obtinguts. L'objectiu final del mètode és recuperar els perfils d'elució i espectres de masses dels metabòlits presents en la mostra analitzada, iguals o el més semblants possible als originals.

Els perfils d'elució i espectres de masses resolts en el procés de MCR-ALS poden no ser versemblants degut a l'existència d'ambigüitats en la resolució. Es diu que hi ha ambigüitat quan parelles diferents de matrius **C** i $\mathbf{S}^\mathbf{T}$ poden reproduir la matriu de dades original **D** amb un ajust òptim. Els dos tipus d'ambigüitats més importants són l'ambigüitat d'intensitat o escala i l'ambigüitat rotacional [129, 132, 143-145]. L'existència de l'ambigüitat d'intensitat o d'escala, comporta l'existència de diferents possibles solucions dels perfils d'elució i espectres de masses purs amb la mateixa forma i la mateixa manca d'ajust, però escalats per un factor desconegut. Com s'ha comentat anteriorment, aquesta ambigüitat es minimitza amb la normalització dels perfils resolts (**C** o $\mathbf{S}^\mathbf{T}$). D'altra banda, existeix

l'ambigüitat rotacional quan diferents parelles de les matrius **C** i $\mathbf{S}^{\mathbf{T}}$ amb perfils diferents, poden reproduir la matriu **D** de manera òptima. Aquest fenomen es dóna quan el sistema no compleix certes condicions de selectivitat i rang local. Resumint aquestes condicions, l'ambigüitat rotacional pot aparèixer quan hi ha dos o més components que es superposen. En les dades de metabolòmica de LC-MS i LC×LC-MS, l'efecte de l'ambigüitat rotacional és baix, degut a l'elevada selectivitat i a la naturalesa dispersa (*sparseness*) de les mesures de MS. Només hi haurà un cert grau d'ambigüitat rotacional en el cas que dos metabòlits fortament coeluïts tinguin en comú algun senyal de *m/z*, com per exemple per dos isòmers estructurals.

MCR-ALS aplicat a múltiples mostres

En un estudi de metabolòmica s'acostuma a analitzar diverses mostres de diferents classes. Sovint, és interesant tractar les dades de totes les mostres simultàniament, ja que així s'obté informació sobre com canvien els perfils metabòlics en cada una de les classes de mostres. Quan les mostres s'analitzen mitjançant LC-MS s'obté una matriu de dades per cadascuna. Per tal d'analitzar-les totes simultàniament, les corresponents matrius es poden organitzar de diferents maneres. Per exemple, en aquesta Tesi, s'han organitzant formant una matriu augmentada en la direcció de les columnes ($\mathbf{D}_{\mathbf{aug}}$). En aquesta matriu augmentada, les diferents matrius individuals s'agrupen segons la direcció dels valors de *m/z* analitzats (columnes). En la Figura 2.21A es mostra una representació d'aquesta matriu augmentada en la direcció de les columnes (*column-wise*). Aquest tipus d'augmentació de matrius s'utilitza quan les diferents mostres d'un estudi s'analitzen mitjançant la mateixa tècnica analítica, ja que d'aquesta manera les seves matrius comparteixen els mateixos valors en les columnes (valors de *m/z* analitzats).

Quan les mostres s'analitzen mitjançant LC×LC-MS, les dades obtingudes es poden organitzar en forma de cub de dades per cadascuna de les mostres analitzades. Tal com s'ha comentat en l'apartat 2.3.1 (*Naturalesa de les dades*), aquest cub es pot també desplegar en forma de matriu de dades augmentada per poder-lo analitzar mitjançant mètodes d'àlgebra lineal emprant models bilineals. En el cas de dades LC×LC-MS, per poder tractar simultàniament diverses mostres alhora, es poden organitzar les dades corresponents en forma de matrius superaugmentades (matriu $\mathbf{B}_{\mathbf{saug}}$, en la Figura 2.21B). En aquesta matriu superaugmentada, les matrius augmentades individuals construïdes per cada una de les mostres s'apilen en la direcció dels valors de *m/z* analitzats (columnes). En la Figura 2.21B es mostra una representació gràfica d'una d'aquestes matrius superaugmentades.

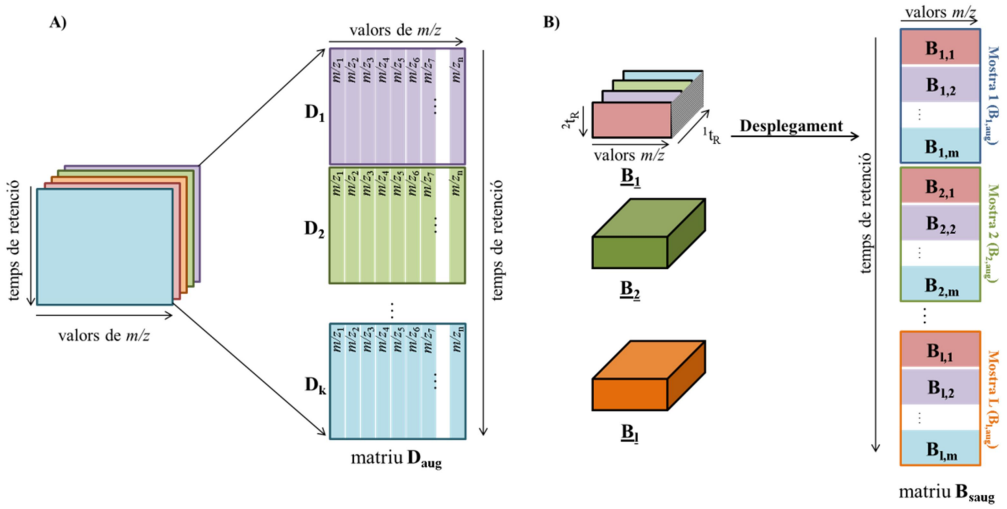


Figura 2.21. A) Augmentació de les matrius de dades corresponents a diferents mostres en la direcció de les columnes (*column-wise*). B) Dades LC×LC-MS ordenades en forma de matriu superaugmentada en la direcció de les columnes (*column-wise*).

Un dels avantatges més importants del mètode MCR-ALS és que permet l'anàlisi conjunta de diverses matrius de dades. Aquesta estratègia permet analitzar totes les mostres d'un experiment alhora i, a més, facilita la resolució del conjunt de dades. Moltes vegades la matriu de dades d'una única mostra no conté informació suficient per a poder resoldre adequadament el sistema químic estudiat i això es pot solucionar, o al menys millorar, quan s'analitzen simultàniament diverses mostres del mateix sistema obtingudes en condicions diferents. L'extensió del model de MCR-ALS aplicat a l'anàlisi de matrius augmentades en la direcció de les columnes es pot expressar de la següent manera en el cas de dades de LC-MS:

$$\mathbf{D_{aug}} = \begin{bmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \mathbf{D_3} \\ \vdots \\ \mathbf{D}_L \end{bmatrix} = \begin{bmatrix} \mathbf{C_1} \\ \mathbf{C_2} \\ \mathbf{C_3} \\ \vdots \\ \mathbf{C}_L \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \\ \mathbf{E_3} \\ \vdots \\ \mathbf{E}_L \end{bmatrix} = \mathbf{C_{aug}} \mathbf{S^T} + \mathbf{E_{aug}} \qquad \text{Equació 2.4}$$

I de la següent manera en el cas de dades de LC×LC-MS:

$$\mathbf{B_{saug}} = \begin{bmatrix} \mathbf{B_{1,1}} \\ \mathbf{B_{1,2}} \\ \mathbf{B_{1,3}} \\ \vdots \\ \mathbf{B}_{L,M} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{1,1}} \\ \mathbf{C_{1,2}} \\ \mathbf{C_{1,3}} \\ \vdots \\ \mathbf{C}_{L,M} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_{1,1}} \\ \mathbf{E_{1,2}} \\ \mathbf{E_{1,3}} \\ \vdots \\ \mathbf{E}_{L,M} \end{bmatrix} = \mathbf{C_{saug}} \mathbf{S^T} + \mathbf{E_{saug}} \qquad \text{Equació 2.5}$$

On la matriu augmentada $\mathbf{D_{aug}}$ conté totes les mostres (L) d'un estudi LC-MS i la matriu superaugmentada $\mathbf{B_{saug}}$ conté totes les modulacions (M) de totes les mostres (L) en el cas d'un estudi LC×LC-MS. La matriu augmentada $\mathbf{D_{aug}}$ es descompon en el producte de la matriu augmentada dels perfils d'elució $\mathbf{C_{aug}}$ amb la matriu d'espectres de masses $\mathbf{S^T}$. A partir de la descomposició bilineal de la

matriu superaugmentada ($\mathbf{B}_{saug}$) s'obté la matriu superaugmentada $\mathbf{C}_{saug}$, que conté els perfils d'elució en les dues dimensions cromatogràfiques. En tots dos casos s'obté també $\mathbf{S}^T$, que conté els espectres de masses purs dels components resolts. En la Figura 2.22 es pot veure la representació gràfica de la resolució d'una matriu $\mathbf{D}_{aug}$ (Figura 2.22A) i d'una matriu $\mathbf{B}_{saug}$ (Figura 2.22B).

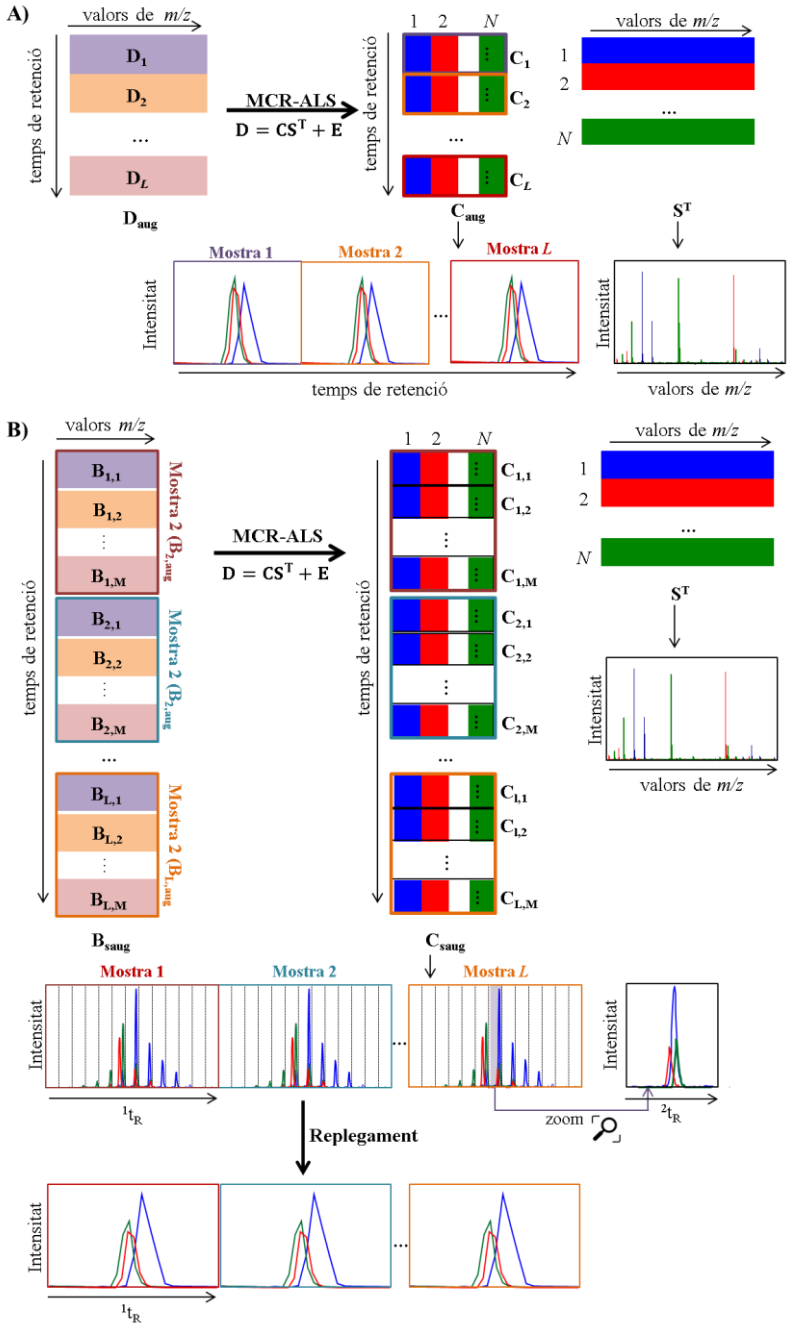


Figura 2.22. Representació gràfica de la resolució pel mètode MCR-ALS de: A) una matriu $\mathbf{D}_{aug}$ de dades de LC-MS que conté *L* mostres i B) una matriu $\mathbf{B}_{saug}$ de dades de LC×LC-MS que conté *L* mostres amb *M* modulacions cadascuna. Exemple de la solució obtinguda per tres components. En el cas de les dades de LC×LC-MS els perfils d'elució a la primera columna s'obtenen replegant els perfils d'elució resolts ($\mathbf{C}_{saug}$).

En la Figura 2.22 es pot observar que en el cas de dades de LC-MS, cada matriu $\mathbf{D_l}$ conté la informació d'una mostra mitjançant aquesta tècnica. En canvi, per les dades de LC×LC-MS, cada matriu de dades individual, $\mathbf{B_{l,m}}$, conté la informació d'una modulació en la segona columna cromatogràfica. De manera que cada mostra està representada per una matriu augmentada $\mathbf{B_{l,aug}}$ formada per un nombre determinat (*M*) de matrius $\mathbf{B_{l,m}}$. En aquest cas, la descomposició bilineal proporcionarà la matriu $\mathbf{C_{saug}}$, a partir de la qual es podran obtenir els perfils d'elució en les dues columnes, tal com es mostra en la Figura 2.22B. La matriu $\mathbf{C_{saug}}$ conté el cromatograma bidimensional desplegat per cadascun dels metabòlits resolts, de manera que cada modulació apareix una al costat de l'altre. En la Figura 2.22B es representa una vista augmentada d'una d'aquestes modulacions. Replegant els perfils continguts en la matriu $\mathbf{C_{saug}}$ es poden obtenir el perfils d'elució a la primera dimensió cromatogràfica de cadascun dels metabòlits resolts, tal com es mostra en la Figura 2.22B.

És important destacar que en els models de MCR descrits en les equacions 2.4 i 2.5, els perfils d'elució cromatogràfics d'un mateix component resolts en les diferents matrius ($\mathbf{C_l}$ i $\mathbf{C_{l,m}}$) no estan forçats a ser iguals. És a dir, els perfils d'elució cromatogràfica d'un determinat compost a les diferents mostres o modulacions poden diferir en posició (temps de retenció), forma i intensitat. Això és molt important perquè permet obviar els problemes de deriva en els temps de retenció entre les diferents injeccions cromatogràfiques (mostres o modulacions), els quals són especialment importants en el cas de LC×LC-MS. D'aquesta manera, no és necessari alinear els pics cromatogràfics entre les diferents mostres. Aquest fet és un avantatge important del mètode MCR respecte altres mètodes, en els que cal aplicar procediments d'alineament i de modelització dels pics cromatogràfics, els quals són complexos i aporten incerteses. Per altra banda, en els models de MCR descrits en les equacions 2.4 i 2.5, els espectres de masses continguts en la matriu $\mathbf{S^T}$ sí que estan forçats a ser iguals en totes les mostres o modulacions. És a dir, l'espectre de masses d'un compost determinat és sempre el mateix, independentment de les condicions experimentals de les diferents mostres.

Generalment, en un estudi de metabolòmica les diferents matrius $\mathbf{D_l}$ o $\mathbf{B_{l,aug}}$ representen les mostres d'un mateix sistema metabòlic sota diferents condicions. Per exemple, en un estudi on es compara l'estat metabòlic d'un organisme determinat (com ara l'arròs) en condicions normals i sota condicions d'estrès (com alta temperatura, salinitat, sequera o exposició a contaminants) les matrius $\mathbf{D_l}$ o $\mathbf{B_{l,aug}}$ es refereixen a l'anàlisi per LC-MS o LC×LC-MS de les mostres control i les mostres tractades. Les matrius $\mathbf{C_{aug}}$ o $\mathbf{C_{saug}}$ contindran els perfils d'elució dels diferents components resolts en les mostres analitzades. A partir

d'aquestes matrius de dades es pot obtenir informació quantitativa, de manera que es podrà observar quins dels components resolts presenten canvis entre les mostres controls i les mostres tractades, indicant, per tant, que pateixen alteracions sota l'estrès estudiat. D'altra banda, la matriu $\mathbf{S^T}$ contindrà els espectres de masses dels components resolts, a partir dels quals es podrà obtenir informació qualitativa i identificar els metabòlits o lípids continguts en les mostres analitzades.

Altres mètodes de resolució de pics, PARAFAC i PARAFAC2.

En el tractament de dades multidireccionals, com les dades de LC×LC-MS o GC×GC-MS, molts estudis de la bibliografia utilitzen el mètode d'anàlisi de factors paral·lels (*parallel factor analysis*, PARAFAC [146]) per a la resolució de pics cromatogràfics. Per aquest motiu en aquesta Tesi s'ha comparat l'ús del PARAFAC i la seva variant PARAFAC2 amb el mètode de MCR-ALS per a l'anàlisi de dades de LC×LC-MS.

El mètode PARAFAC exigeix que les dades compleixin el model trilineal, és a dir, que tots els components continguts en les dades es defineixin per uns perfils únics en els tres modes. En el cas de les dades de LC×LC-MS, aquests tres modes són els perfils d'elució en les dues dimensions cromatogràfiques i l'espectre de masses. Aquest requeriment trilineal restringeix l'ús del PARAFAC a dades on no hi hagi deriva en el temps d'elució ni canvis en la forma dels pics entre dues injeccions cromatogràfiques (modulacions) consecutives [146, 147].

El mètode de PARAFAC es basa en la descomposició de la informació continguda en un cub de dades, $\underline{\mathbf{B}}$, segons el model trilineal següent:

$$\underline{\mathbf{B}} = \sum_{i=1}^{N} \mathbf{f_i} \otimes \mathbf{g_i} \otimes \mathbf{h_i} + \underline{\mathbf{E}} \qquad \text{Equació 2.6}$$

On $\underline{\mathbf{B}}$ representa el cub de dades de LC×LC-MS per una mostra. Aquesta cub $\underline{\mathbf{B}}$ es descompon en el producte de tres factors de contribució utilitzant un nombre reduït de components *N*: $\mathbf{f_i}$,i $\mathbf{g_i}$ descriuen els perfils d'elució en les dues dimensions cromatogràfiques dels metabòlits presents a la mostra analitzada i $\mathbf{h_i}$, conté els espectres de masses associats als metabòlits resolts. Finalment, $\underline{\mathbf{E}}$ conté la part de la variància que no explica el model i $\otimes$ representa el producte extern, que relaciona els diferents factors. En la Figura 2.23 es representa gràficament aquesta descomposició trilineal del cub $\underline{\mathbf{B}}$. Els perfils $\mathbf{f_i}$, $\mathbf{g_i}$ i $\mathbf{h_i}$ es representen agrupats en les matrius **F**, **G** i **H**.

El fet que els perfils d'elució i espectres de masses resolts siguin únics i invariants fa evident que la trilinealitat és una restricció molt forta, que només es compleix en determinades ocasions. No obstant,

permet l'obtenció de solucions úniques i evita la presència de les ambigüitats rotacionals associades als models bilineals, com per exemple el model MCR [146, 147].

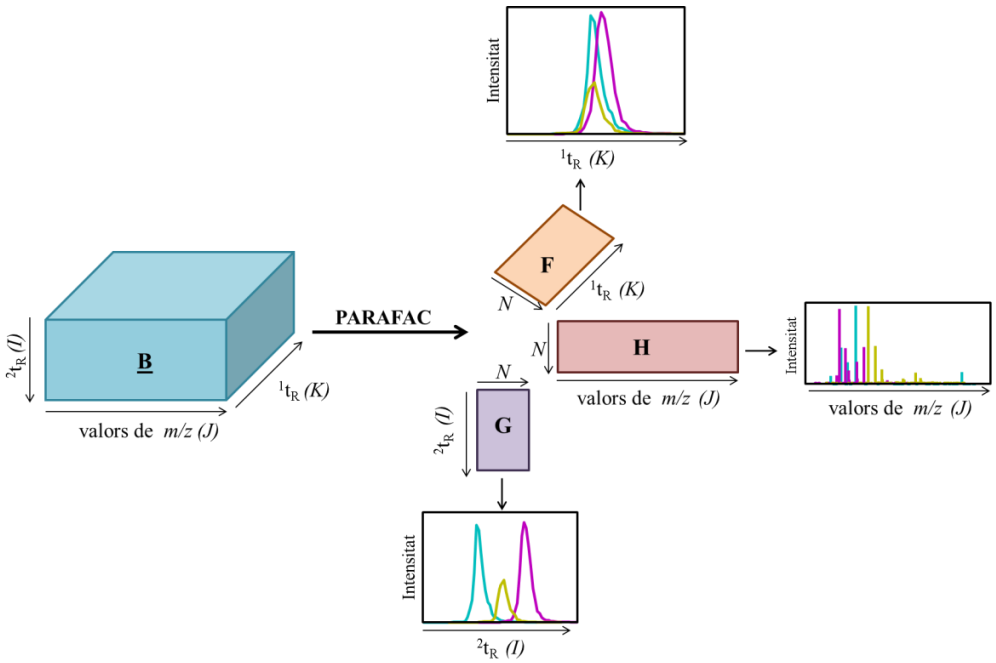


Figura 2.23. Representació gràfica de la resolució d'un cub de dades **C** mitjançant PARAFAC. Exemple de la solució obtinguda per tres components.

L'equació 2.6 es pot resoldre a través de diferents algoritmes i en aquesta Tesi s'ha utilitzat una optimització per ALS. Igual que en cas de MCR, la utilització de restriccions durant el procés d'optimització ajuda a obtenir solucions amb sentit químic. En el cas de les dades de cromatografia multidireccional, el més habitual és aplicar únicament la no-negativitat en les tres direccions [146].

El PARAFAC2 és una variant de PARAFAC desenvolupada per resoldre dades multidireccionals que tinguin petites desviacions en alguna de les tres direccions i que, per tant, no complexin la restricció de trilinealitat. En el cas de les dades de LC×LC-MS, aquestes desviacions poden ser petites derives en els temps de retenció i/o petites diferències en la forma dels pics cromatogràfics entre les diverses modulacions [148, 149].

Per poder controlar aquestes petites desviacions, el PARAFAC2 utilitza un algoritme més flexible que el PARAFAC per a fer la descomposició trilineal de les dades (equació 2.6). Aquest algoritme permet una certa llibertat als perfils d'elució de $\mathbf{g}_i$, deixant que cadascuna de les modulacions tingui els seus propis perfils d'elució. És a dir, per cada modulació (M) es resol una matriu $\mathbf{G}_m$. No obstant, per descompondre trilinealment el cub de dades original ($\underline{\mathbf{B}}$) cal seguir mantenint una solució única [148,

149]. Per complir aquest requeriment, els productes creuats de les diferents matrius $\mathbf{G_m}$ s'han de mantenir constants al llarg de totes les modulacions:

$$\mathbf{G_1^T G_1} = \mathbf{G_2^T G_2} = \cdots = \mathbf{G_m^T G_m} \qquad M = \text{nombre de modulacions} \qquad \text{Equació 2.7}$$

Per acabar, cal comentar que en el mètode de PARAFAC2 les restriccions durant el procés d'optimització no es poden aplicar en el segon mode (perfils d'elució en la segona dimensió cromatogràfica). Per això, el més habitual quan s'utilitza per a resoldre dades de LC×LC-MS és utilitzar la no-negativitat en el perfil d'elució de la primera dimensió cromatogràfica i en els espectres de masses.

Per tal d'estudiar la qualitat dels models de PARAFAC s'ha fet servir com a figura de mèrit la consistència del nucli (*core consistency*), tal com va proposar Bro [150]. Aquest paràmetre avalua si les dades compleixen el model trilineal en la seva descomposició. La consistència del nucli es calcula expressant el model trilineal PARAFAC (equació 2.6) com a un model restringit Tucker3 [151, 152]:

$$\mathbf{B_{aug}} = \mathbf{F J^{(N\times NN)} (G \otimes H)^T} \qquad \text{Equació 2.8}$$

On $\underline{\mathbf{J}}$ és un cub binari amb zeros a tots els elements excepte en la superdiagonal, que conté uns. La matriu $\mathbf{J^{(N\times NN)}}$ correspon al cub $\underline{\mathbf{J}}$ reordenat en una matriu de mida $N\times NN$. Un cop ajustat el model de PARAFAC, la verificació de que les dades tenen una estructura trilineal apropiada es pot obtenir calculant els mínims quadrats del nucli del model de Tucker3 ($\underline{\mathbf{K}}$) d'acord amb la següent equació:

$$\mathbf{B_{aug}} - \mathbf{FK(G \otimes H)^T} \qquad \text{Equació 2.9}$$

Si el model de PARAFAC és vàlid $\underline{\mathbf{J}}$ i $\underline{\mathbf{K}}$ han de ser semblants. En canvi, si les dades no es poden descriure correctament amb un model trilineal o s'han utilitzat massa components (N) per a construir el model de PARAFAC, $\mathbf{K}$ serà diferent de $\mathbf{J}$. La semblança entre aquestes dues matrius s'expressa mitjançant la consistència del nucli segons la següent equació:

$$Consistència\ del\ nucli = 100\left(\frac{1-\sum_{x=1}^{N}\sum_{y=1}^{N}\sum_{z=1}^{N}(k_{def}-j_{def})^2}{\sum_{x=1}^{N}\sum_{y=1}^{N}\sum_{z=1}^{N}j_{def}^2}\right) \qquad \text{Equació 2.10}$$

On j_{def} i k_{def} són els elements dels cubs $\underline{\mathbf{J}}$ i $\underline{\mathbf{K}}$. De tal manera que si les dades tenen una estructura totalment trilineal el valor de la consistència del nucli és del 100%. D'aquesta forma valors de consistència del nucli menors del 100% implicaran un desviament del comportament trilineal de les dades i, en conseqüència, la necessitat d'emprar mètodes bilineals en comptes de trilineals. En el treball de Bro i Kiers del 2003 es poden trobar més detalls sobre el càlcul d'aquest paràmetre [150].

2.3.4. Mètodes quimiomètrics d'exploració, regressió i classificació de dades experimentals metabolòmiques

En l'anàlisi de dades provinents d'estudis metabolòmics s'utilitzen diferents mètodes quimiomètrics per l'exploració, la regressió i la classificació per tal d'extreure la informació d'interès. Seguidament s'expliquen amb més detall els mètodes emprats en aquesta Tesi.

A partir de l'aplicació de diferents mètodes de tractament de dades cromatogràfiques, com per exemple els mètodes de resolució multivariant comentats a l'apartat anterior, s'obtenen les àrees dels perfils cromatogràfics de cadascun dels metabòlits resolts en cada una de les mostres analitzades. Aquesta informació es pot ordenar aleshores en una nova matriu de dades (**A**), que tindrà en les seves files les diverses mostres de l'estudi i en les columnes, els metabòlits resolts. En la Figura 2.24A es mostra un exemple d'aquesta matriu de dades **A**.

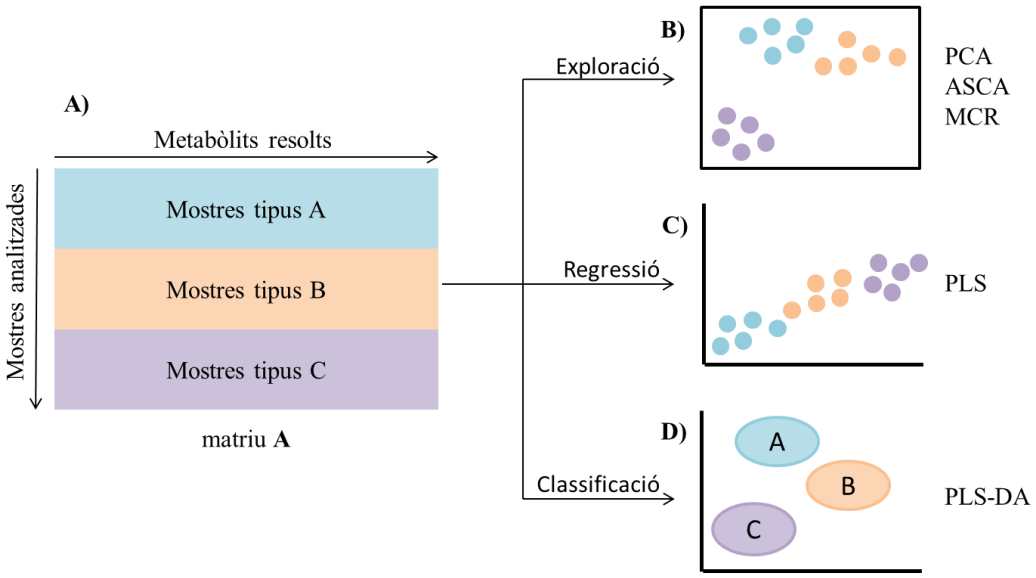


Figura 2.24. Visió general de tres tipus de de mètodes d'anàlisi quimiomètrica de dades que es poden aplicar per extreure informació d'una matriu **A**. A) Representació gràfica de la matriu **A**, que conté els metabòlits presents en diferents tipus de mostres B) Anàlisi exploratòria. C) Anàlisi de regressió i models predictius, D) Anàlisi de classificació i anàlisi discriminant [153].

Aquesta matriu de dades **A** conté les concentracions relatives dels diferents metabòlits en les mostres analitzades, a partir de les quals es pot interpretar el problema experimental d'un determinat estudi. Per exemple, en el cas dels treballs metabolòmics estudiats en aquesta Tesi, en els que s'avalua l'efecte d'un determinat estrès ambiental en l'arròs, la matriu de dades **A** conté la informació sobre quins metabòlits es troben més afectats per aquesta exposició. No obstant, la grandària de les dades metabolòmiques obtingudes i la complexitat dels processos moleculars estudiats, fan que extreure aquesta informació útil

no sigui fàcil. Per aquest motiu, és necessari utilitzar diversos mètodes quimiomètrics a l'hora d'extreure la informació continguda en la matriu **A**.

L'aplicació de mètodes quimiomètrics pot descriure el comportament dels metabòlits mesurats experimentalment a partir de les seves variacions al llarg de les mostres analitzades. Aquests mètodes comprimeixen la informació multidimensional continguda en la matriu **A**, en un nombre reduït de variables noves, les quals expliquen la major part de la variabilitat dels metabòlits originals, així com de les seves relacions. D'aquesta manera, la variància observada es pot analitzar en l'espai de les variables noves (de menor dimensió), la qual cosa pot ajudar a entendre millor el sistema estudiat. Durant la compressió de les dades també s'estableixen vincles amb les variables originals. De tal manera que aquestes es poden recuperar en qualsevol moment i, per tant, la informació original no es perd [153-155].

Així es poden diferenciar tres categories d'anàlisi de dades [153]:

1. Anàlisi exploratòria, la qual proporciona una visió general de les dades per tal de detectar tendències, pautes o agrupacions (Figura 2.24B).
2. Anàlisi de regressió i models predictius, que estableixen una relació quantitativa entre dos blocs de dades (Figura 2.24C).
3. Anàlisi classificatòria i discriminant, que s'utilitza per classificar les mostres en diferents categories (Figura 2.24D).

Anàlisi exploratòria

Anàlisi de components principals (PCA)

L'anàlisi de components principals (*principal component analysis*, PCA) [156, 157] és un mètode d'anàlisi multivariant no supervisat àmpliament utilitzat en diferents àmbits de la ciència, entre ells la metabolòmica. Concretant en la metabolòmica, el PCA té com a objectiu principal extreure i visualitzar la variació sistemàtica dels metabòlits presents en les mostres analitzades [153, 155-157].

Normalment, una gran part de la informació que contenen els metabòlits resolts (variables originals) és redundant (correlacionada amb altres metabòlits) o soroll (no aporta informació útil). El mètode PCA transforma el conjunt de dades originals en un nou conjunt de variables ortogonals (no correlacionades entre si), el qual és més senzill d'interpretar i fa més evident la informació més important. Aquest nou conjunt de variables s'obté a partir de la combinació lineal de les contribucions dels metabòlits originals i explica el màxim de la variància continguda en les dades originals. Malgrat això, cal destacar que la solució obtinguda per PCA no té significat físic i és difícil d'interpretar directament [153, 155-157]. En la

Figura 2.25 es mostra un exemple d'aquesta representació de les dades sobre els nous eixos ortogonals per un estudi metabolòmic amb mostres tractades a dos nivells.

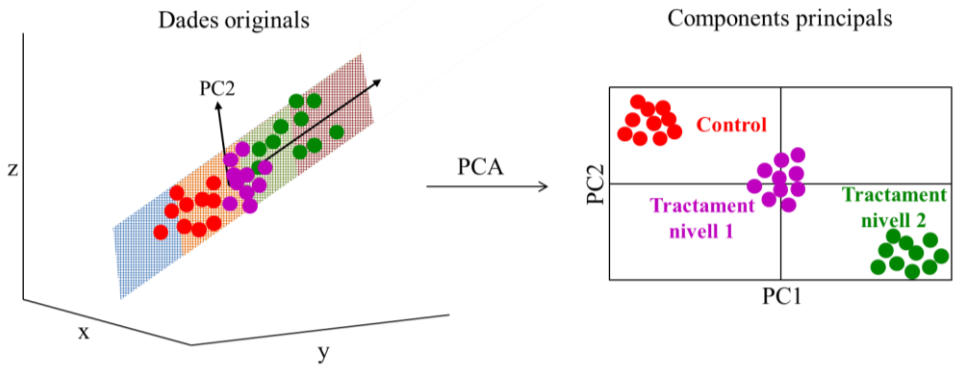


Figura 2.25. Representació de les dades sobre els nous eixos ortogonals obtinguts per PCA.

El mètode PCA resumeix la informació important que es troba continguda en la matriu de dades originals (**A**) i la separa de la part que correspon al soroll. Matemàticament, en el model de PCA la matriu original de dades **A** es descompon, seguint un model bilineal, en el producte de dues matrius ortogonals, **T** i $\mathbf{P}^{\mathbf{T}}$, tal com es mostra en l'equació 2.11 :

$$\mathbf{A} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equació 2.11}$$

On **A** és la matriu de dades original, **T** és la matriu d'*scores* (mapa de les mostres), $\mathbf{P}^{\mathbf{T}}$ defineix els *loadings* (mapa de les variables) i **E** representa la matriu dels residuals (soroll o error experimental). El producte de les matrius d'*scores* i de *loadings* reprodueix la matriu original de dades per un nombre determinat de components *N* (nombre de columnes de **T** i de files de $\mathbf{P}^{\mathbf{T}}$).

Des del punt de vista dels estudis de metabolòmica treballats en aquesta Tesi, els *scores* contenen informació sobre el tipus de mostra (control o tractada). La representació de les mostres en el nou espai definit pels components principals obtinguts es coneix amb el nom de gràfic d'*scores* (*scores plot*). Les coordenades de cada una de les mostres projectades en aquest espai són els valors d'*scores* continguts en la matriu **T** (t_{in}). El gràfic d'*scores* és molt útil, ja que proporciona una visió general de totes les mostres analitzades i de com es relacionen entre elles. Aquestes representacions d'*scores* permeten agrupar les mostres segons la seva similitud i observar les possibles tendències de les mostres, així com detectar la presència de mostres amb valors atípics (*outliers*). Per exemple, en la Figura 2.25 s'observa una tendència al llarg dels dos PCs des de les mostres control fins les mostres sotmeses al nivell alt de tractament.

En el gràfic de *loadings* (*loadings plot*) $\mathbf{P}^{\mathbf{T}}$, es descriu quina és la magnitud de la contribució de cada un des metabòlits originals en cada component descrit pel model PCA. Els metabòlits que tenen valors

grans de *loadings* en el mateix component, covarien. Si tenen el mateix signe covarien positivament, però si tenen signes oposats ho fan negativament (covarien inversament).

Finalment, cal destacar que les direccions del gràfic d'*scores* són les mateixes que les del gràfic de *loadings*. Aquesta última característica és molt útil a l'hora d'interpretar les fonts de variació experimental, ja que es pot fer a partir de la interpretació conjunta dels dos gràfics i així relacionar cada tipus de mostra (control o tractada) amb els metabòlits més destacats [153, 155-157].

Anàlisi de la variància – Anàlisi Simultània de Components (ASCA)

Els treballs de metabolòmica tenen un disseny experimental subjacent, el qual implica diversos factors d'estudi. Per exemple, l'estudi en diferents punts de temps de l'estat del metaboloma d'un organisme (com ara l'arròs) que creix exposat a un determinat contaminant (per exemple metalls pesants), implica dos factors experimentals concrets: el temps i l'exposició. Altres dissenys experimentals comuns en aquest tipus de treballs impliquen diferents nivells d'exposició, diferents tipus d'estrès simultàniament o diferents teixits a estudiar, entre d'altres [158].

L'anàlisi de la variància (*analysis of variace*, ANOVA) és una tècnica estadística utilitzada per analitzar les dades corresponents a un disseny experimental, que permet determinar i quantificar els efectes dels diferents factors experimentals. No obstant, l'ANOVA és una eina univariant, només es pot utilitzar per una única variable mesurada (metabòlit) com a funció del disseny experimental [159, 160]. En aquesta Tesi es treballa amb grans conjunts de dades multivariants, en els quals es poden arribar a resoldre centenars de metabòlits (variables). En aquests casos, cal utilitzar generalitzacions del mètode ANOVA per a dades multivariants.

En els conjunts de dades multivariants s'utilitzen múltiples respostes en funció dels canvis dels factors del disseny experimental. Així, les múltiples respostes es modelen conjuntament amb l'objectiu de trobar les possibles relacions que existeixen entre elles i amb els efectes corresponents que permetin explicar millor els processos desencadenats. En la bibliografia es troben diverses propostes per a l'extensió de l'ANOVA per a dades multivariants [158, 161-165]. La més coneguda d'aquestes extensions és l'ANOVA multivariant (MANOVA). Malauradament, quan el nombre de variables és superior al nombre de mostres, el mètode MANOVA no és adequat, ja que no pot invertir les matrius de covariància de les variables, que són singulars [166]. En els darrers anys s'han proposat alternatives per superar aquestes limitacions com, per exemple, l'anàlisi de la variància-anàlisi de components simultanis (*ANOVA-Simultaneous component analysis*, ASCA) [158, 161, 162], l'anàlisi de la variància-anàlisi de

components principals (ANOVA-PCA) [167] o la MANOVA regularitzada (*regularized MANOVA*, rMANOVA) [168].

En aquesta Tesi, s'ha emprat el mètode ASCA, que és un mètode multivariant d'anàlisi de la variància que combina la capacitat de l'ANOVA per separar les fonts de variació amb els avantatges del PCA per explicar la màxima variància comuna entre les diferents variables. En aquest enfocament, primer s'utilitza l'ANOVA per a descompondre la matriu original de dades (**A**) en diverses matrius additives que caracteritzen cada un dels factors del disseny experimental i l'error experimental (variància no explicada). Seguidament, l'anàlisi de components simultanis (s*imultaneous component analysis,* SCA) s'aplica individualment a cada matriu dels factors experimentals [158, 161, 162].

El model matemàtic de l'ASCA es descriu amb un exemple d'un estudi de metabolòmica d'arròs amb un disseny experimental de dos factors: temperatura ambiental (factor a) i quantitat d'aigua de regadiu (factor b). L'equació del model ANOVA per aquest disseny és:

$$\mathbf{A} = \bar{\mathbf{a}} + \mathbf{A_a} + \mathbf{A_b} + \mathbf{A_{ab}} + \mathbf{E} \qquad \text{Equació 2.12}$$

On **A** és la matriu de dades metabolòmiques (concentracions relatives dels metabòlits en les mostres investigades) original, $\bar{\mathbf{a}}$ és un vector fila que conté les mitjanes de totes les variables de la matriu original per totes les mostres. Les matrius $\mathbf{A_a}$ i $\mathbf{A_b}$ contenen els efectes per separat dels dos factors considerats (temperatura i aigua). La matriu $\mathbf{A_{ab}}$ és la matriu d'interacció entre els dos factors i **E** és la matriu residual, que conté la variància que no s'explica per les matrius de factors ni per la seva interacció. En el cas de les dades de tipus metabolòmic estudiades en aquesta Tesi, la matriu residual conté a més del soroll experimental, la variació biològica natural entre les mostres replicades.

El SCA que s'aplica a cada una de les matrius de factors és un mètode de descomposició bilineal utilitzat per aproximar la informació continguda en la matriu de dades original (**A**) de forma molt semblant al PCA per matrius augmentades en la direcció de les columnes. El mètode de SCA s'empra quan la matriu **A** està formada per més d'un conjunt de mostres, els quals comparteixen les mateixes variables [161]. El model matemàtic SCA és el següent:

$$\mathbf{A_{aug}} = \begin{bmatrix} \mathbf{A_1} \\ \mathbf{A_2} \\ \mathbf{A_3} \\ \vdots \\ \mathbf{A_Q} \end{bmatrix} = \begin{bmatrix} \mathbf{T_1} \\ \mathbf{T_2} \\ \mathbf{T_3} \\ \vdots \\ \mathbf{T_Q} \end{bmatrix} \mathbf{P^T} + \begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \\ \mathbf{E_3} \\ \vdots \\ \mathbf{E_Q} \end{bmatrix} \qquad \text{Equació 2.13}$$

On $\mathbf{A_q}$ són les matrius de cada subconjunt de mostres, $\mathbf{T_q}$ són els *scores* corresponents a cada un d'aquests subconjunts, $\mathbf{P^T}$ és la matriu que conté els *loadings* comuns per tots els subconjunts i $\mathbf{E_q}$ són les

matrius residuals per cada subconjunt. El SCA estima els mateixos *loadings* per tots els subconjunts de mostres, de manera que si les restriccions aplicades a cadascun dels subconjunts no varien, el SCA i el PCA són equivalents [161].

En aplicar SCA sobre cada una de les matrius dels factors, el model matemàtic d'ASCA corresponent a l'ANOVA de l'equació 2.12 per un disseny experimental de tres factors és el següent:

$$\mathbf{A} = \bar{\mathbf{a}} + \mathbf{T_a}\,\mathbf{P_a^T} + \mathbf{T_b}\,\mathbf{P_b^T} + \mathbf{T_{ab}}\,\mathbf{P_{ab}^T} + \mathbf{E} \qquad \text{Equació 2.14}$$

On les matrius $\mathbf{T_a}$, $\mathbf{T_b}$ i $\mathbf{T_{ab}}$ són les matrius dels *scores* de cada submodel i $\mathbf{P^T_a}$, $\mathbf{P^T_b}$ i $\mathbf{P^T_{ab}}$ són les matrius dels l*oadings* d'aquests submodels.

Una conseqüència d'aplicar el mètode SCA directament a les matrius de factors, és que no és possible apreciar la variabilitat natural en el gràfic d'*scores* obtingut per ASCA. Malauradament, aquesta informació és necessària per avaluar la magnitud de les diferències entre els diferents nivells de cada factor en comparació amb la variació natural [169]. Per aquest motiu, l'estimació de la variabilitat dels replicats en l'espai dels components principals de cada factor es determina mitjançant la següent equació:

$$\mathbf{Y_k} = (\mathbf{A_k} + \mathbf{E})\mathbf{P_k} = \mathbf{T_k} + \mathbf{E}\mathbf{P_k} \qquad \text{Equació 2.15}$$

En aquest model, la matriu d'efecte de cada factor més la matriu residual ($\mathbf{A_k} + \mathbf{E}$) es projecta als *loadings* obtinguts per aquest factor concret. La projecció obtinguda ($\mathbf{Y_k}$) conté la variació entre rèpliques de cada nivell del factor *k* en el subespai dels components principals d'aquest factor. $\mathbf{T_k}$ i $\mathbf{P_k}$ són els *scores* i els *loadings* del model de SCA obtingut per $\mathbf{A_k}$ [169]. En la Figura 2.26 es mostra un exemple de l'aplicació d'ASCA en l'estudi metabolòmic d'arròs amb el disseny experimental de dos factors utilitzat com exemple per explicar les equacions 2.12 i 2.14.

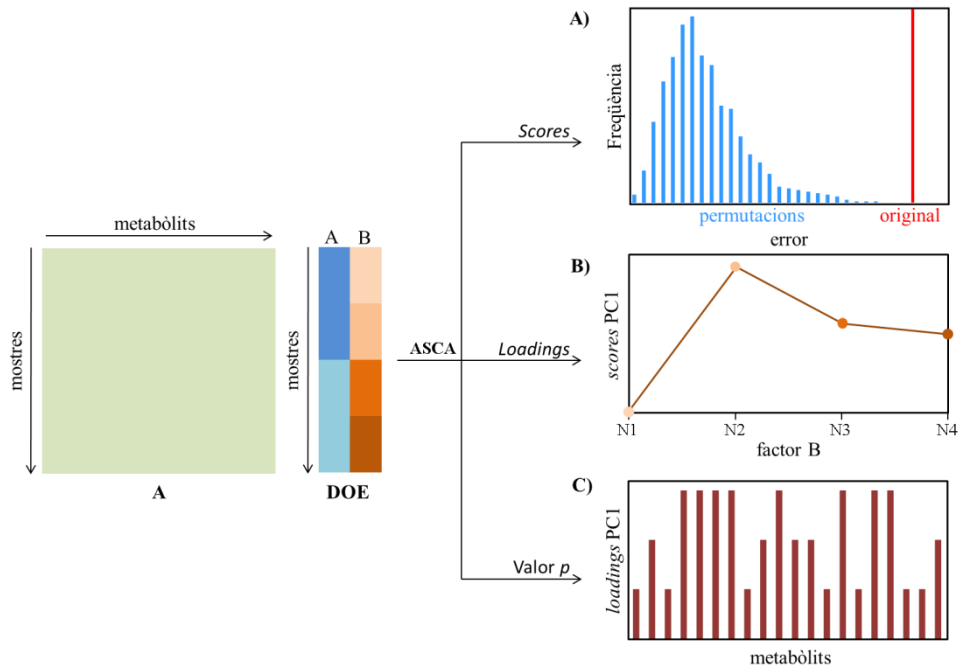


Figura 2.26. Exemple d'aplicació d'ASCA en un estudi metabolòmic amb un disseny experimental (DOE) equilibrat de dos factors. A) Histograma d'errors de les matrius permutades pel factor b, la línia vermella representa la suma de quadrats de la matriu original. B) Gràfic d'*scores* del PC1 pel factor b. C) Gràfic de *loadings* del PC1 pel factor b.

Les dades analitzades en aquesta Tesi s'han obtingut en tots els casos a partir de dissenys experimentals balancejats (*balanced*). Un disseny balancejat és aquell que té el mateix nombre d'observacions replicades per cada nivell dels factors experimentals.

Per tal d'avaluar estadísticament els efectes dels factors investigats i les seves interaccions, s'aplica un test de permutacions. La hipòtesi nul·la (H_0) d'aquest test de permutacions assumeix que els factors considerats no tenen un efecte significatiu. El test s'aplica de manera que es permuten totes les files de les matrius de dades originals de tots els factors (per exemple 10000 permutacions) i se'n torna a calcular el seu error. Seguidament, es construeix un histograma a partir dels errors de totes les matrius permutades per un factor concret. En la Figura 2.26A es mostra un exemple d'aquest histograma. Quan aquest factor té un efecte significatiu en el sistema estudiat, la suma de quadrats de la matriu original és major que en la majoria de les permutacions realitzades. La significació estadística dels factors es quantifica mitjançant el valor *p* (*p*-value). Aquest valor es calcula dividint el nombre de casos en els que l'error de les matrius permutades és major a l'original entre el nombre total de permutacions realitzades. Si el valor *p* és menor que el nivell de significació prefixat es rebutja la H_0 i, per tant, es considera que el factor té un efecte significatiu [162, 170].

L'avantatge de l'ASCA és que a banda d'avaluar la significació dels efectes dels factors del disseny experimental estudiat, també aporta les matrius de *scores* i *loadings* de cada factor. D'una banda, la matriu d'*scores* permet observar el comportament dels diferents tipus de mostres en relació a cada factor (veure Figura 2.26B). Per exemple, si s'estudia el factor temps, el gràfic d'*scores* il·lustra entre quins punts de temps hi ha possibles canvis significatius. D'altra banda, la matriu de *loadings* aporta informació sobre quines de les variables originals tenen més pes en els canvis causat per cada factor [158] (veure Figura 2.26C).

Finalment, cal comentar que el model matemàtic explicat per ASCA només serveix quan es treballa amb un disseny balancejat. Quan el disseny experimental que s'estudia no és balancejat (*nested*), les matrius dels diferents factors ($\mathbf{A_k}$) no són ortogonals i, per tant, l'equació 2.12 no es compleix estrictament. En aquests casos, cal utilitzar l'anàlisi de components principals multinivell (multilevel s*imultaneous componenet analysis* ,MSCA) [163] o l'ASCA+ [165], que són extensions de l'ASCA per a dissenys no equilibrats [163, 165, 171].

MCR-ALS exploratori

El mètode de MCR aplicat sobre la matriu d'àrees dels perfils eluïts dels metabòlits resolts és útil per a extreure informació addicional que ajudi a interpretar les variacions produïdes com a conseqüència del disseny experimental emprat. En aquests casos, la matriu d'àrees **A** es descompon seguint el mateix model bilineal de l'equació 1:

$$\mathbf{A} = \mathbf{M}\mathbf{V}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equació 2.16}$$

On la matriu de les àrees **A** es descompon en el producte de dues matrius noves, **M** i $\mathbf{V}^{\mathbf{T}}$. En aquest cas, la matriu **M** conté informació sobre el comportament de les mostres, mentre que la matriu $\mathbf{V}^{\mathbf{T}}$ aporta informació sobre la influència de les variables en cada component resolt. La matriu **M** aporta informació quantitativa sobre el problema estudiat i conté les concentracions relatives dels diferents components en totes les mostres durant l'experiment. Així cada columna conté l'evolució de cada mostra en el procés estudiat. La matriu $\mathbf{V}^{\mathbf{T}}$, en canvi, aporta informació qualitativa sobre el problema estudiat i permet determinar les característiques dels processos resolts [135]. Per exemple, en el cas de l'estudi en diferents punts de temps de l'estat del metaboloma d'un organisme que creix exposat a un determinat estrès ambiental (per exemple arròs en condicions de sequera), els components resolts per MCR explicaran els canvis que es produeixen en el metaboloma de l'organisme en funció del temps i de l'estrès estudiat. De tal manera que la matriu **M** contindrà la contribució que tenen aquests canvis del metaboloma en les

diverses mostres analitzades, per cadascun dels components. Mentre que la matriu $\mathbf{V}^T$ aportarà informació sobre quins metabòlits són els responsables dels canvis observats en cada component. La Figura 2.27 mostra una representació gràfica d'aquest exemple.

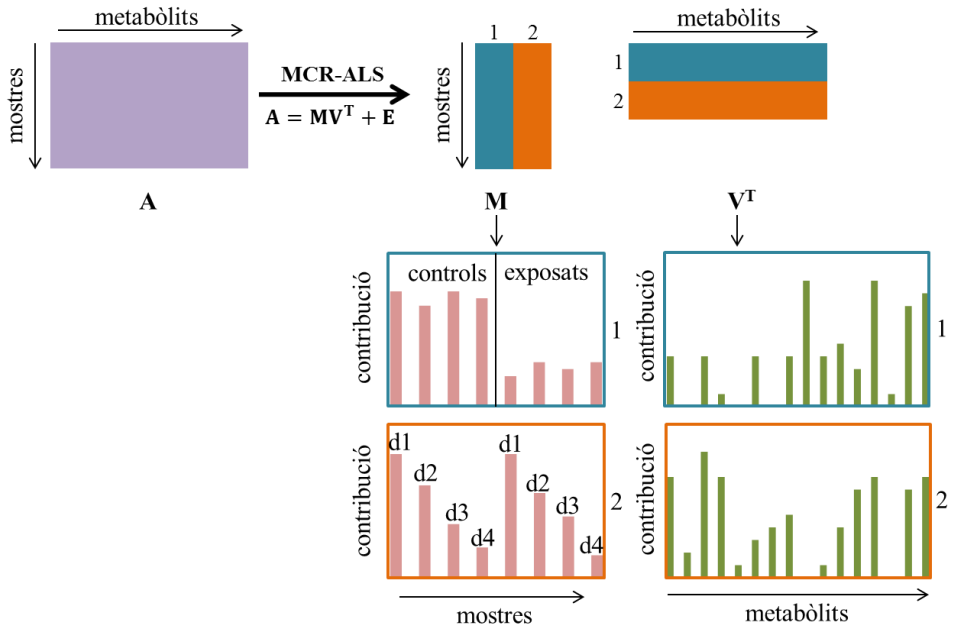


Figura 2.27. Representació gràfica de la resolució d'una matriu **A** pel mètode MCR-ALS. Exemple de l'estudi dels canvis en diferents punts del temps en el metaboloma d'un organisme exposat a un estrès ambiental. El primer component explica els canvis deguts a l'estrès ambiental i el segon component explica els canvis deguts al pas del temps.

Anàlisi de regressió i models predictius

Mínims quadrats parcials (PLS)

En aquesta Tesi s'ha utilitzat l'anàlisi per mínims quadrats parcials (*partial least squares*, PLS) [172-174] com a model predictiu en estudis de relació quantitativa estructura-activitat (*quantitative structure-activity relationship*, QSAR) [175]. El QSAR és el procés pel qual l'estructura química d'un conjunt de molècules es correlaciona quantitativament amb alguna propietat física o química, la seva reactivitat durant un procés o la seva activitat biològica [175, 176]. Així s'ha relacionat el factor de retenció cromatogràfica dels metabòlits analitzats amb la seva estructura química, modelitzada mitjançant descriptors moleculars (MDs) [177]. Aquest tipus concret de models de QSAR en els que es prediu la retenció cromatogràfica de les molècules es coneix amb el nom de relació quantitativa estructura-retenció (*quantitative structure-retention relationship*, QSRR) [178, 179]. En estudis de metabolòmica no dirigits és interesant l'aplicació de models de QSRR, ja que poden aportar informació que pot ajudar a identificar els metabòlits detectats [180, 181].

El PLS és un mètode de regressió lineal multivariant que permet trobar models de correlació entre una matriu **X** que conté un conjunt de variables predictores (MDs que caracteritzen l'estructura dels metabòlits) i un vector **y** que conté els valors resposta a predir (factors de retenció dels metabòlits) [173, 182, 183]. El model PLS treballa amb les variables latents (*latent variables*, LV) de la matriu **X** (variables predictores), les quals són variables subjacents o no explícites que causen la variació experimental observada. En aquest model, es realitza una regressió inversa, en la qual el model calculat relaciona les LVs de **X** amb **y** (variables a predir) segons l'equació 2.17:

$$\mathbf{y} = \mathbf{bX} \qquad \text{Equació 2.17}$$

On **X** és la matriu que conté les variables predictores (MDs), **y** és el vector resposta de variables dependents (factors de retenció) i **b** és un vector que conté els coeficients de regressió calculats durant la construcció del model (calibratge). La finalitat d'aquesta regressió inversa és maximitzar la covariància entre **X** i **y** utilitzant el mínim nombre de LVs. En la Figura 2.28A es mostra la descomposició bilineal que realitza el model PLS dels blocs de dades **X** i **y**, així com la relació entre aquestes mitjançant la matriu de coeficients de pes **W**.

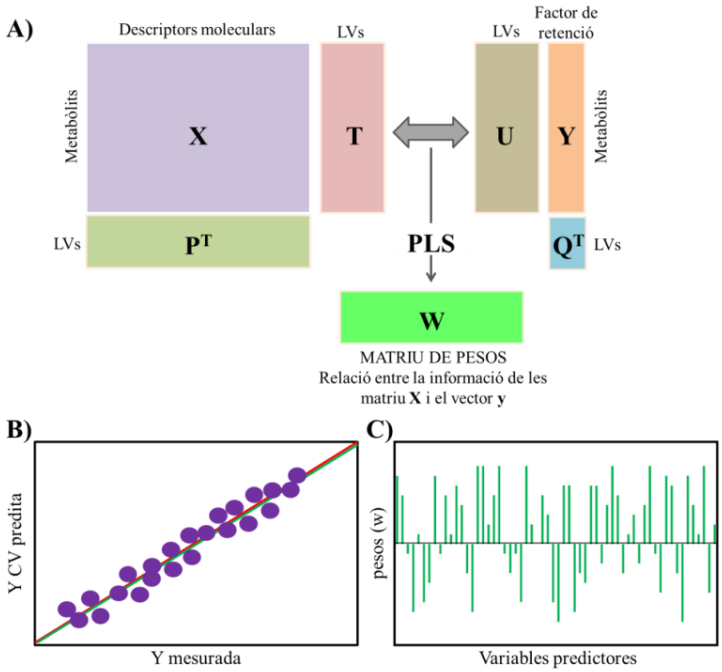


Figura 2.28. A) Representació gràfica de la descomposició de **X** i **y** durant el procés del mètode de PLS, així com de la seva relació a través de la matriu de pesos **W** obtinguda. B) Gràfic de correlació entre els valors predits i mesurats de **y**. C) Gràfic de coeficients de pes, el qual mostra la importància de les variables predictores (matriu **X**) en la predicció de **y**.

Aquesta descomposició bilineal dels dos blocs de dades **X** i **y** es realitza segons les equacions 2.18 i 2.19:

$$\mathbf{y} = \mathbf{Uq} + \mathbf{z}$$ Equació 2.18

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E}$$ Equació 2.19

On **T** i **U** són les matrius d'*scores* de les LVs de **X** i **y** respectivament, **P** i **q** contenen els *loadings* de les variables en aquestes LVs. Finalment, **z** i **E** contenen la variància no explicada pel model PLS (errors) en **y** i **X**, respectivament.

Cal destacar que els s*cores* continguts en la matriu **T** modelen els MDs originals (variables predictores originals, **X)** i, a més, prediuen els factors de retenció (**y**). S'assumeix que **X** i **y** es modelen mitjançant les mateixes LVs. Els *scores* de **X** es construeixen a partir de combinacions lineals de les variables originals amb uns determinats coeficients de pes. De tal manera que, per cada LV, s'obté un vector de coeficients de pes (*weights*), **w**, que descriu la importància de cadascun dels MDs (matriu **X**) en la predicció dels factors de retenció dels metabòlits estudiats (vector **y)**. Els vectors **w** de totes les LVs, s'agrupen en una matriu **W**, la qual reflecteix la covariància entre **X** i **y** i, per tant, s'utilitza pel càlcul del vector de regressió **b** segons les equacions 2.20 i 2.21:

$$\mathbf{b} = \mathbf{X^+ y}$$ Equació 2.20

$$\mathbf{b} = \mathbf{W}(\mathbf{P^T W})^{-1}\,\mathbf{q}$$ Equació 2.21

En la Figura 2.28B es mostra un exemple de la informació que proporciona el mètode de PLS. Igual que en cas del PCA, el gràfic dels *scores* de **X** dóna una visió general de tots els metabòlits estudiats, així com de les seves relacions. D'altra banda, el gràfic dels coeficients de pes (**W**) descriu per cada LV quina és la magnitud de la contribució de cadascun dels MDs de la matriu **X**, en la predicció dels factors de retenció del vector **y** [172-174].

En aquesta Tesi, s'ha utilitzat la validació interna creuada (*cross-validation*, CV) [184, 185] per tal de determinar el nombre de LVs necessari per predir i explicar suficientment bé la variable **y**. Bàsicament, la CV consisteix en dividir el conjunt de dades original en un nombre reduït de subgrups. Seguidament s'analitzen mitjançant el model de PLS tots els subgrups excepte un. Els subgrups analitzats s'anomenen conjunt de dades d'entrenament (*training set*), mentre que el subgrup exclòs de l'anàlisi s'anomena conjunt de dades de prova (*test set*). El conjunt de dades de prova s'utilitza per validar l'anàlisi realitzada. De manera que el model de predicció es construeix utilitzant el conjunt de dades d'entrenament i, a partir d'aquest, es prediuen els valors del conjunt de dades de prova. Finalment, els

valors predits per les dades de prova es comparen amb els valors originals. Aquest procés es repeteix iterativament. A partir de la comparació dels valors predits i els valors originals de les dades de prova es calcula l'error de CV (*root mean square error of cross-validation*, RMSECV) segons la següent equació:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad \text{Equació 2.22}$$

On $\hat{y}_i$ és el valor predit i y_i, el valor original.

Existeixen diferents tipus de validació creuada (CV). En aquesta Tesi s'ha utilitzat la *leave-one-out cross validation* (LOOCV) en els conjunts de dades amb menys de 20 mostres i l'aleatòria en la resta. En la LOOCV es separen les dades de manera que en cada iteració una sola mostra forma el conjunt de dades de prova i tota la resta conforma les dades d'entrenament. En canvi, en la CV aleatòria en cada iteració es divideixen aleatòriament el conjunt de dades d'entrenament i el de prova en grups de mida variable segons el nombre total de mostres.

Anàlisi classificatòria i discriminant

Anàlisi discriminat per mínims quadrats parcials (PLS-DA)

En els estudis de metabolòmica és habitual utilitzar l'anàlisi discriminant per mínims quadrats parcials (*partial least squares– discriminant analysis*, PLS-DA) [186] per a discriminar entre els diferents grups de mostres de l'estudi (controls contra tractades). El PLS-DA és una extensió del PLS en la que el vector de la variable dependent **y** conté valors que descriuen les categories o classes de les mostres analitzades en **X**. Dit d'una altra manera, en el mètode de PLS-DA el que es prediu és si les diverses mostres analitzades pertanyen a una classe determinada [186, 187]. Per exemple, en un estudi on s'observen els canvis que pateix el metaboloma d'un organisme sota la influència d'un determinat estrès ambiental (com ara la presència de contaminats químics o estressants físics com la temperatura o la humitat), la matriu **X** conté la concentració relativa dels metabòlits resolts en cada una de les mostres analitzades. D'altra banda, el vector **y** conté la informació relativa al tipus de mostres analitzades (controls o exposades al contaminant o estressant físic), de manera que una variable y_k serà igual a 0 si la mostra *k* és un control i en canvi, serà igual a 1 si es tracta d'una mostra exposada. En aquest exemple, el model de correlació construït pel PLS-DA prediu si les mostres analitzades són controls o han estat tractades amb el contaminant o estressant físic estudiat. L'aplicació del PLS-DA en aquest exemple es mostra en la Figura 2.29.

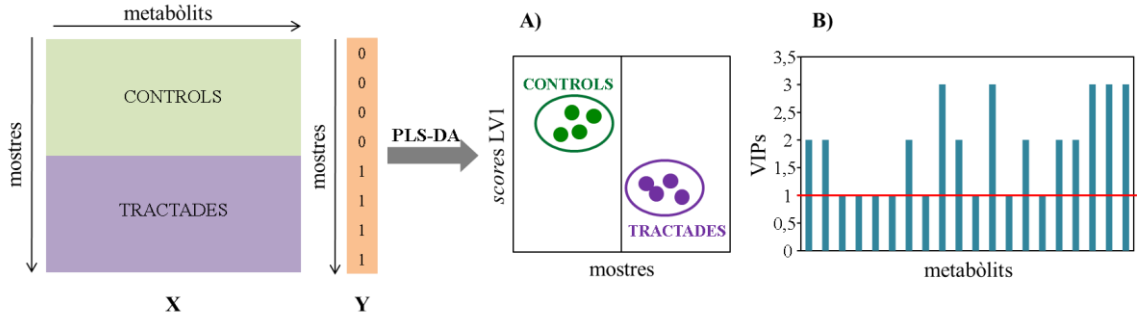


Figura 2.29. Exemple de l'aplicació de PLS-DA en un estudi de metabolòmica per a la diferenciació de les mostres controls de les mostres tractades amb un contaminant. A) Gràfic d'*scores*. B) Gràfic de VIPs.

Igual que el mètode de regressió PLS, el PLS-DA fa l'estimació més eficient de les LVs de la matriu de dades de canvis de concentracions dels metabòlits en les mostres analitzades, **X**, les quals es relacionen de manera òptima amb els canvis observats en el vector de classes (mostres tractades i controls) **y**. El PLS-DA construeix un model que maximitza la covariància entre **X** i **y** utilitzant el mínim nombre de LVs possible. D'aquesta manera, el vector de classes **y** es prediu mitjançant un nombre reduït de factors (LVs) enlloc de utilitzant tots els metabòlits, els quals habitualment contenen informació repetida (estan correlacionats). Igual que en cal del PLS, per cada LV s'obté un vector de coeficients de pes (**w**) que mostra quins són els metabòlits de la matriu **X** que combinen millor per construir els seus *scores* (matriu **T**) [186, 187] de cara a la predicció de les classes, en el vector **y**. En la Figura 2.29A es mostra un exemple del gràfic d'*scores* obtingut pel mètode de PLS-DA.

A més, el PLS-DA també permet deduir quins són els metabòlits de la matriu **X** més importants per aconseguir la discriminació entre mostres. Aquesta característica és especialment interesant en estudis de metabolòmica no dirigida, ja que ajuda a seleccionar quins són els anàlits detectats que presenten diferències significatives entre les mostres controls i les mostres tractades. Aquesta informació s'obté a partir de diversos paràmetres, com el pes dels *loadings*, les variables importants en projecció (VIPs), la *selectivity ratio* i els coeficient de regressió. En aquesta Tesi, s'ha utilitzat el mètode de selecció de variables importants en projecció (*variable importance in projection*, VIP), el qual dóna una mesura de la influència individual de cadascun dels metabòlits (variables) originals de **X** en la construcció del model PLS per a la predicció de **y**.

El mètode de selecció de variables mitjançant els VIPs va ser proposat l'any 1993 per Wold i els seus col·laboradors [183]. Els VIPs, o coeficients VIP (VIP *scores*), són una suma ponderada dels quadrats dels coeficients de pes (vector **w**) obtinguts en la construcció del model PLS, la qual mesura la importància de cada variable predictora en el model PLS. Per aquest motiu, els valors VIPs són una eina

útil per a la selecció de quins metabòlits contribueixen més en l'explicació de la variància del vector **y**. En l'equació 2.23 es mostra com es calcula el valor de VIP per un determinat metabòlit *j*:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} W_{jf}^2 SSY_f J}{SSY_{total} F}} \quad \text{Equació 2.23}$$

$$SSY_f = b_f^2 t_f' t_f \quad \text{Equació 2.24}$$

$$SSY_{total} = \mathbf{b^2 T' T} \quad \text{Equació 2.25}$$

On w_{jf} és el valor del pes del metabòlit *j* en la LV_f. SSY_f és la suma de quadrats de la variància explicada per la LV_f i SSY_{total} és la suma de quadrats de la variància total explicada per **y**, *J* és el nombre total de metabòlits en la matriu **X**, *F* és el nombre de LVs considerades, **T** són els *scores* de la matriu **X** i **b** és el vector de coeficients de regressió del model de PLS. De manera que els VIPs mesuren la contribució de cada metabòlit en funció de la variància explicada per cada LV i del pes de cada metabòlit en cadascuna d'aquestes LVs.

La mitjana del valor dels VIPs al quadrat és igual a 1, per això habitualment s'utilitza aquest valor de la unitat o més gran que la unitat com a criteri de selecció de les variables considerades més importants. Aquest criteri es coneix amb el nom de "més gran que la unitat" [188]. En la Figura 2.29B, es mostra un exemple del gràfic de VIPs obtingut pel mètode de PLS-DA.

La qualitat dels models de PLS-DA s'avalua mitjançant la matriu de confusió diagnòstic del model [189]. Aquesta matriu conté el nombre de vertaders (V) i falsos (F) positius i negatius obtinguts pel model. La matriu de confusió per un model de dues classes s'estructura de la següent manera:

Taula 2.7. Matriu de confusió diagnòstic del model. La classe tractat s'identifica com a positiva (P) i la classe control, com a negativa (N). On VP i VN corresponen a les prediccions fetes correctament, i FP i FN a les fetes incorrectament.

		Classe Calculada	
		Tractat (P)	**Control (N)**
Classe Experimental	**Tractat**	VP	FN
	Control	FP	VN

A partir dels elements de la matriu de confusió es calculen la sensibilitat (equació 2.26) i l'especificitat (equació 2.27) del model i el coeficient de correlació de Matthews (MCC, equació 2.28) [190]:

$$Sensibilitat = \frac{VP}{VP+FN} \quad \text{Equació 2.26}$$

$$Especificitat = \frac{VN}{VN+FP} \quad \text{Equació 2.27}$$

$$\mathrm{MCC} = \frac{\mathrm{VP} \times \mathrm{VN} - \mathrm{FR} \times \mathrm{FN}}{\sqrt{(\mathrm{VP}+\mathrm{FP})(\mathrm{VP}+\mathrm{FN})(\mathrm{VN}+\mathrm{FP})(\mathrm{VN}+\mathrm{FN})}} \quad \text{Equació 2.28}$$

On *VP* és el nombre de vertaders positius, *VN* en nombre de vertaders negatius, *FP* el nombre de falsos positius i *FN* el nombre de falsos positius.

La sensibilitat és la probabilitat de classificar correctament una mostra a la classe que pertany, mentre que l'especificitat és la probabilitat de classificar correctament una mostra que no pertany a la classe. Per últim, el MCC dóna un valor entre -1 i 1 que expressa la qualitat de les classificacions binàries: un coeficient de 1 representa un model de predicció perfecte, 0 una predicció aleatòria i -1 representa un desacord total entre la predicció i l'observació.

Alternativament, la selecció de les variables que presenten diferències significatives entre les mostres control i les mostres tractades també es pot realitzar mitjançant mètodes d'estadística univariant [191, 192]. En aquesta Tesi s'ha utilitzat el procediment d'anàlisi de la variància (ANOVA) d'un factor amb una correcció de Bonferroni per a controlar els falsos positius. Aquesta correcció té en compte que la probabilitat d'obtenir falsos positius augmenta amb el nombre de tests realitzats simultàniament (un per cada metabòlit considerat). La correcció de Bonferroni estableix un valor llindar de probabilitat (α) per a cada metabòlit individual segons:

$$\alpha = P/k \quad \text{Equació 2.29}$$

On *P* és el valor de la probabilitat llindar prèviament establert (habitualment 0,05) i *k* correspon al nombre total de metabòlits considerats. D'aquesta forma, un metabòlit es considera significatiu si el seu valor de *p* és inferior a α [191].

2.4. Referències

1. Haiyuan, Z.;Hao, H.;Cao, D.;Yeona, C.;Shengtao, Z.;Fuqiang, H.;Qin, Z., Integrative System Biology Strategies for Disease Biomarker Discovery, *Combinatorial Chemistry & High Throughput Screening*, 2012, **15**, 286-298.
2. Monteiro, M. S.;Carvalho, M.;Bastos, M. L.;Pinho, P. G. d., Metabolomics Analysis for Biomarker Discovery: Advances and Challenges, *Current Medicinal Chemistry*, 2013, **20**, 257-271.
3. Horgan, R. P.;Kenny, L. C., 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics, *The Obstetrician & Gynaecologist*, 2011, **13**, 189-195.
4. Patti, G. J.;Yanes, O.;Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy, *Nature Reviews Molecular Cell Bioogyl*, 2012, **13**, 263-269.
5. Chadeau-Hyam, M.;Campanella, G.;Jombart, T.;Bottolo, L.;Portengen, L.;Vineis, P.;Liquet, B.;Vermeulen, R. C. H., Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers, *Environmental and Molecular Mutagenesis*, 2013, **54**, 542-557.
6. Fiehn, O.;Kopka, J.;Dörmann, P.;Altmann, T.;Trethewey, R. N.;Willmitzer, L., Metabolite profiling for plant functional genomics, *Nature Biotechnology*, 2000, **18**, 1157-1161.
7. Nicholson, J. K.;Lindon, J. C.;Holmes, E., 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of

biological NMR spectroscopic data, *Xenobiotica; the fate of foreign compounds in biological systems*, 1999, **29**, 1181-1189.
8. Xiao, J. F.;Zhou, B.;Ressom, H. W., Metabolite identification and quantitation in LC-MS/MS-based metabolomics, *TrAC Trends in Analytical Chemistry*, 2012, **32**, 1-14.
9. Griffiths, W. J.;Koal, T.;Wang, Y.;Kohl, M.;Enot, D. P.;Deigner, H.-P., Targeted Metabolomics for Biomarker Discovery, *Angewandte Chemie International Edition*, 2010, **49**, 5426-5445.
10. Nicholson, J. K.;Lindon, J. C., Systems biology: Metabonomics, *Nature*, 2008, **455**, 1054-1056.
11. Srivastava, S., Move Over Proteomics, Here Comes Glycomics, *Journal of Proteome Research*, 2008, **7**, 1799-1799.
12. Phillips, M.;Cataneo, R. N.;Chaturvedi, A.;Kaplan, P. D.;Libardoni, M.;Mundada, M.;Patel, U.;Zhang, X., Detection of an Extended Human Volatome with Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry, *PLOS ONE*, 2013, **8**, e75274.
13. Wenk, M. R., The emerging field of lipidomics, *Nature Reviews Drug Discovery*, 2005, **4**, 594.
14. Checa, A.;Bedia, C.;Jaumot, J., Lipidomic data analysis: Tutorial, practical guidelines and applications, *Analytica Chimica Acta*, 2015, **885**, 1-16.
15. Lam, S. M.;Shui, G., Lipidomics as a Principal Tool for Advancing Biomedical Research, *Journal of Genetics and Genomics*, 2013, **40**, 375-390.
16. Sethi, S.;Brietzke, E., Recent advances in lipidomics: Analytical and clinical perspectives, *Prostaglandins and Other Lipid Mediators*, 2017, **128-129**, 8-16.
17. Yang, K.;Han, X., Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences, *Trends in Biochemical Sciences*, 2016, **41**, 954-969.
18. Fahy, E.;Subramaniam, S.;Brown, H. A.;Glass, C. K.;Merrill Jr, A. H.;Murphy, R. C.;Raetz, C. R. H.;Russell, D. W.;Seyama, Y.;Shaw, W.;Shimizu, T.;Spener, F.;Van Meer, G.;VanNieuwenhze, M. S.;White, S. H.;Witztum, J. L.;Dennis, E. A., A comprehensive classification system for lipids, *Journal of Lipid Research*, 2005, **46**, 839-861.
19. Ohlrogge, J.;Browse, J., Lipid biosynthesis, *Plant Cell*, 1995, **7**, 957-970.
20. Bouhifd, M.;Hartung, T.;Hogberg, H. T.;Kleensang, A.;Zhao, L., Review: Toxicometabolomics, *Journal of Applied Toxicology*, 2013, **33**, 1365-1383.
21. Nicholson, J. K.;Connelly, J.;Lindon, J. C.;Holmes, E., Metabonomics: a platform for studying drug toxicity and gene function, *Nature Reviews Drug Discovery*, 2002, **1**, 153-161.
22. Robinson, A. B.;Robinson, N. E., in *Metabolic Profiling: Methods and Protocols*, ed. T. O. Metz, Humana Press, Totowa, NJ, 2011, DOI: 10.1007/978-1-61737-985-7_1, pp. 1-23.
23. Buescher, J. M.;Moco, S.;Sauer, U.;Zamboni, N., Ultrahigh Performance Liquid Chromatography−Tandem Mass Spectrometry Method for Fast and Robust Quantification of Anionic and Aromatic Metabolites, *Analytical Chemistry*, 2010, **82**, 4403-4412.
24. Lu, W.;Bennett, B. D.;Rabinowitz, J. D., Analytical strategies for LC–MS-based targeted metabolomics, *Journal of Chromatography B*, 2008, **871**, 236-242.
25. Yanes, O.;Tautenhahn, R.;Patti, G. J.;Siuzdak, G., Expanding Coverage of the Metabolome for Global Metabolite Profiling, *Analytical Chemistry*, 2011, **83**, 2152-2161.
26. Khamis, M. M.;Adamko, D. J.;El-Aneed, A., Mass spectrometric based approaches in urine metabolomics and biomarker discovery, *Mass Spectrometry Reviews*, 2017, **36**, 115-134.
27. Rochat, B., From targeted quantification to untargeted metabolomics: Why LC-high-resolution-MS will become a key instrument in clinical labs, *TrAC Trends in Analytical Chemistry*, 2016, **84**, 151-164.
28. Melnik, A. V.;da Silva, R. R.;Hyde, E. R.;Aksenov, A. A.;Vargas, F.;Bouslimani, A.;Protsyuk, I.;Jarmusch, A. K.;Tripathi, A.;Alexandrov, T.;Knight, R.;Dorrestein, P. C., Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples, *Analytical Chemistry*, 2017, **89**, 7549-7559.
29. Robertson, D. G., Metabonomics in Toxicology: A Review, *Toxicological Sciences*, 2005, **85**, 809-822.
30. Kuang, H.;Li, Z.;Peng, C.;Liu, L.;Xu, L.;Zhu, Y.;Wang, L.;Xu, C., Metabonomics Approaches and the Potential Application in Foodsafety Evaluation, *Critical Reviews in Food Science and Nutrition*, 2012, **52**, 761-774.
31. Castro-Puyana, M.;Pérez-Míguez, R.;Montero, L.;Herrero, M., Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability, *TrAC - Trends in Analytical Chemistry*, 2017, **93**, 102-118.
32. Sardans, J.;Peñuelas, J.;Rivas-Ubach, A., Ecological metabolomics: overview of current developments and future challenges, *Chemoecology*, 2011, **21**, 191-225.

33. García-Sevillano, M. Á.;García-Barrera, T.;Gómez-Ariza, J. L., Environmental metabolomics: Biological markers for metal toxicity, *ELECTROPHORESIS*, 2015, **36**, 2348-2365.
34. Viant, M. R., Recent developments in environmental metabolomics, *Molecular BioSystems*, 2008, **4**, 980-986.
35. Meena, K. K.;Sorty, A. M.;Bitla, U. M.;Choudhary, K.;Gupta, P.;Pareek, A.;Singh, D. P.;Prabha, R.;Sahu, P. K.;Gupta, V. K.;Singh, H. B.;Krishanani, K. K.;Minhas, P. S., Abiotic Stress Responses and Microbe-Mediated Mitigation in Plants: The Omics Strategies, *Frontiers in Plant Science*, 2017, **8**.
36. Nakabayashi, R.;Saito, K., Integrated metabolomics for abiotic stress responses in plants, *Current Opinion in Plant Biology*, 2015, **24**, 10-16.
37. Choi, Y. H.;Kim, H. K.;Linthorst, H. J. M.;Hollander, J. G.;Lefeber, A. W. M.;Erkelens, C.;Nuzillard, J.-M.;Verpoorte, R., NMR Metabolomics to Revisit the Tobacco Mosaic Virus Infection in Nicotiana tabacum Leaves, *Journal of Natural Products*, 2006, **69**, 742-748.
38. Cuperlovic-Culf, M.;Rajagopalan, N.;Tulpan, D.;Loewen, M. C., Metabolomics and cheminformatics analysis of antifungal function of plant metabolites, *Metabolites*, 2016, **6**. e31
39. Hong, J.;Yang, L.;Zhang, D.;Shi, J., Plant Metabolomics: An Indispensable System Biology Tool for Plant Science, *International Journal of Molecular Sciences*, 2016, **17**. e767
40. Mazid, M.;Khan, T. A.;Mohammad, F., Role of secondary metabolites in defense mechanisms of plants, *Biology and Medicine*, 2011, **3**, 232-249.
41. Bennett, R. N.;Wallsgrove, R. M., Tansley Review No. 72. Secondary Metabolites in Plant Defence Mechanisms, *The New Phytologist*, 1994, **127**, 617-633.
42. Fukusaki, E.;Kobayashi, A., Plant metabolomics: potential for practical operation, *Journal of Bioscience and Bioengineering*, 2005, **100**, 347-354.
43. Saito, K.;Matsuda, F., Metabolomics for Functional Genomics, Systems Biology, and Biotechnology, *Annual Review of Plant Biology*, 2010, **61**, 463-489.
44. Tian, H.;Lam, S.;Shui, G., Metabolomics, a Powerful Tool for Agricultural Research, *International Journal of Molecular Sciences*, 2016, **17**, 1871.
45. Sumner, L. W.;Lei, Z.;Nikolau, B. J.;Saito, K., Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects, *Natural Product Reports*, 2015, **32**, 212-229.
46. Boyer, J. S., Plant productivity and environment, *Science*, 1982, **218**, 443-448.
47. van Velthuizen, H.;Food;Department, A. O. o. t. U. N. S. D., *Mapping Biophysical Factors that Influence Agricultural Production and Rural Vulnerability*, Food and Agriculture Organization of the United Nations, 2007.
48. Cramer, G. R.;Urano, K.;Delrot, S.;Pezzotti, M.;Shinozaki, K., Effects of abiotic stress on plants: a systems biology perspective, *BMC Plant Biology*, 2011, **11**, 163-163.
49. Debnath, M.;Pandey, M.;Bisen, P. S., An omics approach to understand the plant abiotic stress, *Omics : a journal of integrative biology*, 2011, **15**, 739-762.
50. Obata, T.;Fernie, A. R., The use of metabolomics to dissect plant responses to abiotic stresses, *Cellular and Molecular Life Sciences*, 2012, **69**, 3225-3243.
51. Arbona, V.;Manzi, M.;Ollas, C.;Gómez-Cadenas, A., Metabolomics as a Tool to Investigate Abiotic Stress Tolerance in Plants, *International Journal of Molecular Sciences*, 2013, **14**, 4885.
52. Rodziewicz, P.;Swarcewicz, B.;Chmielewska, K.;Wojakowska, A.;Stobiecki, M., Influence of abiotic stresses on plant proteome and metabolome changes, *Acta Physiologiae Plantarum*, 2014, **36**, 1-19.
53. Singh, B.;Bohra, A.;Mishra, S.;Joshi, R.;Pandey, S., Embracing new-generation 'omics' tools to improve drought tolerance in cereal and food-legume crops, *Biologia Plantarum*, 2015, **59**, 413-428.
54. Shanker, A. K.;Maheswari, M.;Yadav, S. K.;Desai, S.;Bhanu, D.;Attal, N. B.;Venkateswarlu, B., Drought stress responses in crops, *Functional & Integrative Genomics*, 2014, **14**, 11-22.
55. Shanker, A. K.;Maheswari, M.;Yadav, S. K.;Desai, S.;Bhanu, D.;Attal, N. B.;Venkateswarlu, B., Drought stress responses in crops, *Functional & integrative genomics*, 2014, **14**, 11-22.
56. Stocker, T., *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2014.
57. Mittler, R.;Finka, A.;Goloubinoff, P., How do plants feel the heat?, *Trends in Biochemical Sciences*, 2012, **37**, 118-125.
58. Guy, C.;Kaplan, F.;Kopka, J.;Selbig, J.;Hincha, D. K., Metabolomics of temperature stress, *Physiologia plantarum*, 2008, **132**, 220-235.

59. Guy, C.;Kaplan, F.;Kopka, J.;Selbig, J.;Hincha, D. K., Metabolomics of temperature stress, *Physiologia Plantarum*, 2008, **132,** 220-235.
60. Singh, S.;Parihar, P.;Singh, R.;Singh, V. P.;Prasad, S. M., Heavy Metal Tolerance in Plants: Role of Transcriptomics, Proteomics, Metabolomics, and Ionomics, *Frontiers in Plant Science*, 2015, **6**, 1143.
61. Fields, S.;Johnston, M., Whither Model Organism Research?, *Science*, 2005, **307**, 1885-1886.
62. Edison, A. S.;Hall, R. D.;Junot, C.;Karp, P. D.;Kurland, I. J.;Mistrik, R.;Reed, L. K.;Saito, K.;Salek, R. M.;Steinbeck, C.;Sumner, L. W.;Viant, M. R., The Time Is Right to Focus on Model Organism Metabolomes, *Metabolites*, 2016, **6**.
63. Izawa, T.;Shimamoto, K., Becoming a model plant: The importance of rice to plant science, *Trends in Plant Science*, 1996, **1**, 95-99.
64. Schoof, H.;Karlowski, W. M., Comparison of rice and Arabidopsis annotation, *Current Opinion in Plant Biology*, 2003, **6**, 106-112.
65. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, *Nature*, 2000, **408**, 796-815.
66. Goff, S. A.;Ricke, D.;Lan, T. H.;Presting, G.;Wang, R.;Dunn, M.;Glazebrook, J.;Sessions, A.;Oeller, P.;Varma, H.;Hadley, D.;Hutchison, D.;Martin, C.;Katagiri, F.;Lange, B. M.;Moughamer, T.;Xia, Y.;Budworth, P.;Zhong, J.;Miguel, T.;Paszkowski, U.;Zhang, S.;Colbert, M.;Sun, W. L.;Chen, L.;Cooper, B.;Park, S.;Wood, T. C.;Mao, L.;Quail, P.;Wing, R.;Dean, R.;Yu, Y.;Zharkikh, A.;Shen, R.;Sahasrabudhe, S.;Thomas, A.;Cannings, R.;Gutin, A.;Pruss, D.;Reid, J.;Tavtigian, S.;Mitchell, J.;Eldredge, G.;Scholl, T.;Miller, R. M.;Bhatnagar, S.;Adey, N.;Rubano, T.;Tusneem, N.;Robinson, R.;Feldhaus, J.;Macalma, T.;Oliphant, A.;Briggs, S., A draft sequence of the rice genome (Oryza sativa L. ssp. japonica), *Science*, 2002, **296**, 92-100.
67. Vij, S.;Gupta, V.;Kumar, D.;Vydianathan, R.;Raghuvanshi, S.;Khurana, P.;Khurana, J. P.;Tyagi, A. K., Decoding the rice genome, *BioEssays : news and reviews in molecular, cellular and developmental biology*, 2006, **28**, 421-432.
68. Tyagi, A. K.;Khurana, J. P.;Khurana, P.;Raghuvanshi, S.;Gaur, A.;Kapur, A.;Gupta, V.;Kumar, D.;Ravi, V.;Vij, S.;Khurana, P.;Sharma, S., Structural and functional analysis of rice genome, *Journal of genetics*, 2004, **83**, 79-99.
69. Sequencing Project International Rice, G., The map-based sequence of the rice genome, *Nature*, 2005, **436**, 793-800.
70. Gnanamanickam, S. S., in *Biological Control of Rice Diseases*, ed. S. S. Gnanamanickam, Springer Netherlands, Dordrecht, 2009, DOI: 10.1007/978-90-481-2465-7_1, pp. 1-11.
71. Garris, A. J.;Tai, T. H.;Coburn, J.;Kresovich, S.;McCouch, S., Genetic Structure and Diversity in Oryza sativa L, *Genetics*, 2005, **169**, 1631-1638.
72. Allwood, J. W.;Goodacre, R., An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses, *Phytochemical analysis : PCA*, 2010, **21**, 33-47.
73. Biais, B.;Allwood, J. W.;Deborde, C.;Xu, Y.;Maucourt, M.;Beauvoit, B.;Dunn, W. B.;Jacob, D.;Goodacre, R.;Rolin, D.;Moing, A., 1H NMR, GC−EI-TOFMS, and Data Set Correlation for Fruit Metabolomics: Application to Spatial Metabolite Analysis in Melon, *Analytical Chemistry*, 2009, **81**, 2884-2894.
74. Hummel, J.;Selbig, J.;Walther, D.;Kopka, J., in *Metabolomics: A Powerful Tool in Systems Biology*, eds. J. Nielsen and M. C. Jewett, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, DOI: 10.1007/4735_2007_0229, pp. 75-95.
75. Kind, T.;Wohlgemuth, G.;Lee, D. Y.;Lu, Y.;Palazoglu, M.;Shahbaz, S.;Fiehn, O., FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, *Analytical chemistry*, 2009, **81**, 10038-10048.
76. Gika, H. G.;Theodoridis, G. A.;Plumb, R. S.;Wilson, I. D., Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics, *Journal of Pharmaceutical and Biomedical Analysis*, 2014, **87**, 12-25.
77. Ettre, L. S.;Sakodynskii, K. I., M. S. Tswett and the discovery of chromatography I: Early work (1899–1903), *Chromatographia*, 1993, **35**, 223-231.
78. Fanali, S.;Haddad, P. R.;Poole, C.;Riekkola, M.-L., *Liquid chromatography: fundamentals and instrumentation*, Elsevier, 2017.
79. Fanali, S.;Haddad, P.;Poole, C.;Schoenmakers, P.;Lloyd, D., *Liquid chromatography: fundamentals and instrumentation*, Elsevier Inc., 2013.
80. Snyder, L. R.;Kirkland, J. J.;Glajch, J. L., *Practical HPLC Method Development*, Wiley & Sons Inc., 2012.

81. Jandera, P., Stationary and mobile phases in hydrophilic interaction chromatography: a review, *Analytica Chimica Acta*, 2011, **692**, 1-25.
82. Buszewski, B.;Noga, S., Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique, *Analytical and Bioanalytical Chemistry*, 2012, **402**, 231-247.
83. Alpert, A. J., Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds, *Journal of Chromatography A*, 1990, **499**, 177-196.
84. Tolstikov, V. V.;Fiehn, O., Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry, *Analytical Biochemistry*, 2002, **301**, 298-307.
85. Cubbon, S.;Antonio, C.;Wilson, J.;Thomas-Oates, J., Metabolomic applications of HILIC–LC–MS, *Mass Spectrometry Reviews*, 2010, **29**, 671-684.
86. Spagou, K.;Tsoukali, H.;Raikos, N.;Gika, H.;Wilson, I. D.;Theodoridis, G., Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies, *Journal of Separation Science*, 2010, **33**, 716-727.
87. Idborg, H.;Zamani, L.;Edlund, P.-O.;Schuppe-Koistinen, I.;Jacobsson, S. P., Metabolic fingerprinting of rat urine by LC/MS: Part 1. Analysis by hydrophilic interaction liquid chromatography–electrospray ionization mass spectrometry, *Journal of Chromatography B*, 2005, **828**, 9-13.
88. Greco, G.;Letzel, T., Main interactions and influences of the chromatographic parameters in HILIC separations, *Journal of chromatographic science*, 2013, **51**, 684-693.
89. Ikegami, T.;Tomomatsu, K.;Takubo, H.;Horie, K.;Tanaka, N., Separation efficiencies in hydrophilic interaction chromatography, *Journal of Chromatography A*, 2008, **1184**, 474-503.
90. Nováková, L.;Solichová, D.;Solich, P., Advantages of ultra performance liquid chromatography over high-performance liquid chromatography: Comparison of different analytical approaches during analysis of diclofenac gel, *Journal of Separation Science*, 2006, **29**, 2433-2443.
91. Stoll, D. R.;Carr, P. W., Two-Dimensional Liquid Chromatography: A State of the Art Tutorial, *Analytical Chemistry*, 2017, **89**, 519-531.
92. Carr, P. W.;Davis, J. M.;Rutan, S. C.;Stoll, D. R., Principles of Online Comprehensive Multidimensional Liquid Chromatography, *Advances in chromatography*, 2012, **50**, 139-235.
93. Dugo, P.;Cacciola, F.;Kumm, T.;Dugo, G.;Mondello, L., Comprehensive multidimensional liquid chromatography: Theory and applications, *Journal of Chromatography A*, 2008, **1184**, 353-368.
94. Apffel, J. A.;Alfredson, T. V.;Majors, R. E., Automated on-line multi-dimensional high-performance liquid chromatographic techniques for the clean-up and analysis of water-soluble samples, *Journal of Chromatography A*, 1981, **206**, 43-57.
95. Pandohee, J.;G. Stevenson, P.;Zhou, X.-R.;J.S. Spencer, M.;A.H. Jones, O., Multi-Dimensional Liquid Chromatography and Metabolomics, Are Two Dimensions Better Than One?, *Current Metabolomics*, 2015, **3**, 10-20.
96. Marriott, P. J.;Ze-ying, W.;Schoenmakers, P., Nomenclature and Conventions in Comprehensive Multidimensional Chromatography–An Update, *LCGC Europe*, 2012, **25**, 266-275.
97. Snyder, L. R.;Kirkland, J. J.;Dolan, J., *Introduction to Modern Liquid Chromatography*, Wiley, 2010.
98. Vanhoenacker, G.;Vandenheede, I.;David, F.;Sandra, P.;Sandra, K., Comprehensive two-dimensional liquid chromatography of therapeutic monoclonal antibody digests, *Analytical and Bioanalytical Chemistry*, 2015, **407**, 355-366.
99. Sorensen, M.;Harmes, D. C.;Stoll, D. R.;Staples, G. O.;Fekete, S.;Guillarme, D.;Beck, A., Comparison of originator and biosimilar therapeutic monoclonal antibodies using comprehensive two-dimensional liquid chromatography coupled with time-of-flight mass spectrometry, *mAbs*, 2016, **8**, 1224-1234.
100. Jandera, P.;Staňková, M.;Hájek, T., New zwitterionic polymethacrylate monolithic columns for one- and two-dimensional microliquid chromatography, *Journal of Separation Science*, 2013, **36**, 2430-2440.
101. Holcapek, M.;Ovcacikova, M.;Lisa, M.;Cifkova, E.;Hajek, T., Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, *Analytical and Bioanalytical Chemistry*, 2015, **407**, 5033-5043.
102. Li, M.;Tong, X.;Lv, P.;Feng, B.;Yang, L.;Wu, Z.;Cui, X.;Bai, Y.;Huang, Y.;Liu, H., A not-stop-flow online normal-/reversed-phase two-dimensional liquid chromatography–quadrupole time-of-flight mass spectrometry method for comprehensive lipid profiling of human plasma from atherosclerosis patients, *Journal of Chromatography A*, 2014, **1372**, 110-119.

103. Pandohee, J.;Stevenson, P. G.;Conlan, X. A.;Zhou, X.-R.;Jones, O. A. H., Off-line two-dimensional liquid chromatography for metabolomics: an example using Agaricus bisporus mushrooms exposed to UV irradiation, *Metabolomics*, 2015, **11**, 939-951.
104. Wang, S.;Zhou, L.;Wang, Z.;Shi, X.;Xu, G., Simultaneous metabolomics and lipidomics analysis based on novel heart-cutting two-dimensional liquid chromatography-mass spectrometry, *Analytica Chimica Acta*, 2017, **966**, 34-40.
105. Dettmer, K.;Aronov, P. A.;Hammock, B. D., Mass spectrometry-based metabolomics, *Mass spectrometry reviews*, 2007, **26**, 51-78.
106. Wilm, M., Principles of electrospray ionization, *Molecular & cellular proteomics : MCP*, 2011, **10**, M111.009407.
107. Kebarle, P.;Verkerk, U. H., Electrospray: from ions in solution to ions in the gas phase, what we know now, *Mass spectrometry reviews*, 2009, **28**, 898-917.
108. Ernst, M.;Silva, D. B.;Silva, R. R.;Vencio, R. Z. N.;Lopes, N. P., Mass spectrometry in plant metabolomics strategies: from analytical platforms to data acquisition and processing, *Natural Product Reports*, 2014, **31**, 784-806.
109. El-Aneed, A.;Cohen, A.;Banoub, J., Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers, *Applied Spectroscopy Reviews*, 2009, **44**, 210-230.
110. de Hoffmann, E., Tandem mass spectrometry: A primer, *Journal of Mass Spectrometry*, 1996, **31**, 129-137.
111. Marshall, A. G.;Hendrickson, C. L., High-resolution mass spectrometers, *Annual review of analytical chemistry (Palo Alto, Calif.)*, 2008, **1**, 579-599.
112. Makarov, A., Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis, *Analytical Chemistry*, 2000, **72**, 1156-1162.
113. Hu, Q.;Noll, R. J.;Li, H.;Makarov, A.;Hardman, M.;Graham Cooks, R., The Orbitrap: a new mass spectrometer, *Journal of Mass Spectrometry*, 2005, **40**, 430-443.
114. Zeaiter, M.;Rutledge, D., in *Comprehensive Chemometrics*, eds. R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00074-0, pp. 121-231.
115. Tautenhahn, R.;Böttcher, C.;Neumann, S., Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*, 2008, **9**, 504.
116. Gorrochategui, E.;Jaumot, J.;Lacorte, S.;Tauler, R., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC - Trends in Analytical Chemistry*, 2016, **82**, 425-442.
117. Stolt, R.;Torgrip, R. J. O.;Lindberg, J.;Csenki, L.;Kolmert, J.;Schuppe-Koistinen, I.;Jacobsson, S. P., Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data, *Analytical Chemistry*, 2006, **78**, 975-983.
118. Barclay, V. J.;Bonner, R. F.;Hamilton, I. P., Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data Set Compression, *Analytical Chemistry*, 1997, **69**, 78-90.
119. Trygg, J.;Kettaneh-Wold, N.;Wallbäcks, L., 2D wavelet analysis and compression of on-line industrial process data, *Journal of Chemometrics*, 2001, **15**, 299-319.
120. Daubechies, I., *Ten lectures on wavelets*, SIAM, 1992.
121. Trygg, J.;Gabrielsson, J.;Lundstedt, T., in *Comprehensive Chemometrics*, eds. R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00097-1, pp. 1-8.
122. Bijlsma, S.;Bobeldijk, I.;Verheij, E. R.;Ramaker, R.;Kochhar, S.;Macdonald, I. A.;van Ommen, B.;Smilde, A. K., Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation, *Analytical Chemistry*, 2006, **78**, 567-574.
123. Gurden, S. P.;Westerhuis, J. A.;Bro, R.;Smilde, A. K., A comparison of multiway regression and scaling methods, *Chemometrics and Intelligent Laboratory Systems*, 2001, **59**, 121-136.
124. Smith, C. A.;Want, E. J.;O'Maille, G.;Abagyan, R.;Siuzdak, G., XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, *Analytical Chemistry*, 2006, **78**, 779-787.
125. Sarpe, V.;Schriemer, D. C., Supporting metabolomics with adaptable software: design architectures for the end-user, *Current Opinion in Biotechnology*, 2017, **43**, 110-117.
126. Mahieu, N. G.;Genenbacher, J. L.;Patti, G. J., A roadmap for the XCMS family of software solutions in metabolomics, *Current Opinion in Chemical Biology*, 2016, **30**, 87-93.
127. Shao, X.;Pang, C.;Su, Q., A novel method to calculate the approximate derivative photoacoustic spectrum using continuous wavelet transform, *Fresenius' Journal of Analytical Chemistry*, 2000, **367**, 525-529.

128. Tautenhahn, R.;Cho, K.;Uritboonthai, W.;Zhu, Z.;Patti, G. J.;Siuzdak, G., An accelerated workflow for untargeted metabolomics using the METLIN database, *Nature Biotechnoogy*, 2012, **30**, 826-828.
129. Tauler, R.;Kowalski, B.;Fleming, S., Multivariate curve resolution applied to spectral data from multiple runs of an industrial process, *Analytical Chemistry*, 1993, **65**, 2040-2047.
130. Tauler, R.;Barceló, D., Multivariate curve resolution applied to liquid chromatography—diode array detection, *TrAC Trends in Analytical Chemistry*, 1993, **12**, 319-327.
131. Tauler, R., Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**, 133-146.
132. Tauler, R.;Smilde, A.;Kowalski, B., Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *Journal of Chemometrics*, 1995, **9**, 31-58.
133. Jaumot, J.;Gargallo, R.;de Juan, A.;Tauler, R., A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, *Chemometrics and Intelligent Laboratory Systems*, 2005, **76**, 101-110.
134. de Juan, A.;Tauler, R., Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution, *Analytica Chimica Acta*, 2003, **500**, 195-210.
135. de Juan, A.;Tauler, R., Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications, *Critical Reviews in Analytical Chemistry*, 2006, **36**, 163-176.
136. de Juan, A.;Jaumot, J.;Tauler, R., Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, *Analytical Methods*, 2014, **6**, 4964-4976.
137. Golub, G. H.;Reinsch, C., in *Linear Algebra*, eds. J. H. Wilkinson, C. Reinsch and F. L. Bauer, Springer Berlin Heidelberg, Berlin, Heidelberg, 1971, DOI: 10.1007/978-3-662-39778-7_10, pp. 134-151.
138. Windig, W.;Guilment, J., Interactive self-modeling mixture analysis, *Analytical Chemistry*, 1991, **63**, 1425-1432.
139. Windig, W.;Gallagher, N. B.;Shaver, J. M.;Wise, B. M., A new approach for interactive self-modeling mixture analysis, *Chemometrics and Intelligent Laboratory Systems*, 2005, **77**, 85-96.
140. Windig, W.;Stephenson, D. A., Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach, *Analytical Chemistry*, 1992, **64**, 2735-2742.
141. Lawson, C.L.; Hanson, R.J., *Solving Least Squares Problems*. Classics in Applied Mathematics, ed. SIAM. doi.org/10.1137/1.9781611971217
142. Bro, R.;De Jong, S., A fast non-negativity-constrained least squares algorithm, *Journal of chemometrics*, 1997, **11**, 393-401.
143. Manne, R., On the resolution problem in hyphenated chromatography, *Chemometrics and Intelligent Laboratory Systems*, 1995, **27**, 89-94.
144. Tauler, R., Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, *Journal of Chemometrics*, 2001, **15**, 627-646.
145. Gargallo, R.;Jaumot, J.;Tauler, R., Noise propagation and error estimations in multivariate curve resolution alternating least squares using resampling methods, *Journal of Chemometrics*, 2004, **18**, 327-340.
146. Bro, R., PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, 1997, **38**, 149-171.
147. Smilde, A.;Bro, R.;Geladi, P., *Multi-way analysis: applications in the chemical sciences*, John Wiley & Sons Ltd., West Sussex, England, 2004.
148. Bro, R.;Andersson, C. A.;Kiers, H. A. L., PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, *Journal of Chemometrics*, 1999, **13**, 295-309.
149. Kiers, H. A. L.;ten Berge, J. M. F.;Bro, R., PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, *Journal of Chemometrics*, 1999, **13**, 275-294.
150. Bro, R.;Kiers, H. A. L., A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, 2003, **17**, 274-286.
151. Tucker, L. R., Some mathematical notes on three-mode factor analysis, *Psychometrika*, 1966, **31**, 279-311.
152. Smilde, A. K.;Tauler, R.;Henshaw, J. M.;Burgess, L. W.;Kowalski, B. R., Multicomponent Determination of Chlorinated Hydrocarbons Using a Reaction-Based Chemical Sensor. 3. Medium-Rank Second-Order Calibration with Restricted Tucker Models, *Analytical Chemistry*, 1994, **66**, 3345-3351.

153. Trygg, J.;Gullberg, J.;Johansson, A. I.;Jonsson, P.;Moritz, T., Chemometrics in metabolomics—an introduction, Plant Metabolomics, 2006, **57**, 117-128.
154. Wold, S., Chemometrics; what do we mean with it, and what do we want from it?, *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**, 109-115.
155. Eriksson, L.;Andersson, P. L.;Johansson, E.;Tysklind, M., Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD), *Molecular Diversity*, 2006, **10**, 169-186.
156. Esbensen, K. H.;Geladi, P., in *Comprehensive Chemometrics*, eds. R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00043-0, pp. 211-226.
157. Wold, S.;Esbensen, K.;Geladi, P., Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.
158. Smilde, A. K.;Jansen, J. J.;Hoefsloot, H. C.;Lamers, R. J.;van der Greef, J.;Timmerman, M. E., ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, *Bioinformatics*, 2005, **21**, 3043-3048.
159. Ahrens, H., Searle, S. R.: Linear Models. John Wiley & Sons, Inc., New York-London-Sydney-Toronto 1971. XXI, 532 S. $9.50, *Biometrische Zeitschrift*, 1974, **16**, 78-79.
160. Lundstedt, T.;Seifert, E.;Abramo, L.;Thelin, B.;Nyström, Å.;Pettersen, J.;Bergman, R., Experimental design and optimization, *Chemometrics and Intelligent Laboratory Systems*, 1998, **42**, 3-40.
161. Jansen, J. J.;Hoefsloot, H. C. J.;van der Greef, J.;Timmerman, M. E.;Westerhuis, J. A.;Smilde, A. K., ASCA: analysis of multivariate data obtained from an experimental design, *Journal of Chemometrics*, 2005, **19**, 469-481.
162. Hoefsloot, H. C. J.;Vis, D. J.;Westerhuis, J. A.;Smilde, A. K.;Jansen, J. J., in *Comprehensive Chemometrics*, 2010, vol. 2, pp. 453-472.
163. Jansen, J. J.;Hoefsloot, H. C. J.;van der Greef, J.;Timmerman, M. E.;Smilde, A. K., Multilevel component analysis of time-resolved metabolic fingerprinting data, *Analytica Chimica Acta*, 2005, **530**, 173-183.
164. Jansen, J. J.;Bro, R.;Hoefsloot, H. C. J.;van den Berg, F. W. J.;Westerhuis, J. A.;Smilde, A. K., PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data, *Journal of Chemometrics*, 2008, **22**, 114-121.
165. Thiel, M.;Féraud, B.;Govaerts, B., ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs, *Journal of Chemometrics*, 2017, **31**, e2895-n/a.
166. Stohle, L.;Wold, S., Multivariate analysis of variance (MANOVA), *Chemometrics and Intelligent Laboratory Systems*, 1990, **9**, 127-141.
167. Harrington, P. d. B.;Vieira, N. E.;Espinoza, J.;Nien, J. K.;Romero, R.;Yergey, A. L., Analysis of variance–principal component analysis: A soft tool for proteomic discovery, *Analytica Chimica Acta*, 2005, **544**, 118-127.
168. Engel, J.;Blanchet, L.;Bloemen, B.;van den Heuvel, L. P.;Engelke, U. H.;Wevers, R. A.;Buydens, L. M., Regularized MANOVA (rMANOVA) in untargeted metabolomics, *Anal Chim Acta*, 2015, **899**, 1-12.
169. Zwanenburg, G.;Hoefsloot, H. C. J.;Westerhuis, J. A.;Jansen, J. J.;Smilde, A. K., ANOVA–principal component analysis and ANOVA–simultaneous component analysis: a comparison, *Journal of Chemometrics*, 2011, **25**, 561-567.
170. Vis, D. J.;Westerhuis, J. A.;Smilde, A. K.;van der Greef, J., Statistical validation of megavariate effects in ASCA, *BMC Bioinformatics*, 2007, **8**.
171. Langsrud, Ø.;J∅rgensen, K.;Ofstad, R.;Næs, T., Analyzing Designed Experiments with Multiple Responses, *Journal of Applied Statistics*, 2007, **34**, 1275-1296.
172. Wold, H., in *Multivariate Analysis.*, Academic Press, 1966, DOI: citeulike-article-id:8609111, pp. 391-420.
173. Wold, S.;Sjöström, M.;Eriksson, L., PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**, 109-130.
174. Geladi, P.;Kowalski, B. R., Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, 1986, **185**, 1-17.
175. Todeschini, R.;Consonni, V.;Gramatica, P., in *Comprehensive Chemometrics*, eds. R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00007-7, pp. 129-172.
176. Yousefinejad, S.;Hemmateenejad, B., Chemometrics tools in QSAR/QSPR studies: A historical perspective, *Chemometrics and Intelligent Laboratory Systems*, 2015, **149**, 177-204.

177. Hutter, M. C., Molecular Descriptors for Chemoinformatics (2nd ed.). By Roberto Todeschini and Viviana Consonni, *ChemMedChem*, 2010, **5**, 306-307.
178. Kaliszan, R., QSRR: quantitative structure-(chromatographic) retention relationships, *Chemical reviews*, 2007, **107**, 3212-3246.
179. Kaliszan, R., in *Liquid Chromatography*, eds. P. R. Haddad, C. F. Poole, P. Schoenmakers and D. Lloyd, Elsevier, Amsterdam, 2013, DOI: https://doi.org/10.1016/B978-0-12-415807-8.00017-1, pp. 385-405.
180. Taraji, M.;Haddad, P. R.;Amos, R. I. J.;Talebi, M.;Szucs, R.;Dolan, J. W.;Pohl, C. A., Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures, *Journal of Chromatography A*, 2017, **1486**, 59-67.
181. Creek, D. J.;Jankevics, A.;Breitling, R.;Watson, D. G.;Barrett, M. P.;Burgess, K. E. V., Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction, *Analytical Chemistry*, 2011, **83**, 8703-8710.
182. Höskuldsson, A., PLS regression methods, *Journal of Chemometrics*, 1988, **2**, 211-228.
183. Wold, S.;Johansson, A.;Cochi, M., in *3 D QSAR in Drug Design. Theory, Methods and Applications*, ESCOM Science Publishers, Leiden, 1993, pp. 523-550.
184. Jackson, J. E., in *A User's Guide to Principal Components*, John Wiley & Sons, Inc., 2004, DOI: 10.1002/0471725331.ch13, pp. 301-318.
185. Burnham, A. J.;MacGregor, J. F.;Viveros, R., Latent variable multivariate regression modeling, *Chemometrics and Intelligent Laboratory Systems*, 1999, **48**, 167-180.
186. Barker, M.;Rayens, W., Partial least squares for discrimination, *Journal of Chemometrics*, 2003, **17**, 166-173.
187. Gromski, P. S.;Muhamadali, H.;Ellis, D. I.;Xu, Y.;Correa, E.;Turner, M. L.;Goodacre, R., A tutorial review: Metabolomics and partial least squares-discriminant analysis-a marriage of convenience or a shotgun wedding, *Analytica Chimica Acta*, 2015, **879**, 10-23.
188. Chong, I.-G.;Jun, C.-H., Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, 2005, **78**, 103-112.
189. Szymańska, E.;Saccenti, E.;Smilde, A. K.;Westerhuis, J. A., Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, *Metabolomics*, 2012, **8**, 3-16.
190. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et biophysica acta*, 1975, **405**, 442-451.
191. Vinaixa, M.;Samino, S.;Saez, I.;Duran, J.;Guinovart, J. J.;Yanes, O., A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data, *Metabolites*, 2012, **2**, 775-795.
192. Saccenti, E.;Hoefsloot, H. C. J.;Smilde, A. K.;Westerhuis, J. A.;Hendriks, M. M. W. B., Reflections on univariate and multivariate analysis of metabolomics data, *Metabolomics*, 2014, **10**, 361-374.

Capítol 3

Estudis de metabolòmica no dirigida: estratègies analítiques i de tractament de dades

3.1. Introducció

Els estudis de metabolòmica consten de 6 etapes principals, tal com es mostra en la Figura 3.1:

1) Definició dels objectius i disseny de l'experiment.
2) Preparació, tractament i recollida de la mostra.
3) Extracció i anàlisi instrumental dels metabòlits.
4) Tractament de les dades.
5) Identificació dels metabòlits.
6) Interpretació biològica dels resultats obtinguts.

Totes les etapes requereixen un disseny minuciós per tal d'assegurar l'èxit de la investigació realitzada. Aquest capítol es centra en les etapes 3, 4 i 5, les quals representen els reptes analítics i quimiomètrics més importants en un estudi de metabolòmica.

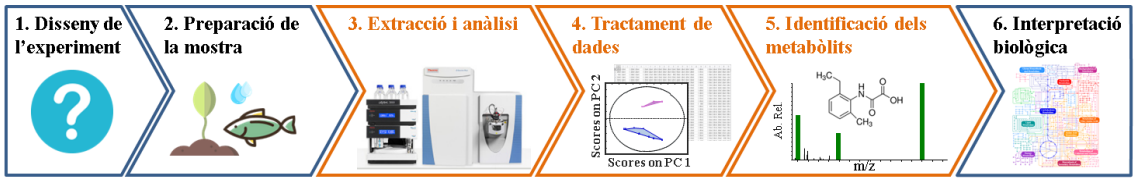


Figura 3.1. Diagrama representatiu de les diverses etapes dels estudis metabolòmics. Es destaquen en color taronja les etapes tractades en aquest capítol.

Els organismes vegetals contenen milers de metabòlits d'una gran diversitat química. Per aquest motiu, l'estratègia analítica utilitzada ha de ser sensible, selectiva i amb una alta capacitat de separació. En aquesta Tesi, s'ha utilitzat la LC-MS, que compleix aquests tres requisits [1-3]. Un punt clau en les anàlisis LC és l'elecció del tipus de fase estacionària més adient per a l'estudi a realitzar. Degut a la naturalesa polar de la majoria dels metabòlits, les columnes del mode HILIC han destacat en els darrers anys com una bona opció per a la seva separació cromatogràfica [4, 5]. Des de la seva aparició durant els anys 90, s'han desenvolupat molts tipus de fases estacionàries modificades amb lligands disponibles pel mode HILIC com, per exemple, amida, amina, zwitteriònica o mixtes, la qual cosa dificulta l'elecció de la més adequada per a un estudi metabolòmic concret. Per ajudar a entendre el seu funcionament, en aquesta Tesi s'han avaluat quatre columnes HILIC diferents pel seu ús en estudis de metabolòmica no dirigida (Publicació 1). A més, també s'han estudiat els efectes de diferents factors experimentals que influeixen en la separació en mode HILIC, com són el pH i la força iònica de la fase mòbil. Per tal de facilitar aquesta avaluació s'han utilitzat diverses eines quimiomètriques, com el PCA, l'ASCA i el PLS, per explorar i modelitzar el comportament de les diferents fases estacionàries.

En segon lloc, les dades obtingudes en estudis metabolòmics són molt complexes ja que es registren milions de senyals per cada mostra analitzada i els arxius de dades generats són de l'ordre de gigabytes (Gb) per mostra. Conseqüentment, les etapes de compressió i processament de dades són especialment importants [6, 7]. En aquest capítol es comparen dues estratègies de tractament de dades enfocades a superar els reptes associats a les dades metabolòmiques per MS. D'una banda, s'ha utilitzat el programa XCMS, que és l'eina de tractament de dades més emprada actualment en estudis de metabolòmica no dirigida [8, 9]. D'altra banda, s'ha avaluat l´ús del mètode MCR-ALS [10, 11] per a la detecció de possibles biomarcadors en estudis de metabolòmica no dirigida. En aquest capítol, per tal de comparar de forma adequada aquests dos enfocaments, el mateix conjunt de dades metabolòmiques s'ha processat amb les dues metodologies, XCMS i MCR-ALS (Publicació 2).

Idealment, en els estudis metabolòmics cal emprar patrons de referència per confirmar la identificació dels metabòlits. Malauradament, en aquests estudis és habitual resoldre centenars de compostos, de manera que comprar i analitzar estàndards autèntics per a cada possible metabòlit no és pràctic a nivell de temps i diners [12, 13]. A més, en molts casos els patrons dels metabòlits no estan disponibles comercialment. Existeixen bases de dades públiques, com per exemple METLIN [14] o HMDB [15], que ajuden en la identificació dels metabòlits, ja que contenen la massa exacta i els espectres de MS/MS teòrics (*in-silico*) d'un gran nombre de metabòlits, els quals permeten la seva comparació amb dades experimentals. Desafortunadament, aquesta informació no està disponible per a tots el metabòlits [12, 13]. En la darrera part d'aquest capítol, s'avalua la utilització de models QSRR per a superar els reptes relacionats amb la identificació dels metabòlits (resultats no publicats). Aquests models de QSRR permeten relacionar el temps de retenció dels metabòlits amb la seva estructura, caracteritzada a partir de descriptors moleculars coneguts (MDs) [16]. Aquests models poden predir amb precisió els temps de retenció dels metabòlits candidats, la qual cosa pot ser una informació molt útil a l'hora de completar la seva identificació [17-19].

Aquest capítol inclou les següents publicacions:

- **Publicació 1:** *Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches*. M. Navarro-Reig, E. Ortiz-Villanueva, R. Tauler, J. Jaumot. Metabolites 7 (2017), 54.

  En aquest article s'avaluen quatre columnes HILIC (amida, BEH amida, zwitteriònica i mode mixt diol) en diferents condicions experimentals (pH i força iònica de la fase mòbil) per a

l'anàlisi d'una mescla de 54 metabòlits de diferents famílies (nucleòsids, aminoàcids, sucres, àcids orgànics i altres). Els resultats d'aquestes anàlisis cromatogràfiques es comparen utilitzant PCA, ASCA i PLS.

- **Publicació 2:** *Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies*. M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler. Analytical and Bioanalytical Chemistry 407 (2015), 8835-8847.

  En aquest article es comparen els procediments XCMS i el MCR-ALS quan s'utilitzen en el processament d'un mateix conjunt de dades de metabolòmica no dirigida, sobre els efectes de la contaminació de cadmi i coure en els metabòlits de l'arròs.

3.2. Publicació 1

Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches.

M. Navarro-Reig, E. Ortiz-Villanueva, R. Tauler, J. Jaumot.

Metabolites 7 (2017), 54.

metabolites

MDPI

Article

Modelling of Hydrophilic Interaction Liquid Chromatography Stationary Phases Using Chemometric Approaches

Meritxell Navarro-Reig, Elena Ortiz-Villanueva, Romà Tauler and Joaquim Jaumot *

Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain; meritxell.navarro@idaea.csic.es (M.N.-R.); elena.ortiz@idaea.csic.es (E.O.-V.); rtaqam@idaea.csic.es (R.T.)
* Correspondence: joaquim.jaumot@idaea.csic.es; Tel.: +34-934-006-100

Received: 15 July 2017; Accepted: 21 October 2017; Published: 24 October 2017

Abstract: Metabolomics is a powerful and widely used approach that aims to screen endogenous small molecules (metabolites) of different families present in biological samples. The large variety of compounds to be determined and their wide diversity of physical and chemical properties have promoted the development of different types of hydrophilic interaction liquid chromatography (HILIC) stationary phases. However, the selection of the most suitable HILIC stationary phase is not straightforward. In this work, four different HILIC stationary phases have been compared to evaluate their potential application for the analysis of a complex mixture of metabolites, a situation similar to that found in non-targeted metabolomics studies. The obtained chromatographic data were analyzed by different chemometric methods to explore the behavior of the considered stationary phases. ANOVA-simultaneous component analysis (ASCA), principal component analysis (PCA) and partial least squares regression (PLS) were used to explore the experimental factors affecting the stationary phase performance, the main similarities and differences among chromatographic conditions used (stationary phase and pH) and the molecular descriptors most useful to understand the behavior of each stationary phase.

Keywords: hydrophilic interaction liquid chromatography (HILIC); non-targeted metabolomics; stationary phase; chemometrics

1. Introduction

Over the past decade, metabolomics has received considerable attention from the scientific community. Metabolomics aims to screen endogenous small molecules (metabolites) present in biological samples, providing a direct measure of the phenotypic state of an organism [1–3]. There are two main metabolomic strategies: targeted and non-targeted. The targeted approach is focused on the investigation of a specific metabolic pathway and, therefore, in the analysis of a reduced and known set of compounds. In contrast, non-targeted metabolomics aims to screen the entire metabolite content of biological samples containing compounds with different physical and chemical properties [4,5].

Due to its high-resolution power, sensitivity and accuracy of *m/z* detection, liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) has become the analytical platform most used in metabolomics studies. Reversed phased liquid chromatography (RPLC) is useful for the separation of the more hydrophobic compounds, such as lipids. However, RPLC is not recommended for the analysis of some of the most usual metabolite families characterized as being polar and hydrophilic compounds [2,6–9]. In the analysis of polar compounds, and in particular in the metabolomics field, hydrophilic interaction liquid chromatography (HILIC) has become a valuable alternative to RPLC, due to its ability to separate these more hydrophilic compounds [10,11]. Different types of HILIC stationary phases, such as amide, amine, mixed-mode diol and zwitterionic, have

been employed previously in metabolomics studies [11–13]. This variety of HILIC stationary phases allows the separation of metabolites of different properties. Nevertheless, this diversity also makes the selection of the most suitable HILIC stationary phase for a particular study more challenging [10,11,13]. Moreover, the retention mechanism in HILIC has been demonstrated to be more complex than in RPLC mode [14]. In HILIC mode, the separation is primarily achieved due to the partition of analytes between the mobile phase and the hydrophilic layer adsorbed at the surface of the stationary phase. Also, it is proposed that other electrostatic interactions, such as ion exchange interactions, may also contribute to the retention mechanisms. These electrostatic interactions vary among the different types of HILIC stationary phases, making their comparison and selection more difficult [15,16].

Chemometric tools can help in addressing the challenge of finding the most suitable HILIC stationary phase for the analysis of a complex mixture of polar compounds [17]. On the one hand, multivariate data exploratory methods, such as principal component analysis (PCA) [18], can be used to investigate the behavior of HILIC stationary phases and understanding retention mechanisms. Other multivariate statistical methods, like ANOVA-simultaneous component analysis (ASCA) [19], can help evaluating the statistical significance of the experimental factors involved in a non-targeted metabolomics study, such as when different HILIC stationary phases and mobile phase conditions are assessed. On the other hand, building models linking the physicochemical properties of compounds (molecular descriptors, MDs) and their chromatographic behavior (retention factors) could help to get a different insight into the HILIC stationary phase performance [11,20–29]. In addition, these models could also be used to predict the behavior of a chromatographic system and, therefore, can be used to predict the chromatographic retention of unknown compounds and, in some cases, provide additional information to support their identification [11,23,24,30].

In a previous work, different types of HILIC stationary phases have already been evaluated for metabolomics studies [17]. However, in that preliminary work, only 12 metabolites were considered, which simplified the analysis considerably. Moreover, the detection was performed using diode array detector (DAD), which is not the standard in complex studies in the metabolomics field. In contrast, here, a mixture of 54 metabolites is analyzed using LC-HRMS. Four different types of HILIC stationary phases commonly used in metabolomics research have been evaluated under various experimental conditions using different chemometric tools for their application to non-targeted studies. First, PCA and ASCA were applied to explore the general behavior of the considered stationary phases and to evaluate the statistical significance of the three experimental factors discussed in this work (stationary phase, pH and ionic strength). Then, partial least squares regression (PLS) models based on MDs computed from molecular structures of different metabolite families were calculated for evaluating the chromatographic behavior of the four HILIC stationary phases considering the retention factor of the analyzed metabolites.

2. Results and Discussion

2.1. Determination of Retention Factors

A mixture of 54 metabolites from different families (see Table 1) was analyzed using four different stationary phases (Ethylene Bridged Hybrid (BEH) amide, amide, zwitterionic and mixed-mode diol), working with mobile phase at three pH values (acid, moderately acid and neutral) and two ionic strengths (low and high). Each chromatographic condition was injected twice, giving a total of 48 chromatographic runs. The regions of interest (ROI) strategy was used to arrange the 48 chromatographic runs in LC-MS data matrices. Then, each matrix was evaluated to automatically find the *m/z* value of each metabolite and provide their retention time in each chromatogram. Lastly, retention factor (k) of each metabolite in each chromatographic run was calculated using Equation (1) (see Materials and methods section for more details).

Table 1. Metabolites contained in the analyzed mixture.

Metabolite Families					
Nucleosides	**Amino Acids**		**Sugars**	**Organic Acids**	**Others**
1-methyladenosine	1-methyl-L-histidine	L-citrulline	D(−)-ribose	citric acid	hypoxanthine
2′′-*O*-methylcytidine	3-methyl-L-histidine	L-glutamic acid	glucose	ketoglutaric acid	L-carnitine
2-thiocytidine	4-hydroxy-L-proline	L-histidine	trehalose	pimelic acid	serotonin
5-methylcytidine	5-hydroxylysine	L-homocystine	mannitol	succinic acid	tryptamine
cytidine	β-alanine	L-isoleucine	-	creatine	-
guanosine	creatinine	L-leucine	-	-	-
inosine	cysteine	L-methionine	-	-	-
pseudouridine	L-(−)-proline	L-ornithine	-	-	-
ribothymidine	L-(+)-arginine	L-serine	-	-	-
uridine	L-(+)-cystathionine	L-threonine	-	-	-
-	L-(+)-lysine	L-tryptophan	-	-	-
-	L-2-aminoadipic acid	L-valine	-	-	-
-	L-2-amino-n-butyric acid	taurine	-	-	-
-	L-alanine	sarcosine	-	-	-
-	L-anserine	L-aspartic acid	-	-	-
-	L-carnosine	-	-	-	-

Finally, a matrix, **D**, containing the retention factors of metabolites at each chromatographic condition was built up. This matrix had a number of rows equal to the number of chromatographic runs performed (48 runs) and a number of columns equal to the number of analyzed metabolites (54 compounds). Figure 1 shows the obtained retention factors for the 54 metabolites in the 48 chromatographic runs. The visual inspection of the obtained retention factors already allowed the differentiation of the evaluated chromatographic conditions. For instance, the retention factors of all metabolites in BEH amide are shorter than in the rest of the stationary phases. However, direct evaluation of the HILIC stationary phases behavior is not straightforward. For this reason, matrix **D** was first evaluated using two chemometric exploratory methods, PCA and ASCA, to get a deeper insight into the effects of experimental factors on the retention behavior of metabolites.

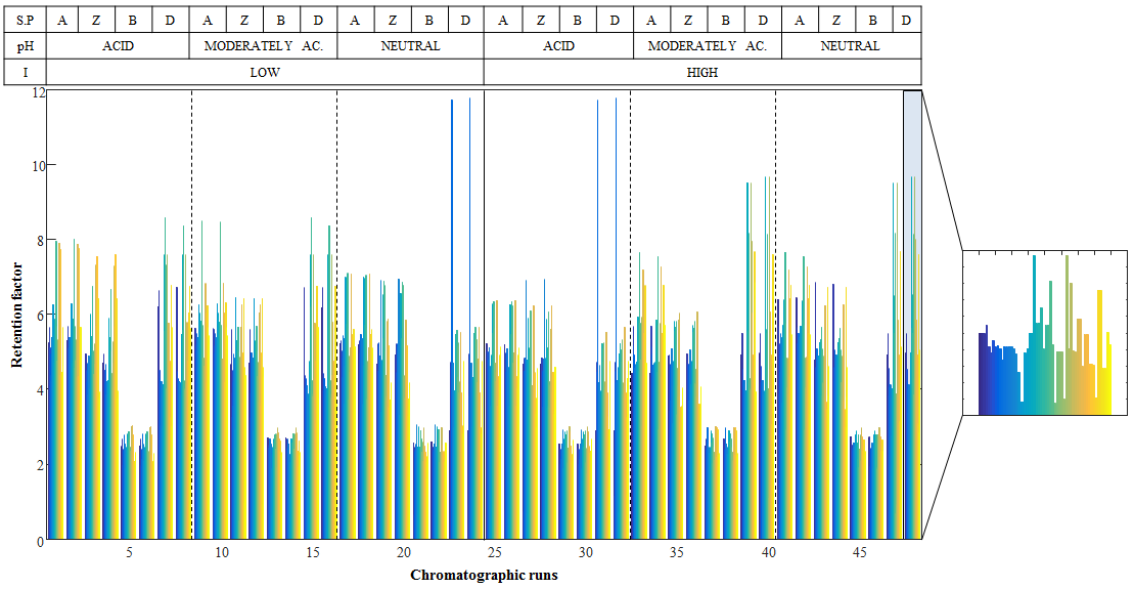


Figure 1. Retention factors for the 54 metabolites in the 48 chromatographic runs (matrix **D**). The table on the top shows the chromatographic conditions of each sample: (A) indicates the chromatographic runs perfumed using amide, (Z) zwitterionic, (B) BEH amide and (D) mixed-mode diol stationary phases. As an example, a zoomed view of chromatographic run 48 is depicted.

2.2. Evaluation of HILIC Stationary Phases Behavior

PCA was applied to matrix **D** in order to explore the behavior of the chromatographic systems studied in this work (four different stationary phases and mobile phase at three different pH conditions and two different ionic strengths). PCA results indicated that stationary phase was the most critical factor to be considered. In Figure 2a, the PCA scores plot shows that samples analyzed with the four stationary phases were differentiated. Moreover, PC1 distinguished BEH amide stationary phase (XBridge™ Amide) samples (cyan triangles in Figure 2a), which appeared on the left side of PC1, from the rest of the samples. Additionally, PC2 distinguished mixed-mode diol stationary phase (Acclaim™ Mixed-Mode HILIC-1) samples (green squares in Figure 2a), with large positive PC2 scores values, from the rest of the samples. Amide (TSK-Gel Amide-80) and zwitterionic (ZIC-HILIC) stationary phases showed a similar behavior since their samples appeared close to each other in the scores plot. Furthermore, samples analyzed with the three different pH values were also clearly distinguished (see Figure S1 in Supplementary Information), whereas, samples analyzed at different ionic strengths (low and high) could not be differentiated by the PCA scores plot. These results were consistent with the results obtained in the previous authors' work about HILIC stationary phases [17].

Statistical significance of the three experimental factors considered in this work (stationary phase, pH and ionic strength) and their interactions were assessed by applying ASCA to matrix **D**. Results showed that both stationary phase and pH had statistically significant effects (p-value of 0.0001). On the contrary, the ionic strength effects were not significant (p-value of 0.2). Moreover, the interaction of three factors (stationary phase $\times$ pH, stationary phase $\times$ ionic strength and pH $\times$ ionic strength) was also found to be statically significant (p-value of 0.0001). Hence, from the combination of ASCA and PCA results, the two most relevant factors were defined as the stationary phase and the pH of the aqueous solvent.

Figure 2b shows the ASCA principal component scores (at the mean level) for each HILIC stationary phase. In this scores plot, some trends could be observed. For example, PC1 distinguished the mixed-mode diol stationary phase with a large negative scores value, whereas the two amides and the zwitterionic stationary phases showed a similar positive PC1 scores value. PC2 also differentiated the BEH amide stationary phase with a negative scores value. In contrast, amide, zwitterionic and mixed-mode diol stationary phases had a similar positive PC2 scores value. Finally, it should be mentioned that, as observed in PCA results, amide and zwitterionic stationary phases had similar PC1 and PC2 scores values. Therefore, these two stationary phases showed a similar behavior.

ASCA loadings were useful to know which variables (metabolite retention factors) were the most important to distinguish the stationary phases. Figure 2c,d show the ASCA loadings plot for PC1 and PC2, respectively. For instance, six amino acids (L-(−)-proline, L-valine, L-methionine, L-tyrosine, L-homocysteine and L-anserine) showed a higher loadings value in PC1 (Figure 2c). Consequently, these amino acids were useful to distinguish mixed-mode diol stationary phase from the other three (amide, BEH amide and zwitterionic). In the case of PC2 (Figure 2d), pimelic and citric organic acids showed the highest loadings values. Therefore, these two organic acids appeared as important to differentiate BEH amide from the rest of the stationary phases.

In general, PCA and ASCA results coincided showing that the most important factors were the HILIC stationary phase and the pH of the aqueous solvent. In addition, some facts related to the stationary phase behavior can be highlighted. For instance, the zwitterionic stationary phase showed an intermediate behavior between the two amide stationary phases.

The next step in this work was to find relationships between the observed chromatographic retention observed and the physicochemical properties of metabolites using their molecular descriptors. In addition, since PCA and ASCA results showed that the ionic strength of the mobile phase was not a significant factor in metabolomics studies, these new PLS models were only assessed considering chromatographic runs done at low ionic strength.

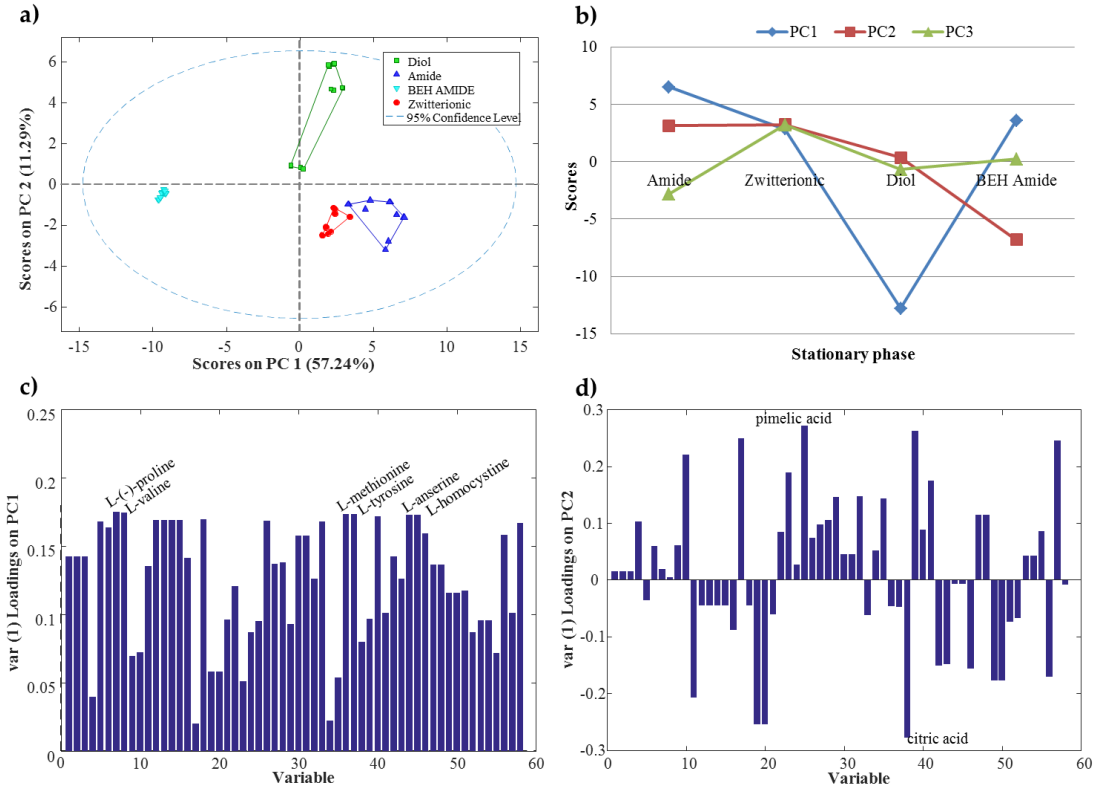


Figure 2. (**a**) Principal component analysis (PCA) scores plot of samples classified according to the stationary phase used in the chromatographic system; (**b**) ANOVA simultaneous component analysis (ASCA) principal component scores for the stationary phase factor; (**c**) ASCA PC1 loadings for the stationary phase factor; (**d**) ASCA PC2 loadings for the stationary phase factor.

2.3. Exploratory Relationship between Physicochemical Properties and HILIC Chromatographic Retention

PLS models were independently built for the 12 chromatographic studied systems. They were obtained by the combination of the four stationary phases (BEH amide, amide, zwitterionic and mixed-mode diol) and of the three mobile phases at different pH values (acid, moderately acid and neutral). These models were generated using the experimental retention factors of the 54 metabolites contained in the mixture (a **y** vector for each condition) and their corresponding molecular descriptors (MDs) organized in an **X** matrix. A preliminary selection step was performed over the whole set of MDs available from PCLIENT to reduce the total number and finally consider 844 of them (see Materials and Methods section for more details).

PLS modelling of the retention factor obtained for each condition using metabolite molecular descriptors required a reduced number of latent variables (between 2 and 3) to explain most of the variance of each **y** vector (between 80% and 95%). These models did not show an accuracy enough to be used for the prediction of the retention factors of unknown compounds. However, as these PLS models explain a major part of the retention factor variance, the exploration of scores and loading plots can provide additional insight into the HILIC behavior. First, the analysis of scores plots could allow confirming the differentiation between groups of samples. More interestingly, the analysis of loadings plots could provide information regarding the molecular descriptors related to this differentiation, and could give additional information to known the main physicochemical properties involved in the HILIC retention mechanisms. As an example, Figure 3 shows the scores plots obtained for the amide stationary phase at the three studied pH values of the mobile phases. In these plots, some interesting trends can be observed. First, the differentiation between nucleosides and the rest of the metabolites present in the mixture. In the three cases, a clear group with all nucleosides is visible. Amino acids are

the metabolite family with the most compounds in the mixture. These amino acids are spread along the first latent variable. However, in the three pH conditions, differentiation between two groups can be detected (Figure 3a). An inspection of metabolites forming these two groups allowed observing that the left group was composed of metabolites with a molecular weight lower than 130 Da, whereas the right group was composed of metabolites with a molecular weight larger than 130 Da. Regarding the other families of metabolites present in the mixture (i.e., sugars or organic acids), metabolites were grouped but they were overlapping with amino acids. Similar trends can be observed when considering the other chromatographic conditions.

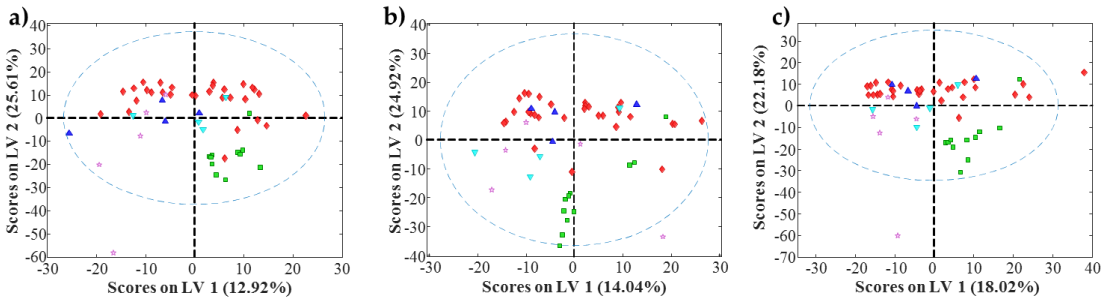


Figure 3. Partial Least Squares (PLS) scores plot for amide stationary phase. (**a**) Acid pH; (**b**) Moderately acid pH; (**c**) Neutral pH. Red diamonds are amino acids (♦), green squares are nucleosides (■), blue triangles are organic acids (▲), purple stars are sugars (★) and cyan triangles are others (▼).

The evaluation of which MDs allowed obtaining this retention factor modelling is more appealing. The identification of MDs can be performed using the variable importance on projection (VIP) scores obtained for each PLS model as a feature selection tool allowing to identify which variables (MDs) were the most descriptive of every chromatographic system studied in this work. Table 2 shows the twenty most relevant MDs for each PLS model and their VIP scores values. MDs that appeared to be important in more than one chromatographic system are shown in bold letters. The most repeated MDs values present in Table 2 were the 3D-MoRSE values like the molecular representation of structures based on electronic diffraction, Mor02p, Mor10p, Mor15p, Mor08u, Mor19u, Mor24u, Mor31u, Mor01m, Mor06m, Mor10m, Mor 10v, Mor15v, Mor08e, Mor24e and Mor31e. All of them are geometrical descriptors that codify the 3D molecular structure [31–34]. Some other geometrical descriptors (HTcxp, RTe and RTm) also seemed relevant. Geometrical descriptors are calculated from the coordinates of the molecule atoms, interatomic distances and distances from a specific origin. They are geometrical descriptors of the molecular size, shape, symmetry and atom distribution [33,34]. Moreover, these descriptors are weighted by ionization potential, electronegativity, polarizability and molecular mass [11]. Different topological descriptors (J, Lop, MWC03, MWC04, EEig10r, ESpm01d, ESpm07d, ESpm09d, ESpm13d, ESpm14d and ESpm14r) were also highlighted as relevant for the modelling of the studied chromatographic systems. These topological descriptors are numerical quantifiers of molecular topology that are sensitive to one or more structural properties, such as size, shape, symmetry or branching, and can also include chemical information about atom type and bond multiplicity [34–37]. Autocorrelation descriptors (GATS2e, GATS2p, GATS2m, GATS2v, MATS4e, MATS4p, MATS4m, MATS4v, MATS7v, RTm and ATS3m) also appeared to be significant in almost all the studied systems. These descriptors encode both molecular structure and physicochemical properties of a molecule (molecular mass, van der Waals volume, electronegativity or polarizability) [34,38,39]. Constitutional descriptors and molecular properties, like the number of double bonds (nDB) and the number of ratable bonds (RBN), the unsaturation index (Ui) and the octanol-water partition coefficient (ALOGP and MLOGP) also were relevant in the obtained PLS models. Finally, MDs related to connectivity had significant VIP values for the studied systems. These MDs are called BCUT (Burden eigenvalue descriptors) descriptors (BELe3) [34,40] and Randic connectivity indexes (VRv2 and VRp2) [34,41]. In this case, the selection of different descriptors related to the Sanderson electronegativity can be

mentioned giving a preliminary insight into the interaction mechanism between metabolites and stationary phases.

Table 2. Variable importance on projection (VIP) scores of the twenty most important molecular descriptors (MD) for each chromatographic system.

BEH Amide Acid		BEH Amide Moderately Acid		BEH Amide Neutral		Amide Acid		Amide Moderately Acid		Amide Neutral	
MD	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**
EEig10r	13.4	**RTe**	9.62	**Mor24u**	8.80	G2p	12.3	Mor02e	8.27	**Mor31e**	13.8
MWC03	11.5	**BELe3**	7.92	**Mor24e**	8.62	G2m	11.5	ESpm02x	7.89	**Mor31u**	13.8
Mor31e	9.76	**BLTD48**	7.92	RDF080v	7.95	G2u	11.1	**Mor31u**	6.76	**Mor08u**	9.85
GATS2v	8.72	**BLTA96**	7.89	**GATS2e**	6.10	G2v	11.1	**Mor31e**	6.70	**Mor08e**	8.85
GATS2p	8.66	**Mor08u**	7.84	**BELe3**	5.64	G2e	11.0	**Mor10m**	6.61	EPS1	7.04
Mor31u	8.56	**Mor24u**	7.19	**BLTA96**	5.61	**Mor31e**	9.29	**ESpm09d**	5.87	**Mor06m**	6.08
Mor02p	8.50	**Lop**	6.69	**BLTD48**	5.61	BELe2	9.05	**GATS2e**	5.81	GATS6v	6.00
GATS2m	8.33	G3v	6.63	ESpm09x	5.00	**Mor02p**	8.74	**Mor06m**	5.74	**Mor15v**	5.79
ESpm05d	8.08	**Mor08e**	6.59	**ESpm14d**	4.980	**Mor31u**	8.45	RDF020p	5.54	Mor08p	5.64
Mor06m	8.04	G3p	6.57	Mor19e	4.87	BEHp1	7.32	CIC0	5.41	**Mor15p**	5.50
MATS4m	7.45	G3u	6.39	R4e	4.76	G1u	6.99	**Mor08u**	5.39	**Mor10m**	5.40
MATS4e	6.74	G3s	6.29	**TIC4**	4.73	G1e	6.64	WA	5.32	RDF080m	5.31
MATS4v	6.44	G3e	6.27	**Mor08u**	4.71	G1m	6.60	**Mor08e**	5.08	**GATS2v**	5.00
Mor16e	5.39	**Mor10p**	6.05	**TIC3**	4.59	G1p	6.48	R5e	4.98	**Mor10v**	4.64
GATS4v	5.35	**Mor24e**	6.02	AAC	4.56	G1v	6.45	**RTm**	4.80	**RTm**	4.54
HVcpx	5.33	G3m	5.91	IC0	4.56	**MWC03**	6.21	**Mor01m**	4.74	Mor08v	4.52
SP03	5.23	**Mor10v**	5.54	piID	4.49	BELp2	6.10	VDA	4.68	ESpm14x	4.48
Mor17e	5.16	RDF075m	5.39	ESpm02u	4.41	**Mor10m**	5.25	RDF070u	4.64	**RBN**	4.47
MATS4p	4.92	**VRp2**	5.37	**Mor19u**	4.41	**HVcpx**	5.22	**Mor19u**	4.30	**ALOGP**	4.32
GATS2e	4.71	**VRv2**	5.37	ESpm12r	4.41	**MATS4m**	5.15	**RBN**	4.27	**Mor19u**	4.32
Zwitterionic Acid		**Zwitterionic Moderately Acid**		**Zwitterionic Neutral**		**Diol Acid**		**Diol Moderately Acid**		**Diol Neutral**	
MD	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**	**MD**	**VIP**
MATS7v	9.20	**EEig10r**	7.99	**MATS4m**	1.11	**ESpm09d**	1.50	G(O..O)	9.65	**Ui**	6.84
Mor24u	8.03	L3u	7.07	**MATS4e**	1.00	ESpm02r	1.42	IC3	8.52	**ESpm07d**	6.54
HTv	7.87	Mor18e	6.69	**MATS4v**	9.75	ESpm06d	1.12	**ATS3m**	7.82	**J**	6.49
Mor24e	7.42	**TIC4**	6.65	**Mor31e**	7.65	**ESpm13d**	7.21	T(O..O)	7.66	**ESpm01d**	6.41
BELe3	6.72	**TIC3**	6.40	**Mor31u**	7.51	**RTe**	7.08	nO	7.64	**nDB**	6.13
BLTD4	6.52	TIC5	6.22	GATS1m	7.05	MWC05	7.05	EEig10d	7.46	**MWC04**	6.05
BELe3	6.51	**Mor01m**	6.22	**EEig10r**	6.36	GNar	6.88	**J**	7.25	Mor23u	5.44
BLTA96	6.48	**Lop**	5.92	**MATS4p**	6.24	ATS1p	6.30	**ESpm14r**	7.12	Mor23e	5.28
Mor08e	5.52	ESpm11x	5.49	ESpm05u	6.11	**ATS3m**	6.26	**MWC04**	6.61	**ESpm14r**	5.21
Mor10p	5.29	**GATS2e**	5.26	HATS3u	6.09	**ESpm14d**	6.25	**ESpm01d**	6.50	**ESpm13d**	4.77
Mor15p	5.29	**Mor31e**	5.22	Jhetm	5.75	**MLOGP**	5.83	**nDB**	6.41	RDF040m	4.67
MLOGP	5.11	VRv1	5.17	ADDD	5.33	**GATS2e**	5.78	H0p	6.20	GGI2	4.65
Lop	5.09	**Mor31u**	4.96	**MLOGP**	5.31	MATS6p	5.23	**ESpm07d**	6.16	SPAN	4.65
Mor10v	5.05	ATS2p	4.95	**GATS2p**	5.23	**MATS7v**	5.21	AMW	5.93	**ATS3m**	4.62
Mor31e	4.98	**Mor24u**	4.95	**Mor15p**	5.08	ATS1m	5.19	X1sol	5.90	**RARS**	4.61
Mor15v	4.87	L3e	4.92	Mor07u	4.95	**RARS**	4.91	AECC	5.79	EEig09d	4.61
VRp2	4.74	ICR	4.82	**GATS2v**	4.92	**Mor24e**	4.85	**Mor02p**	5.73	QXXv	4.58
VRv2	4.74	ESpm08u	4.80	MATS6e	4.78	**GATS2m**	4.78	**Ui**	5.68	Mor21u	4.21
Mor31u	4.63	ESpm10x	4.72	**GATS2m**	4.77	CIC1	4.78	HDcpx	5.08	**ALOGP**	4.17
Mor03u	4.60	**Mor08e**	4.64	**RBN**	4.75	L2e	4.67	**ESpm13d**	5.03	EEig10x	4.12

Note: Bold format is used to highlight those MDs appearing in more than one chromatographic system.

A deep analysis of the identified molecular descriptors and their relationships with the significant experimental factors (stationary phase and pH value) allowed finding some interesting trends. Figure 4a shows a Venn diagram showing the MDs for each stationary phase considering all pH values. In this plot, the different behavior of the diol stationary phase can be observed. Most MDs (38) were unique, and only some of them appeared as relevant to other stationary phases. This difference in the behavior of the mixed-mode diol stationary phase can be explained by the mixed chemistry of the surface with a hydrophobic alkyl chain with a diol group. These dual properties allow the use of this stationary phase for both RP and HILIC separations but, from our results, modelling using a PLS model approach was more difficult. In addition, despite the ionic strength factor not being significant (using ASCA), the mixed-mode diol stationary phase seemed to be more affected than the other stationary phases, which could also be related to worse modelling. When considering amide,

BEH amide and zwitterionic stationary phases, more similarities in the identified MDs were observed. In addition, from these results, and in accordance with PCA, the zwitterionic stationary phase seemed to have an intermediate behavior between the two amide stationary phases. Finally, evaluation of the MDs identified for the different stationary phases at different pH values showed that most of MDs were unique for a particular condition. However, more similarities can be observed between the moderately acid and neutral pH values (especially in the case of the diol stationary column, confirming its different behavior).

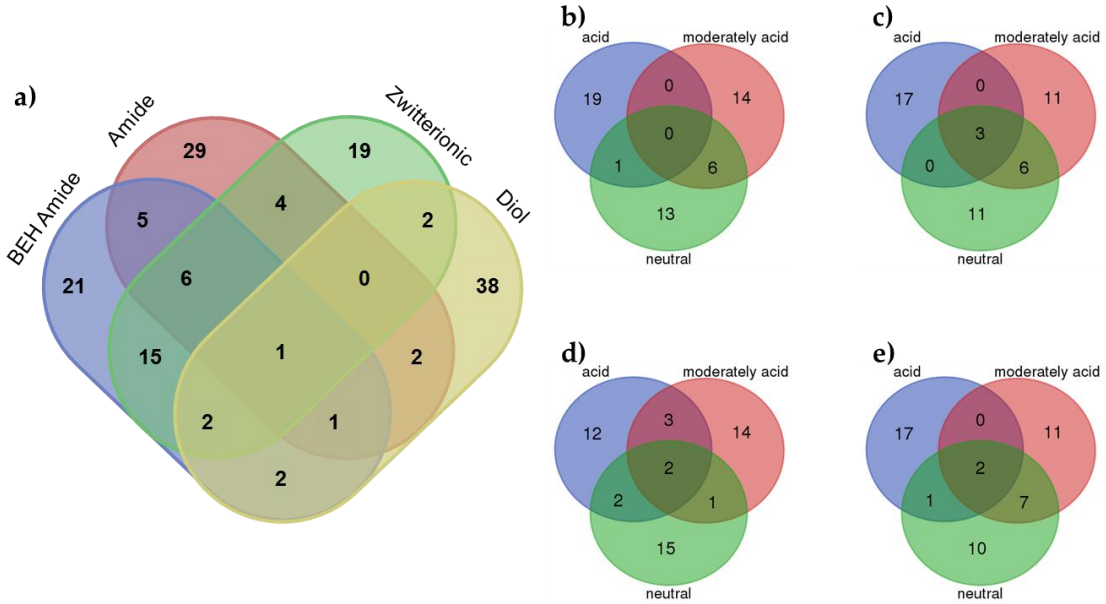


Figure 4. Venn diagrams of the VIPs MDs at (**a**) each stationary phase considering all pH values; (**b**) BEH Amide at each pH condition; (**c**) Amide at each pH condition; (**d**) Zwitterionic at each pH condition and (**e**) Diol at each pH condition.

3. Materials and Methods

3.1. Chemicals and Reagents

Acetic acid (≥95.0%), formic acid (≥95.0%), ammonia (25%), LC-MS water and Acetonitrile (ACN, LC-MS grade) were obtained from Merck (Darmstadt, Germany). Ammonium acetate (≥99.0%) was supplied by Sigma-Aldrich (St. Louis, MO, USA).

A mixture of 54 metabolites was used to evaluate the HILIC stationary phases behavior. Table 1 shows the metabolites contained in the analyzed mixture. All standards were purchased from Sigma-Aldrich (St. Louis, MO, USA). Nucleosides and amino acids were from two mix solutions provided by Sigma-Aldrich (St. Louis, MO, USA). A summary of the polarity of the analyzed metabolites is shown in Table S1 in the supplementary information.

Standard stock solutions (1000 µg mL^{-1}) of metabolite mixture were prepared by dissolving an appropriate amount of each metabolite in water and stored at −20 °C until their use. Working standard solutions (20 µg mL^{-1}) were prepared by diluting the stock solution in ACN:H2O (1:1).

3.2. Instrumentation

The metabolite mixture was analyzed using an Acquity UHPLC system (Waters, Milford, MA, USA) for the chromatographic separation, equipped with a quaternary pump, an autosampler and a column oven. The mass spectrometer was a triple quadrupole detector (TQD, Waters, Milford, MA, USA) equipped with an electrospray (ESI) ionization source in negative and positive modes. The mass acquisition range was set to 90–1000 *m/z*.

Four different HILIC stationary phases (BEH amide, amide, zwitterionic and mixed-mode diol) were evaluated (properties summarized in Table 3). The elution gradient was performed using solvent A (acetonitrile) and solvent B (ammonium acetate buffer solution). Chromatographic conditions used for each column are also detailed in Table 3. In order to reproduce the most used chromatographic conditions in metabolomics studies, the experiments were performed using solvent B at three different pH values: acidic (3.0 adjusted with formic acid), moderately acidic (5.5 adjusted with acetic acid) and neutral (7.0 adjusted with ammonia). The pH measurements were performed at 25 °C using an Orion Star A111 pH meter (Thermo Scientific, Waltham, MA, USA), before the additions of the organic solvent. Moreover, two ionic strengths in the aqueous phase were also compared: low (5.0 mM) and high (25 mM).

The metabolite mixture was analyzed with the four stationary phases working with solvent B at the three pH values and the two ionic strengths. Each condition was injected twice giving an experimental design with a total number of 48 chromatographic runs.

3.3. Data Analysis

3.3.1. Retention Factor Determination

Figure 5 shows a complete picture of the data analysis strategy from the raw MS data to the chemometric modelling. First, raw chromatographic data files (in .raw format) were converted to the standard CDF format by Databridge function of MassLynx^{TM} v 4.1 software (Waters, Milford, MA, USA). Then, these data files were imported into the MATLAB environment (Release 2015b, The Mathworks Inc., Natick, MA, USA) by using *mzcdfread* and *mzcdf2peak* functions of the MATLAB Bioinformatics Toolbox (4.3.1.version). LC-MS data were then arranged and aligned according to their m/z in a data matrix, containing retention times in the rows and selected *m/z* values in the columns. Here, this data matrix was built up using the previously proposed regions of interest (ROI) strategy [42,43]. The ROI approach selects the most relevant mass traces, which are those *m/z* values whose intensity signals are higher than a fixed signal-to-noise ratio threshold and appear a number of times consecutively in the time dimension. These mass traces are searched among all the chromatographic and spectral data. The obtained vectors, containing the intensity of the found ROIs at each time point, are reorganized into a matrix grouping ROIs among all the retention times. The final *m/z* values of each ROI are calculated as the mean of all m/z values obtained for that particular ROI. In this work, the parameters for the implementation of this ROI approach are the signal-to-noise ratio threshold (set at 0.1% of the maximum MS signal intensity), the mass accuracy of the mass spectrometer (set at 0.5 Da/e for the TQD MS analyzer used in this work) and the minimum number of consecutive retention times to be considered as a chromatographic peak (set at 25). More details on how this strategy works are given in previous works [1]. Finally, an ROI matrix was obtained for positive and negative ionization modes for each of the 72 chromatographic runs of the present study.

Every ROI matrix corresponding to each chromatographic run was then evaluated to automatically find the *m/z* value of each metabolite and provide their retention time in each chromatogram. Lastly, retention factor (k) of each metabolite in each chromatographic run was calculated using their retention times (t_R) and the dead time (t_0, theoretically obtained from the dead volume) as follows:

$$k = \frac{t_R - t_0}{t_0} \quad (1)$$

Table 3. HILIC column specifications and chromatographic separation conditions used during the analysis.

Column Specifications				Chromatographic Separation Conditions	
Name	**Manufacturer**	**Stationary Phase**	**Dimensions**	**Flow (mL·min^{-1})**	**Elution Gradient (A: Acetonitrile; B: Water with Ammonium Acetate)**
XBridge^{TM} Amide	Waters (Milford, MA, USA)	BEH amide	150 × 4.6 mm^2 i.d., 5 µm	0.15	0–4 min, at 5% B; 4–34 min, from 5% to 70% B; 34–42 min, at 70% B; and 42–44 min, at 5%B
TSK Gel Amide-80	Tosoh Bioscience (Tokyo, Japan)	Amide	250 × 2.0 mm^2 i.d., 5 µm	0.15	0–3 min, at 5% B; 3–27 min, from 5% to 70% B; 27–30 min, at 70% B; and 30–32 min, at 5%B
ZIC-HILIC	SeQuant (Umeå, Sweden)	Zwitterionic	250 × 2.1 mm^2 i.d., 5 µm	0.15	0–3 min, at 5% B; 3–27 min, from 5% to 70% B; 27–30 min, at 70% B; and 30–32 min, at 5%B
Acclaim^{TM} Mixed-Mode HILIC-1	Thermo Scientific (Sunnyvale, CA, USA)	Mixed-mode diol	150 × 2.1 mm^2 i.d., 5 µm	0.15	0–2 min, at 5% B; 2–16 min, from 5% to 70% B; 16–20 min, at 70% B; and 20–22 min, at 5%B

Metabolites **2017**, *7*, 54

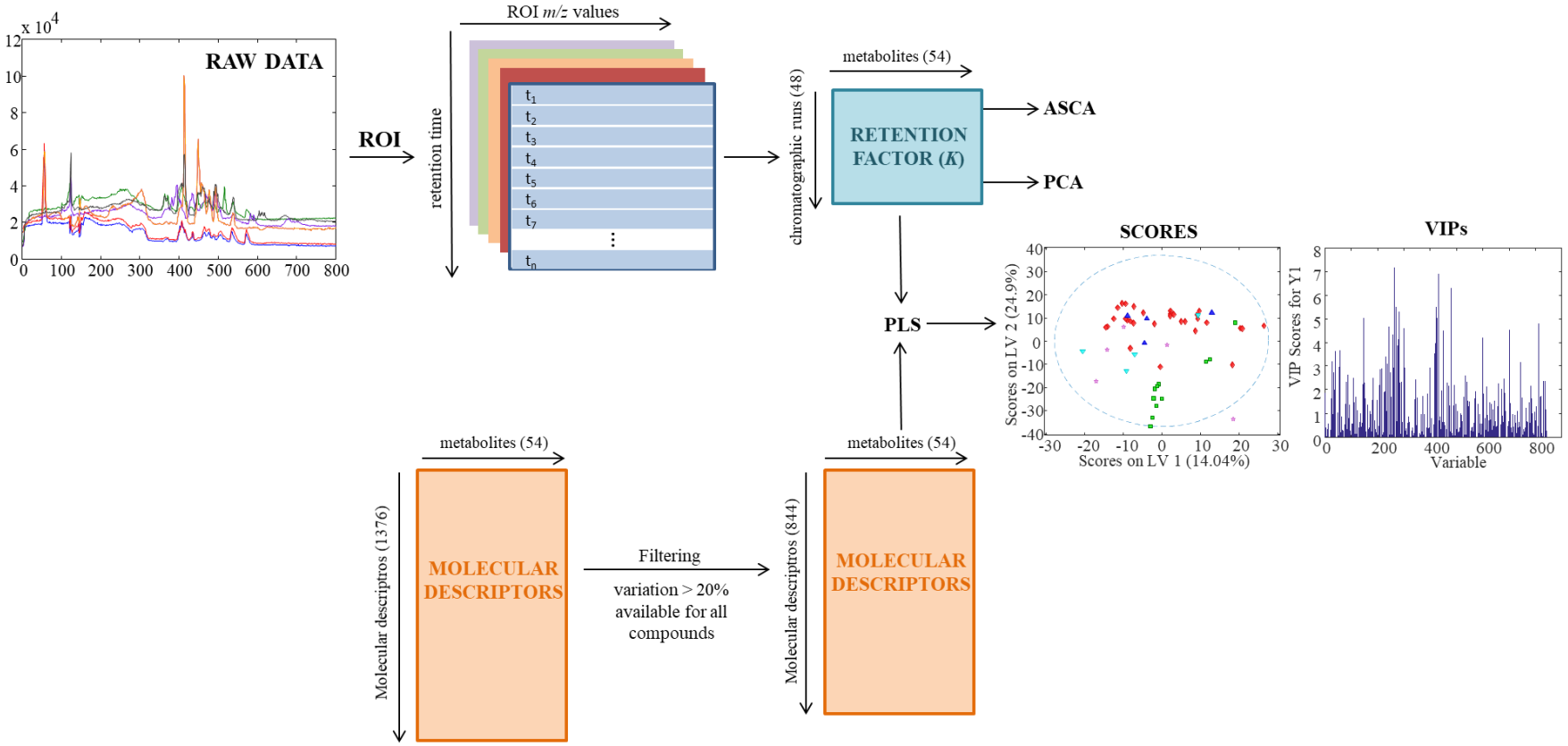


Figure 5. Scheme of data analysis strategy.

3.3.2. Molecular Descriptors Determination

Canonical SMILES representations for the standard metabolites were retrieved from PubChem [44] and HMDB [45] databases. These SMILES were input into the PCLIENT software to calculate molecular descriptors (MDs). PCLIENT software can calculate more than 3000 MDs that are divided into 25 logical blocks. Here 1376 MDs were calculated including constitutional, topological, geometrical, electrostatic, physical, shape, and quantum chemical descriptors. Details of MDs calculation can be found in the Handbook of Molecular Descriptors [34]. PCLIENT software is available online at http://www.vcclab.org. When 3D atom coordinates were needed for parameter calculation, they were obtained using CORINA software (Molecular Networks GmbH, Nürnberg, Germany).

To reduce the number of MDs descriptors with a percentage of variation (calculated dividing the standard deviation of the MD values by their mean) lower than 20% were excluded. Also, those descriptors not available for all compounds were removed. After this reduction, 844 MDs were obtained for further analysis.

3.3.3. Evaluation of HILIC Chromatographic Performance

The retention factors of the 54 metabolites in each HILIC stationary phase at different chromatographic conditions were used to investigate the behavior of the chromatographic systems studied in this work by explorative chemometric methods.

The behaviors of the chromatographic systems studied in this work were evaluated using the retention factors of the 54 metabolites. The retention factor data matrix **D** (containing the retention factor of 54 metabolites at the 48 chromatographic runs) was evaluated using diverse chemometrics exploratory methods: principal component analysis (PCA) [18], ANOVA-simultaneous component analysis (ASCA) [19] and partial least squares regression (PLS).

PCA [18] compresses the information of the original variables into a smaller number of uncorrelated variables known as principal components [18]. In this work, PCA was applied to evaluate the relationships between the experimental conditions studied: stationary phase, pH and ionic strength. Therefore, matrix **D** was analyzed and information about the chromatographic runs and metabolite distribution were obtained in scores and loadings, respectively.

ASCA [19] is a multivariate analysis of variance method that combines the capacity of ANOVA to separate variance sources with the advantages of simultaneous component analysis (SCA, a generalization of PCA for the situation where the same variables have been measured in multiple conditions) [46]. In this work, ASCA was applied to statistically assess the significance of experimental factors in the experimental design: stationary phase, pH and ionic strength. ASCA was performed on a well-balanced experimental design, and 10,000 permutations were used for the permutation test [47]. More details about the ASCA method can be found in the work of Smilde [19] and Jansen [48]. Data were autoscaled prior to applying PCA and centered before applying ASCA.

Finally, PLS regression was used to explore the relationships between obtained retention factors for each chromatographic condition and molecular descriptors (MDs). PLS [49–51] is a multivariate linear regression model used to find correlation models between predictor variables (**X** data matrix) and response values to be predicted (**y** vector). In this work, PLS is used as a regression analysis method to build a model to link the determined retention factor of metabolites (arranged in a vector **y**) using their MDs (arranged in matrix **X**) and to investigate the most influential MDs in the regression. In this work, the optimum number of latent variables for each model was selected using leave-one-out cross-validation.

PLS also provides information about the most relevant variables for achieving the retention factors modelling. For instance, variables importance in projection (VIP) scores can be used for that purpose [51]. According to the common use, variables with a VIP score greater than 1 were important [52]. In this work, these VIP variables corresponded to those MDs that allowed a better description of the retention factor for each considered metabolite.

PCA, ASCA, and PLS were performed using PLS Toolbox 8.0 (Eigenvector Research Inc., Wenatchee, WA, USA) working under MATLAB (The Mathworks, Natick, MA, USA).

4. Conclusions

Results obtained in the assessment of the behavior of HILIC chromatographic stationary phases by means of a variety of chemometric methods showed that the two most important factors to be considered in metabolic studies are the stationary phase and the pH of the aqueous solvent. Moreover, BEH amide and mixed-mode diol stationary phases behaved rather differently compared to amide and zwitterionic phases, which performed similarly. ASCA loadings were useful to know which metabolites were the most important to distinguish the stationary phases. Amino acids appeared to be useful to distinguish the mixed-mode diol stationary phase, while organic acids seemed to distinguished BEH amide from the rest of the stationary phases. In addition, exploratory PLS models allowed linking the retention of metabolites at different chromatographic conditions with molecular descriptors defining their physicochemical properties. Again, a similar behavior was observed for amide and zwitterionic stationary phases whereas the mixed-mode diol stationary phase showed a different performance.

Finally, the obtained PLS models could be considered as a starting point in a more comprehensive work for modelling and prediction of chromatographic retention factors of metabolites in different HILIC stationary phases. However, building up these models requires bigger metabolite datasets with a larger number of compounds for each of the metabolite families. Moreover, efforts should be made in performing a comprehensive external validation of the models. Future work should address these issues.

Supplementary Materials: The following are available online at www.mdpi.com/2218-1989/7/4/54/s1, Figure S1: PCA scores plot of samples classified according to the pH of the mobile phase. Table S1: Metabolites' logP values obtained from PCLIENT software.

Acknowledgments: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement Number 320737.

Author Contributions: All authors conceived and designed the experiments; M.N.-R. and E.O.-V. performed the chromatographic experiments; M.N.-R. performed the chemometric data analysis; M.N.-R., R.T. and J.J. discussed and interpreted results; M.N.-R. and J.J. drafted the manuscript. All authors revised the article and gave final approval of the submitted version.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends Anal. Chem.* **2016**, *82*, 425–442. [CrossRef]
2. Rochat, B. From targeted quantification to untargeted metabolomics: Why LC-high-resolution-MS will become a key instrument in clinical labs. *TrAC Trends Anal. Chem.* **2016**, *84*, 151–164. [CrossRef]
3. Jorge, T.F.; Rodrigues, J.A.; Caldana, C.; Schmidt, R.; van Dongen, J.T.; Thomas-Oates, J.; António, C. Mass spectrometry-based plant metabolomics: Metabolite responses to abiotic stress. *Mass Spectrom. Rev.* **2016**, *35*, 620–649. [CrossRef] [PubMed]
4. Horvatovich, P.L.; Bischoff, R. Current technological challenges in biomarker discovery and validation. *Eur. J. Mass Spectrom.* **2010**, *16*, 101–121. [CrossRef] [PubMed]
5. Monteiro, M.S.; Carvalho, M.; Bastos, M.L.; De Pinho, P.G. Metabolomics analysis for biomarker discovery: Advances and challenges. *Curr. Med. Chem.* **2013**, *20*, 257–271. [CrossRef] [PubMed]
6. Madji Hounoum, B.; Blasco, H.; Emond, P.; Mavel, S. Liquid chromatography-high-resolution mass spectrometry-based cell metabolomics: Experimental design, recommendations, and applications. *TrAC Trends Anal. Chem.* **2016**, *75*, 118–128. [CrossRef]

7. Naz, S.; Moreira Dos Santos, D.C.; García, A.; Barbas, C. Analytical protocols based on LC-MS, GC-MS and CE-MS for nontargeted metabolomics of biological tissues. *Bioanalysis* **2014**, *6*, 1657–1677. [CrossRef] [PubMed]
8. Cubbon, S.; Antonio, C.; Wilson, J.; Thomas-Oates, J. Metabolomic applications of HILIC–LC–MS. *Mass Spectrom. Rev.* **2010**, *29*, 671–684. [CrossRef] [PubMed]
9. Spagou, K.; Tsoukali, H.; Raikos, N.; Gika, H.; Wilson, I.D.; Theodoridis, G. Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies. *J. Sep. Sci.* **2010**, *33*, 716–727. [CrossRef] [PubMed]
10. Tang, D.Q.; Zou, L.; Yin, X.X.; Ong, C.N. HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS. *Mass Spectrom. Rev.* **2016**, *35*, 574–600. [CrossRef] [PubMed]
11. Taraji, M.; Haddad, P.R.; Amos, R.I.J.; Talebi, M.; Szucs, R.; Dolan, J.W.; Pohl, C.A. Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures. *J. Chromatogr. A* **2017**, *1486*, 59–67. [CrossRef] [PubMed]
12. Guo, Y. Recent progress in the fundamental understanding of hydrophilic interaction chromatography (HILIC). *Analyst* **2015**, *140*, 6452–6466. [CrossRef] [PubMed]
13. Hendrickx, S.; Adams, E.; Cabooter, D. Recent advances in the application of hydrophilic interaction chromatography for the analysis of biological matrices. *Bioanalysis* **2015**, *7*, 2927–2945. [CrossRef] [PubMed]
14. Alpert, A.J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J. Chromatogr. A* **1990**, *499*, 177–196. [CrossRef]
15. Jandera, P. Stationary and mobile phases in hydrophilic interaction chromatography: A review. *Anal. Chim. Acta* **2011**, *692*, 1–25. [CrossRef] [PubMed]
16. Buszewski, B.; Noga, S. Hydrophilic interaction liquid chromatography (HILIC)—A powerful separation technique. *Anal. Bioanal. Chem.* **2012**, *402*, 231–247. [CrossRef] [PubMed]
17. Ortiz-Villanueva, E.; Navarro-Reig, M.; Jaumot, J.; Tauler, R. Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: Resolving complex mixtures of metabolites. *Anal. Method.* **2017**, *9*, 774–785. [CrossRef]
18. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
19. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef] [PubMed]
20. Creek, D.J.; Jankevics, A.; Breitling, R.; Watson, D.G.; Barrett, M.P.; Burgess, K.E.V. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction. *Anal. Chem* **2011**, *83*, 8703–8710. [CrossRef] [PubMed]
21. Falchi, F.; Bertozzi, S.M.; Ottonello, G.; Ruda, G.F.; Colombano, G.; Fiorelli, C.; Martucci, C.; Bertorelli, R.; Scarpelli, R.; Cavalli, A.; et al. Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: A useful tool for metabolite identification. *Anal. Chem.* **2016**, *88*, 9510–9517. [CrossRef] [PubMed]
22. Yousefinejad, S.; Hemmateenejad, B. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 177–204. [CrossRef]
23. Goryński, K.; Bojko, B.; Nowaczyk, A.; Buciński, A.; Pawliszyn, J.; Kaliszan, R. Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. *Anal. Chim. Acta* **2013**, *797*, 13–19. [CrossRef] [PubMed]
24. Zisi, C.; Sampsonidis, I.; Fasoula, S.; Papachristos, K.; Witting, M.; Gika, H.G.; Nikitas, P.; Pappa-Louisi, A. QSRR modeling for metabolite standards analyzed by two different chromatographic columns using multiple linear regression. *Metabolites* **2017**, 7. [CrossRef] [PubMed]
25. Randazzo, G.M.; Tonoli, D.; Hambye, S.; Guillarme, D.; Jeanneret, F.; Nurisso, A.; Goracci, L.; Boccard, J.; Rudaz, S. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. *Anal. Chim. Acta* **2016**, *916*, 8–16. [CrossRef] [PubMed]

26. Kritikos, N.; Tsantili-Kakoulidou, A.; Loukas, Y.L.; Dotsikas, Y. Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure–retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction. *J. Chromatogr. A* **2015**, *1403*, 70–80. [CrossRef] [PubMed]
27. Park, S.H.; Haddad, P.R.; Talebi, M.; Tyteca, E.; Amos, R.I.J.; Szucs, R.; Dolan, J.W.; Pohl, C.A. Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model. *J. Chromatogr. A* **2017**, *1486*, 68–75. [CrossRef] [PubMed]
28. Aicheler, F.; Li, J.; Hoene, M.; Lehmann, R.; Xu, G.; Kohlbacher, O. Retention time prediction improves identification in nontargeted lipidomics approaches. *Anal. Chem.* **2015**, *87*, 7698–7704.
29. Wolfer, A.M.; Lozano, S.; Umbdenstock, T.; Croixmarie, V.; Arrault, A.; Vayer, P. UPLC–MS retention time prediction: A machine learning approach to metabolite identification in untargeted profiling. *Metabolomics.* **2015**, *12*, 8. [CrossRef]
30. Cao, M.; Fraser, K.; Huege, J.; Featonby, T.; Rasmussen, S.; Jones, C. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **2015**, *11*, 696–706. [CrossRef] [PubMed]
31. Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graph. Model.* **2014**, *54*, 194–203. [CrossRef] [PubMed]
32. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030–1037. [CrossRef]
33. Schuur, J.H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344. [CrossRef]
34. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2008.
35. Balaban, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404. [CrossRef]
36. Balaban, A.T. Chemical graphs-XXXIV. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355–375. [CrossRef]
37. Gozalbes, R.; Doucet, J.P.; Derouin, F. Application of topological descriptors in QSAR and drug design: History and new trends. *Curr. Drug Targets Infect. Disord.* **2002**, *2*, 93–102. [CrossRef] [PubMed]
38. Hollas, B. An analysis of the autocorrelation descriptor for molecules. *J. Math. Chem.* **2003**, *33*, 91–101. [CrossRef]
39. Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: Perception, autocorrelation descriptor and sar studies. *Eur. J. Med. Chem.* **1984**, *19*, 66–70.
40. Burden, F.R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.-Act. Relationsh.* **1997**, *16*, 309–314. [CrossRef]
41. Randić, M. Molecular shape profiles. *J. Chem. Inform. Comp. Sci.* **1995**, *35*, 373–382. [CrossRef]
42. Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*. [CrossRef] [PubMed]
43. Stolt, R.; Torgrip, R.J.O.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S.P. Second-order peak detection for multicomponent high-resolution LC/MS data. *Anal. Chem.* **2006**, *78*, 975–983. [CrossRef] [PubMed]
44. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. [CrossRef] [PubMed]
45. Wishart, D.S.; Knox, C.; Guo, A.C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D.D.; Psychogios, N.; Dong, E.; Bouatra, S.; et al. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610. [CrossRef] [PubMed]
46. Jansen, J.J.; Hoefsloot, H.C.J.; Van Der Greef, J.; Timmerman, M.E.; Smilde, A.K. Multilevel component analysis of time-resolved metabolic fingerprinting data. *Anal. Chim. Acta* **2005**, *530*, 173–183. [CrossRef]
47. Vis, D.J.; Westerhuis, J.A.; Smilde, A.K.; van der Greef, J. Statistical validation of megavariate effects in ASCA. *BMC Bioinform.* **2007**, *8*. [CrossRef] [PubMed]

48. Jansen, J.J.; Hoefsloot, H.C.J.; Van Der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chem.* **2005**, *19*, 469–481. [CrossRef]
49. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]
50. Wold, H. *Multivariate Analysis*; Academic Press: New York, NY, USA, 1966; pp. 391–420.
51. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chem. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
52. Chong, I.G.; Jun, C.H. Performance of some variable selection methods when multicollinearity is present. *Chem. Intell. Lab. Syst.* **2005**, *78*, 103–112. [CrossRef]

Informació Suplementària a la Publicació 1

Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches.

M. Navarro-Reig, E. Ortiz-Villanueva, R. Tauler, J. Jaumot.

Metabolites 7 (2017), 54.

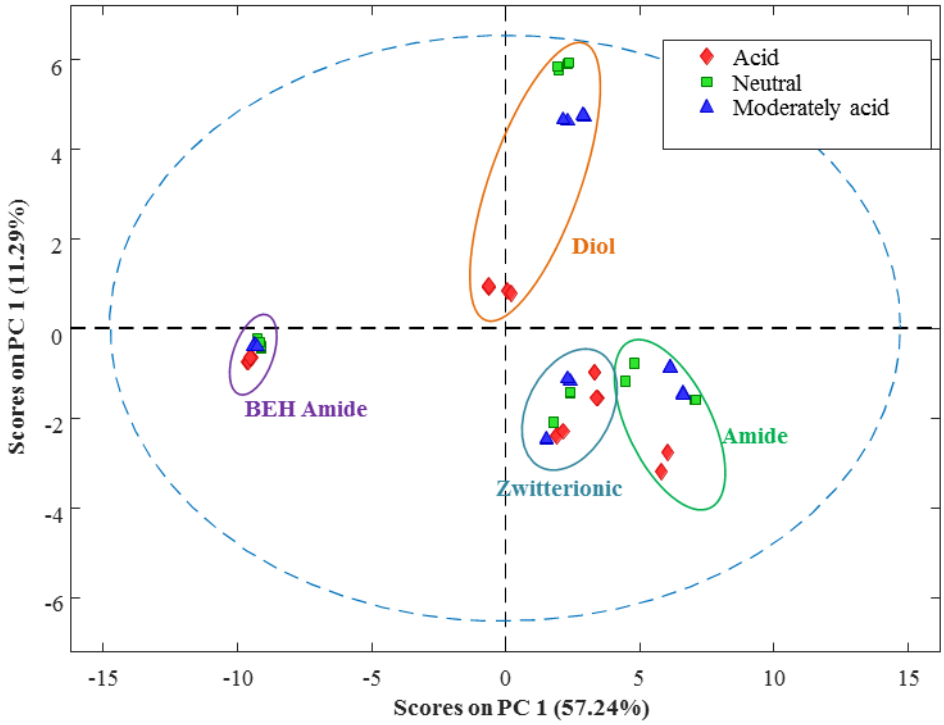


Figure S1.PCA scores plot of samples classified according to the pH of the mobile phase.

Table S1. Metabolites' logP values obtained from PCLIENT software.

Metabolites	logP
Nucleosides	
1-methyladenosine	-0.054
2"-O-methylcytidine	-0.872
2-thiocytidine	-0.427
5-methylcytidine	-0.872
cytidine	-1.176
guanosine	-0.808
inosine	-0.752
pseudouridine	-1.988
ribothymidine	-1.278
uridine	-1.582
Amino Acids	
1-methyl-L-histidine	-2.704
3-methyl-L-histidine	-2.704
4-hydroxy-L-proline	-1.04
5-hydroxylysine	-3.273
β-alanine	-2.918
creatinine	-0.339
cysteine	-5.555
L-(-)-proline	-0.232
L-(+)-arginine	-2.934
L-(+)-cystathionine	-5.202
L-(+)-lysine	-2.485
L-2-aminoadipic acid	-2.568
L-2-amino-n-butyric acid	-2.465
L-alanine	-2.918
L-anserine	-0.75
L-carnosine	-1.054
L-citrulline	-3.34
L-glutamic acid	-2.946
L-histidine	-3.057
L-homocystine	-4.869
L-isoleucine	-1.677
L-leucine	-1.677
L-methionine	-2.055
L-ornithine	-2.863
L-serine	-3.726
L-threonine	-3.272
L-tryptophan	-1.267
L-valine	-2.055
taurine	-1.731
sarcosine	-0.703
L-aspartic acid	-3.356

Sugars	
D(-)-ribose	-2.089
glucose	-2.483
trehalose	-3.898
mannitol	-2.497
Organic acids	
citric acid	-1.169
ketoglutaric acid	-0.828
pimelic acid	0.787
succinic acid	-0.353
creatine	-0.735
Others	
hypoxanthine	0.582
L-carnitine	-3.601
serotonin	1.006
tryptamine	1.574

3.3. Publicació 2

Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies.

M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler.

Analytical and Bioanalytical Chemistry 407 (2015), 8835-8847.

Anal Bioanal Chem (2015) 407:8835–8847
DOI 10.1007/s00216-015-9042-2

RESEARCH PAPER

Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies

Meritxell Navarro-Reig[1] · Joaquim Jaumot[1] · Alejandro García-Reiriz[1,2] · Romà Tauler[1]

Received: 30 July 2015 / Revised: 7 September 2015 / Accepted: 10 September 2015 / Published online: 24 September 2015

Abstract The comprehensive analysis of untargeted metabolomics data acquired using LC-MS is still a major challenge. Different data analysis tools have been developed in recent years such as XCMS (various forms (X) of chromatography mass spectrometry) and multivariate curve resolution alternating least squares (MCR-ALS)-based strategies. In this work, metabolites extracted from rice tissues cultivated in an environmental test chamber were subjected to untargeted full-scan LC-MS analysis, and the obtained data sets were analyzed using XCMS and MCR-ALS. These approaches were compared in the investigation of the effects of copper and cadmium exposure on rice tissue (roots and aerial parts) samples. Both methods give, as a result of their application, the whole set of resolved elution and spectra profiles of the extracted metabolites in control and metal-treated samples, as well as the values of their corresponding chromatographic peak areas. The effects caused by the two considered metals on rice samples were assessed by further chemometric analysis and statistical evaluation of these peak area values. Results showed that there was a statistically significant interaction between the considered factors (type of metal of treatment and tissue). Also, the discrimination of the samples according to both factors was possible. A tentative identification of the most discriminant metabolites (biomarkers) was assessed. It is finally concluded that both XCMS- and MCR-ALS-based strategies provided similar results in all the considered cases despite the completely different approaches used by these two methods in the chromatographic peak resolution and detection strategies. Finally, advantages and disadvantages of using these two methods are discussed.

Keywords Metabolomics · LC-MS · XCMS · MCR-ALS · Cu · Cd · Rice · Sample discrimination

Electronic supplementary material The online version of this article (doi:10.1007/s00216-015-9042-2) contains supplementary material, which is available to authorized users.

✉ Joaquim Jaumot
joaquim.jaumot@idaea.csic.es

[1] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[2] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina

Introduction

Metabolomics can be defined as the exhaustive profiling study of all metabolites contained in an organism. It is known that external perturbations imposed on organisms can produce changes in their metabolome. These perturbations can be environmental changes; physical, abiotic, or nutritional stresses; mutation; and transgenic events [1–3]. Therefore, metabolomics is a powerful approach to study molecular mechanisms and metabolic pathways implicated in the response to different perturbations and in the organism defense strategies against them. Over the last decade, data processing has been a challenge in untargeted metabolomics due to the extreme complexity of the experimental data sets, especially in the case of combining a MS detector with chromatographic techniques such as LC or GC. As a consequence, software programs for automated processing of data have been introduced, such as MetAlign [4], MZmine [5], or XCMS [6], among others.

In the last years, XCMS has become a favorite method among the metabolomic community for feature detection,

and it has been used for a broad range of applications. In brief, XCMS is a tool dedicated to chromatographic feature detection which includes automatic processing of huge size full-scan LC-MS data and estimates candidate metabolites by using peak detection and retention time correction algorithms and methods. For each proposed candidate, XCMS gives *p* value (statistical test comparing the integrated peak areas of this candidate in control versus treated samples) and fold change (defined as the ratio of the integrated peak areas of the treated samples versus the control samples) [7, 8].

MCR-ALS is also a popular chemometric method used for the resolution of pure contributions in unresolved mixtures [9]. MCR-ALS is used in a wide variety of applications as, for instance, the resolution of overlapped chromatographic peaks in environmental samples. MCR-ALS has been recently proposed as an alternative approach to detect potential biomarkers in untargeted metabolomics studies [10]. MCR-ALS decomposes the experimental LC-MS data matrix into their factor contributions which can be assigned to the chromatographic elution profiles and to the mass spectra of each resolved component. The main difference between these two approaches lies in peak detection and resolution. While XCMS identifies each feature characterized by its retention time and a unique *m*/*z* value, MCR-ALS resolves mathematical components characterized by their elution profiles and mass spectra (with more than one possible MS feature assigned to the same elution profile) [6, 10]. With the aim of comparing these two approaches, in the present work, the same metabolomic data set was processed by means of XCMS and by MCR-ALS, and further evaluated by using other chemometric methods for exploration and discrimination purposes. The proposed untargeted metabolomic approach has been used to assess the effects of cadmium and copper treatment on Japanese rice.

Plants are complex organisms exposed to a set of abiotic and biotic stresses [11]. One of these abiotic stresses is the pollution by toxic metals present in the environment. These metals can be found as constituents of the Earth's crust and geological processes, but human activities, such as mining, agriculture, and a wide range of industrial activities, can drastically alter their geochemical cycles and distribution on earth surface [12, 13]. These anthropogenic activities caused that the level of some of these toxic metals in the environment increased notably in recent years. Although the discovery of adverse health effects resulting from toxic metals has caused the decrease of emissions in most of the developed countries during the last century, there are still some metals like cadmium, whose emissions increased during the twentieth century, due to its large industrial use and reduced recycling [14]. Among toxic metals, cadmium and copper have been listed on the priority list of hazardous materials by the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) in 2013 [13]. These two pollutants are readily absorbed by roots and rapidly translocate to the aerial parts of plants [15]. Since diet is the primary source of exposure to these metals for the general population, intensive research has been performed on the accumulation of these pollutants in edible plants [13]. In this work, Japanese rice (*Oryza sativa japonica* Nipponbare) has been used as a target organism because it is one of the model organisms frequently used in plant metabolomics and is also an edible plant [1, 3, 13].

The metabolomic study presented in this work considers two categorical factors related to the metal exposure: rice tissue sample analyzed (root or aerial part) and metal of treatment (Cd or Cu). Metabolomic datasets commonly use statistical experimental designs, where different dose groups, multiple time points, diverse sample groups, or various subjects are simultaneously investigated [16, 17]. For this reason, comprehensive data analysis methods able to deal with this type of complex designs are required. In this work, statistical evaluation of the different investigated effects has been performed using different multivariate data analysis methods such as ANOVA-simultaneous component analysis (ASCA), principal component analysis (PCA), and partial least-squares discriminant analysis (PLS-DA).

Experimental

Reagents

Cadmium chloride hydrate (≥98.0 %), copper(II) sulfate pentahydrate (≥98.0 %), and ammonium acetate (≥98.0 %) were from Sigma-Aldrich (Steinheim, Germany). HPLC-grade water, acetonitrile (≥99.8 %), and methanol (≥99.8 %) were supplied by Merck (Darmstadt, Germany). Chloroform was obtained from Carlo Erba (Peypin, France). Piperazine-*N*, *N*′-bis(2-ethanesulfonic acid) (PIPES) (≥99.0 %) was used as internal standard (Sigma-Aldrich, Steinheim, Germany).

Solutions containing 10, 50, and 100 μM of cadmium (Cd) and copper (Cu) were prepared weekly by diluting a 1000 μM stock solution of these metals. Stock solutions were prepared weekly by dissolution of the appropriate amounts of cadmium chloride hydrate and copper(II) sulfate salts. All the solutions were stored at 6 °C until their use.

Water used for plant watering, for preparing cadmium and copper solutions, and during the extraction procedure was purified using an Elix 3 coupled to a Milli-Q system (Millipore, Belford, MA, USA), and filtered through a 0.22-μm nylon filter integrated into the Milli-Q system.

Plant growth, stress treatment, and metabolite extraction

Oryza sativa japonica Nipponbare seeds, obtained from the Center for Research in Agricultural Genomics (CRAG) at Autonomous University of Barcelona, were incubated for

2 days at 30 °C in a wet environment. After this period, seeds were planted in 3.0×3.0 cm individual pots and grown on an Environmental Test Chamber MLR-352H (Panasonic®) for 22 days under white fluorescent light. Temperature, relative humidity, and light long-day conditions at the chamber were set as described in the Electronic supplementary material (ESM) Table S1. During the first 10 days of growth, rice plants were watered with Milli-Q water three times a week. Since then, plant treated samples were subjected to irrigation water containing different concentrations of Cd and Cu, whereas the plant control samples were watered with milli-Q water until harvest. Metal concentrations used for stressing rice plants were 10, 50, and 1000 μM, and for every concentration, two trays containing 18 pots were used. The lower concentration was set to 10 μM, in agreement with the lowest reported metal concentration producing noticeable changes in plants [13, 18, 19], and the higher concentration was set to 1000 μM because it is the highest metal concentration inducing changes in plants without causing their death [13, 18, 19]. In order to avoid differences in the growth of individual plants, the position of the trays inside the chamber was changed daily following a random design, and the volume of irrigation water was controlled and set at 200 mL per tray. After harvest, roots and aerial part were separated and, immediately, metabolism was quenched by freezing at liquid nitrogen temperature. Samples were stored at −80 °C until extraction.

Before extraction, aerial parts and roots were ground under liquid nitrogen to a fine powder and lyophilized overnight until dryness. Metabolite extraction was carried out by dispersing 40 mg of the dried tissue in 1 mL of MeOH in a 2.0-mL Eppendorf tube. Then, the mixture was vortexed for 1 min and sonicated for 10 min; this step was repeated twice. After centrifuging for 20 min at 14,100×*g*, a 750-μL aliquot of the supernatant was transferred to a 1.5-mL Eppendorf tube. Then, 500 μL of chloroform and 400 μL of water were added. After that, the mixture was vortexed for 1 min, incubated for 15 min at −4 °C, and centrifuged for 20 min at 14,100×*g*. Finally, a 750-μL aliquot of aqueous fraction was transferred to a 1.5-mL Eppendorf tube, evaporated to dryness under nitrogen gas, and reconstituted with 450 μL of acetonitrile/water (1:1 *v*/*v*). For internal standard quantification, 50 μL of 50 mg/L solution of the internal standard (PIPES) was added to the extract. For each tray, two replicates were done. All of the extracts were stored at −80 °C until analyzed and were filtered through 0.2-μm nylon filters before injection (Pall Life Sciences, Port Washington, NY, USA).

HPLC-MS analysis

Chromatographic separation was performed on an Acquity UHPLC system (Waters, Milford USA), equipped with a quaternary pump, an autosampler, and a column oven. An HILIC TSK gel Amide-80 column (250×2.0 mm^2 i.d., 5 μm) with a 2.0 mm×1 cm i.d. guard column of the same material provided by Tosoh Bioscience (Tokyo, Japan) was used for analytical separation of metabolites. Elution gradient was performed using solvent A (acetonitrile) and solvent B (ammonium acetate 3 mM at pH 5.5, adjusted with acetic acid) as follows: 0–3 min, isocratic gradient at 5 % B; 3–27 min, linear gradient from 5 to 70 % B; 27–30 min, isocratic gradient at 70 % B; 30–32 min back to the initial conditions at 5 % B; and from 32 to 40 min, at 5 % B. The mobile phase flow rate was 0.15 mL/min and the injection volume was 5 μL.

The mass spectrometer was an LCT Premier XE-time-of-flight (TOF) analyzer (Waters, Milford USA) equipped with an electrospray (ESI) as ionization source in negative and positive modes. Nitrogen (purity >99.98 %) was used as cone and desolvation gas at flow rates of 50 and 600 L/h, respectively. Desolvation temperature was set to 350 °C, and electrospray voltages were set to 3.0 kV (positive mode) and to 2.2 kV (negative mode). The mass acquisition range was 90–1000 *m*/*z*.

Data analysis

Waters raw chromatographic data files (.raw format) were converted to the standard CDF format by the Databridge function of MassLynx™ v 4.1 software (Waters, USA).

These data files were then imported into the MATLAB environment (release 2014b; The Mathworks Inc., Natick, MA, USA) by using the MATLAB Bioinformatics Toolbox (version 4.3.1) and in-house built routines. Finally, every LC-MS-analyzed rice sample gave a data matrix containing the acquired retention times on the rows and the detected *m*/*z* values on the columns. In order to facilitate calculations, the total number of columns (i.e., *m*/*z* values) was reduced by using a binning approach (grouping mass values into a number of bins within a particular *m*/*z* range, in this case 0.05 amu). Every analyzed sample gave a data matrix with 1020 rows (retention time from 0 to 40 min) and 18,200 columns (from 90 to 1000 amu at 0.05 resolution). In the case of XCMS, raw chromatographic data files in CDF format were directly imported into MetaboNexus bioinformatics platform [20] without applying the binning approach.

Peak areas analysis

Two different methodologies were used and compared for the calculation of chromatographic peak areas: XCMS and MCR-ALS. In order to ascertain the effect of the treatment with the two metals, chromatographic peak areas obtained using any of these two methods were analyzed using PCA, PLS-DA, and ASCA. Before applying these chemometric methods, peak areas were autoscaled (mean-centered and scaled by their

standard deviation) to give equal weight (scale) to each one of the detected features.

XCMS XCMS approach allows an automatic processing of data for feature detection and calculation of chromatographic peak areas [6]. A typical XCMS analysis starts with the application of the *centWave* data processing algorithm which basically consists of two main steps. First, dominant mass spectra features are identified in this domain by using the so-called regions of interest (ROIs). In these identified ROIs, the presence of a chromatographic peak is denoted by a signal which at a particular *m/z* value has intensity over a particular preselected threshold value. The second step is the identification and modeling of chromatographic peaks by means of a wavelet transformation and a Gaussian shape curve fitting approach. Then, non-relevant features are dismissed by considering only those that are present in more than a certain percentage of all the samples (commonly 50 %). Finally, chromatographic peaks of the same component in different samples are aligned by means, for instance, of the *obiwarp* algorithm [21]. For more detailed information about the XCMS algorithm see the work of Smith [6] and Tautenhahn [22].

In this work, MetaboNexus bioinformatics platform [20] has been used to import and pre-process raw chromatographic data files in CDF format. MetaboNexus pre-processing platform relies on the XCMS package in R language environment and it provides a dashboard of controls to handle pre-processing in an intuitive manner with available pre-sets for different instruments. In this work, the settings were manually adjusted starting with the pre-sets corresponding to an HPLC/Q-TOF analyzer. The optimization of these parameters in our particular case was not straightforward. For this reason, the full analysis was repeated using different combinations of the parameters with variations from the default settings, and the results were compared to decide which the best parameters were. Finally, the *centWave* algorithm was employed as a feature detection method using 30 ppm as the maximal tolerated *m/z* deviation in consecutive scans, and allowing chromatographic peak widths ranging from 10 to 60 s. The number of peaks across samples of intensity higher than 1000 was fixed to 5, and the signal-to-noise threshold was set to 10. Peak integration was carried out using a Mexican hat approach considering a minimum difference in *m/z* for peaks with overlapping retention time of −0.0025. Regarding chromatographic peak alignment, the *obiwarp* algorithm [21] was selected for retention time correction. Grouping parameters were set to 5 s for the bandwidth of Gaussian smoothing kernel to apply to the peak density chromatogram whereas the width of overlapping *m/z* slices to use for creating peak density chromatograms and grouping peaks across samples was set to 0.025 amu. Finally, the minimum percentage of samples at where the same peaks need to be present in at least one sample class was set to 70 %.

The final output is a data table that contains the selected features (identified by their exact *m/z* values) in the rows, and the area of these features for each sample in the columns. Finally, sample areas were normalized by using the area of the PIPES internal standard.

Multivariate curve resolution by alternating least squares MCR-ALS is a chemometric method used for the resolution of pure contributions in unresolved mixtures [9]. MCR-ALS can be used to resolve a wide variety of datasets from different research fields, like hyphenated and multidimensional chromatographic systems, -omics data, process analysis, spectroscopic images, environmental data tables, etc., as it has been already described in the literature [23–25].

In this work, MCR-ALS has been used to resolve the elution and mass spectra profiles of the metabolites obtained in the full-scan untargeted LC-MS analysis of the rice sample extracts before and after metal treatment. MCR-ALS decomposes every individual experimental dataset arranged in a data matrix according to the following bilinear model:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

Where **D** (size *IxJ*) represents the experimental LC-MS data matrix (from a single rice sample) in which the rows are the MS spectra at all retention times (i=1,…I), and the columns are the chromatograms at all *m/z* channels (j=1,…J). According to Eq. 1, **D** matrix is decomposed into the product of two factor matrices, **C** and $\mathbf{S}^{\mathbf{T}}$, that corresponds respectively to the matrix of the resolved elution profiles, **C** (size *IxN*), and to the matrix of their corresponding mass spectra, $\mathbf{S}^{\mathbf{T}}$ (size *NxJ*). *N* represents the total number of resolved components considering during MCR-ALS analysis. **E** matrix (size *IxJ*) contains the residuals not explained by the model using the *N* considered components.

This data analysis strategy can be easily extended to the simultaneous analysis of several samples. For instance, in the case of this work, a total number of 128 samples have been simultaneously considered: 16 aerial part rice control samples, 24 aerial part rice samples treated with Cd at different concentrations (10, 50, and 1000 μM), 24 aerial part rice samples treated with Cu at different concentrations (10, 50, and 1000 μM), 16 root control rice samples, 24 root rice samples treated with Cd at different concentrations (10, 50, and 1000 μM), and 24 root rice samples treated with Cu at different concentrations (10, 50, and 1000 μM). All these samples were arranged in a single column-wise augmented data matrix ($\mathbf{D}_{\mathbf{aug}}$), containing the 128 individual data matrices ($\mathbf{D}_{\mathbf{x}}$ where x=1,…128), one for each rice sample, settled one on the top of

the other. This long column-wise augmented matrix ($\mathbf{D_{aug}}$) is also decomposed using a bilinear model such as:

$$\mathbf{D_{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_{128} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_{128} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_{128} \end{bmatrix}$$
$$= \mathbf{C_{aug}S^T} + \mathbf{E_{aug}} \quad (2)$$

Before applying MCR-ALS, every individual data matrix from every sample ($\mathbf{D_x}$) was normalized dividing by the chromatographic peak area of the internal standard (PIPES) in the considered sample. In order to accelerate and reduce memory requirements of MCR-ALS calculations, every chromatogram of each sample was divided into six separate chromatographic time windows. In this way, a final number of six column-wise augmented data matrices, each one corresponding to one of the six chromatographic time windows, were independently analyzed by MCR-ALS.

MCR-ALS solves Eq. 1 ($\mathbf{D}$, single data matrix case) or Eq. 2 ($\mathbf{D_{aug}}$, augmented data matrix case) starting with an initial guess of the number of components needed to explain sufficiently well the considered matrix by its singular value decomposition (SVD) [26]. This number should be large enough to include most of the metabolites extracted from the rice samples giving a significant MS chromatographic signal and, also, to consider possible background and solvent signal contributions. Next, an initial estimate of either $\mathbf{C}$ or $\mathbf{S^T}$ factor matrices should be provided. For instance, this estimate can be obtained from the purest MS signals in the data set gathered by a variable detection method such as SIMPLISMA [27]. Then, the estimation of $\mathbf{C}$ or $\mathbf{S^T}$ factor matrices is performed by means of an alternating least squares optimization under preselected constraints. In the case of this work, the applied constraints were only non-negativity (elution and spectra profiles) and spectral normalization (equal height) [28]. Resolved mass spectra in $\mathbf{S^T}$ are enforced to be the same for the common constituents in the different analyzed samples, whereas elution profiles resolved in $\mathbf{C_{aug}}$ matrix are allowed to be different for each one of the samples ($\mathbf{C_x}$ (x=1,…128)), as shown in Eq. 2. It is important to point out here that in the MCR-ALS approach, there is no need to correct for the unavoidable changes in the elution profiles of the same metabolite in different samples and chromatographic runs, for instance in their retention times (peak shifting) and in their profile shapes, differently to what occurs with XCMS (see above). This is a clear advantage of MCR-LS compared to XCMS, as it will be discussed later because it fits better with the natural behavior of LC-MS data. It is also important to emphasize that in this case of LC-MS data, MCR-ALS results will be not affected by uncertainties due to rotational ambiguities since MS spectral resolution is very high and also due to the large number of simultaneously analyzed individual rice sample data matrices (up to 128), which gives very robust results. Finally, from MCR-ALS results, a data table that contains the resolved components in the rows and the peak area of these components for each sample in the columns is obtained.

Discrimination and evaluation of metal effects on rice samples by chemometric analysis of metabolite concentration (peak areas) changes

Once metabolite relative concentration changes were estimated by XCMS and MCR-ALS, various methods were used to evaluate them and to discriminate samples according to metal treatment. PCA [29], PLS-DA [30], and ASCA [31] were used for this purpose and are briefly described in the ESM.

In this work, PCA has been used to explore the behavior of tissue (aerial part or root) samples when they were treated either with Cd or Cu metals. PLS-DA has been used to discriminate between root and aerial parts of rice samples, and between samples treated either by Cd or Cu (Venetian blind cross-validation was used to assess the results of the PLS-DA model). Finally, ASCA analysis (performed on a well-balanced experimental design) allowed the interpretation of the sources of the experimental variance: the tissue sample (root or aerial part) and the metal of treatment (Cd or Cu). Possible interactions between these two factors have been also evaluated. In order to assess the statistical significance of the considered factors, a permutation test can be done. In this work, the number of permutations for each model was set to 10,000.

Chemometric software

PCA, PLS-DA, and ASCA were performed by using PLS Toolbox 7.8 (Eigenvector Research Inc., Wenatche, WA, USA) working under MATLAB (The Mathworks, Natick, MA, USA). MCR-ALS was carried out using MCR-ALS toolbox freely available at www.mcrals.info. XCMS was performed using MetaboNexus interactive data analysis platform [20].

Results and discussion

Two categorical factors (tissue sample and metal of treatment) were considered. First, the statistical significance of these two factors was evaluated by means of ASCA analysis of LC-MS peak areas from MCR-ALS resolved components and from XCMS detected features. In addition, the possibility of differentiating samples according to the experimental factors was studied using an exploratory analysis of LC-MS peak areas

using PCA and PLS-DA. Table 1 shows a summary of the data matrices considered in this work.

Assessment of the effects of experimental factors

As is described below, in all the cases studied here, XCMS and MCR-ALS data analysis gave comparable results. Furthermore, results confirmed that the treatment with cadmium and copper had a significant effect on all treated samples. XCMS was applied to the entire full-scan chromatograms, and the output was a table containing 1627 features. Not all of these detected features could be considered an independent metabolite since some of them could be adducts or isotopic masses coming from the same metabolite, and others could be assigned to electric signals or other noise contributions. This table was used to build $\mathbf{D}_{\mathbf{xcms\_areas}}$ matrix (see Table 1). This number of XCMS features detected could be reduced by using additional applications such as, for instance, CAMERA [32]. This software facilitates compound annotation and identification by considering adducts and isotopic peaks which reduce the final number of features detected by XCMS. However, in this work, this was not used since this selection was performed by chemometric analysis (see below). MetaboNexus allowed a fast analysis of the data set. The entire full-scan chromatograms were resolved in approximately 10 min. However, the bottleneck of XCMS approach was the selection of the optimal pre-processing parameters that required testing several combinations of the feature detection and peak alignment parameters, which increase the total time of analysis considerably.

MCR-ALS was used to analyze the six column-wise augmented data matrices described in the methods section. Figure 1 shows a TIC chromatogram of an aerial part sample with the selected chromatographic regions highlighted. Between 25 and 30 MCR-ALS components were resolved for each column-wise augmented data matrix with a minimum explained variance of 98 % (an example of an SVD used to select the number of components considered in the resolution of each window is shown in ESM Fig. S1). This selection considered a sufficiently large number of components to include all significant metabolite contributions and other interfering MS signal contributions (instrumental background, solvent, etc.) that could appear among the resolved components. A total number of 165 MCR-ALS components (resolved peaks) were needed to explain sufficiently the data variance. In this case, 29 from the 165 resolved components were associated with noisy signals, like background and solvent contributions, and hence, they were ignored in subsequent analyses. Finally, 136 components were used to build the $\mathbf{D}_{\mathbf{mcr\_areas}}$ matrix (see Table 1).

Figure 2 shows the final results for one of these 136 components resolved in the simultaneous analysis of 32 aerial part rice samples, 24 treated with copper at three concentrations (10, 50, and 1000 μM), and 8 control samples (not treated with copper). This example shows a component resolved in the second chromatographic region, from 5.1 to 7.1 min. A total of 30 components were resolved in this particular interval. Elution profile of this resolved component in the different aerial part samples is depicted in Fig. 2a, where a clear decrease is observed in the height and area of the resolved chromatographic peaks in this profile between control and Cu-treated samples. However, no clear distinction is observed between the heights or areas of these peaks for the different levels of metal, probably reflecting that at the lowest concentration of the metal tested in this study, the profile of this particular metabolite is already significantly affected, without any further change of it at higher Cu concentration. Figure 2b shows the corresponding resolved mass spectrum for this component. A high intense mass signal was found at *m*/*z* 515.10 with two lower intensity signals (see inset) at *m*/*z* 516.10 and 517.10, which are isotopic contributions. Some other small signals at different *m*/*z* values are present also in this spectrum (see other insets). They may be possible adducts of the main component, or also they can be very minor

Table 1 Design of experimental data matrices

Matrix	Size[a]	Description	Uses
$\mathbf{D}_{\mathbf{mcr\_areas}}$	128×136	Full data set	Evaluation of the effect of tissue sample, of metal of treatment and of their interaction by ASCA
$\mathbf{D}_{\mathbf{xcms\_area}}$	128×1627		
$\mathbf{Cd}_{\mathbf{mcr\_areas}}$	64×136	Roots and aerial part samples treated with Cd	Evaluation of the effect of tissue sample by PCA and PLS-DA
$\mathbf{Cd}_{\mathbf{xcms\_areas}}$	64×1627		
$\mathbf{Cu}_{\mathbf{mcr\_areas}}$	64×136	Roots and aerial part samples treated with Cu	
$\mathbf{Cu}_{\mathbf{xcms\_areas}}$	64×1627		
$\mathbf{L}_{\mathbf{mcr\_areas}}$	48×136	Aerial part samples treated with Cd or Cu	Evaluation of the effect of metal of treatment by PCA and PLS-DA
$\mathbf{L}_{\mathbf{xcms\_areas}}$	48×1627		
$\mathbf{R}_{\mathbf{mcr\_areas}}$	48×136	Roots samples treated with Cd or Cu	
$\mathbf{R}_{\mathbf{xcms\_areas}}$	48×1627		

Subscripts indicate the kind of data: *-mcr_areas* is the data matrix containing areas of the MCR-ALS resolved components; *-xcms_areas* is the data matrix containing peak areas obtained by XCMS

[a] Size: number of samples×136 MCR-ALS resolved components or 1627 XCMS detected features

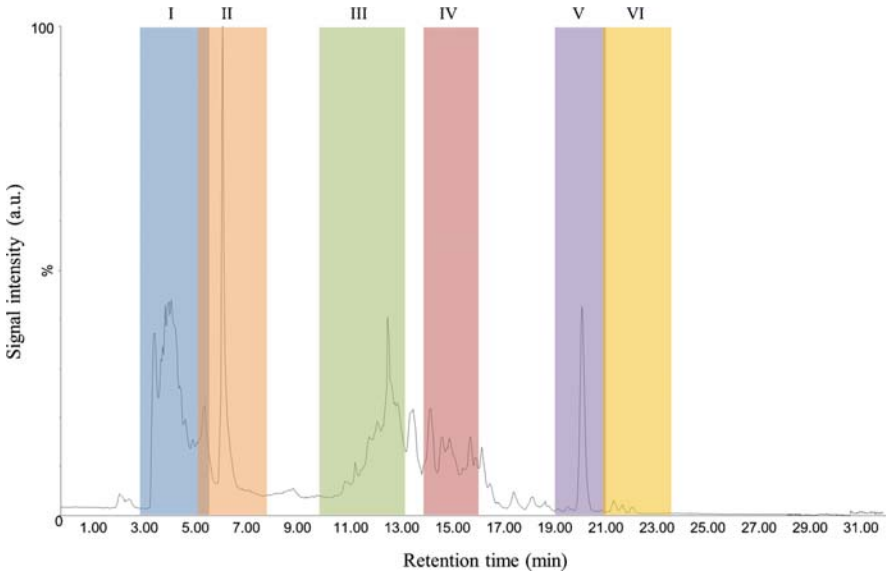


Fig. 1 Example of a TIC chromatogram of an aerial part sample with the selected important chromatographic regions highlighted

strongly coeluted metabolites, not resolved by MCR-ALS in an independent component due to their very low variance contribution during the ALS optimization. The resolution time for MCR-ALS was about 15 min for each chromatographic window. Although the full MCR-ALS approach workflow (mass spectral binning, selection of chromatographic time windows, resolution and selection of relevant components) required a larger amount of time. However, most of these steps did not require user participation, and so, automation and parallel computation could reduce the total time of analysis significantly.

Chromatographic information provided by XCMS and MCR-ALS strategies was rather similar. As mentioned before, XCMS algorithm detects features in all the cases that at a particular *m*/*z* value the signal is higher than a given threshold. Some of these features can be assigned to artifacts (instrumental noise, background and solvent contributions) and in other cases a single metabolite can give multiple features due to the

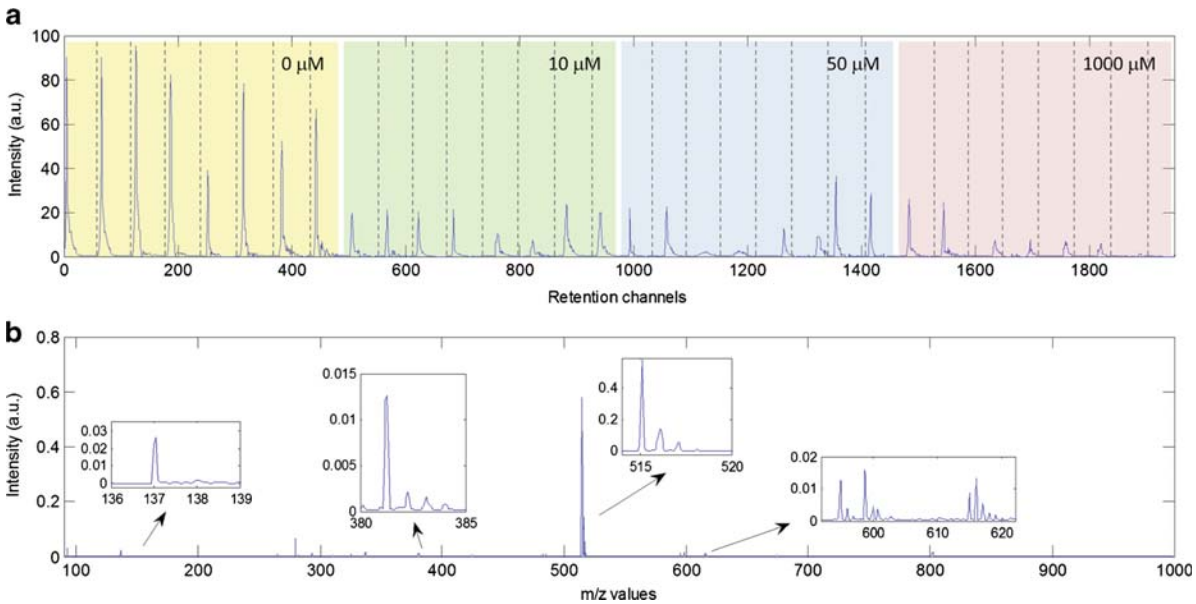


Fig. 2 Example of a resolved MCR component: aerial part rice samples treated with copper. **a** Resolved elution profiles, **b** resolved mass spectra. *Insets*: zoom in different *m*/*z* regions. *Dashed lines* separate each individual sample

detection of isotopic peaks or adducts. These two facts reduced the total number of independent features related to metabolites. The opposite case was found when considering the MCR-ALS resolved components. In this case, the final number of resolved components and, subsequently, the size of the $\mathbf{D}_{\mathbf{mcr\_areas}}$ matrix were significantly lower than the size of the $\mathbf{D}_{\mathbf{xcms\_areas}}$ matrix. This fact is caused because MCR-ALS resolves the components according to their elution profiles and their corresponding mass spectra. Every MCR-ALS resolved component will be associated with an elution profile and a mass spectrum which, however, may include various features at different *m*/*z* values. In this case, these features could be easily assigned to isotopic peaks of the nominal *m*/*z* values or different adducts of the same metabolite. Therefore, a single MCR-ALS resolved component provides information on several features related to the same metabolite grouped together. A more difficult situation could be found when two metabolites present a very strongly overlapped elution profile. If the number of components resolved by MCR-ALS resolution is too low, these two metabolites could appear as being resolved in the same component and, therefore, the resolved MS spectrum will have the features of the two compounds. Then, the interpretation of the resolved mass spectrum can be more difficult. Nevertheless, this information is still provided by MCR-ALS despite the smaller dimension of the $\mathbf{D}_{\mathbf{mcr\_areas}}$ matrix compared to $\mathbf{D}_{\mathbf{xcms\_areas}}$. It is possible that some information retrieved by the XCMS approach can be left in the residuals of the MCR-ALS model and lost for further analysis. This could be the case of rather small signals that contributes to a small quantity of the total variance of the data. Nevertheless, it is important to perform a deep study on those samples to differentiate if these little signals were due to the samples (and their treatment) or the previously described artifacts (background, solvent contributions).

Finally, when the use of the two approaches is compared, it may be argued that the XCMS workflow, as implemented in MetaboNexus, is more straightforward and that the final results can be reached faster. However, as stated above, the selection of the optimal pre-processing parameters in XCMS is a very critical aspect of the obtained results, and they should be properly optimized in order to have meaningful results. On the contrary, the MCR-ALS approach is more robust, although not so easy to use for the non-experienced user and, in principle, more time consuming. This robustness is demonstrated by the fact that satisfactory results can also be obtained when dealing with lower resolution data. For instance, MCR-ALS has successfully overcome inherent difficulties of GC-MS and CE-MS data analysis, such as multiple MS signals for the same metabolite due to derivatization in GC-MS and large migration time shifts between samples in CE-MS [33, 34].

As a result, $\mathbf{D}_{\mathbf{mcr\_areas}}$ and $\mathbf{D}_{\mathbf{xcms\_areas}}$ peak area data matrices were studied using an ASCA two-factor model with interaction considering the two factors, the sample tissue (root or aerial part) and the metal of treatment (Cd or Cu). Balanced experimental design considered 32 samples for each combination of the factors: root-Cd, root-Cu, aerial-Cd, and aerial-Cu. Results showed that both factors had a statistically significant effect (p=0.0001). Interaction between the kind of metal and the tissue analyzed was significant (p<0.050) in both cases (XCMS and MCR-ALS data). These results would indicate that the effects of each one of these two factors (type of metal and rice tissue) were dependent on the levels of the other.

Discrimination of rice samples due to metal treatment

Once peak areas were obtained with both XCMS and MCR-ALS, discrimination of samples according to the effects produced by the two metals (Cu and Cd) was investigated. With this aim, PCA and PLS-DA were applied to LC-MS peak areas of the different features or components obtained with both methodologies, XCMS and MCR-ALS, respectively. In both cases, results were analogous.

Tissue sample effect

Samples treated with Cd and samples treated with Cu were first studied separately (matrices $\mathbf{Cd}_{\mathbf{mcr\_areas}}$, $\mathbf{Cu}_{\mathbf{mcr\_areas}}$, $\mathbf{Cd}_{\mathbf{xcms\_areas}}$, $\mathbf{Cu}_{\mathbf{xcms\_areas}}$ described in Table 1**)**. For both treatments, root samples were discriminated from aerial part samples by both PCA and PLS-DA. On the other hand, samples were also distinguished according to the metal treatment (Cd or Cu) when ASCA scores plot was considered (32 samples for each factor). Table 2 summarizes the figures of merit of the PLS-DA models.

As example of these results, Fig. 3a shows the scores plot for Cu treatment when PLS-DA was applied to MCR-ALS resolved peak areas, where root samples were discriminated from aerial part of rice samples by LV1, explaining 36 % of X-data variance. Figure 3b depicts the same results for Cd treatment, where root samples were discriminated from aerial part samples by LV1, explaining 45 % of X-data variance. As a conclusion, for both metal treatments, roots were clearly distinguished from aerial parts of rice samples with a Mathew's correlation coefficient (MCC) equal to 1.0, either for XCMS and for MCR-ALS results (analogous XCMS scores plots were obtained and shown in the ESM).

Type of metal effect

In order to evaluate more specifically the effect of metal treatment, root and aerial parts of rice samples were evaluated separately (matrices $\mathbf{L}_{\mathbf{mcr\_areas}}$, $\mathbf{R}_{\mathbf{mcr\_areas}}$, $\mathbf{L}_{\mathbf{xcms\_areas}}$, $\mathbf{R}_{\mathbf{xcms\_areas}}$ described in Table 1**)**. For both tissues, samples treated with Cd were separated from samples treated with Cu by both PCA and PLS-DA. Samples treated with different metals were also discriminated when ASCA scores plots were considered.

Table 2 Figures of merit of the PLS-DA models

Matrix	Number of latent variables	X-variance explained	Y-variance explained	R^2 CV	MCC
$\mathbf{Cd}_{\mathbf{mcr\_areas}}$	2	50.23	81.19	0.54	1.000
$\mathbf{Cu}_{\mathbf{mcr\_areas}}$	2	51.70	94.61	0.86	1.000
$\mathbf{Cd}_{\mathbf{xcms\_areas}}$	2	37.23	94.76	0.85	1.000
$\mathbf{Cu}_{\mathbf{xcms\_areas}}$	2	50.22	96.82	0.95	1.000
$\mathbf{L}_{\mathbf{mcr\_areas}}$	2	41.57	91.59	0.66	1.000
$\mathbf{R}_{\mathbf{mcr\_areas}}$	2	49.67	77.61	0.59	1.000
$\mathbf{L}_{\mathbf{xcms\_areas}}$	2	40.47	96.77	0.94	1.000
$\mathbf{R}_{\mathbf{xcms\_areas}}$	2	49.19	91.60	0.82	1.000

As an example, Fig. 4a gives the scores plot for aerial parts of rice samples when PLS-DA was applied to resolved MCR-ALS peak areas, where samples treated with Cd were discriminated from samples treated with Cu by LV1, explaining 32 % of X-data variance. The scores plot for root samples is shown in Fig. 4b, where samples treated with Cd were distinguished from samples treated with Cu by LV1, explaining 44 % of X-data variance. For both tissues, samples exposed to Cd were clearly separated from samples exposed to Cu with a Mathew's correlation coefficient (MCC) equal to 1.0 (analogous XCMS results and scores plots were obtained, see ESM). Quality parameters of the obtained PLS-DA results are reported in Table 2. It can be seen that MCC values are close to one in all cases, although slightly better R^2 were obtained for XCMS data.

Identification of possible metabolite biomarkers of the effects of metals (Cu and Cd) on rice tissues

This study is concluded by the identification of some of the most relevant *m*/*z* values associated with the peak areas obtained using both strategies (XCMS and MCR-ALS). Variable importance on projection (VIP) scores (see ESM for the mathematical definition) were used to detect the most important features in each case.

Metabolomic interpretation of the changes caused by the different factors requires the identification of those variables (metabolites) having a VIP score higher than the average (greater than 1 rule), using for instance high-resolution MS combined with tandem MS. In this study and for brevity, only the six metabolites with the highest VIP scores for each PLS-DA model were considered. In order to select these six metabolites, the *m*/*z* values associated with the peak areas with the highest VIP scores were used. For instance, Fig. 5 gives the PLS-DA VIP scores plot for Cd-treated samples from peak areas obtained by MCR-ALS and XCMS ($\mathbf{Cd}_{\mathbf{mcr\_areas}}$ and $\mathbf{Cd}_{\mathbf{xcms\_areas}}$ matrices described in Table 1), showing the most relevant features (MCR-ALS resolved peaks or XCMS features, respectively) for discriminating between roots and aerial part rice samples. Results confirmed again that most of the *m*/*z* values obtained from XCMS peak areas were the same than those retrieved from the MCR-ALS resolved components.

The same procedure was also applied to the samples treated with Cu ($\mathbf{Cu}_{\mathbf{mcr\_areas}}$ and $\mathbf{Cu}_{\mathbf{xcms\_areas}}$ matrices) and it was also applied to the matrices containing only root samples or

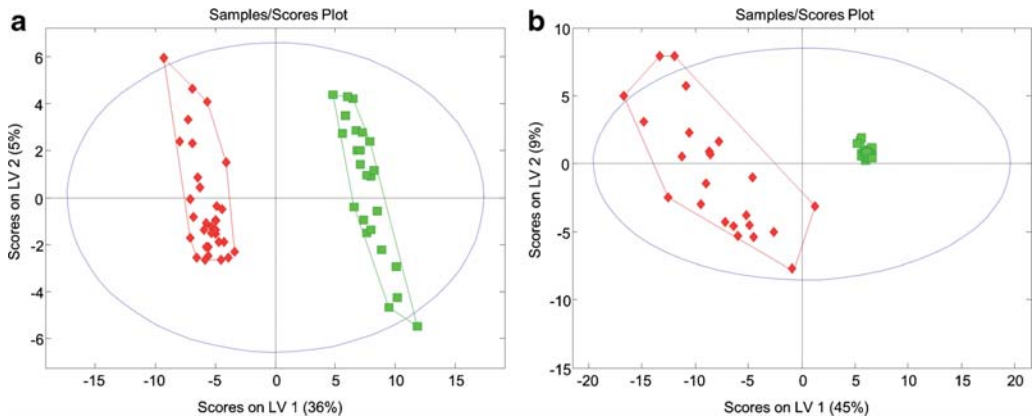


Fig. 3 PLS-DA results for tissue sample factor and MCR-ALS data. *Red diamonds* (♦) are aerial part samples and *green squares* (■) are root samples. **a** Scores plot for Cu treatment. **b** Scores plot for Cd treatment

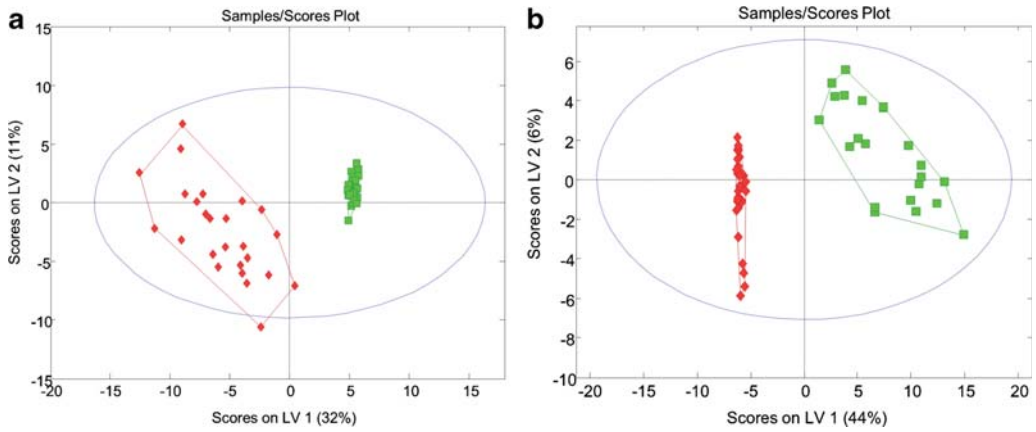


Fig. 4 PLS-DA results for metal of treatment factor and MCR-ALS data. *Red diamonds* (♦) are Cd samples and *green squares* (■) are Cu samples. **a** Scores plot for aerial part samples. **b** Scores plot for root samples

only aerial part samples ($\mathbf{L}_{\mathbf{mcr\_areas}}$, $\mathbf{R}_{\mathbf{mcr\_areas}}$, $\mathbf{L}_{\mathbf{xcms\_areas}}$, and $\mathbf{R}_{\mathbf{xcms\_areas}}$ matrices). Most of the relevant *m*/*z* values obtained using XCMS approach were almost the same than those obtained by MCR-ALS. In some cases, MCR-ALS resolved components gave more than one mass signal, for instance, peak number 100 in Fig. 5a with 337.09 and 199.80 *m*/*z* values. As mentioned before, these additional mass signals usually correspond to isotopic peaks and/or adducts of the same metabolite, but it could also correspond to two metabolites with extremely highly overlapped elution profiles (embedded peaks) which had been resolved in the same MCR-ALS component.

At this point, some characteristics of each method regarding metabolite identification can be discussed. MCR-ALS has the advantage of resolving the mass spectrum corresponding to every elution profile simultaneously. In this mass spectra, several features can be detected such as isotopic peaks, adducts eluted at the same retention time, and some solvent/background contributions. It is also possible that these different features can come from various metabolites eluting at the same retention times. This can be considered a drawback of the MCR-ALS since it makes their identification more laborious and time consuming. However, we think that these steps can be handled reasonably well, and new tools to do this in a more general and accurate way need to be developed. In the case of XCMS, the excessive number of detected features can be reduced by using appropriate complementary software (such as the previously mentioned CAMERA). However, in our experience, the detection of metabolites with similar behavior is not so straightforward such as in MCR-ALS where different metabolites in the same resolved component can be easily identified by their different MS signals.

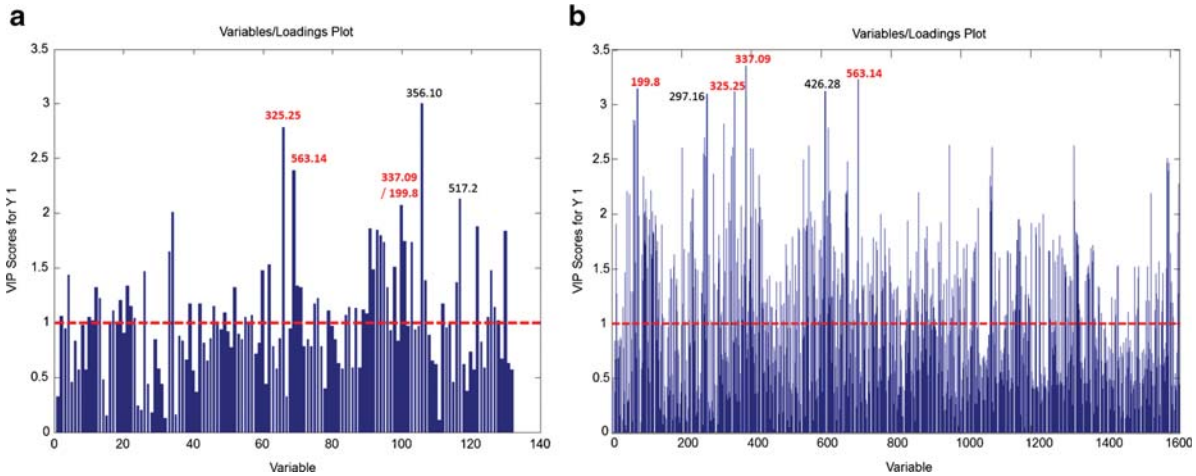


Fig. 5 PLS-DA VIP scores for samples treated with Cd. **a** Results for peak areas obtained with MCR-ALS. **b** Results for peak areas obtained with XCMS. *Numbers* indicate the *m*/*z* values associated with the peak areas with a higher VIP score. *Red* numbers indicate those that are equal for MCR-ALS and XCMS

Despite the main goal of this work was not the detailed biological interpretation of the observed metabolite changes, the tentative identification of some metabolites was attempted to confirm the reliability of the compared methodologies. In order to do this, *m*/*z* values of resolved peaks were compared with *m*/*z* values obtained from public databases such as MassBank [35], Metlin [36], and HMDB [37]. Since MS spectra of the components resolved by MCR-ALS were obtained from binned data, their accuracy is low (0.05 amu) and they cannot be used for a good identification of the metabolites. Therefore, once one particular component has been also identified by their precise elution profile, the retention time of its peak maximum in a given sample can be easily estimated and its accurate mass spectrum recovered from the original raw data file using MassLynx™ v 4.1 software. On the contrary, when XCMS was used, no binning approach was necessary. Therefore, the *m*/*z* accurate value of the feature detected was directly used for biomarker identification. Results of these tentative identifications are shown in Table 3.

From the results shown in Table 3, a preliminary biochemical interpretation can be attempted. For example, most of the metabolites shown in Table 3 are glycosides, such as 4-methylumbelliferyl glucoside, or other metabolites related to their biosynthesis, such as 6‴-*O*-sinapoylsaponarin or stearyl citrate [38]. Glycosides are very abundant in plants, and they have been reported to be altered by metal exposure in *Arabidopsis thaliana* [1, 39]. Glycosides have been also reported to form strong metal complexes, which would explain why their concentration changes so much by metal treatment [40, 41].

Conclusions

Both XCMS and MCR-ALS appeared to be powerful methodologies for metabolomic studies. These two approaches gave very similar results for all the experimental metabolomics systems investigated in the present work. XCMS resolved the entire full-scan chromatogram very fast, allowing a direct calculation of chromatographic peak areas for every *m*/*z* value giving a statistically significant MS signal, resulting in a huge number of features. In contrast, MCR-ALS led to a lower number of features since isotopic peaks and metabolites with very similar chromatographic profiles may be described in the same resolved elution profile with a MS spectrum including more than one *m*/*z* value. A comparison between the two approaches showed that XCMS was more straightforward for non-experienced users and required a smaller amount of processing time. However, MCR-ALS was more robust since it allowed working with more complex data (due to large time shifts, background contributions, and lower mass resolution) and did not require the optimization of parameters for each LC-MS instrument.

Chemometric evaluation of the peak areas for the candidate metabolites obtained by XCMS and MCR-ALS approaches allowed the statistical assessment of the effects caused by the metal exposure. ASCA statistical evaluation showed that there was a significant interaction between analyzed tissue sample and the type of metal of treatment. Further exploration by PCA and PLS-DA showed that metal exposure allowed the discrimination of rice samples considering both studied factors (type of metal and type of tissue). The identification of principal molecular biomarkers related to metal treatment

Table 3 Tentative identification of molecular markers using relevant *m*/*z* values considering the VIP scores

Highest mass ion	Retention time (min)	Proposed metabolite	Adduct	Adduct mass	Error *m*/*z* (ppm)	Factor
199.0392	26.87	Camalexin	M−H	199.0408	8.2	Part of the plant
325.2400	6.33	Avocadyne 4-acetate	M−H	325.2384	4.8	
337.0895[a]	6.33	4-Methylumbelliferyl glucoside	M−H	337.0929	10.1	
563.1365[a]	20.21	Apigenin 7-*O*-[beta-D-apiosyl-(1→2)-beta-D-glucoside]	M−H	563.1406	7.2	
799.1999[a]	20.42	6‴-*O*-Sinapoylsaponarin	M−H	799.2091	11.5	
593.2798	16.75	7,8-Dihydrovomifoliol 9-[rhamnosyl-(1→6)-glucoside]	M+Hac-H	593.2815	2.7	
279.0528[a]	27.52	7-Hydroxy-2-methyl-4-oxo-4H-1-benzopyran-5-carboxylic acid	M+Hac-H	279.0510	6.4	Metal of Treatment
329.0898[a]	6.1	1-*O*-Vanilloyl-beta-D-glucose	M−H	329.0878	6.1	
330.09[a]	19.25	Koenimbine	M+K-2H	330.0902	5.6	
331.0554	19.51	*S*-7-Methylthioheptylhydroximoyl-L-cysteine	M+K-2H	331.0558	1.2	
481.2549	6.17	Stearyl citrate	M+K-2H	481.2573	4.9	
446.1357[a]	28.6	(*R*)-2-Hydroxy-7,8-dimethoxy-2H-1,4-benzoxazin-3(4H)-one 2-glucoside	M+Hac-H	446.1304	11.8	

[a] Metabolites relevant also in the ASCA analysis

appears to be an important issue in plant biology studies to assess their metabolic changes. A preliminary evaluation of what metabolites were affected by metal exposure highlighted that glycoside family of compounds were much affected. Further work is pursued to perform a more exhaustive identification of metabolites by using other MS technologies, such as higher resolution MS instruments and MS/MS approaches. The combination of these advanced MS technologies with chemometric procedures will facilitate a more complete identification of molecular markers related to rice exposure to metals and the improvement of knowledge about metabolic changes.

Acknowledgments The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 320737. Also, recognition from the Catalan government (grant 2014 SGR 1106) is acknowledged. JJ acknowledges a CSIC JAE-Doc contract cofounded by the FSE, and AGR thanks CONICET for a fellowship.

Conflict of interest The authors declare that they have no competing interests.

References

1. Fukusaki E, Kobayashi A (2005) Plant metabolomics: potential for practical operation. J Biosci Bioeng 100(4):347–354
2. Xiao JF, Zhou B, Ressom HW (2012) Metabolite identification and quantitation in LC-MS/MS-based metabolomics. TrAC Trends Anal Chem 32:1–14
3. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13(1):11–29
4. Lommen A (2009) Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem 81(8):3079–3086
5. Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395
6. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78(3):779–787
7. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G (2015) Bioinformatics: the next frontier of metabolomics. Anal Chem 87(1):147–156
8. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS online: A web-based platform to process untargeted metabolomic data. Anal Chem 84(11):5035–5039
9. Jaumot J, de Juan A, Tauler R (2015) MCR-ALS GUI 2.0: new features and applications. Chemometr Intell Lab 140:1–12
10. Farrés M, Piña B, Tauler R (2014) Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS. Metabolomics 11:210–224
11. Cramer GR, Urano K, Delrot S, Pezzotti M, Shinozaki K (2011) Effects of abiotic stress on plants: a systems biology perspective. BMC Plant Biol 11:163
12. D'Alessandro A, Taamalli M, Gevi F, Timperio AM, Zolla L, Ghnaya T (2013) Cadmium stress responses in Brassica juncea: hints from proteomics and metabolomics. J Proteome Res 12(11): 4979–4997
13. Villiers F, Ducruix C, Hugouvieux V, Jarno N, Ezan E, Garin J, Junot C, Bourguignon J (2011) Investigating the plant response to cadmium exposure by proteomic and metabolomic approaches. Proteomics 11(9):1650–1663
14. Järup L (2003) Hazards of heavy metal contamination. Brit Med Bull 68(1):167–182
15. Ahsan N, Nakamura T, Komatsu S (2012) Differential responses of microsomal proteins and metabolites in two contrasting cadmium (Cd)-accumulating soybean cultivars under Cd stress. Amino Acids 42(1):317–327
16. Antti H, Ebbels TMD, Keun HC, Bollard ME, Beckonert O, Lindon JC, Nicholson JK, Holmes E (2004) Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. Chemometr Intell Lab 73(1 SPEC. ISS): 139–149
17. Johnson HE, Lloyd AJ, Mur LAJ, Smith AR, Causton DR (2007) The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. Metabolomics 3(4):517–530
18. Aina R, Labra M, Fumagalli P, Vannini C, Marsoni M, Cucchi U, Bracale M, Sgorbati S, Citterio S (2007) Thiol-peptide level and proteomic changes in response to cadmium toxicity in Oryza sativa L. roots. Environ Exp Bot 59(3):381–392
19. Roth U, Von Roepenack-Lahaye E, Clemens S (2006) Proteome changes in Arabidopsis thaliana roots upon exposure to Cd 2+. J Exp Bot 57(15):4003–4013
20. Huang SM, Toh W, Benke PI, Tan CS, Ong CN (2014). MetaboNexus: an interactive platform for integrated metabolomics analysis. Metabolomics 10:1084–1093
21. Prince JT, Marcotte EM (2006) Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem 78(17):6140–6152
22. Tautenhahn R, Bottcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 9:504
23. De Juan A, Jaumot J, Tauler R (2014) Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Anal Methods 6(14):4964–4976
24. de Juan A, Tauler R (2007) Factor analysis of hyphenated chromatographic data. Exploration, resolution and quantification of multicomponent systems. J Chromatogr A 1158(1–2):184–195
25. Ruckebusch C, Blanchet L (2013) Multivariate curve resolution: a review of advanced and tailored applications and challenges. Anal Chim Acta 765:28–36
26. Golub GH, Loan CFV (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore
27. Windig W, Guilment J (1991) Interactive self-modeling mixture analysis. Anal Chem 63(14):1425–1432
28. Tauler R, Maeder M, de Juan A (2010) Multiset data analysis: extended multivariate curve resolution. In: Brown SD, Tauler R, Walczak, B (ed) Comprehensive Chemometrics, vol 2. Elsevier B.V., Amsterdam, pp 473–505
29. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab 2(1–3):37–52
30. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17(3):166–173
31. Jansen JJ, Hoefsloot HCJ, Van Der Greef J, Timmerman ME, Westerhuis JA, Smilde AK (2005) ASCA: analysis of multivariate data obtained from an experimental design. J Chemom 19(9):469–481

32. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 84(1):283–289
33. Ortiz-Villanueva E, Jaumot J, Benavente F, Piña B, Sanz-Nebot V, Tauler R (2015) Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling. Electrophoresis 36:2324–2335
34. Schmidtke LM, Blackman JW, Clark AC, Grant-Preece P (2013) Wine metabolomics: objective measures of sensory properties of semillon from GC-MS profiles. J Agric Food Chem 61(49):11957–11967
35. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45(7):703–714
36. Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G (2012) An accelerated workflow for untargeted metabolomics using the METLIN database. Nat Biotechnol 30(9):826–828
37. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, de Souza A, Zuniga A, Dawe M, Xiong Y, Clive D, Greiner R, Nazyrova A, Shaykhutdinov R, Li L, Vogel HJ, Forsythei I (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37(Suppl 1):D603–D610
38. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(D1):D109–D114
39. Sun X, Zhang J, Zhang H, Ni Y, Zhang Q, Chen J, Guan Y (2010) The responses of Arabidopsis thaliana to cadmium exposure explored via metabolite profiling. Chemosphere 78(7):840–845
40. Satterfield M, Brodbelt JS (2001) Structural characterization of flavonoid glycosides by collisionally activated dissociation of metal complexes. J Am Soc Mass Spectrom 12(5):537–549
41. Gyurcsik B, Nagy L (2000) Carbohydrates as ligands: coordination equilibria and structure of the metal complexes. Coordin Chem Rev 203(1):81–149

Informació Suplementària a la Publicació 2

Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies.

M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler.

Analytical and Bioanalytical Chemistry 407 (2015), 8835-8847.

Description of chemometric methods: PCA, PLS-DA and ASCA

PCA is probably the more often used multivariate data analysis method. It compresses the information of the original variables into a smaller number of uncorrelated variables known as principal components. These principal components are built as linear combinations of the original variables and they retain most of the valuable information about the experimental data variance without overlapping information among them due to the application of orthogonality constraints [1].

PLS-DA [2] is a multivariate regression method oriented to discriminate among different groups of samples. In this method, peak areas of every sample (**X**, predictor variables) were correlated with the vector describing the metal treatment or the rice tissue type class membership (**y**, predicted variable). In case of multiclass (more than two) discrimination, a dummy matrix of zeros and ones (**Y**, predicted variable) is built up [3]. In addition to the class membership discrimination of the samples, PLS-DA provides information about which are the most important variables for achieving this discrimination. One of the common ways to do this is, for instance, using the variable importance on projection (VIP) scores [4]. VIP scores are a weighted sum of squared PLS variable weights which measure the importance of each predictor variable by giving a score value for each variable and rank them according to their significance in the projection used by the PLS model [4]. In this way, the higher the VIP score of a particular variable is the more importance of this variable for the sample discrimination. VIP scores for a certain variable, *j*, are defined as:

$$VIP_j = \sqrt{m \sum_{k=1}^{p} b_k^2 w_{jk}^2 \Big/ \sum_{k=1}^{p} b_k^2} \qquad \text{Equation 1}$$

Where m corresponds to the total number of variables, p is the number of latent variables, w_{jk} is the j-th element of vector w_k and b_k is the regression weight for the k-th latent variable. The average of squared VIP score is equal to 1, therefore the 'greater than one' rule is commonly used as variable selection criteria: variables with a VIP score greater than 1 can be considered important for a given model [5].

The quality of PLS-DA models can be assessed by using the Mathews correlation coefficient (MCC), which measures the quality of binary classifications of belonging or not to a particular class MCC can vary from -1 to 1, with values closer to 1 indicating better predictions and values closer to -1 showing worse predictions [6].

Finally, ASCA is a multivariate analysis of variance method that combines the power of ANOVA to separate variance sources with the advantages of simultaneous component analysis (SCA) to the modelling of the individual separate effect matrices. SCA is a generalization of PCA for the situation where the same variables have been measured in multiple conditions [7]. ASCA is especially useful for the analysis of complex multivariate datasets containing an underlying experimental design, and it allows for a natural interpretation of the variance induced by the different factors in the design [8].

In ASCA, ANOVA is firstly performed on the raw data matrix, which is decomposed into the sum of different data matrices characterising the variance caused by each one of the considered factors, plus a residual matrix containing the unexplained variance. Then, SCA is applied to each ANOVA factor matrix individually [9,8]. For a more detailed description of ASCA procedure see the work of Smilde [8] and Jansen [9]. In order to check the statistical significance of the effects of the investigated factors and their interactions, a permutation test can be performed in which the null hypothesis (H_0) assumes that there is no effect of the considered factor. More details regarding the

statistical assessment of ASCA results by using a permutation test can be found at the work of Vis *et. al.* [10].

References

1. Wold S, Esbensen K, Geladi P (1987). Chemometrics Intell Lab Syst 2 (1-3):37-52.
2. Barker M, Rayens W (2003). J Chemometr 17 (3):166-173.
3. Geladi P, Kowalski BR (1986). Anal Chim Acta 185 (C):1-17.
4. Wold S, Johansson A, Cocchi M (1993) PLS-partial least squares projections to latent structures. In: Kubiny H (ed) 3D QSAR in Drug Design, vol Theory Methods and Applications. ESCOM Science Publishers, Leiden, pp 583-618.
5. Chong IG, Jun CH (2005). Chemometrics Intell Lab Syst 78 (1):103-112.
6. Matthews BW (1975). BBA - Prot Struct 405 (2):442-451.
7. Jansen JJ, Hoefsloot HCJ, Van Der Greef J, Timmerman ME, Smilde AK (2005). Anal Chim Acta 530 (2):173-183.
8. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME (2005). Bioinformatics 21 (13):3043-3048.
9. Jansen JJ, Hoefsloot HCJ, Van Der Greef J, Timmerman ME, Westerhuis JA, Smilde AK (2005). J Chemometr 19 (9):469-481.
10. Vis DJ, Westerhuis JA, Smilde AK, van der Greef J (2007). BMC Bioinformatics 8:322.

Table S1. Temperature, relative humidity and light long-day conditions at the growth chamber.

TIME (HOUR)	TEMPERATURE (ºC)	LIGHT [µmol·m^{-2}·s^{-1}]	RELATIVE HUMIDITY (%)
06:00-07:00	22	25	67
07:00-08:00	23	40	69
08:00-09:00	24	60	71
09:00-10:00	26	160	73
10:00-16:00	28	275	75
16:00-17:00	28	160	73
17:00-18:00	27	60	71
18:00-19:00	26	40	69
19:00-20:00	24	25	67
20:00-06:00	22	0	65

Figure S1: SVD plot showing the results of the analysis of the third window (number of components considered for the MCR-ALS analysis was set to 27).

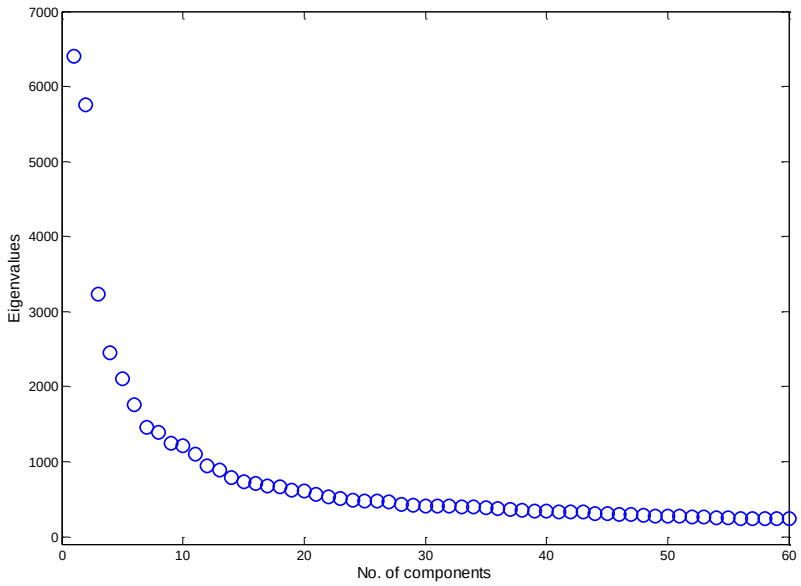


Figure S2: PLS-DA results for tissue sample factor and XCMS data. Red diamonds (♦) are aerial part samples and green squares (■) are root samples. A) Scores plot for Cu treatment. B) Scores plot for Cd treatment.

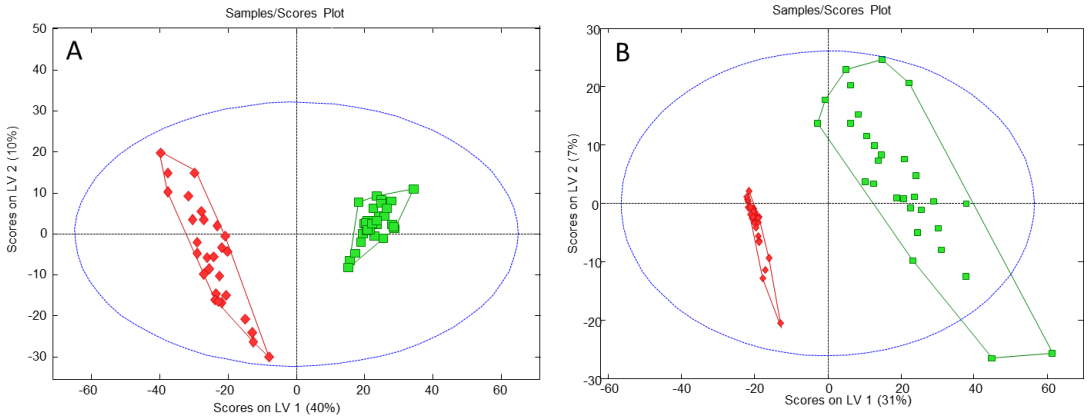
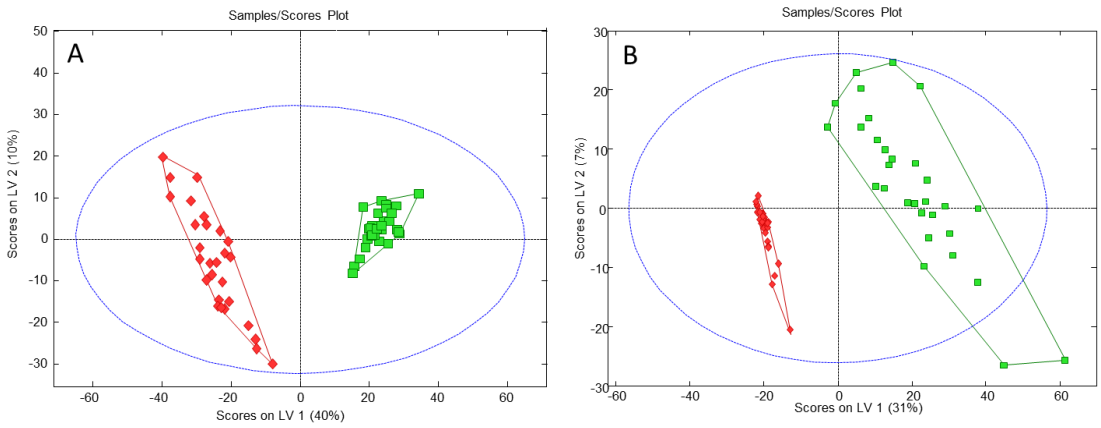


Figure S3: PLS-DA results for metal of treatment factor and XCMS data. Red diamonds (♦) are Cd samples and green squares (■) are Cu samples. A) Scores plot for aerial part samples. B) Scores plot for root samples.



3.4. Discussió conjunta dels resultats

A continuació es presenten els resultats obtinguts en el treball inclòs en aquest capítol.

3.4.1. Fases estacionàries HILIC en anàlisis metabolòmiques

En la primera publicació presentada es van avaluar tres factors experimentals que influeixen en l'anàlisi de metabòlits mitjançant LC-MS: tipus de fase estacionària HILIC, pH de la fase mòbil i força iònica de la fase mòbil. Amb aquest objectiu es va analitzar una mescla de 54 metabòlits de diverses famílies (nucleòsids, aminoàcids, sucres, àcids orgànics i altres, veure Taula 1 en la publicació 1) en 24 condicions cromatogràfiques diferents: quatre tipus de fase estacionària (amida, BEH amida, zwitteriònica i mode mixt diol), tres condicions de pH (àcid, moderadament àcid i neutre) i dos nivells de força iònica (baix i alt). La Figura 3.2 mostra un resum del disseny experimental utilitzat en aquest estudi. En un treball previ realitzat en el grup de recerca en el que s'ha dut a terme aquesta tesi, es va demostrar que l'acetonitril és el millor solvent orgànic en les anàlisis de metabolòmica mitjançant HILIC [20]. Per aquest motiu, en aquest treball, no es va considerar el tipus de solvent orgànic utilitzat en la fase mòbil com a factor d'estudi.

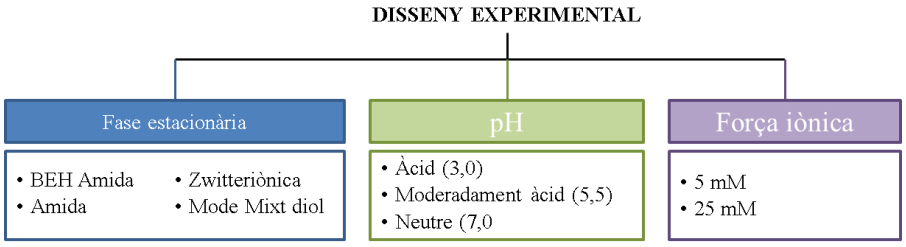


Figura 3.2. Disseny experimental utilitzat en l'estudi de l'efecte de diferents factors experimentals en la separació de metabòlits en HILIC.

Els factors de retenció dels 54 metabòlits obtinguts experimentalment en les diferents condicions cromatogràfiques estudiades es van agrupar en una matriu de dades **D**. L'aplicació de PCA i ASCA a aquesta matriu va permetre determinar que els factors més influents en l'anàlisi de metabòlits són el tipus de fase estacionària i el pH (nivell de significació *p* igual o menor a 0,0001 utilitzant 10000 permutacions). En canvi, les mostres analitzades a diferents forces iòniques no es distingien per cap dels dos primers components principals i tampoc en ASCA aquest factor mostrava un efecte significatiu (nivell de significació *p* igual a 0,2 utilitzant 1000 permutacions). Concretant en el tipus de fase estacionària a utilitzar, l'aplicació de PCA i ASCA va permetre avaluar el comportament de les 4 columnes HILIC utilitzades. En els gràfics d'*scores* dels dos models (Figures 2A i 2B de la publicació 1),

es va observar que la columna HILIC de mode mixt diol era la que es comportava de forma més diferent de la resta i que la columna HILIC zwitteriònica tenia un comportament entremig de les dues columnes HILIC amida. Aquesta diferenciació s'aprecia en la Figura 3.3, en la qual es representen les diferents anàlisis cromatogràfiques sobre el PC2.

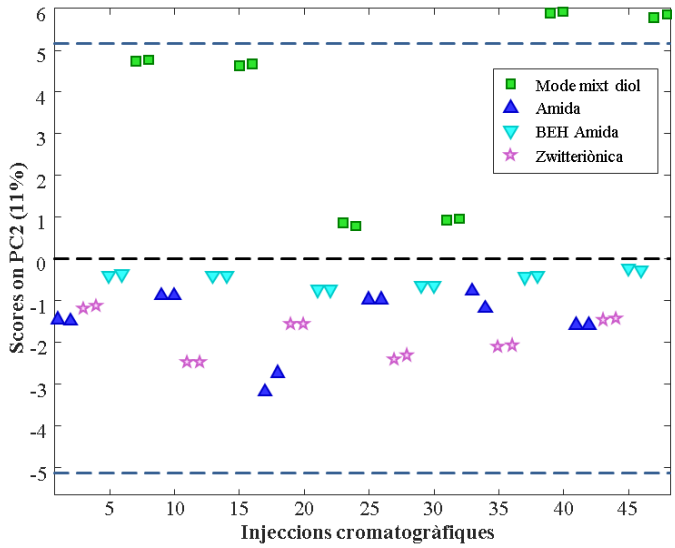


Figura 3.3. Gràfic d'*scores* del PC2 amb les injeccions cromatogràfiques classificades segons el tipus de fase estacionària utilitzada. Les 12 mostres de cada tipus de fase estacionària corresponen a les injeccions cromatogràfiques realitzades per duplicat en tres condicions de pH (3,0, 5,5 i 7,0) i dues de força iònica (5 mM i 25 mM).

Aquests resultats confirmen l'estudi anterior realitzat en l'anàlisi de 12 metabòlits per LC-DAD [20], en el qual s'estudiaven cinc fases estacionàries HILIC diferents (amida, BEH amida, amina, zwitteriònica i mode mixt diol) en diverses condicions experimentals (tipus de solvent orgànic, pH i força de ionització de la fase mòbil). A més, les tendències observades en aquest treball coincidien amb les observades en estudis similars de la bibliografia [21-27]. El factor experimental més important en estudis de metabolòmica per HILIC-MS és el tipus de fase estacionària utilitzat. Es confirma també que la fase estacionària de mode mixt diol és la que es comporta d'un mode més diferent a la resta. Aquesta diferència en el comportament de la fase estacionària de mode mixt diol s'explica per les propietats químiques duals de la seva superfície, que conté una cadena hidrofòbica amb un grup diol. En canvi, les superfícies de la resta de columnes estudiades només presenten propietats hidrofíliques, donades pels grups amida i sulfobetaïna. Això indica que en una anàlisi de metabòlits polars és més recomanable utilitzar les fases estacionàries amida o zwitteriònica. En canvi, si s'analitzen simultàniament metabòlits polars i apolars, és més recomanable considerar la fase estacionària de mode mixt diol.

Per acabar aquest estudi d'avaluació dels efectes del pH i el tipus de fase estacionària sobre la separació cromatogràfica dels metabòlits, es van construir diversos models quimiomètrics de PLS, relacionant el factor de retenció de cada metabòlit amb els seus descriptors moleculars, MDs [16]. Els MDs caracteritzen l'estructura dels metabòlits i es van calcular a partir de les seves representacions tipus

SMILES [28] i utilitzant el software PCLIENT (accessible a la web http://www.vcclab.org/). El resultat més important d'aquests models de PLS va ser la selecció dels MDs més descriptius per cada condició cromatogràfica (veure Taula 2 de la publicació 1), els quals permetien conèixer quines són les principals propietats fisicoquímiques involucrades en el mecanisme de retenció en les columnes tipus HILIC. Per exemple, destaca que els descriptors més influents en les quatre columnes van ser els geomètrics i els topològics, els quals descriuen la mida, la forma i la simetria de les molècules. A més, la gran majoria d'aquests descriptors estaven corregits segons la polaritat. Això explica que el mecanisme de retenció prioritari en les quatre columnes és la partició hidrofílica, la qual cosa és lògica perquè tres d'elles són neutres (amides i mode mixt diol) i l'altra és zwitteriònica. Tal com es mostra en la Taula 3.1, cal destacar que en la fase estacionària zwitteriònica els descriptors de connectivitat presentaven més influència que en les altres tres, la qual cosa indica que les interaccions electrostàtiques tenen major influència en el mecanisme de retenció de la columna ZIC que en la de la resta de columnes estudiades. Finalment, la comparació dels MDs més importants per cada condició cromatogràfica estudiada està d'acord amb els resultats obtinguts a partir dels models PCA i ASCA: la columna de mode mixt diol tenia un comportament clarament diferenciat del de la resta de columnes i la columna zwitteriònica presentava un comportament entremig de les dues amides.

Taula 3.1. MDs de connectivitat trobats per cada fase estacionària.

BEH Amida	Amida	Zwitteriònica	Mode Mixt Diol
BELe3	BELe2	BELe2	GGI2
VRp2	BEHp1	BELe3	
VRv2	BELp2	VRp2	
		VRv1	
		VRv2	

BEH (*highest eigenvalue of Burden matrix*), BEL (*lowest eigenvalue of Burden matrix*), GGI (*topological charge index*), VR (*Randic-type eigenvector-based index*).

3.4.2. Resolució de metabòlits per mètodes quimiomètrics

Existeixen diferents estratègies de tractament de dades que permeten realitzar la compressió i el processament dels grans conjunts de dades obtinguts en els estudis metabolòmics. En la segona publicació d'aquest capítol es van comparar els resultats obtinguts amb els mètodes XCMS i MCR-ALS, en el tractament de les dades corresponents a un estudi no dirigit dels efectes del Cd i del Cu en el metabolisme de l'arròs.

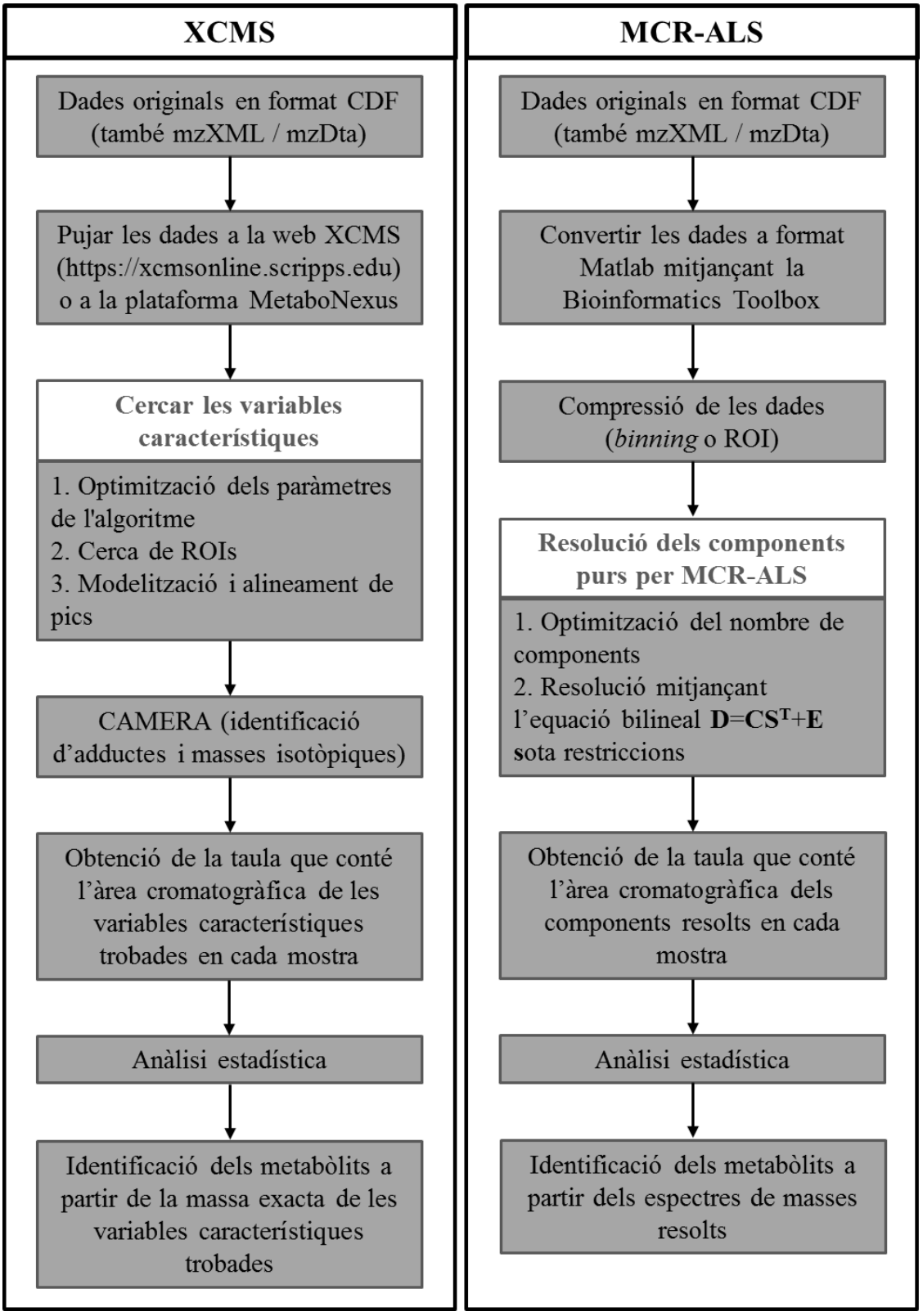


Figura 3.4. Procediment de treball de les estratègies de tractament de dades basades en XCMS i en MCR-ALS.

En l'esquema de la Figura 3.4 es resumeixen les etapes dels procediments de treball de les dues metodologies estudiades. La principal diferència entre les dues estratègies es troba en la concepció i utilització de les variables anomenades característiques (*features*) i de la corresponent detecció i resolució

dels pics cromatogràfics. D'una banda, el XCMS identifica les variables característiques (*features*) a partir d'uns valors determinats de *m/z* i de temps de retenció. Aquesta identificació es porta a terme mitjançant dues etapes principals. La primera busca i selecciona els senyals de *m/z* més importants utilitzant l'estratègia de la cerca de ROIs. La segona etapa, consisteix en modelitzar els senyals detectats donant-los forma de pic gaussià i en el seu alineament entre diferents cromatogrames. En canvi, en la metodologia basada en MCR-ALS es resolen els constituents de la mostra a partir dels seus perfils d'elució i espectres de masses. La resolució dels primers no requereix cap tipus de modelització ni d'alineament cromatogràfic. Els espectres de masses poden estar constituïts per múltiples senyals *m/z* (*features*) d'intensitat variable. Abans de processar les dades mitjançant MCR-ALS és convenient realitzar-ne la seva compressió. En la publicació 2 aquesta compressió es va dur a terme amb el procediment de *binning* però, si no es vol perdre resolució espectral, també es pot realitzar mitjançant la cerca de ROIs, tal com es descriu en el treball de Gorrochategui [29], fent servir el procediment anomenat ROIMCR [29].

Per a un usuari no expert, el XCMS pot semblar més fàcil d'utilitzar, ja que està més automatitzat i, a més, el seu temps de computació és relativament curt. Per aquest motiu ha esdevingut la metodologia de tractament de dades estàndard en estudis de metabolòmica no dirigida mitjançant espectrometria de masses. Es troba disponible tant per executar en l'entorn R com en format en línia a internet (https://xcmsonline.scripps.edu), en el qual els resultats obtinguts es poden processar directament per mètodes quimiomètrics d'exploració, classificació i discriminació, com el PCA o el PLS-DA. A més, també està connectat amb la base de dades pública METLIN [14], la qual cosa resulta molt útil a l'hora de fer una primera identificació temptativa dels metabòlits. No obstant, aquesta estratègia presenta alguns inconvenients. Per exemple, cal optimitzar els paràmetres per a cada tipus d'instrument (error de massa, amplada de pic, relació senyal/soroll, diferència de massa entre pics solapats, tipus d'alineament de pics i desviació permesa en el temps de retenció), la qual cosa implica un exercici de prova i error que pot requerir molt més temps. En canvi, en l'estratègia basada en MCR-ALS moltes d'aquestes etapes no són necessàries, és més flexible i proporciona una sèrie d'avantatges, sobretot un cop es té experiència en el seu funcionament i possibilitats, tal com s'indica a continuació a partir dels resultats obtinguts en la seva aplicació per un cas particular.

En la publicació 2, el nombre de variables característiques detectades amb XCMS (1627) va ser molt superior al nombre de components resolts per MCR-ALS (165). Això es deu a que l'espectre de masses d'un sol metabòlit conté més d'un senyal de *m/z* (adductes i masses isotòpiques). És a dir, cada

component de MCR-ALS resolt s'associa a més d'una variable característica (*feature*) identificada per XCMS. El nombre elevat de variables característiques detectades es pot reduir mitjançant l'aplicació de programes complementaris com el CAMERA, que identifica els pics isotòpics i els adductes [30], però afegeix una etapa addicional a l'anàlisi. El fet de que en MCR-ALS no sigui necessari aquest pas representa un avantatge d'aquesta estratègia respecte el XCMS. Cal tenir en compte, no obstant, que l'aplicació apropiada de la metodologia basada en MCR-ALS haurà de considerar un nombre de components MCR-ALS prou elevat que permeti explicar de forma suficient la variància de les dades.

També cal comentar que no tots els components resolts ni les variables característiques poden associar-se directament a metabòlits, ja que algunes representen contribucions del soroll instrumental o del solvent. En aquest sentit cal remarcar que el MCR-ALS permet avaluar de manera més ràpida i fàcil quins dels components resolts representen metabòlits i quines contribucions són soroll de fons. Aquesta diferenciació es realitza fàcilment a partir de l'observació dels perfils d'elució i espectres corresponents a aquestes contribucions. Quan aquestes no tenen sentit des d'un punt de vista físic, poden ja descartar-se directament.

Finalment, cal esmentar que l'avantatge més important de l'aproximació basada en MCR-ALS en comparació al procediment basat en XCMS és que no requereix una etapa de modelització ni alineament dels pics cromatogràfics. D'una banda, la modelització dels pics a través de la utilització de procediments de transformada d'ondetes [31] i d'ajust amb corbes de tipus gaussià que es realitza en XCMS complica considerablement el processament de les dades i, a més, pot emmascarar la seva naturalesa original. Totes les variables característiques seleccionades es modelen en forma de pic gaussià i això pot portar a una interpretació errònia dels senyals detectats. Per exemple, senyals procedents del soroll de fons poden interpretar-se com a possibles metabòlits. Per altra banda, quan s'analitzen diverses mostres i es comparen els seus cromatogrames, el mètode XCMS realitza una etapa addicional d'alineament dels pics cromatogràfics, que pot produir també errors i dificultats en la seva anàlisi. En canvi, el mètode MCR-ALS descriu directament els perfils d'elució i espectres de masses originals per a cadascun dels components seleccionats associats als constituents de la mostra. Així, el procediment basat en MCR-ALS permet obviar els problemes de deriva en els temps de retenció entre les diferents mostres analitzades cromatogràficament, ja que els perfils d'elució resolts per un mateix component en les diferents mostres no són forçats a aparèixer als mateixos temps d'elució ni a tenir la mateixa forma. Aquesta robustesa de la metodologia basada en MCR-ALS s'ha demostrat en molts treballs anteriors, ja que permet obtenir

resultats satisfactoris quan s'utilitza pel tractament de dades amb molta menys resolució, tant cromatogràfiques [10, 29] com espectrofotomètriques, en l'estudi de reaccions i processos químics de naturalesa molt diversa [32-34]. Així per exemple, en el cas de la cromatografia el procediment MCR-ALS facilita la resolució de les dificultats associades a l'anàlisi de dades de GC-MS, que contenen múltiples senyals de *m/z* per a un mateix metabòlit a causa de la seva derivatització [35]. També en el cas de l'anàlisi de dades d'electroforesi capil·lar acoblada a espectrometria de masses (CE-MS), on es presenten desviacions grans en els temps de migració, el procediment MCR-ALS pot emprar-se sense dificultats [36]. A mode de resum, en la Taula 3.2 es mostren els principals avantatges i inconvenients de les dues metodologies de tractament de dades estudiades.

Taula 3.2. Avantatges i inconvenients de les dues metodologies de tractament de dades estudiades.

	Avantatges	**Inconvenients**
XCMS	Automatitzat. Temps de computació curt. Apte per usuaris no experts. Disponible en format web. Connectat a METLIN.	Optimització dels paràmetres per a cada tipus d'instrument. Requereix l'ús de programaris complementaris per identificar els senyals de massa que pertanyen al mateix metabòlit. Modelització dels pics. Alineament de pics.
MCR-ALS	Robust. Resolució dels espectres de masses complets de cada metabòlit. No requereix modelització ni alineament de pics.	Difícil d'utilitzar per usuaris no experts. Temps de computació més llarg. Optimització del nombre de components.

L'avaluació mitjançant PCA, ASCA i PLS-DA de les matrius de dades obtingudes a partir de les dues metodologies ($\mathbf{D}_{\mathbf{mcr\_areas}}$ i $\mathbf{D}_{\mathbf{xcms\_areas}}$, veure Taula 1 de la publicació 2) va portar als mateixos resultats en tots els casos. La distinció de les diferents classes de mostres observada en els models de PCA va ser molt semblant; la significança dels factors obtinguda per ASCA va ser la mateixa; les mostres controls i les tractades es van poder diferenciar per PLS-DA amb un coeficient de correlació de Matthews (MCC, descrit a la introducció de la present Tesi, secció 2.3.4, [37]) igual a 1 en tots els casos i les masses seleccionades com a rellevants van ser gairebé totes iguals. Aquests resultats indiquen que les dues metodologies permeten extreure la informació més important dels conjunts de dades de metabolòmica no dirigida.

En resum, els resultats obtinguts en la publicació 2 van demostrar que les dues estratègies estudiades són útils per a solucionar el repte del tractament de dades de metabolòmica no dirigida.

3.4.3.Identificació de metabòlits a partir de les seves propietats cromatogràfiques

El tercer aspecte avaluat en aquest capítol és l'aplicació de mètodes quimiomètrics per ajudar en la identificació dels metabòlits a partir de les seves propietats cromatogràfiques (factors de retenció). Així, es proposa l'ús de models de QSRR que relacionin l'estructura i les propietats fisicoquímiques dels metabòlits amb el seu factor de retenció, seguint l'exemple d'estudis previs de la bibliografia que treballen tant amb tot el metaboloma com amb una família de metabòlits concreta [12, 13, 17-19, 38-43]. Amb aquest objectiu, es va realitzar un estudi preliminar utilitzant models de PLS com els descrits en la publicació 1. Els models de QSRR es van construir relacionant els descriptors moleculars, MDs (obtinguts mitjançant el software PCLIENT), dels 54 metabòlits amb els factors de retenció obtinguts en les 12 condicions cromatogràfiques avaluades amb models de PLS en la publicació 1 (quatre tipus de fase estacionària i tres condicions de pH). En aquesta ocasió, abans de construir els models de PLS es va utilitzar un algoritme genètic (GA) per preseleccionar aquells MDs que proporcionen millors prediccions pels 12 sistemes cromatogràfics considerats. El GA és una tècnica de selecció de variables que identifica el subconjunt de variables més útils per a construir un model de regressió acurat i precís [44, 45]. Per exemple, en aquest treball el GA es va utilitzar per seleccionar els MDs que permetien fer la millor predicció dels factors de retenció dels metabòlits. La capacitat predictiva de cada MD es va avaluar per l'error en la validació creuada (RMSECV) dels models generats, que es defineix com:

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{k_i-\hat{k}_i}{k_i}\right)^2}{n}} \times 100(\%) \qquad \text{Equació 3.1}$$

On $\hat{k}_i$ és el valor del factor de retenció de cadascun dels 54 metabòlits predit per la validació creuada i k_i, n'és el valor original.

A més, es va realitzar una validació externa per avaluar la qualitat dels models de GA-PLS obtinguts i així garantir la capacitat predictiva dels models de QSRR creats. Aquesta validació externa es va dur a terme utilitzant com a conjunt de validació externa (*test set*) els factors de retenció experimentals de 14 metabòlits juntament amb els seus MDs corresponents. En la Taula 3.3 s'indiquen els metabòlits presents en les mostres (mescles) emprades per a la validació externa. Cal destacar que les dades de validació externa no es van utilitzar durant la selecció de MDs per GA ni en la construcció dels models de PLS. A més, les dades del conjunt d'entrenament (mescla de 54 metabòlits descrita en la Taula 1 de la publicació 1) es van enregistrar amb HPLC-TQD, mentre que les del conjunt de validació externa es van obtenir

mitjançant HPLC-TOF. El fet de que els dos conjunts de dades s'enregistressin amb diferents instruments va fer que la validació externa fos més robusta.

Taula 3.3. Metabòlits presents en la mescla de validació externa.

Aminoàcids	Sucres	Àcids orgànics	Altres
fenilalanina	adonitol	àcid màlic	monofosfat d'adenosina (AMP)
tirosina	Inositol	àcid úric	dopamina
metionina sulfona	glucosa 6-fosfat	àcid itacònic	uracil
	fructosa 1,6-bisfosfat	àcid piroglutàmic	

La qualitat dels models de predicció finals es va avaluar mitjançant l'error de predicció (RMSEP) definit com:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{k_i-\hat{k}_i}{k_i}\right)^2}{n}} \times 100(\%) \qquad \text{Equació 3.2}$$

On k_i, i $\hat{k}_i$ són els valors original i predit del factor de retenció de cadascun dels metabòlits de les dades de validació externa.

La Taula 3.4 resumeix els resultats obtinguts en els models de QSRR. Aquesta taula mostra el nombre de MDs seleccionats pel GA, que va variar entre 197 o 288. També conté els errors de RMSECV i RMSEP expressats en percentatge. D'una banda, els valors de RMSECV van variar entre 6% i 18% per les fases estacionàries BEH amida, amida i zwitteriònica. En canvi, per la fase estacionària de mode mixt diol aquest error va ser més elevat, entre 22% i 26%. D'altra banda, els valors de RMSEP es trobaven entre 7% i 23% per les dues amides i la zwitteriònica, mentre que la fase estacionària de mode mixt diol va presentar uns errors bastants més elevats, entre 31% i 42%. Aquests resultats indiquen que els models predictius obtinguts de RMSECV i RMSEP per les fases estacionàries BEH amida, amida i zwitteriònica tenen una magnitud similar i uns valors acceptables. Per tant, aquests models es podrien utilitzar per a predir els factors de retenció dels metabòlits desconeguts. En canvi, els models obtinguts per la fase estacionària de mode mixt diol mostraven una predicció pitjor. Un cop més, aquesta fase estacionària va mostrar un comportament diferent respecte les altres 3 fases estacionàries. Aquesta diferència en el comportament de la fase estacionària de mode mixt diol s'explica per les propietats químiques duals de la seva superfície, que conté una cadena hidrofòbica amb un grup diol. Aquestes propietats duals permeten l'ús d'aquesta fase estacionària tant en mode de RP com en HILIC. No obstant, els resultats exposats en aquest capítol indiquen que les propietats cromatogràfiques d'aquest tipus de columna són difícils de predir a partir de la seva modelització mitjançant models de QSRR.

Taula 3.4. Resultats dels models de QSRR generats.

Sistema cromatogràfic	Nº de MDs seleccionats pel GA	RMSECV	RMSEP
BEH amida àcid	223	18%	23%
BEH amida moderadament àcid	210	7%	7%
BEH amida neutre	202	6%	13%
Amida àcid	199	12%	20%
Amida moderadament àcid	224	14%	23%
Amida neutre	216	11%	11%
Zwitteriònica àcid	202	13%	23%
Zwitteriònica moderadament àcid	197	13%	19%
Zwitteriònica neutre	206	18%	18%
Mode mixt diol àcid	244	26%	40%
Mode mixt diol moderadament àcid	288	22%	42%
Mode mixt diol neutre	209	26%	31%

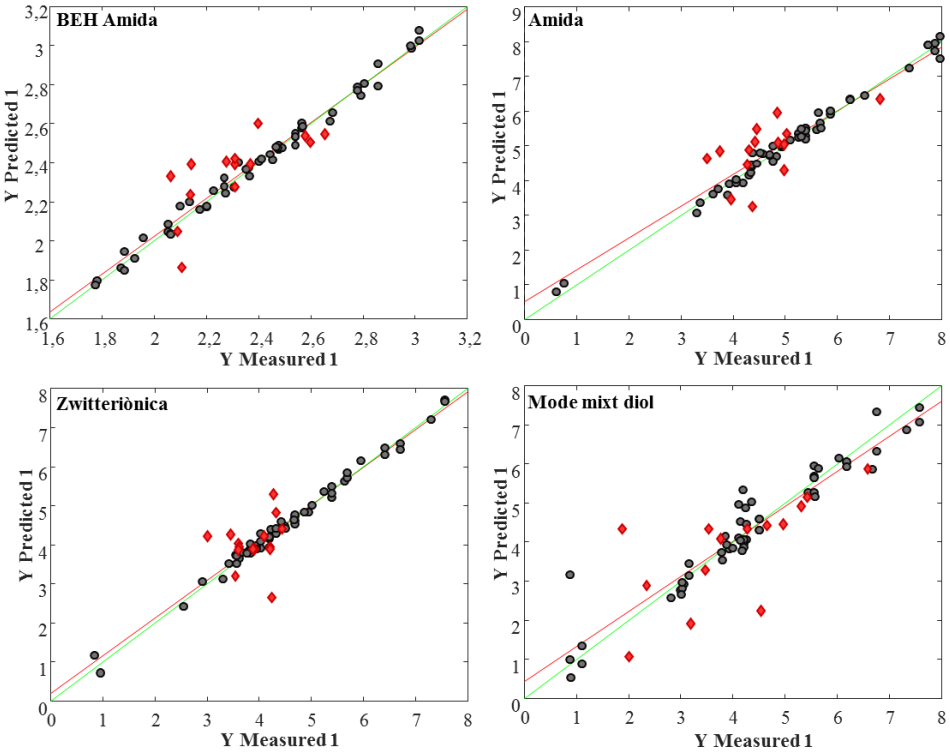


Figura 3.5. Gràfics de correlació entre els factors de retenció mesurats i predits per les 4 fases estacionàries a pH moderadament àcid. Els cercles grisos (●) corresponen als metabòlits de les dades d'entrenament i els diamants vermells (♦) als metabòlits de les dades de validació externa.

La Figura 3.5 mostra els gràfics de correlació entre els factors de retenció mesurats i predits per les quatre fases estacionàries a pH moderadament àcid. En aquesta figura s'aprecia que els factors de retenció dels metabòlits dels conjunts de validació externa i entrenament es trobaven força ben estimats al llarg de tot el rang de factors de retenció. A més, també es pot observar com en el cas de la fase estacionària de

mode mixt diol, la predicció dels metabòlits a partir de les dades de validació externa era bastant pitjor que per la resta de fases estacionàries.

Tenint en compte els resultats de la Taula 3.4 i la Figura 3.5, és clar que els models obtinguts es podrien millorar. Una possible millora seria la repetició dels models amb un nombre més gran de metabòlits en les mescles d'entrenament i validació externa, tal com es demostra en el treball de F. Aicheler [42], en el qual es comparen els models obtinguts variant el nombre de compostos inclosos en les dues mescles. Una altra solució podria ser realitzar l'optimització de la selecció de MDs mitjançant el GA, tal com es proposa en el treball de M. Taraji [17]. En el treball mencionat, l'optimització aplicada va permetre reduir l'error RMSEP un 30%. Aquesta optimització consisteix en repetir 10 vegades el procediment de GA i seleccionar només aquells MDs que apareixen en cada iteració. Una última proposta per a millorar els models obtinguts és utilitzar mètodes estadístics més potents per a seleccionar els metabòlits presents en la mescla de validació externa de manera que siguin el més representatius possibles dels metabòlits de la mescla d'entrament [46-48].

Els resultats obtinguts suggereixen que els models de QSRR poden proporcionar informació útil en estudis de metabolòmica no dirigida, ja que poden ajudar a identificar les propietats cromatogràfiques de metabòlits desconeguts i la seva comparació amb possibles candidats. L'aplicació d'un model de QSRR en les condicions cromatogràfiques utilitzades en un determinat estudi, pot reduir el nombre de metabòlits candidats a ser assignats a un determinat valor de *m/z* i, en alguns casos, identificar de manera unívoca el compost. Per tal de mostrar la potencial utilitat dels models de QSRR, aquests es van aplicar a dos exemples d'identificació de metabòlits de la publicació 2.

En la separació cromatogràfica de la publicació 2 es van utilitzar la mateixa fase estacionària amida i el mateix gradient d'elució que en la publicació 1, treballant a pH moderadament àcid i força iònica baixa. De tal manera que el model QSRR creat per la fase estacionària amida a pH moderadament àcid resultava útil per a predir els factors de retenció dels metabòlits a assignar en la publicació 2.

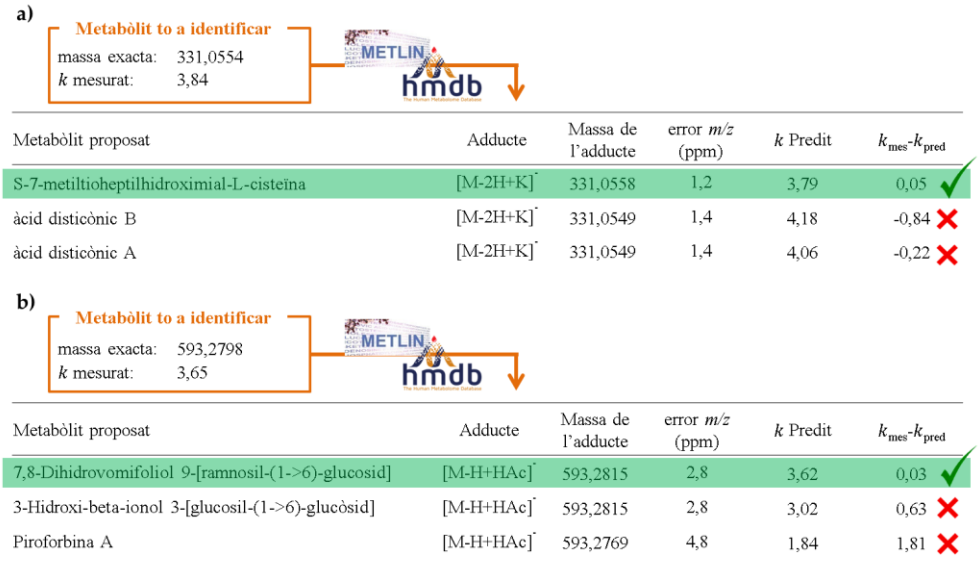
En la Taula 3 de la publicació 2 es mostra la identificació temptativa d'alguns biomarcadors potencials dels efectes del Cd i del Cu en el metabolisme de l'arròs. Aquesta identificació temptativa es va fer mitjançant la comparació de la massa exacta mesurada dels metabòlits amb valors de massa teòrics inclosos en bases de dades públiques (METLIN [14], HMDB [15]). Aquestes bases de dades públiques solen suggerir més d'un metabòlit que coincideix amb el valor de massa mesurat. En aquest punt, el model predictiu de QSRR pot ajudar a fer estimacions dels seus temps de retenció cromatogràfica i

comparar-los amb els obtinguts experimentalment. D'aquesta manera es pot reduir significativament el nombre de metabòlits potencials.

La Figura 3.6 mostra els valors de massa exacta mesurats, els factors de retenció i els metabòlits candidats pels dos exemples. Els factors de retenció de tots els candidats es van predir utilitzant els seus MDs i el model de QSRR creat. En el cas de la Figura 3.6A, el metabòlit que calia identificar tenia un factor de retenció de 3,84 i un valor de massa exacta de 331,2798. Les bases de dades públiques (METLIN [14] i HMDB [15]) van incloure tres metabòlits que podrien associar-se a la massa mesurada amb un error de massa tolerable (inferior a 5 ppm). Tanmateix, el factor de retenció predit del primer candidat (S-7-metiltioheptilhidroximial-L-cisteïna) era molt més semblant al factor de retenció mesurat que el factor de retenció predit per als altres dos candidats (àcid disticònic B i àcid disticònic A). Per tant, el model QSRR va ajudar a identificar el metabòlit com S-7-metiltioheptilhidroximial-L-cisteïna.

En la Figura 3.6B, el metabòlit que calia identificar tenia un factor de retenció de 3,65 i un valor de massa mesurat de 593,2798. De nou, segons les bases de dades públiques, aquesta massa es podria associar a tres candidats amb un error inferior a 5 ppm. No obstant això, el factor de retenció predit del primer candidat (7,8-Dihidrovomifoliol 9-[ramnosil-(1->6)-glucòsid]) era gairebé idèntic al factor de retenció mesurat. En canvi, els factors de retenció predits dels altres dos candidats (3-Hidroxi-beta-ionol 3-[glucosil-(1->6)-glucòsid] i Piroforbina A) van ser significativament diferents al factor mesurat. Per tant, el metabòlit desconegut va ser identificat com 7,8-Dihidrovomifoliol 9-[ramnosil-(1->6)-glucòsid]. En aquest segon cas, cal destacar que els dos primers metabòlits candidats eren isòmers, i per tant, la seva identificació a partir de la seva massa exacta no era possible. Els resultats obtinguts mostren la possible utilitat dels models QSRR en l'etapa d'identificació de metabòlits, ja que permet l'associació dels metabòlits d'interès amb un dels candidats proposats per les bases de dades públiques.

Malgrat tot, la completa identificació analítica dels metabòlits sempre requereix l´ús addicional de patrons de referència i/o de l'obtenció dels seus espectres de MS/MS. En aquest punt, els models de QSRR faciliten la confirmació de la identificació dels metabòlits, ja que redueixen el nombre de estàndards a comprar i comparar.



Metabòlit proposat	Adducte	Massa de l'adducte	error *m/z* (ppm)	*k* Predit	k_{mes}-k_{pred}
S-7-metiltioheptilhidroximial-L-cisteïna	[M-2H+K]⁻	331,0558	1,2	3,79	0,05 ✓
àcid disticònic B	[M-2H+K]⁻	331,0549	1,4	4,18	-0,84 ✗
àcid disticònic A	[M-2H+K]⁻	331,0549	1,4	4,06	-0,22 ✗

Metabòlit proposat	Adducte	Massa de l'adducte	error *m/z* (ppm)	*k* Predit	k_{mes}-k_{pred}
7,8-Dihidrovomifoliol 9-[ramnosil-(1->6)-glucosid]	[M-H+HAc]⁻	593,2815	2,8	3,62	0,03 ✓
3-Hidroxi-beta-ionol 3-[glucosil-(1->6)-glucòsid]	[M-H+HAc]⁻	593,2815	2,8	3,02	0,63 ✗
Piroforbina A	[M-H+HAc]⁻	593,2769	4,8	1,84	1,81 ✗

Figura 3.6. Identificació de metabòlits desconeguts mitjançant el factor de retenció predit pel model de QSRR. La lletra *k* representa factor de retenció.

3.5. Referències

1. Allwood, J. W.;Goodacre, R., An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses, *Phytochemical analysis : PCA*, 2010, **21**, 33-47.
2. Zhou, B.;Xiao, J. F.;Tuli, L.;Ressom, H. W., LC-MS-based metabolomics, *Molecular bioSystems*, 2012, **8**, 470-481.
3. Jurowski, K.;Kochan, K.;Walczak, J.;Barańska, M.;Piekoszewski, W.;Buszewski, B., Analytical Techniques in Lipidomics: State of the Art, *Critical Reviews in Analytical Chemistry*, 2017, **47**, 418-437.
4. Buszewski, B.;Noga, S., Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique, *Analytical and Bioanalytical Chemistry*, 2012, **402**, 231-247.
5. Jandera, P., Stationary and mobile phases in hydrophilic interaction chromatography: a review, *Analytica Chimica Acta*, 2011, **692**, 1-25.
6. Bouhifd, M.;Hartung, T.;Hogberg, H. T.;Kleensang, A.;Zhao, L., Review: Toxicometabolomics, *Journal of Applied Toxicology*, 2013, **33**, 1365-1383.
7. Patti, G. J.;Yanes, O.;Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy, *Nature Reviews Moecularl Cell Biology*, 2012, **13**, 263-269.
8. Smith, C. A.;Want, E. J.;O'Maille, G.;Abagyan, R.;Siuzdak, G., XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, *Analytical Chemistry*, 2006, **78**, 779-787.
9. Tautenhahn, R.;Patti, G. J.;Rinehart, D.;Siuzdak, G., XCMS online: A web-based platform to process untargeted metabolomic data, *Analytical Chemistry*, 2012, **84**, 5035-5039.
10. Farrés, M.;Piña, B.;Tauler, R., Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS, *Metabolomics*, 2014, **11**, 210-224.
11. Jaumot, J.;de Juan, A.;Tauler, R., MCR-ALS GUI 2.0: New features and applications, *Chemometric Intelligence and Laboratory Systems*, 2015, **140**, 1-12.
12. Creek, D. J.;Jankevics, A.;Breitling, R.;Watson, D. G.;Barrett, M. P.;Burgess, K. E. V., Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction, *Analytical Chemistry*, 2011, **83**, 8703-8710.
13. Falchi, F.;Bertozzi, S. M.;Ottonello, G.;Ruda, G. F.;Colombano, G.;Fiorelli, C.;Martucci, C.;Bertorelli, R.;Scarpelli, R.;Cavalli, A.;Bandiera, T.;Armirotti, A., Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification, *Analytical Chemistry*, 2016, **88**, 9510-9517.

14. Tautenhahn, R.;Cho, K.;Uritboonthai, W.;Zhu, Z.;Patti, G. J.;Siuzdak, G., An accelerated workflow for untargeted metabolomics using the METLIN database, *Nature Biotechnology*, 2012, **30**, 826-828.
15. Wishart, D. S.;Knox, C.;Guo, A. C.;Eisner, R.;Young, N.;Gautam, B.;Hau, D. D.;Psychogios, N.;Dong, E.;Bouatra, S.;Mandal, R.;Sinelnikov, I.;Xia, J.;Jia, L.;Cruz, J. A.;Lim, E.;Sobsey, C. A.;Shrivastava, S.;Huang, P.;Liu, P.;Fang, L.;Peng, J.;Fradette, R.;Cheng, D.;Tzur, D.;Clements, M.;Lewis, A.;de souza, A.;Zuniga, A.;Dawe, M.;Xiong, Y.;Clive, D.;Greiner, R.;Nazyrova, A.;Shaykhutdinov, R.;Li, L.;Vogel, H. J.;Forsythei, I., HMDB: A knowledgebase for the human metabolome, *Nucleic Acids Research*, 2009, **37**, D603-D610.
16. Hutter, M. C., Molecular Descriptors for Chemoinformatics (2nd ed.). By Roberto Todeschini and Viviana Consonni, *ChemMedChem*, 2010, **5**, 306-307.
17. Taraji, M.;Haddad, P. R.;Amos, R. I. J.;Talebi, M.;Szucs, R.;Dolan, J. W.;Pohl, C. A., Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures, *Journal of Chromatography A*, 2017, **1486**, 59-67.
18. Goryński, K.;Bojko, B.;Nowaczyk, A.;Buciński, A.;Pawliszyn, J.;Kaliszan, R., Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds, *Analytica Chimica Acta*, 2013, **797**, 13-19.
19. Cao, M.;Fraser, K.;Huege, J.;Featonby, T.;Rasmussen, S.;Jones, C., Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, *Metabolomics*, 2015, **11**, 696-706.
20. Ortiz-Villanueva, E.;Navarro-Reig, M.;Jaumot, J.;Tauler, R., Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites, *Analytical Methods*, 2017, **9**, 774-785.
21. Van Dorpe, S.;Vergote, V.;Pezeshki, A.;Burvenich, C.;Peremans, K.;De Spiegeleer, B., Hydrophilic interaction LC of peptides: columns comparison and clustering, *Journal of separation science*, 2010, **33**, 728-739.
22. Periat, A.;Debrus, B.;Rudaz, S.;Guillarme, D., Screening of the most relevant parameters for method development in ultra-high performance hydrophilic interaction chromatography, *Journal of chromatography. A*, 2013, **1282**, 72-83.
23. Kawachi, Y.;Ikegami, T.;Takubo, H.;Ikegami, Y.;Miyamoto, M.;Tanaka, N., Chromatographic characterization of hydrophilic interaction liquid chromatography stationary phases: hydrophilicity, charge effects, structural selectivity, and separation efficiency, *Journal of chromatography. A*, 2011, **1218**, 5903-5919.
24. Kumar, A.;Heaton, J. C.;McCalley, D. V., Practical investigation of the factors that affect the selectivity in hydrophilic interaction chromatography, *Journal of Chromatography A*, 2013, **1276**, 33-46.
25. Schuster, G.;Lindner, W., Comparative characterization of hydrophilic interaction liquid chromatography columns by linear solvation energy relationships, *Journal of Chromatography A*, 2013, **1273**, 73-94.
26. Bajad, S. U.;Lu, W.;Kimball, E. H.;Yuan, J.;Peterson, C.;Rabinowitz, J. D., Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry, *Journal of chromatography. A*, 2006, **1125**, 76-88.
27. Sampsonidis, I.;Witting, M.;Koch, W.;Virgiliou, C.;Gika, H. G.;Schmitt-Kopplin, P.;Theodoridis, G. A., Computational analysis and ratiometric comparison approaches aimed to assist column selection in hydrophilic interaction liquid chromatography–tandem mass spectrometry targeted metabolomics, *Journal of Chromatography A*, 2015, **1406**, 145-155.
28. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31-36.
29. Gorrochategui, E.;Jaumot, J.;Lacorte, S.;Tauler, R., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC - Trends in Analytical Chemistry*, 2016, **82**, 425-442.
30. Kuhl, C.;Tautenhahn, R.;Böttcher, C.;Larson, T. R.;Neumann, S., CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets, *Analytical Chemistry*, 2012, **84**, 283-289.
31. Shao, X.;Pang, C.;Su, Q., A novel method to calculate the approximate derivative photoacoustic spectrum using continuous wavelet transform, *Fresenius' Journal of Analytical Chemistry*, 2000, **367**, 525-529.

32. Platikanov, S.;Rodriguez-Mozaz, S.;Huerta, B.;Barceló, D.;Cros, J.;Batle, M.;Poch, G.;Tauler, R., Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements, *Journal of Environmental Management*, 2014, **140**, 33-44.
33. de Juan, A.;Tauler, R., *Journal*, 2016, **30**, 5-51. Multivariate Curve Resolution-Alternating Least Squares for Spectroscopic Data, in Data Handling in Science and Technology.
34. Gómez-Canela, C.;Bolivar-Subirats, G.;Tauler, R.;Lacorte, S., Powerful combination of analytical and chemometric methods for the photodegradation of 5-Fluorouracil, *Journal of Pharmaceutical and Biomedical Analysis*, 2017, **137**, 33-41.
35. Garreta-Lara, E.;Campos, B.;Barata, C.;Lacorte, S.;Tauler, R., Metabolic profiling of Daphnia magna exposed to environmental stressors by GC–MS and chemometric tools, *Metabolomics*, 2016, **12**.
36. Ortiz-Villanueva, E.;Jaumot, J.;Benavente, F.;Pina, B.;Sanz-Nebot, V.;Tauler, R., Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling, *Electrophoresis*, 2015, **36**, 2324-2335.
37. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et biophysica acta*, 1975, **405**, 442-451.
38. Zisi, C.;Sampsonidis, I.;Fasoula, S.;Papachristos, K.;Witting, M.;Gika, H. G.;Nikitas, P.;Pappa-Louisi, A., QSRR modeling for metabolite standards analyzed by two different chromatographic columns using multiple linear regression, *Metabolites*, 2017, **7**.
39. Randazzo, G. M.;Tonoli, D.;Hambye, S.;Guillarme, D.;Jeanneret, F.;Nurisso, A.;Goracci, L.;Boccard, J.;Rudaz, S., Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification, *Analytica Chimica Acta*, 2016, **916**, 8-16.
40. Kritikos, N.;Tsantili-Kakoulidou, A.;Loukas, Y. L.;Dotsikas, Y., Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure–retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction, *Journal of Chromatography A*, 2015, **1403**, 70-80.
41. Park, S. H.;Haddad, P. R.;Talebi, M.;Tyteca, E.;Amos, R. I. J.;Szucs, R.;Dolan, J. W.;Pohl, C. A., Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model, *Journal of Chromatography A*, 2017, **1486**, 68-75.
42. Aicheler, F.;Li, J.;Hoene, M.;Lehmann, R.;Xu, G.;Kohlbacher, O., Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches, *Analytical Chemistry*, 2015, **87**, 7698-7704.
43. Wolfer, A. M.;Lozano, S.;Umbdenstock, T.;Croixmarie, V.;Arrault, A.;Vayer, P., UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling, *Metabolomics*, 2015, **12**, 8.
44. Niazi, A.;Leardi, R., Genetic algorithms in chemometrics, *Journal of Chemometrics*, 2012, **26**, 345-351.
45. Holland, J. H., *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, MIT Press, Cambridge, MA, USA, 1992.
46. Bolboacă, S. D., Assessment of random assignment in training and test sets using generalized cluster analysis technique, *Applied Medical Informatics*, 2010, **28**, 9-14.
47. Martin, T. M.;Harten, P.;Young, D. M.;Muratov, E. N.;Golbraikh, A.;Zhu, H.;Tropsha, A., Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling?, *Journal of Chemical Information and Modeling*, 2012, **52**, 2570-2578.
48. Golbraikh, A.;Tropsha, A., QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology, *Journal of Chemical Information and Computer Sciences*, 2003, **43**, 144-154.

Capítol 4

Estudi dels efectes de diversos estressants ambientals sobre l'arròs

4.1. Introducció

En aquest capítol es mostren les aplicacions de la metabolòmica i la lipidòmica no dirigides realitzades durant aquesta Tesi utilitzant les estratègies d'anàlisi i tractament de dades presentades en el capítol anterior. En primer lloc, es presenta un estudi de metabolòmica no dirigida sobre els efectes de metalls pesants (cadmi i coure) en el creixement de l'arròs. En segon lloc, s'exposa un estudi de lipidòmica no dirigida per avaluar els efectes de diferents factors estressants ambientals (alta temperatura i estrès hídric) en el creixement de l'arròs.

Els factors ambientals estressants estudiats en aquesta Tesi han estat estressos abiòtics, que són aquells que alteren els nivells òptims d'energia (llum), aigua, carboni i minerals que necessiten els organismes vegetals per a créixer i desenvolupar-se. Així, els factors abiòtics estudiats en aquesta Tesi són l'estrès hídric (relacionat amb la falta d'aigua o sequera), l'augment de temperatura i la presència de contaminants al sòl [1-3].

Els estressos relacionats amb unes condicions ambientals extremes són els factors que més limiten el creixement i el desenvolupament de les plantes [4-6]. Entre aquests factors, l'augment de la temperatura ambiental i la sequera són els que han generat més controvèrsia durant la última dècada, ja que han estat àmpliament tractats en els estudis sobre el canvi climàtic [7, 8]. Aquests estudis estimen que a finals d'aquest segle l'escalfament global produirà un augment de fins a 3,1ºC en la temperatura global de l'aire a la superfície terrestre [7, 9]. A més, també preveuen que els episodis de sequera seran més freqüents i severs [7].

D'altra banda, els metalls pesants són els principals contaminants presents en el sòl perjudicials per a les plantes [2, 6, 10]. Aquests metalls són constituents naturals de l'escorça terrestre i s'originen a partir de diversos processos geològics. Els metalls a nivell traça són essencials per a diversos processos metabòlics dels organismes, com per exemple la modificació de proteïnes o la fixació del nitrogen [10]. Tot i així, en les últimes dècades, la ràpida industrialització i les tècniques d'agricultura moderna han incrementat la concentració de metalls pesants en el medi ambient fins a nivells que resulten tòxics pels animals i les plantes [10]. Entre els metalls pesants, el cadmi (Cd) i el coure (Cu) han estat inclosos en la llista del 2015 de materials perillosos de la Llei de Responsabilitat, Compensació i Responsabilitat Ambiental Integral (CERCLA, *Comprehensive Environmental Response, Compensation, and Liability Act*) de l'Agencia de Protecció Ambiental d'Estats Units (EPA, *United States Environmental Protection*

Agency) [11]. A més, tots dos metalls s'absorbeixen molt fàcilment per les arrels de les plantes, des d'on són ràpidament transportats a la resta de parts de l'organisme [12, 13].

Aquests factors estressants s'han estudiat en les següents publicacions:

- **Publicació 3:** *Metabolomic analysis of the effects of cadmium and copper treatment in Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation*. M. Navarro-Reig, J. Jaumot, B. Piña, E. Moyano, M.T. Galceran, R. Tauler. Metallomics 9 (2017), 660-675.

  En aquest article s'estudien els efectes de l'exposició a concentracions relativament elevades de metalls pesants (cadmi i coure) en el metaboloma de l'arròs. Amb aquest objectiu, l'arròs es va sotmetre a diferents concentracions dels dos metalls en l'aigua de regadiu (des de 10 fins a 1000 µM). Les mostres d'arròs es van analitzar mitjançant LC acoblada a un espectròmetre de masses quadrupol-Orbitrap (Q-Exactive), el qual va permetre obtenir els espectres de fragmentació de tots els ions enregistrats (espectres AIF). Es va aplicar el mètode de MCR-ALS per resoldre els perfils d'elució i espectres de masses dels metabòlits detectats i es va identificar quins eren els metabòlits més afectats pels metalls pesants.

- **Publicació 4:** *Untargeted lipidomic evaluation of hydric and heat stresses on rice growth.* M. Navarro-Reig, R. Tauler, G. Iriondo-Frias, J. Jaumot. Enviat.

  En aquest article s'han avaluat els efectes de diferents estressos ambientals (calor i sequera) en el lipidoma de l'arròs. Amb aquest objectiu l'arròs es va sotmetre a un creixement sota condicions adverses de temperatura ambiental (3ºC superior a l'òptima) i a diferents nivells d'escassetat d'aigua. Els extractes de lípids es van analitzar per LC acoblada a un espectròmetre de masses de temps de vol (TOF). Els resultats obtinguts van permetre trobar quins lípids es trobaven més alterats a partir de l'aplicació de diferents eines quimiomètriques com el PCA, el PLS-DA, l'ASCA i el mètode MCR-ALS.

4.2. Publicació 3

Metabolomic analysis of the effects of cadmium and copper treatment in Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation.

M. Navarro-Reig, J. Jaumot, B. Piña, E. Moyano, M.T. Galceran, R. Tauler.

Metallomics 9 (2017), 660-675.

Metallomics

PAPER

View Article Online
View Journal | View Issue

Cite this: *Metallomics*, 2017, 9, 660

Metabolomic analysis of the effects of cadmium and copper treatment in *Oryza sativa* L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation†

Meritxell Navarro-Reig,[a] Joaquim Jaumot,[a] Benjamín Piña,[a] Encarnación Moyano,[b] Maria Teresa Galceran[b] and Romà Tauler*[a]

Received 25th November 2016,
Accepted 27th February 2017

DOI: 10.1039/c6mt00279j

rsc.li/metallomics

While the knowledge of plant metabolomes has increased in the last few years, their response to the presence of toxicants is still poorly understood. Here, we analyse the metabolomic changes in Japanese rice (*Oryza sativa var. Japonica*) upon exposure to heavy metals (Cd(II) and Cu(II)) in concentrations from 10 to 1000 μM. After harvesting, rice metabolites were extracted from aerial parts of the plants and analysed by HPLC (HILIC TSK gel amide-80 column) coupled to a mass spectrometer quadrupole-Orbitrap (Q-Exactive). Full scan and all ion fragmentation (AIF) mass spectrometry modes were used during the analysis. The proposed untargeted metabolomics data analysis strategy is based on the application of the multivariate curve resolution alternating least squares (MCR-ALS) method for feature detection, allowing the simultaneous resolution of pure chromatographic profiles and mass spectra of all metabolites present in the analysed rice extracts. All-ion fragmentation data were used to confirm the identification of MCR-ALS resolved metabolites. A total of 112 metabolites were detected, and 97 of them were subsequently identified and confirmed. Pathway analysis of the observed metabolic changes suggested an underlying similarity of the responses of the plant to Cd(II) and Cu(II), although the former treatment appeared to be the more severe of the two. In both cases, secondary metabolism and amino acid-, purine-, carbon- and glycerolipid-metabolism pathways were affected, in a pattern consistent with reduction in plant growth and/or photosynthetic capacity and with induction of defence mechanisms to reduce cell damage.

Significance to metallomics

The assessment of the effects caused in crop plants by exposure to pollutants (*i.e.* metals) is of key importance. In this work, metabolic changes caused by the presence of different concentrations of cadmium and copper in irrigation water were evaluated taking Japanese rice (*Oryza sativa* L.) as a model organism. From the results, the affected pathways could be determined to allow a biological interpretation regarding the effects on plant growth and defense.

Introduction

Heavy metals are important constituents of the Earth's crust and geological processes. Some natural processes, for instance erosion of the underground geological material, and emissions from volcanoes or forest fires, contribute to their presence in the environment. Nevertheless, during the last century, some anthropogenic activities such as mining, industry or agriculture, have altered heavy metal distribution on the earth surface, increasing significantly the levels of these pollutants in the environment. Due to the scarcity of arable land around the planet, especially in industrialized countries, contaminated soils must be used by farmers and their heavy metal pollution may cause several environmental problems and risks to human health, including contamination of edible plants.[1–3] Plants are affected by heavy metal pollutants because they are easily

[a] *Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain. E-mail: roma.tauler@idaea.csic.es*
[b] *Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6mt00279j

absorbed by roots and readily translocated to the aerial parts of the plants.[1] Plant responses to heavy metal stress include numerous defense mechanisms and stress-inducible reactions, such as synthesis or metal-binding peptides, antioxidative mechanisms produced by the generation of reactive oxygen species and formation of highly active signaling compounds.[4,5]

Biomarker discovery appears as a powerful tool to study the effects of heavy metal pollution in plants. A biomarker can be defined as a substance used as an indicator of a biological state that can cause alterations in organisms.[6] The identification of these biomarkers is useful to study pathogenic processes, risk and progression of diseases and risk assessment and toxicity of pollutants.[7,8] When biomarkers are specific to a given pollutant, they can be used as a monitoring index for long-term monitoring studies.[6–8] It is clear that the need for sensitive and specific biomarkers is increasing and much effort is being made in the research related to its discovery. Omics sciences (genomics, proteomics, transcriptomics and metabolomics) can provide an untargeted, global knowledge of organisms' physiology and of their responses to environmental inputs and, consequently, they are really useful in biomarker research. In particular, the use of metabolomics in biomarker discovery is based on the assumption that biotic and abiotic stress causes disruptions of biochemical pathways leading to a metabolic fingerprint characteristic of each stressor or group of stressors.[7,9] There are two principal metabolomics approaches: targeted and untargeted. The first one is only focused on analysing selected molecular classes, whereas untargeted metabolomics aims to screen the entire metabolite content of biological samples. In this work, the untargeted approach is used because it enables the simultaneous profiling of the largest number of metabolites present in the sample. Also, untargeted metabolomics provides the possibility of finding which metabolites show changes in their concentration under a treatment and elucidate previously unexplored biological pathways.[7]

Considering that the samples analysed in plant metabolomics can be diverse and complex, the analytical techniques used in this field should have high separation power. The most frequently utilized techniques have been nuclear magnetic resonance (NMR) and separation techniques (gas chromatography (GC), liquid chromatography (LC) and capillary electrophoresis (CE)) coupled to mass spectrometry (MS). NMR is less sensitive than MS-based techniques, and its ability to detect low abundance metabolites is limited. For this reason, MS-based techniques are nowadays generally the chosen option.[10–12] In untargeted metabolomics studies, the high complexity of the samples and the presence of a wide variety of metabolites, many of them at low concentrations and with very different physicochemical properties, make metabolite identification and determination really challenging.[6,13] MS-based techniques allow a putative identification of the metabolites by comparing the m/z of the molecular ion with the theoretical metabolite exact mass. To ensure complete identification of the metabolites, currently high-resolution mass spectrometry (HRMS) and tandem mass spectrometry (MS/MS), as well as comparison with standards are used. In this work, the selected methodology is LC-HRMS based on hydrophilic interaction liquid chromatography (HILIC). This strategy allows the analysis of a broad range of compounds without the need for chemical derivatization.[13] For the analysis of untargeted compounds HRMS and all-ion fragmentation (AIF) without selecting precursor ions provide accurate mass and structural information of the unknown metabolites.

Plant metabolomics is growing as an essential System Biology Tool in plant science and, in particular, for crop-enhancing projects. Plants produce extremely complex metabolomes, with large numbers of metabolites presenting a huge variety of structures and relative abundances. These metabolites play important roles in plant growth, development, and response to environment changes. At the same time, metabolome composition constitutes the chemical base of crop yield and quality, determining the nutritional properties both for humans and livestock.[14] Rice (*Oryza sativa* L.) is a plant of remarkable alimentary and economic importance, being one of the cereals most consumed by the world population.[15,16] The cultivar Japonica *Nipponbare* was selected for this work, as it is one of the best-known rice varieties, with a relatively short period of growth and easy genetic modification.[5,11]

Untargeted metabolomics, particularly from plants, produces extremely complex datasets, the processing of which constitutes a crucial step of the whole analytical process. Multivariate data analysis tools allow for a reliable evaluation of metabolite concentration changes. In a previous work, the authors demonstrated the usefulness of two different untargeted metabolomics data analysis approaches for the study of Japanese rice under heavy metal stress.[17] The main objective of this previous work was the preliminary evaluation and validation of the proposed data analysis tool. In contrast, the goal of the present work is focused on the analytical identification and confirmation of the metabolites of Japanese rice showing altered concentrations due to cadmium and copper treatments. For this purpose, an untargeted LC-HRMS metabolomic approach has been used. Multivariate curve resolution alternating least squares (MCR-ALS) has been employed to resolve and detect the most important metabolites. Then, all-ion fragmentation (AIF) has been applied to confirm metabolite identifications, allowing a biological interpretation of the main pathways affected.

Experimental

Reagents

Cadmium chloride hydrate (≥98.0%), copper(II) sulphate pentahydrate (≥98.0%), ammonium acetate (≥98.0%) and LC-MS grade water, acetonitrile (≥99.9%), methanol (≥99.9%) and acetic acid were supplied by Sigma-Aldrich (Steinheim, Germany). Chloroform was obtained from Carlo Erba (Peypin, France). Piperazine-*N*,*N*′-bis(2-ethanesulfonic acid) (PIPES) (≥99.0%) was used as an internal standard (Sigma-Aldrich, Steinheim, Germany).

1000 μM stock solutions of cadmium (Cd(II)) and copper (Cu(II)) were prepared weekly by the dissolution of appropriate amounts of cadmium chloride hydrate and copper(II) sulphate salts. Solutions containing 10, 50 and 100 μM metals were prepared weekly by diluting the 1000 μM stock solutions. All these solutions were stored at 6 °C until their use.

Water used for plant watering, for preparing cadmium and copper solutions, and during the extraction procedure was purified using an Elix 3 coupled to a Milli-Q system (Millipore, Belford, MA, USA), and filtered through a 0.22 μm nylon filter integrated into the Milli-Q system.

Plant growth, stress treatment and metabolite extraction

Plant growth, stress treatment and metabolite extraction were performed using a procedure previously described.[17] First, *Oryza sativa var. Japonica Nipponbare* seeds, obtained from the Centre for Research in Agricultural Genomics (CRAG), were incubated for two days at 30 °C in a wet environment. After this period, plants were grown on an Environmental Test Chamber MLE-352H (Panasonic®) for 22 days under white fluorescent light. Temperature, relative humidity, and light long-day conditions at the chamber were set as described in Fig. S1 in the ESI†.

During the first ten days of growth, rice plants were watered with Milli-Q water three times per week. After that, the treated plant samples were subjected to irrigation water containing different concentrations of Cd(II) or Cu(II), whereas the control plant samples were watered only with Milli-Q water until harvest. Metal concentrations were 10, 50, 100 and 1000 μM. After harvest, the aerial parts of the rice samples were frozen at liquid nitrogen temperature in order to quench metabolism. Samples were stored at −80 °C until extraction.

Before extraction, the aerial parts of the rice samples were ground to a fine powder using a liquid nitrogen mortar and lyophilized for 24 h until dryness. Metabolite extraction was carried out by dispersing 40 mg of the dried tissue in 1 mL of MeOH in a 2.0 mL Eppendorf tube. Then, the mixture was vortexed for 1 min and sonicated for 10 min; this step was repeated twice. After centrifuging for 20 min at 14 100 × *g*, a 750 μL aliquot of the supernatant was transferred to a 1.5 mL Eppendorf tube. Then, 500 μL of chloroform and 400 μL of water were added. After that, the mixture was vortexed for 1 min, incubated for 15 min at −4 °C, and centrifuged for 20 min at 14 100 × *g*. Finally, the aqueous fraction was transferred to a 1.5 mL Eppendorf tube, evaporated to dryness under nitrogen gas, and reconstituted with 450 μL of acetonitrile/water (1 : 1 v/v). For internal standard quantification, 50 μL of 50 mg L^{-1} solution of the internal standard (PIPES) were added to the extract. All of the extracts were stored at −80 °C until they were analyzed. Before injection, the samples were filtered through 0.2 μm nylon filters (Pall Life Sciences, Port Washington, NY, USA).

LC-MS analysis

Chromatographic separation was carried on an Accela UHPLC system (Thermo Scientific, Hemel Hempstead, UK) using the method described elsewhere.[17] The LC column was an HILIC TSK gel amide-80 column (250 × 2.0 mm i.d., 5 μm) with a guard column (10 × 2.0 mm i.d., 5 μm) of the same material provided by Tosoh Bioscience (Tokyo, Japan). Since the main aim was to separate metabolites, which are highly polar molecules, the use of HILIC columns was recommended for their analysis. An elution gradient was produced using solvent A (acetonitrile) and solvent B (acetic acid:ammonium acetate buffer 3 mM at pH 5.5) as follows: 0–3 min, isocratic gradient at 5% B; 3–27 min, linear gradient from 5 to 70% B; 27–30 min, isocratic gradient at 70% B; 30–32 min back to the initial conditions at 5% B; and from 32 to 40 min, at 5% B. The mobile phase flow rate was 0.15 mL min^{-1} and the injection volume was 5 μL.

A Q-Exactive (Thermo Fisher Scientific, Hemel Hempstead, UK) equipped with a quadrupole-Orbitrap mass analyser was used as a mass spectrometer. The ionization source employed was a heated electrospray (HESI) in negative ion mode. Mass spectra were acquired in profile mode at a resolution of 70 000 FWHM (full width half maximum) at *m*/*z* 400. Working parameters were as follows: electrospray voltage, 3.0 kV; sheath gas flow rate, 25 arbitrary units (a.u.); auxiliary gas flow rate, 10 a.u.; heated capillary temperature, 300 °C; S-lens level, 60%; automatic gain control (AGC), 3×10^6; and the maximum injection time was set at 200 ms with two microscans/scan. The full scan mass range was from *m*/*z* 90 to 1000. All ion fragmentation (AIF) was also performed with a normalized collision energy (NCE) of 35 eV.

Analysis of metal content in rice samples

After harvesting, the aerial parts and roots of the treated samples were lyophilized for 24 h until dryness and ground to a fine powder. After that, the plants were subjected to Teflon digestion. In the case of the aerial parts, 3 mL of HNO_3 and 1 mL of H_2O_2 were added to 100 mg of sample, and then the mixture was digested in a Teflon reactor for three days. The digestion of root samples is explained in the ESI.† Finally, the samples were diluted with 30 mL of water. Metal concentrations were determined using inductively coupled plasma-mass spectrometry (ICP-MS). An Agilent 7500ce model ICP-MS system (Agilent, Santa Clara, CA, USA) from the scientific and technological centers of the University of Barcelona (CCiTUB) was used for the analysis. The RF power was set at 1550 W, isotopes used during the analysis were Cu^{63} and Cd^{111} and the internal standard was Rh.

Data analysis strategy

Thermo Fisher raw chromatographic data files (raw format) were converted to the standard CDF format by the FileConverter function of Xcalibur™ 2.2.44 software (Thermo Scientific, Hemel Hempstead, UK). These data files were then imported into the MATLAB environment (release 2014b, The Mathworks Inc, Natick, MA, USA) by using the appropriate functions of the MATLAB Bioinformatics Toolbox (4.3.1.version) and in-house built routines. Each sample was represented by a data matrix containing the acquired retention times on 1020 rows (from 0 to 40 min) and the detected *m*/*z* values on 38 000 columns from 90 to 1000 *m*/*z*. Since in the Q-Exactive mass spectrometer, the mass spectra were acquired at high resolution in the Orbitrap mass analyser, the obtained matrices contained information from 90 to 1000 *m*/*z* with an accuracy of ±0.0001. The high amount of information obtained makes it difficult to process the dataset. To reduce the computer storage requirements and facilitate calculations, the total number of columns (*i.e.* *m*/*z* values) was reduced by using a binning approach (grouping

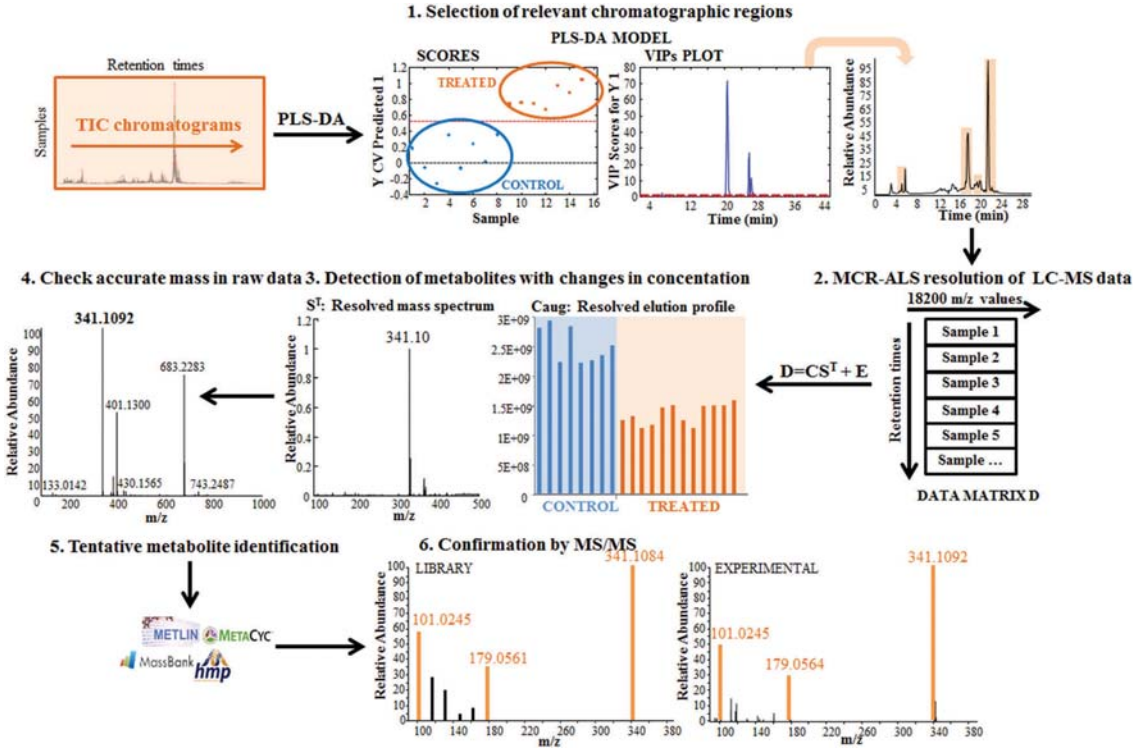


Fig. 1 MCR-ALS based untargeted metabolomics data analysis workflow.

mass values into a number of bins within a particular m/z range, in this case 0.05). Therefore, the final data matrix for each sample had 1020 rows (retention time from 0 to 40 min) and 18 200 columns (from m/z 90 to 1000 with a resolving power of 0.05 m/z units).[18]

The data analysis strategy is shown in Fig. 1. The first step was the identification of chromatographic regions able to discriminate between control and treated samples by using partial least squares-discriminant analysis (PLS-DA) to evaluate total ion current (TIC) chromatograms. Only the PLS-DA regions showing a high discrimination power between samples were then analysed by means of MCR-ALS, which resolved the pure elution profiles and mass spectra of the metabolites present in the analysed sample. The accurate mass of the MCR-ALS resolved mass spectra was verified by checking the HPLC-HRMS raw data. Then, metabolites were identified using their exact mass and confirmed by their AIF mass spectra. Finally, the statistical significance of the metabolite concentration changes between sample groups was assessed for the identified metabolites.

Selection of relevant chromatographic regions

The first step in metabolite identification was the selection of relevant chromatographic regions where metabolite concentration changes can be expected. This selection was performed applying PLS-DA to the TIC chromatograms of the analysed rice samples. Prior to PLS-DA analysis, TIC chromatograms were normalized by dividing them by the peak area of the internal standard (PIPES) in the considered sample. Then, baseline correction (using a weighted least squares method)[19] and peak alignment (using the correlation optimized warping (COW) method)[20] were performed. Finally, TIC chromatograms were only mean-centered (autoscaling was not used to avoid giving baseline noisy regions too much influence in the model). PLS-DA[21] is a multivariate regression method oriented to discriminate among different groups of samples. The TIC chromatograms of every sample (**X**, predictor variables) were related with a vector describing the class membership (**y**, predicted variable).[22,23] In this case, for every metal class, the following were selected: control samples (class 0) and samples treated with Cd(II) or Cu(II) (class I). Apart from discriminating among different groups of samples, PLS-DA also provides information about which are the most important variables (in this case, chromatographic retention times) for discrimination. For instance, variable importance on projection (VIP) scores can be used for this purpose.[23] VIP scores are a weighted sum of squared PLS variables which measure the importance of each predictor variable to the final PLS model.[23] The "greater than one" rule is often used as a variable selection criterion, because the average of squared VIP scores is equal to 1. This means that variables with a VIP score greater than 1 can be considered relevant for sample discrimination.[24] In the particular case described in this work, control samples were distinguished from treated samples

(10, 50, 100 and 1000 μM). Metabolites of interest are those that have different concentrations in control samples compared to treated samples. Therefore, VIP plots should spotlight chromatographic regions (retention times) that allow discriminating between samples because of the differences in the concentrations of their metabolites. PLS-DA results were assessed by leave-one-out cross-validation.[25] PLS Toolbox 7.8 working under MATLAB was used in all calculations.

Multivariate curve resolution by alternating least squares

Elution profiles and mass spectra of the metabolites present in every selected chromatographic region were resolved by the MCR-ALS method. MCR-ALS is a powerful chemometric tool used for the investigation and resolution of pure component contributions in unresolved mixtures. This method has been applied to a great variety of examples from different fields, such as hyphenated and multidimensional chromatographic systems, -omics data sets or spectroscopic images, among others.[15,26–29] In the case of untargeted LC-MS metabolomics studies, MCR-ALS can resolve a large number of overlapped chromatographic peaks, obtaining both the chromatogram and the mass spectrum of the components (ideally a single metabolite). It is especially helpful in untargeted metabolomics analysis, where there is no previous knowledge about the chemical compounds (metabolites) present in the analysed sample.[12,17,18,30–32]

MCR-ALS decomposes experimental data sets arranged in a data matrix according to the following bilinear model:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

where $\mathbf{D}$ (size $I \times J$) represents the experimental LC-MS data set (*i.e.* a single rice sample) organized in a data matrix in which its rows are the MS spectra at all retention times ($i = 1, \ldots, I$), and columns are the chromatograms at all m/z values ($j = 1, \ldots, J$). According to eqn (1), the $\mathbf{D}$ matrix is decomposed into the product of two-factor matrices, $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$, which are respectively the matrix of resolved elution profiles, $\mathbf{C}$ (size $I \times N$), and the matrix of their corresponding mass spectra, $\mathbf{S}^{\mathbf{T}}$ (size $N \times J$). N represents the number of resolved components using the MCR-ALS method. The matrix $\mathbf{E}$ (size $I \times J$) contains the residuals not explained by the model using the N considered components.[15,33] This data analysis strategy was extended to the simultaneous analysis of several samples. For instance, in the case of this work, different augmented data matrices (one for each metal treatment and selected chromatographic region), each with a total number of 40 data matrices (samples) were analysed. For instance, the augmented data matrix for the cadmium treated samples contained 8 control samples and 32 samples treated with Cd(II) at four different concentration levels (10, 50, 100 and 1000 μM), and with 8 replicates for each level. Analogously, an augmented data matrix for copper treated samples contained 8 control samples and 32 samples treated with Cu(II), at four different concentration levels (10, 50, 100 and 1000 μM). In both cases, these 40 samples were arranged in a single column-wise augmented data matrix ($\mathbf{D}_{\mathbf{aug}}$), which contains the individual data matrices ($\mathbf{D}_x$ where $x = 1, \ldots, 40$), one for each rice sample, and settled one on the top of the other. This column-wise augmented data matrix ($\mathbf{D}_{\mathrm{aug}}$) can be decomposed by the MCR-ALS method using the same bilinear model:

$$\mathbf{D}_{\mathrm{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_{40} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_{40} \end{bmatrix} \mathbf{S}^{\mathbf{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_{40} \end{bmatrix} = \mathbf{C}_{\mathrm{aug}}\mathbf{S}^{\mathbf{T}} + \mathbf{E}_{\mathrm{aug}} \quad (2)$$

According to eqn (2), resolved mass spectra ($\mathbf{S}^{\mathbf{T}}$) were forced to be the same for the common components (metabolites) in the different analysed rice samples (chromatographic runs). However, the elution profiles of the same component resolved in the column-wise augmented concentration matrix $\mathbf{C}_{\mathbf{aug}}$ were allowed to be different for each one of the analysed rice samples ($\mathbf{C}_x$ ($x = 1, \ldots, 40$)), as shown in eqn (2).[15,33] MCR-ALS solves eqn (1) ($\mathbf{D}$, single data matrix case) or eqn (2) ($\mathbf{D}_{\mathbf{aug}}$, augmented data matrix case) starting with an estimation of the number of components of the considered matrix by the singular value decomposition (SVD) method.[34] This number is only an initial approximation, as the final number of components is decided by taking into account the data fitting results and the reliability of the resolved profiles. An initial estimate of $\mathbf{S}^{\mathbf{T}}$ for the iterative optimization should be provided, which can be easily obtained by selection of the purest variables using a method based on the SIMPLISMA approach.[35,36] Then, the ALS constrained optimization procedure is carried out. The application of constraints provides chemical meaning to the pure mathematical solution. In the case of this work, the applied constraints were non-negativity (for chromatographic elution and spectra profiles of every component) and normalization of the spectra profile of every component (equal height).[33,37,38]

Before applying MCR-ALS, individual data matrices of each rice sample ($\mathbf{D}_x$) were normalized by dividing the individual values of each of them by the value of the chromatographic peak area of the internal standard (PIPES) of the considered sample. MCR-ALS analysis was carried out using the MCR-ALS toolbox freely available at www.mcrals.info.

Metabolite identification

After MCR-ALS analysis, matrices $\mathbf{C}_{\mathbf{aug}}$ and $\mathbf{S}^{\mathbf{T}}$ from MCR-ALS results were evaluated to detect the metabolites whose concentrations change due to the metal treatment of rice plants. The evaluation of the elution profiles resolved in $\mathbf{C}_{\mathbf{aug}}$ for each component allowed the determination of the pure metabolites that showed a significant change in their concentrations between the different groups of samples (control and metal treated samples). For instance, Fig. 2A shows an example of the peak areas and the MS spectrum resolved for a component when control and Cd(II) treated samples were considered. In this case, Cd(II) produced a significant decrease in the peak areas of metabolites in comparison with the control samples. Therefore, from the corresponding resolved mass spectrum of this component (see Fig. 2B), it was possible to estimate the mass of the diagnostic ion associated to the metabolite causing the observed differences in

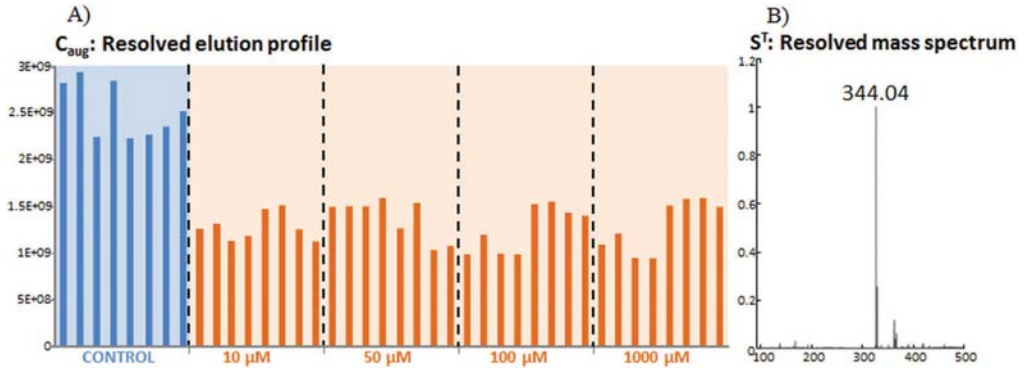


Fig. 2 An example of detection of pure metabolites showing a change in their concentration between control and Cd(II) treated samples. (A) Area of the resolved elution profiles of a particular metabolite for each sample derived from $\mathbf{c}_{aug,i}$. (B) Resolved mass spectrum of the particular metabolite ($\mathbf{s}^T_i$ in eqn (2)).

the area of the resolved chromatogram when the control and the treated samples are compared. Since HESI is a soft ionization source, in most of the cases, the mass of the diagnostic ion corresponds to the mass of the deprotonated metabolite. However, it is now possible to deduce the accurate mass (four decimal figures) of the selected metabolite looking for the HRMS Q-Exactive Orbitrap raw mass spectrum of the considered metabolite. This allowed an initial identification (elemental composition) of the selected metabolite to be performed with a relative mass error lower than 5 ppm, which fulfils Directive 2002/657/CE mass spectrometric detection performance criteria and requirements.[39] This tentative identification of the metabolites changing concentration due to heavy metal stress was performed by comparing the accurate mass of the MCR-ALS resolved metabolites (after checked in the HRMS Q-Exactive Orbitrap raw data) with the theoretical exact mass values included in public databases such as MassBank,[40] Metlin,[41] HMDB,[42] and MetaCyc.[43]

The last step in metabolite identification was the confirmation of the hypothesized chemical structure. This confirmation was performed by comparing the experimental AIF mass spectra with the corresponding theoretical MS/MS spectrum of the suspected metabolites (checked from MassBank[40] and the library of National Institute of Standards and Technology (NIST)[44] databases). In order to achieve the complete identification of the found metabolites, the requirements of the previously mentioned Directive 2002/657/CE[39] were also followed. To comply with this Directive, accurate mass measurements of diagnostic and product ions were required to have a relative error lower than 5 ppm. Furthermore, four identification points (IPs) were mandatory for positive identification. Since mass spectra were acquired with a resolution higher than 20 000 FWHM, the HRMS precursor ion and two products earn 2 and 2.5 IPs respectively, which made the achievement of the four mandatory IPs possible.

Candidate metabolites were searched on-line in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database,[45] specifically selecting *O. sativa*-annotated compounds when necessary.[46] The same strategy was followed to classify the identified metabolites into KEGG metabolic pathways.

Statistical assessment

The statistical significance of the concentration changes of the identified metabolites between sample groups was finally assessed. With this aim, the chromatographic peak areas of these metabolites were integrated for every sample. These peak areas can be obtained from MCR-ALS or using Xcalibur™ 2.2.44 software (Thermo Scientific, Hemel Hempstead, UK). In the case of this work, the original software was used and the accurate mass values obtained from MCR-ALS were extracted. The obtained values were normalized by diving by the value of the chromatographic peak area of the internal standard (PIPES) of the considered sample. One-way ANOVA was applied to the normalized areas for each metabolite with a level of significance equal to 0.05. The statistical significance of the enrichment in members from different KEGG metabolic pathways was tested using the hypergeometric distribution. In both cases, Bonferroni *post hoc* tests were applied to correct for multiple comparisons. IBM® SPSS® Statistics 22.0.0.1 software was used for ANOVA calculations. Network analyses were performed in R software using the *igraph* package.[47]

Results and discussion

Detection and identification of rice metabolites from an untargeted study is a challenging task owing to the complexity of the data, the wide range of existing metabolites and the fact that there is no prior knowledge about the chemical compounds (metabolites) contained in the analysed samples. For this reason, we propose the combined use of powerful analytical approaches with chemometric strategies. In this work PLS-DA and MCR-ALS chemometric methods have been used to seek rice metabolites changing their concentrations due to metal treatment from an untargeted study.

Assessment of metal treatment in plants

Table 1 gives the Cd(II) and Cu(II) concentrations in the aerial part samples analysed using the ICP-MS method explained in the section entitled ''Analysis of metal content in rice samples'' (results for root samples are shown in Table S1 in the ESI†).

Table 1 Results obtained in the determination of Cd and Cu in the aerial parts of the analysed rice samples

	Content (μg g^{-1} dry weight)	
Treatment	Cu(II)	Cd(II)
Control 1	18.1 ± 0.7	0.09 ± 0.02
Control 2	14.0 ± 0.6	<LOD
10 μM Cu	18.64 ± 0.04	0.27 ± 0.04
50 μM Cu	19 ± 1	<LOD
100 μM Cu	22.9 ± 0.1	<LOD
1000 μM Cu	26.0 ± 0.4	<LOD
10 μM Cd	23 ± 2	1.8 ± 0.1
50 μM Cd	17 ± 1	3.78 ± 0.09
100 μM Cd	22.0 ± 0.7	5.4 ± 0.2
1000 μM Cd	14.7 ± 0.8	6.3 ± 0.9

Limits of detection were 1.75 μg g^{-1} for Cu and 0.18 μg g^{-1} for Cd. Three replicates were analysed.

Results show that Cu(II) was already detected in the rice control samples, whereas the Cd(II) concentration was negligible in these control samples (close to the limit of detection of the technique, 0.18 μg g^{-1}). The concentration of Cd(II) in the aerial parts of the 1000 μM treated samples was 70 times bigger than in control samples, whereas the concentration of Cu(II) in the aerial parts of the 1000 μM treated samples was only 1.9 times larger than in control samples. Therefore, it seems clear that Cd(II) accumulation in the aerial parts of rice plants was much higher than Cu(II) accumulation.

Analysis of chromatographic regions

Untargeted metabolomics chromatographic data can be extremely complex containing both useful and meaningless information for the purpose of the study. In order to simplify the analysis, a first selection of the chromatographic regions of interest (those regions with metabolites changing their concentration) is performed. As described in the Experimental section (see section entitled "Selection of relevant chromatographic regions"), PLS-DA of the TIC chromatograms and VIPs plot were used for this purpose. Five chromatographic regions were selected for Cd(II) treatment (from 2.2 min to 7.2 min, from 12.2 min to 20.35 min, from 20.35 min to 22.3 min, from 22.3 min to 25.2 min and from 25.2 to 28.4 min) and six for Cu(II) treatment (from 2.2 min to 4.8 min, from 4.8 min to 7.7 min, from 12.2 min to 19.5 min, from 23.1 min to 25.2 min and from 25.2 to 28.4 min). For each one of these chromatographic regions, a column-wise augmented data matrix was built for each metal treatment and analysed separately by MCR-ALS as detailed in the section entitled "Multivariate curve resolution by alternating least squares".

Between 15 and 30 MCR-ALS components were resolved for each one of the total 11 chromatographic regions (5 for Cd(II) and 6 for Cu(II)) with explained variances (R^2) larger of 98%. The total number of MCR-ALS components used to explain data variance and patterns of all these chromatographic regions were 115 for Cd(II) treatment and 100 for Cu(II) treatment. Nevertheless, not all of these MCR-ALS resolved components were assigned to individual metabolites. Some of them did not correspond to the true metabolite chromatographic peaks but to other noisy chromatographic contributions, such as background and solvent signals. Despite the high complexity of the untargeted LC-MS data set, due to the strong overlap among metabolite elution profile at the same retention times, MCR-ALS could properly resolve a large number of metabolites from the investigated rice samples.

Fig. 3 is an example of results of the application of MCR-ALS to the resolution of two strongly coeluted metabolites corresponding to the first selected chromatographic region (elution times between 2.2 and 7.2 min). Fig. 3A depicts the elution profiles of two strongly coeluted metabolites successfully resolved after MCR-ALS analysis. In the case of the upregulated example (red elution profile), there was an increase in the chromatographic peak heights from control samples to treated samples (with their maximum height at intermediate Cd(II) concentration levels). In the case of the downregulated example (blue elution profile), there was a decrease in the chromatographic peak height when considering control and treated samples and the peak practically disappeared for the 1000 μM Cd(II) treated samples. Fig. 3B gives the mass spectra of the two resolved metabolites and the m/z values of their diagnostic ions were used for their preliminary identification. In this case, these m/z values are not accurate mass values because MCR-ALS was applied to compressed data using the binning approach. The accurate mass of selected metabolites was checked afterwards in the experimental high-resolution LC-MS raw data, which conserved full precision and accuracy of four decimal points obtained by Q-Exactive (quadrupole-Orbitrap). For instance, in the case of upregulated metabolite (red mass spectrum), the diagnostic ion appeared at 319.00 m/z and had the accurate mass value of 319.0466 m/z in the original raw data. In the case of downregulated metabolite (blue mass spectrum), the MCR-ALS diagnostic ion at 193.05 m/z had an accurate mass value of 193.0509 m/z in the original raw data.

Once the 11 column-wise augmented data matrices (from the 11 different chromatographic regions) were analysed by MCR-ALS, the peak areas of the resolved elution profiles were statistically evaluated as described in the experimental section (see section entitled "Metabolite identification"). The main aim was to detect the possible metabolites whose concentrations change significantly due to the metal treatment. Fig. 4A depicts a Venn diagram showing the total number of metabolites detected by the MCR-ALS approach varying their concentration upon metal treatment. A total number of 77 candidates were exclusively found for Cd(II) treatment, 20 solely for Cu(II) treatment and 15 of them were common for both treatments.

Metabolite identification

The accurate mass of the candidate metabolite components whose elution profile peak areas changed was compared with the exact mass values included in public databases such as MassBank,[40] Metlin,[41] HMDB,[42] and MetaCyc.[43] As an example, the tentative identification of the two MCR-ALS components shown in Fig. 3 is explained here in more detail. In the case of the upregulated metabolite (depicted in red in Fig. 3a and b), the accurate mass of the observed diagnostic ion was 319.0466 Da. This mass could be tentatively assigned with 2.1 ppm of relative

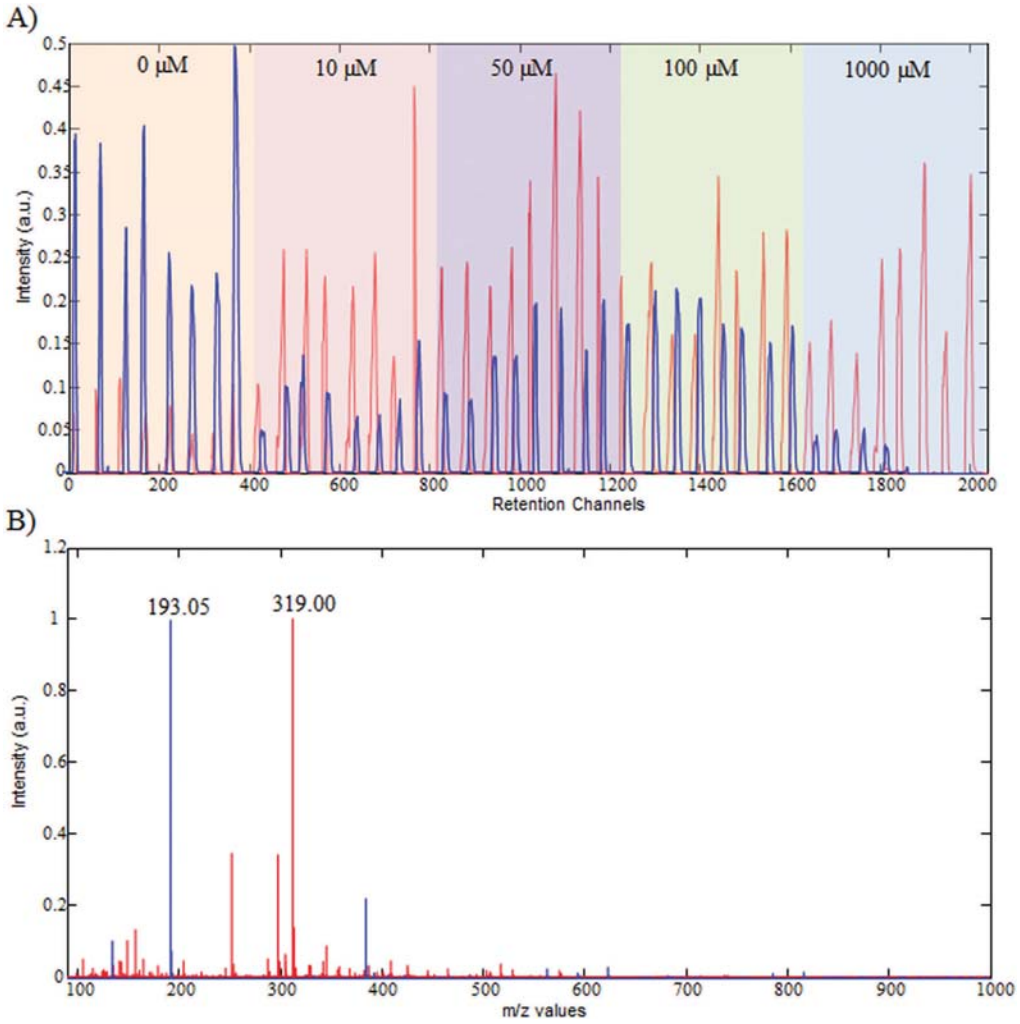


Fig. 3 An example of MCR-ALS results obtained in the analysis of LC-HRMS data from 40 rice samples under Cd(II) treatment at 5 concentration levels. (A) Resolved elution profiles obtained for two coeluted metabolites. Red profile corresponds to a metabolite showing lower concentrations in the control than in the treated samples (upregulated metabolite), and blue profile is for one metabolite showing higher concentrations in the control than in the treated samples (downregulated metabolite). (B) Resolved mass spectra for the upregulated metabolite (in red) and the downregulated metabolite (in blue).

error, to dihydromyricetin, whose deprotonated molecule ($[M - H]^-$) has an exact mass of 319.0459 Da. In the case of the downregulated metabolite (depicted in blue in Fig. 3a and b), the accurate mass of the observed diagnostic ion was 193.0509 Da, which was tentatively assigned to ferulic acid (deprotonated molecule exact mass: 193.0501 Da) with a relative mass error of 4.1 ppm. A total number of 112 metabolites could be tentatively identified with a relative mass error lower than 5.0 ppm.

A further step in metabolite identification and confirmation was the comparison of the obtained experimental AIF mass spectra with their theoretical MS/MS spectra. As mentioned in the experimental section, entitled ''LC-MS analysis'', the AIF scan mode of Q-Exactive (Orbitrap) allowed the product ion spectra of all detected ions to be obtained without the pre-selection of their precursor ions in the quadrupole. The AIF scan mode allowed metabolite identification through the product ion mass spectrum without the need to inject samples twice.

Fig. 5 depicts an example of how this AIF confirmation was performed. Fig. 5A represents the experimental AIF high-resolution mass spectrum corresponding to the chromatographic peak at 22.24 minutes. At this retention time, the corresponding full scan showed an ion at *m*/*z* 341.1092, which was tentatively identified as trehalose with 2.3 ppm of relative mass error. Fig. 5B corresponds to the MS/MS spectrum of trehalose obtained from a database (Massbank library), which shows the precursor ion at 341.1084 *m*/*z* and several product ions. Among these ions, the most intense were at 101.0245 and 179.0561 *m*/*z*. The experimental spectrum showed two major product ions at 101.0245 and 179.0564 *m*/*z*, which could be correlated with the corresponding theoretical product ions with less than 1.7 ppm of relative mass error.

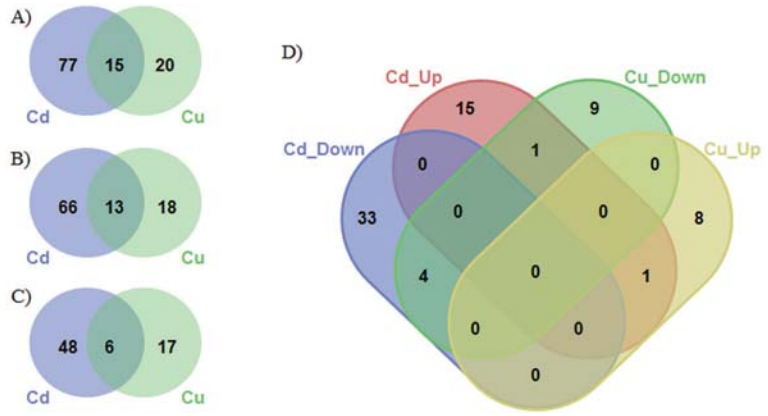


Fig. 4 Venn diagram showing the number of metabolites selected for Cd(II) treatment (purple), for Cu(II) treatment (green) and commonly for both treatments. (A) Metabolites detected after MCR-ALS analysis. (B) Metabolites identified and confirmed after AIF analysis. (C) Metabolites showing a statistically significant change. (D) Statistically significant upregulated and downregulated metabolites for Cd(II) and Cu(II) treatment.

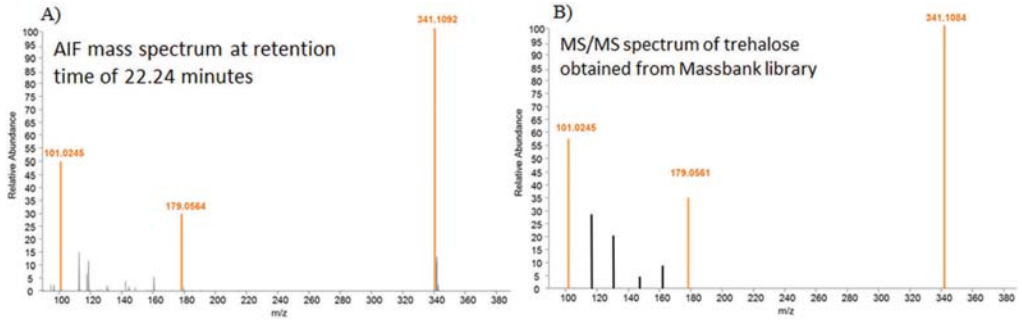


Fig. 5 Identification of trehalose. (A) AIF mass spectrum at a retention time of 22.24 minutes. (B) MS/MS spectrum of trehalose obtained from the Massbank library (Compound ID PR100542).

Taking all this information into account, trehalose was identified with 4.5 IPs (2 IPs earned for the HRMS precursor ion and 2.5 for the two product ions) and, therefore, the identification criteria recommended by Directive 2002/657/CE were fully accomplished.

Fig. 4B depicts a Venn diagram of finally identified and confirmed metabolites. As it can be seen, from the total number of 77 tentatively identified metabolites for Cd(II) treatment, 66 were confirmed. From the total number of 20 metabolites detected for Cu(II) treatment, 18 were verified. Moreover, from the total number of 15 metabolites present in both metal treatments, 13 were ratified. In summary, a total number of 97 metabolites were confirmed, which is 87% of the metabolites primarily selected by MCR-ALS analysis. Furthermore, most of the identified metabolites had 4.5 IPs of identification, accomplishing the requirements of Directive 2002/657/CE.[39] This is an important result of this study which represents an improvement in comparison to the results obtained in our previous work about Japanese rice under Cd(II) and Cu(II) stress,[17] in which the confirmation of the metabolite identification was not possible due to instrumental limitations (no accurate measurement of product ions was available).

Tables S2–S4 in the ESI† show the final AIF identification results. On the one hand, Table S2 (ESI†) gives the common metabolites obtained by both metal (Cd(II) and Cu(II)) treatments. On the other hand, Tables S3 and S4 (ESI†) give the metabolites identified only for Cd(II) and for Cu(II) metal treatment, respectively. These tables show the accurate mass of the diagnostic ions of the selected metabolites, the name of the identified metabolites, the ions assigned to diagnostic ions, the relative mass errors and the product ions used for confirmation. As it was expected, diagnostic ions corresponding to most of the identified metabolites were the deprotonated ion ($[M - H]^-$). Nevertheless, some metabolites also generated other ions frequently observed in LC-MS. For example, the loss of a water molecule ($[M - H_2O - H]^-$) or the formation of adduct ions with mobile phase components ($[M - H + HAc]^-$) or metal cations in the LC-MS system ($[M - 2H + K]^-$) or ($[M - 2H + Na]^-$). It should be highlighted that in all cases, the relative error in the accurate mass value was lower than 5 ppm, accomplishing

Directive 2002/657/CE requirements.[39] Furthermore, a minimum of two product ions were formed and detected for the major part of these identified metabolites, achieving therefore the four mandatory IPs for positive identification. Only 27 of the total 112 metabolites did not attain this recommendation: for 15 of them fragment ions with an *m*/*z* value lower than 90 were predicted, but they could not be detected under the mass scanning range used in this study. Also, 12 others showed only one predicted product ion. These 27 metabolites have been marked on Tables S2–S4 (ESI†) with an asterisk, (*). All these results confirm the potential of the proposed untargeted approach as a tool to gather the list of the most interesting metabolites showing changes due to the stress conditions of the study, in this case, Cd(II) and Cu(II) of the rice samples.

Finally, one-way ANOVA was applied to the chromatographic peak areas of the identified metabolites in order to investigate which of these metabolites presented statistically significant differences between sample groups treated by metal ions. A total number of 48 metabolites showed a significant change only for Cd(II) treatment, 17 exclusively for Cu(II) treatment and 6 commonly for both metal treatments, with a *p*-value lower than 0.05. Tables 2–4 show the metabolites that showed significant changes. On the one hand, Table 2 displays the 6 common metabolites changing significantly with both metal (Cd(II) and Cu(II)) treatments. On the other hand, Tables 3 and 4 give the metabolites showing a significant change only for Cd(II) and for Cu(II) metal treatment, respectively. Tables 2–4 give *p*-values (obtained after a Bonferroni *post hoc* test) for all those metabolites that had significant variation in their concentrations (peak areas), the value of the fold change for 10, 50, 100 and 1000 μM treated samples and if the identified metabolite is up or down regulated. The major part of the metabolites presented a similar fold change for the four levels of treatment (10, 50, 100 and 1000 μM). This result probably reflects that at the lowest concentration of Cd(II) and Cu(II) tested in this study (10 μM) the metabolites were already significantly affected, without any further modification at higher metal concentrations (50, 100 and 1000 μM). However, in some cases, appreciable differences between treatment levels were observed. For instance, in the case of synaptic acid for copper treatment (Table 2) the fold change decreased gradually when the level of treatment increased, which means that the concentration of synaptic acid in rice is lower when the concentration of Cu(II) increased. Another relevant example is the case of *o*-feruloylquinic acid for cadmium (Table 3), where the fold changes were similar for the three lower levels of treatment (10, 50 and 100 μM) but it was almost doubled for 1000 μM treated samples, indicating a large effect of this high metal concentration. Another interesting point is that 11 metabolites showing a significant change for Cd(II) metal treatment (Table 3), presented a fold change at 100 μM level treatment lower than at the other three levels. For instance, this is observed for glutamine or 1-*O*-vanilloyl-beta-D-glucose. This different behaviour might indicate an adaptive plant response at this intermediate concentration for this reduced number of

Table 2 Metabolites presenting statistically significant differences between sample groups treated with Cd(II) and Cu(II)

						Cd-treatment						Cu-treatment					
									Fold change control *vs.*						Fold change control *vs.*		
Exact mass	Compound name	KEGG	Ion assignment	Rel. mass error (ppm)	AIF (*m*/*z*)	Corrected *p*-value	Up/down-regulated	10	50	100	1000	Corrected *p*-value	Up/down-regulated	10	50	100	1000
305.0184	UMP	C00105	$[M - H_2O - H]^-$	3.1	111.0201/211.0014	5.461×10^{-3}	Down	0.24	0.44	0.42	0.67	3.57×10^{-5}	Down	0.26	0.23	0.19	0.16
223.0611	Synaptic acid	C00482	$[M - H]^-$	0.4	208.0379/93.0347/121.0297/ 149.0246/164.0481/193.0146	5.115×10^{-3}	Up	1.89	1.84	2.18	1.65	7.90×10^{-3}	Down	0.79	0.54	0.41	0.38
328.0457	Cyclic-AMP	C00575	$[M - H]^-$	1.5	134.0054/107.0504/192.9911	3.407×10^{-3}	Down	0.24	0.44	0.35	0.61	3.15×10^{-5}	Down	0.25	0.20	0.15	0.13
157.0508	2-Isopropylmaleate	C02631	$[M - H]^-$	0.9	127.0039/112.0531/123.0453	5.97×10^{-3}	Up	2.40	2.58	1.07	3.34	3.07×10^{-3}	Up	1.45	1.66	3.18	1.52
327.2186	2,3-Dinor-8-iso prostaglandin F1alpha	C14795	$[M - H]^-$	2.7	310.2150/293.2122/125.0972	1.49×10^{-4}	Down	0.18	0.51	0.32	0.46	2.94×10^{-4}	Down	0.20	0.13	0.18	0.09
481.2577	Stearyl citrate	—	$[M - 2H + K]^-$	0.7	190.0149/174.0171/268.2727	1.430×10^{-3}	Down	0.04	0.12	0.07	0.13	9.12×10^{-5}	Down	0.52	0.29	0.18	0.17

Table 3 Metabolites presenting statistically significant differences between sample groups treated with Cd(II)

Exact mass	Compound name	KEGG	Ion assignment	Rel. mass error (ppm)	AIF (*m*/*z*)	Corrected *p*-value	Up/down-regulated	Fold change control *vs.* 10	50	100	1000
176.9363	Diphosphate	C00013	$[M - H]^-$	2.1	133.0146/114.9489	2.21×10^{-4}	Down	0.49	0.45	0.78	0.35
145.0143	2-Oxoglutarate[a]	C00026	$[M - H]^-$	0.5	101.0246	8.11×10^{-3}	Up	1.17	1.32	1.15	1.05
145.0619	Glutamine	C00064	$[M - H]^-$	0.2	127.0513/128.0355/109.0409	3.603×10^{-3}	Up	3.77	3.02	1.64	4.33
104.0355	Serine[a]	C00065	$[M - H]^-$	1.9	—	1.90×10^{-2}	Up	2.20	1.96	1.45	2.46
171.0068	Glycerol 3-phosphate	C00093	$[M - H]^-$	2.1	96.9697/152.9597	9.85×10^{-6}	Down	0.39	0.39	0.48	0.51
259.0225	Glucose 1-phosphate	C00103	$[M - H]^-$	0.1	96.9155/138.9805/181.0511/240.9522	1.36×10^{-2}	Down	0.65	0.59	0.66	0.59
118.0511	Threonine[a]	C00188	$[M - H]^-$	1.0	—	6.53×10^{-3}	Up	2.78	2.46	1.38	3.47
105.0195	Glycerate[a]	C00258	$[M - H]^-$	1.5	—	3.31×10^{-2}	Down	0.98	0.62	1.21	0.81
299.0990	D-Ribulose	C00309	$[2M - H]^-$	2.1	149.0459/133.0509	6.11×10^{-3}	Down	0.29	0.64	0.46	0.44
671.4685	PA(16:0/18:2(9*Z*,12*Z*))	C00416	$[M - H]^-$	4.2	391.2256/255.2329/433.2364/409.2363/415.2257/279.2329	1.20×10^{-2}	Down	0.32	0.43	0.67	0.29
173.0458	Shikimate	C00493	$[M - H]^-$	1.2	93.0347/99.0453/111.0453/137.0245/155.0352	1.82×10^{-2}	Down	0.87	0.61	1.14	0.60
133.0143	Malic acid[a]	C00711	$[M - H]^-$	0.6	115.0038	3.04×10^{-3}	Down	0.77	0.88	0.93	0.47
213.1498	(−)-Menthone	C00843	$[M - H + HAc]^-$	0.9	111.0453/152.9959/138.0199	1.72×10^{-2}	Down	0.95	0.51	1.17	0.64
127.0517	5,6-Dihydrothymine[a]	C00906	$[M - H]^-$	2.7	—	4.08×10^{-3}	Up	3.36	2.66	1.58	3.69
344.0407	Cyclic-GMP	C00942	$[M - H]^-$	1.7	133.0151/150.0422	5.01×10^{-4}	Down	0.23	0.42	0.34	0.53
239.0773	Galactose	C00984	$[M - H + HAc]^-$	0.1	179.0565/161.0458	1.32×10^{-2}	Down	1.01	0.92	0.99	0.95
311.1353	4-Hydroxybutanoate[a]	C00989	$[3M - H]^-$	1.7	—	2.86×10^{-3}	Up	2.01	3.41	2.81	3.18
244.0227	3-Phosphoserine	C01005	$[M - H + HAc]^-$	0.4	184.0021/96.9697	1.89×10^{-2}	Up	1.95	1.86	1.65	1.51
99.0089	4-Fumaryl-acetoacetate	C01061	$[M - 2H]^{2-}$	1.7	100.0123/154.0273/112.0123/98.0011	1.76×10^{-2}	Down	1.06	1.06	1.59	0.98
341.1092	Trehalose	C01083	$[M - H]^-$	2.3	101.0245/113.0246/119.0305/143.0349/161.0456	3.68×10^{-3}	Down	0.87	0.77	0.81	0.69
113.0610	2-Hydroxycyclohexan-1-one[a]	C01147	$[M - H]^-$	1.6	99.0452	5.40×10^{-3}	Down	0.50	0.46	0.80	0.46
193.0509	Ferulic acid	C01494	$[M - H]^-$	1.1	134.0374/149.0609/178.0274	9.92×10^{-4}	Down	0.35	0.45	0.54	0.49
371.0990	Syringin	C01533	$[M - H]^-$	1.8	209.0820/433.1508	3.81×10^{-2}	Up	2.13	2.14	1.95	1.93
157.0369	Allantoin	C01551	$[M - H]^-$	1.1	129.0196/114.0311	1.63×10^{-3}	Up	4.65	4.17	2.56	5.78
337.0932	Columbamine	C01795	$[M - H]^-$	0.8	275.0952/289.1108	1.97×10^{-2}	Down	1.09	0.77	1.60	0.87
128.0356	5-Oxoproline[a]	C01879	$[M - H]^-$	2.1	—	3.89×10^{-3}	Up	3.61	2.97	1.13	4.56
367.1041	*O*-Feruloylquinic acid	C02572	$[M - H]^-$	1.9	202.1086/122.0375/199.0615/174.0561/129.0559/192.0432	1.29×10^{-2}	Up	1.82	1.83	1.63	2.67
319.0466	Dihydromyricetin	C02906	$[M - H]^-$	2.1	194.0284/124.0168/177.0196/160.0169/143.0139/107.0140	1.49×10^{-3}	Down	0.59	0.58	1.10	0.65
365.1092	*cis*-3,4-Leucopelargonidin	C03648	$[M - H + HAc]^-$	4.9	109.02960/137.02453/289.07368	1.46×10^{-4}	Down	0.67	0.42	1.00	0.40
274.0120	6-Pyruvoyltetrahydropterin	C03684	$[M - 2H + K]^-$	1.2	192.0903/162.0489	2.36×10^{-2}	Down	0.66	0.66	0.34	0.60
623.1642	Apigenin 7-*O*-beta-D-glucoside	C04608	$[M - H + HAc]^-$	3.9	432.1024/431.0988/270.0420/269.0454/239.0345/151.0040	9.80×10^{-4}	Down	0.47	0.57	1.28	0.72
337.0570	5-Amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	C04677	$[M - H]^-$	4.5	96.96978/125.04167	9.75×10^{-3}	Down	0.73	0.74	1.19	0.68
563.1421	Apigenin 7-*O*-[beta-D-apiosyl-(1->2)-beta-D-glucoside]	C04858	$[M - H]^-$	2.6	269.0666/431.0988	1.04×10^{-3}	Down	0.52	0.49	1.12	0.67
315.0717	3′-Hydroxy-*N*-methyl-(*S*)-coclaurine	C05202	$[M - H]^-$	1.4	108.02183/122.0375/191.0955	5.19×10^{-3}	Up	1.32	0.87	1.12	1.18
313.1145	(2*R*)-1-*O*-beta-D-Galactopyranosylglycerol	C05401	$[M - H + HAc]^-$	1.7	259.0945/236.0954/191.0509/162.0499	1.43×10^{-5}	Down	0.25	0.56	0.37	0.36
191.0560	L-Quinic acid[a]	C06746	$[M - H]^-$	0.6	96.96971	3.96×10^{-3}	Down	0.76	0.50	0.99	0.51
134.0374	Medicarpin	C10503	$[M - 2H]^{2-}$	0.5	269.1031/254.0956	1.18×10^{-2}	Down	0.49	0.45	0.66	0.48
815.2290	Cyanidin-3-*O*-rutinoside-5-*O*-β-D-glucoside	C12646	$[M - H + HAc]^-$	4.8	93.0347/177.0197/163.0616/179.0565/149.0457/133.0508	1.80×10^{-6}	Down	0.22	0.25	0.86	0.24
251.1037	2,6-Dihydroxy-*N*-methylmyosmine	C16151	$[M - H + HAc]^-$	0.1	109.01709/157.03167/176.67197	3.29×10^{-3}	Down	0.38	0.53	0.49	0.63
329.0874	1-*O*-Vanilloyl-beta-D-glucose	C20470	$[M - H]^-$	1.1	179.0565/149.0457/165.0406/93.0347	2.92×10^{-2}	Up	9.52	7.42	0.93	61.69

Table 3 (*continued*)

Exact mass	Compound name	KEGG	Ion assignment	Rel. mass error (ppm)	AIF (*m*/*z*)	Corrected *p*-value	Up/down-regulated	Fold change control *vs.* 10	50	100	1000
793.5193	1-18:1-2-16:0-monogalactosyldiacylglycerol	—	$[M - 2H + K]^-$	4.9	255.2329/295.2643	4.63×10^{-4}	Down	0.46	0.67	0.54	0.36
227.1290	6(*E*)-8-Oxogeraniol	—	$[M - H + HAc]^-$	0.4	167.1079/152.0845/137.0609/151.1009	3.28×10^{-3}	Down	0.44	0.50	0.66	0.42
799.2141	6‴-*O*-Sinapoylsaponarin	—	$[M - H]^-$	4.8	92.0256/131.0352/142.0053/756.1915	4.51×10^{-3}	Down	0.28	0.30	1.68	0.66
277.0333	Caffeoylmalic acid	—	$[M - H_2O - H]^-$	4.9	134.0351/278.0457/227.0337/186.0163/141.0182/116.0107/178.0240	2.68×10^{-5}	Down	0.36	0.43	0.51	0.34
785.2200	Kaempferol 3-(2*G*-apiosylrobinobioside)	—	$[M - H + HAc]^-$	4.9	93.0347/177.0198/131.0351/147.0303/117.0194/101.0245	3.24×10^{-5}	Down	0.35	0.28	0.85	0.28
347.0599	Maclurin 3-*C*-(2″-galloyl-6″-*p*-hydroxybenzoyl-glucoside)	—	$[M - 2H]^{2-}$	2.5	695.1251/93.0347/109.0296/124.0166/125.0245/136.0166	2.50×10^{-8}	Up	1000.0	1000.0	1000.0	1000.0
331.0554	*S*-7-Methylthioheptylhydroximoyl-L-cysteine	—	$[M - 2H + K]^-$	1.1	293.0990/130.0876/277.0839/278.0763/219.0772	3.88×10^{-3}	Up	9.23	7.69	3.23	9.55
385.0547	Shoyuflavone A[a]	—	$[M - H]^-$	4.7	133.0144	4.91×10^{-2}	Down	0.72	0.43	0.82	0.39

[a] Metabolites identified with less than four IPs.

metabolites. Further work should be done to corroborate this hypothesis, focusing particularly on the pathways related to these compounds.

Analysis of biological effects

A subset of 71 metabolites was identified as significantly affected by at least one of the treatments (see Venn diagram in Fig. 4D), with 60 of them annotated as *bona fide O. sativa* metabolites in KEGG (Tables 2–4). KEGG pathway analysis combining the two treatments detected 11 *O. sativa* pathways with at least four metabolites significantly affected by either Cd(II) or Cu(II) exposure (Table 5). Hypergeometric distribution analyses indicated that the number of affected metabolites was higher than expected by a random distribution (metabolite enrichment analysis, asterisks in Table 5) for many of these pathways, although the analysis is limited by the difficulty in determining which metabolites could be detected in our analyses. In any case, secondary metabolism, glycerolipid and glycerophospholipid metabolism, and some components of the carbon and amino acid metabolism seem to be affected by at least one of the treatments (Table 5).

Network analyses show the mutual correlations between affected metabolites by either treatment and the associated metabolic pathways (Fig. 6). The graph shows five distinct functional groups, related to amino acid, purine, and glycerolipid metabolism, together with a more general, secondary metabolism group (Fig. 6). The graph also shows that metabolites related to amino acid metabolism became mainly elevated in both Cd(II)- and Cu(II)-treated samples (marked in red and orange in Fig. 6, respectively), whereas the concentrations of purine- and glycerolipid-related metabolites (including glycerophospholipids) were mainly reduced by both treatments (blue, green and purple characters in Fig. 6). It also shows that, although the overlap between significantly changed metabolites by the two treatments is relatively small (Fig. 4D), both treatments affected similar pathways in essentially the same direction (red and orange on one side, blue, green and purple on the other, Fig. 6). Note that only one out of the 60 metabolites, synaptic acid, showed a divergent response to both exposures, increasing its concentration upon Cd exposure and decreasing it upon Cu exposure (Table 2, labelled in black in Fig. 6). The increase in amino acid metabolism (with some metabolites also annotated in the carbon metabolism pathway, Table 5) can be explained as a detoxification mechanism of plants and as a protection mechanism of cell constituents. On the other hand, the decrease in nucleotide and lipid concentrations may be related to a reduced growth rate and/or photosynthetic activity. Treated plants were significantly small and yellowish compared to controls (see Fig. S2 in ESI†), probably as a result of oxidative stress associated with metal poisoning. The same growth limitations may be related to the general decrease on secondary metabolism (Fig. 6). Finally, and considering metabolites not annotated in KEGG, the data shows changes in concentrations of glycosides (6‴-*O*-sinapoylsaponarin, kaempferol 3-(2*G*-apiosylrobinobioside) and maclurin 3-*C*-(2⁗-galloyl-6⁗-*p*-hydroxybenzoyl-glucoside), Table 3), as well as of stearyl citrate (Table 2), related to their biosynthesis. Glycosides

Table 4 Metabolites presenting statistically significant differences between sample groups treated with Cu(II)

Exact mass	Compound name	KEGG	Ion assignment	Rel. mass error (ppm)	AIF (*m*/*z*)	Corrected *p*-value	Up/down-regulated	Fold change control *vs.* 10	50	100	1000
175.0251	Ascorbic acid	C00072	$[M - H]^-$	1.8	113.0246/157.0367/115.07661/139.0880/130.28988	2.686×10^{-5}	Up	5.62	7.40	36.71	22.14
134.0474	Adenine	C00147	$[M - H]^-$	1.5	106.0664/107.0365	3.97×10^{-3}	Down	0.52	0.52	0.43	0.42
266.0895	Adenosine	C00212	$[M - H]^-$	0.0	134.0474/107.0365	1.37×10^{-3}	Down	0.52	0.45	0.41	0.40
409.2365	1-Palmitoylglycerol 3-phosphate	C00416	$[M - H]^-$	1.0	152.9959/255.2329/171.0065	1.35×10^{-4}	Down	0.27	0.25	0.19	0.17
433.2365	LPA(0:0/18:2(9Z,12Z))	C00416	$[M - H]^-$	1.1	278.2207/153.9993/169.9923/416.2294	2.20×10^{-5}	Down	0.43	0.33	0.14	0.14
316.1173	Glycerophosphocholine[a]	C00670	$[M - H + HAc]^-$	2.0	199.0341	4.36×10^{-3}	Up	1.46	1.47	1.37	1.62
226.9966	3-Dehydroquinate	C00944	$[M - 2H + K]^-$	1.0	127.0403/171.0302/109.0295/153.0194/125.0245/189.0199/117.0348	1.260×10^{-4}	Up	1.80	1.65	1.62	2.51
267.1084	2′-Deoxyribose	C01801	$[2M - H]^-$	0.5	133.0509/103.0403	5.405×10^{-4}	Up	24.43	22.25	30.03	14.04
431.2209	6-Aminopenicillanate	C02954	$[2M - H]^-$	4.4	215.1078/171.0066	4.893×10^{-5}	Down	0.45	0.31	0.14	0.14
179.0387	Methyl-5-thio-D-ribose	C03089	$[M - H]^-$	2.2	161.0403/144.0305/130.0229/116.0072/102.0279	7.104×10^{-3}	Up	2.03	1.85	2.21	1.15
102.0563	GABA[a]	C03665	$[M - H]^-$	2.4	—	2×10^{-3}	Down	0.49	0.50	0.49	0.56
312.0954	5-(3′-Carboxy-3′-oxopropenyl)-4,6-dihydroxypicolinate	C05641	$[M - H + HAc]^-$	1.3	99.0088/151.0262/219.0161	1.936×10^{-3}	Down	0.48	0.42	0.34	0.34
299.0776	(1*R*,6*R*)-6-Hydroxy-2-succinylcyclohexa-2,4-diene-1-carboxylate	C05817	$[M - H + Hac]^-$	1.1	239.0561/137.0246/101.0246/222.0532/194.0588	3.914×10^{-7}	Up	0.80	1.36	5.65	5.64
242.0800	Pantothenol	C05944	$[M - 2H + K]^-$	0.2	102.0562/126.0938	1.60×10^{-3}	Up	1.44	1.45	1.34	1.62
336.0875	*S*-(Hydroxymethyl)glutathione	C14180	$[M - H]^-$	1.3	320.0678/319.0844/305.0689/277.0738/185.0571	3.201×10^{-7}	Up	3.65	5.67	13.53	9.87
503.2417	PG(18:4(6Z,9Z,12Z,15Z)/0:0)	—	$[M - H]^-$	1.4	152.9959/90.0350	1.990×10^{-5}	Down	0.70	0.41	0.24	0.24
459.2337	Fusicoplagin A[a]	—	$[M - 2H + Na]^-$	4.8	347.2593	4.90×10^{-5}	Down	0.11	0.03	0.07	0.03

[a] Metabolites identified with less than four IPs.

Table 5 KEGG pathway analysis of rice metabolites with significant changes in concentration upon Cd(II) or Cu(II) exposure

KEGG pathway	Description	Metabolites	Input	Hit	Total[a]	Total hits	*p*-value[b]
Cd exposure							
osa01100	Metabolic pathways	C00026, C00064, C00065, C00093, C00103, C00105, C00188, C00258, C00309, C00416, C00482, C00493, C00984, C00989, C01005, C01061, C01083, C01494, C03684, C03692, C04677, C05202	40	23	2558	1629	0.092
osa01110	Biosynthesis of secondary metabolites	C00026, C00065, C00093, C00103, C00188, C00258, C00416, C00482, C00493, C00843, C01083, C01494, C01795, C02631, C02906, C03648, C04677, C05202, C10503	40	19	2558	648	0.001**
osa01230	Biosynthesis of amino acids	C00026, C00064, C00065, C00188, C00493, C01005	40	6	2558	147	0.018*
osa00230	Purine metabolism	C00064, C00212, C00147, C00575, C00942, C04677	40	5	2558	105	0.017*
osa02010	ABC transporters	C00064, C00065, C00093, C00188, C01083	40	5	2558	95	0.012*
osa00561	Glycerolipid metabolism	C00093, C00258, C00416, C03692, C05401	40	5	2558	120	0.027*
osa01200	Carbon metabolism	C00026, C00065, C00258, C00989, C01005	40	4	2558	57	0.009**
osa00970	Aminoacyl-tRNA biosynthesis	C00064, C00065, C00188, C01005	40	4	2558	62	0.012*
osa00630	Glyoxylate and dicarboxylate metabolism	C00026, C00064, C00065, C00258	40	4	2558	62	0.012*
osa00260	Glycine, serine and threonine metabolism	C00065, C00188, C00258, C01005	40	4	2558	57	0.009**
Cu exposure							
osa01100	Metabolic pathways	C00072, C00105, C00147, C00212, C00416, C00482, C00681, C00944, C03089, C05817	18	10	2558	1629	0.146
osa01110	Biosynthesis of secondary metabolites	C00072, C00416, C00482, C00681, C00944, C02631, C05817	18	7	2558	648	0.086
osa00230	Purine metabolism	C00416, C00670, C00681	18	3	2558	105	0.03*
osa00564	Glycerophospholipid metabolism	C00147, C00212, C00575	18	3	2558	43	0.003**

[a] Total *O. sativa* KEGG annotations. [b] Hypergeometric distribution. * Metabolite enrichment analysis.

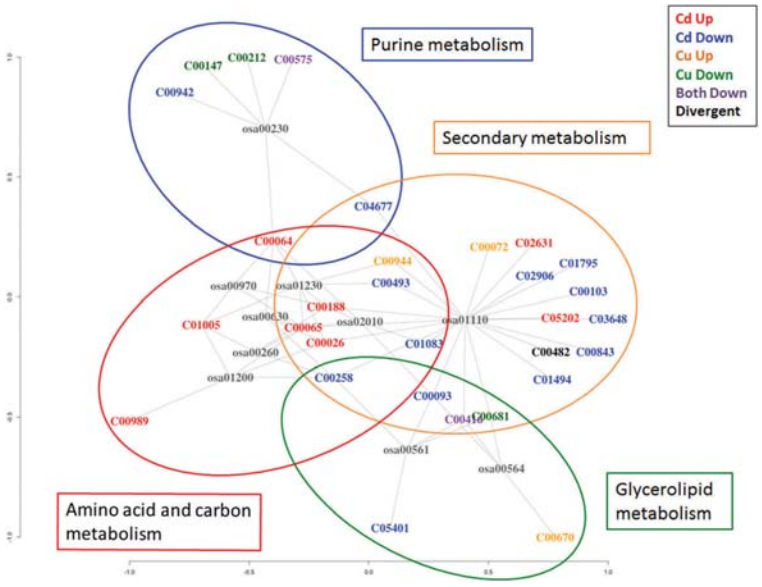


Fig. 6 Network analysis showing the mutual correlations between affected metabolites by both treatments and the associated metabolic pathways. Compounds and pathways are represented by their KEGG IDs, CXXXXX numbers for compounds (Table 3), and osaXXXXX labels for metabolic pathways (Table 4). Red and orange labels correspond to compounds whose concentrations increased upon Cd and Cu treatments, respectively; blue and green ones represent those whose concentration decreased. Purple labels indicate compounds whose concentrations decreased in both treatments, whereas the single compound with a contradictory response is shown in black (see the text). Ovals approximately delimit clusters of compounds and pathways affecting specific biological functions, as depicted in the graph.

are very abundant in plants and they form strong complexes with metals (Cd(II) and Cu(II)), which explains the significant changes observed in their concentrations under metal treatment. This type of glycoside alteration has also been observed in our

previous work[17] and also in other studies and plants, such as in *Arabidopsis thaliana*.[11]

Conclusions

The identification and confirmation of metabolites whose concentrations change due to the effects of cadmium and copper treatment in Japanese rice were assessed. The MCR-ALS chemometric procedure was first applied to LC-HRMS data with the aim of resolving elution and spectra profiles of metabolites present in the analysed rice samples. Metabolites whose elution profile peak areas were changed by metal treatment were then further identified from the accurate mass values of the corresponding diagnostic ions. High-resolution mass spectrometric detection combined with MCR-ALS data treatment has proven to be a powerful tool for untargeted metabolomics studies where no previous knowledge about what metabolites are affected by the investigated treatment was available. High-resolution AIF data analysis allowed the confirmation of metabolite identity. Most of the identified metabolites had the mandatory score of 4.5 IPs of identification, accomplishing Directive 2002/657/CE requirements.

ANOVA statistical assessment of chromatographic peak areas revealed that concentration changes of 54 metabolites were statistically significant under Cd(II) treatment and that concentration changes of 23 metabolites were statistically significant under Cu(II) treatment. Cd(II) treatment appeared more severe that Cu(II) treatment as it caused more significant changes in rice plant metabolism. However, the affected pathways (secondary metabolism and amino acid-, purine-, carbon- and glycerolipid-metabolism) are essentially the same, suggesting an underlying similarity of the responses of the plant to both divalent cations. These responses are consistent with a reduction in plant growth and/or photosynthetic capacity and with the induction of defence mechanisms to reduce cell damage.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 320737. The authors would like to thank Center for Research in Agricultural Genomics (CRAG) for kindly supplying Japanese rice seeds.

Notes and references

1 L. Sebastiani, A. Francini, S. Romeo, A. Ariani and A. Minnocci, in *Approaches to Plant Stress and their Management*, ed. R. K. Gaur and P. Sharma, Springer India, New Delhi, 2014, pp. 267–279.

2 F. Villiers, C. Ducruix, V. Hugouvieux, N. Jarno, E. Ezan, J. Garin, C. Junot and J. Bourguignon, Investigating the plant response to cadmium exposure by proteomic and metabolomic approaches, *Proteomics*, 2011, **11**, 1650–1663.

3 A. D'Alessandro, M. Taamalli, F. Gevi, A. M. Timperio, L. Zolla and T. Ghnaya, Cadmium stress responses in *Brassica juncea*: hints from proteomics and metabolomics, *J. Proteome Res.*, 2013, **12**, 4979–4997.

4 X. Sun, J. Zhang, H. Zhang, Y. Ni, Q. Zhang, J. Chen and Y. Guan, The responses of *Arabidopsis thaliana* to cadmium exposure explored *via* metabolite profiling, *Chemosphere*, 2010, **78**, 840–845.

5 R. Aina, M. Labra, P. Fumagalli, C. Vannini, M. Marsoni, U. Cucchi, M. Bracale, S. Sgorbati and S. Citterio, Thiol-peptide level and proteomic changes in response to cadmium toxicity in *Oryza sativa* L. roots, *Environ. Exp. Bot.*, 2007, **59**, 381–392.

6 H. Zhang, H. Hu, C. Deng, Y. Chun, S. Zhou, F. Huang and Q. Zhou, Integrative system biology strategies for disease biomarker discovery, *Comb. Chem. High Throughput Screening*, 2012, **15**, 286–298.

7 M. S. Monteiro, M. Carvalho, M. L. Bastos and P. G. De Pinho, Metabolomics analysis for biomarker discovery: advances and challenges, *Curr. Med. Chem.*, 2013, **20**, 257–271.

8 H. Kuang, Z. Li, C. Peng, L. Liu, L. Xu, Y. Zhu, L. Wang and C. Xu, Metabonomics approaches and the potential application in foodsafety evaluation, *Crit. Rev. Food Sci. Nutr.*, 2012, **52**, 761–774.

9 P. L. Horvatovich and R. Bischoff, Current technological challenges in biomarker discovery and validation, *Eur. J. Mass Spectrom.*, 2010, **16**, 101–121.

10 M. Commisso, P. Strazzer, K. Toffali, M. Stocchero and F. Guzzo, Untargeted metabolomics: an emerging approach to determine the composition of herbal products, *Comput. Struct. Biotechnol. J.*, 2013, **4**, e201301007, DOI: 10.5936/csbj.201301007.

11 E. Fukusaki and A. Kobayashi, Plant metabolomics: potential for practical operation, *J. Biosci. Bioeng.*, 2005, **100**, 347–354.

12 E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot and R. Tauler, Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling, *Electrophoresis*, 2015, **36**, 2324–2335.

13 J. F. Xiao, B. Zhou and H. W. Ressom, Metabolite identification and quantitation in LC-MS/MS-based metabolomics, *TrAC, Trends Anal. Chem.*, 2012, **32**, 1–14.

14 J. Hong, L. Yang, D. Zhang and J. Shi, Plant metabolomics: an indispensable system biology tool for plant science, *Int. J. Mol. Sci.*, 2016, **17**, 767.

15 J. Jaumot, A. de Juan and R. Tauler, MCR-ALS GUI 2.0: new features and applications, *Chemom. Intell. Lab. Syst.*, 2015, **140**, 1–12.

16 J. K. Kim, S.-Y. Park, S.-H. Lim, Y. Yeo, H. S. Cho and S.-H. Ha, Comparative metabolic profiling of pigmented rice (*Oryza sativa* L.) cultivars reveals primary metabolites are correlated with secondary metabolites, *J. Cereal Sci.*, 2013, **57**, 14–20.

17 M. Navarro-Reig, J. Jaumot, A. García-Reiriz and R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, *Anal. Bioanal. Chem.*, 2015, **407**, 8835–8847.

18 E. Gorrochategui, J. Casas, C. Porte, S. Lacorte and R. Tauler, Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells, *Anal. Chim. Acta*, 2015, **854**, 20–33.

19 P. H. C. Eilers, Parametric Time Warping, *Anal. Chem.*, 2004, **76**, 404–411.

20 G. Tomasi, F. Van Den Berg and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, *J. Chemom.*, 2004, **18**, 231–241.

21 M. Barker and W. Rayens, Partial least squares for discrimination, *J. Chemom.*, 2003, **17**, 166–173.

22 P. Geladi and B. R. Kowalski, Partial least-squares regression: a tutorial, *Anal. Chim. Acta*, 1986, **185**, 1–17.

23 S. Wold, A. Johansson and M. Cocchi, in *3D QSAR in Drug Design: Theory Methods and Applications*, ed. H. Kubiny, ESCOM Science Publishers, Leiden, 1993, pp. 583–618.

24 I. G. Chong and C. H. Jun, Performance of some variable selection methods when multicollinearity is present, *Chemom. Intell. Lab. Syst.*, 2005, **78**, 103–112.

25 B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig and R. S. Koch, *PLS_Toolbox 3.5 for use with Matlab*, Eigenvector Research Inc, Wenatchee, WA, 2005.

26 A. de Juan and R. Tauler, Factor analysis of hyphenated chromatographic data. Exploration, resolution and quantification of multicomponent systems, *J. Chromatogr. A*, 2007, **1158**, 184–195.

27 A. De Juan, J. Jaumot and R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, *Anal. Chim. Acta*, 2014, **6**, 4964–4976.

28 A. de Juan and R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications, *Crit. Rev. Anal. Chem.*, 2006, **36**, 163–176.

29 C. Ruckebusch and L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, *Anal. Chim. Acta*, 2013, **765**, 28–36.

30 M. Farrés, B. Piña and R. Tauler, Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS, *Metabolomics*, 2014, **11**, 210–224.

31 F. Puig-Castellví, I. Alfonso, B. Pinã and R. Tauler, 1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis, *Sci. Rep.*, 2016, **6**, 30982, DOI: 10.1038/srep30982.

32 C. Bedia, N. Dalmau, J. Jaumot and R. Tauler, Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors, *Environ. Res.*, 2015, **140**, 18–31.

33 R. Tauler, Multivariate curve resolution applied to second order data, *Chemom. Intell. Lab. Syst.*, 1995, **30**, 133–146.

34 G. H. Golub and C. F. V. Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore, 3rd edn, 1996.

35 W. Windig and J. Guilment, Interactive self-modeling mixture analysis, *Anal. Chem.*, 1991, **63**, 1425–1432.

36 W. Windig and D. A. Stephenson, Self-modeling mixture analysis of second-derivative near-infrared spectral data using the simplisma approach, *Anal. Chem.*, 1992, **64**, 2735–2742.

37 A. De Juan, S. C. Rutan and R. Tauler, in *Comprehensive Chemometrics*, ed. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2010, vol. 2, pp. 325–344.

38 R. Tauler, M. Maeder and A. de Juan, in *Comprehensive Chemometrics*, ed. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2010, vol. 2, pp. 473–505.

39 Implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, 2002, Directive 2002/657/CE.

40 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences, *J. Mass Spectrom.*, 2010, **45**, 703–714.

41 R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti and G. Siuzdak, An accelerated workflow for untargeted metabolomics using the METLIN database, *Nat. Biotechnol.*, 2012, **30**, 826–828.

42 D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. de souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel and I. Forsythei, HMDB: a knowledgebase for the human metabolome, *Nucleic Acids Res.*, 2009, **37**, D603–D610.

43 R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang and P. D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, *Nucleic Acids Res.*, 2014, **42**, D459–D471.

44 J. J. Jansen, H. C. J. Hoefsloot, J. Van Der Greef, M. E. Timmerman, J. A. Westerhuis and A. K. Smilde, ASCA: analysis of multivariate data obtained from an experimental design, *J. Chemom.*, 2005, **19**, 469–481.

45 M. Kanehisa, S. Goto, Y. Sato, M. Furumichi and M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res.*, 2012, **40**, D109–D114.

46 T. Kind, M. Scholz and O. Fiehn, How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry, *PLoS One*, 2009, **4**, e5440, DOI: 10.1371/journal.pone.0005440.

47 R_Development_Core_Team, *A language and environment for statistical computing*, R Foundation for Statistical Computing Publ, Venna, Austria, 2005.

Informació Suplementària a la Publicació 3

Metabolomic analysis of the effects of cadmium and copper treatment in Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation.

M. Navarro-Reig, J. Jaumot, B. Piña, E. Moyano, M.T. Galceran, R. Tauler.

Metallomics 9 (2017), 660-675.

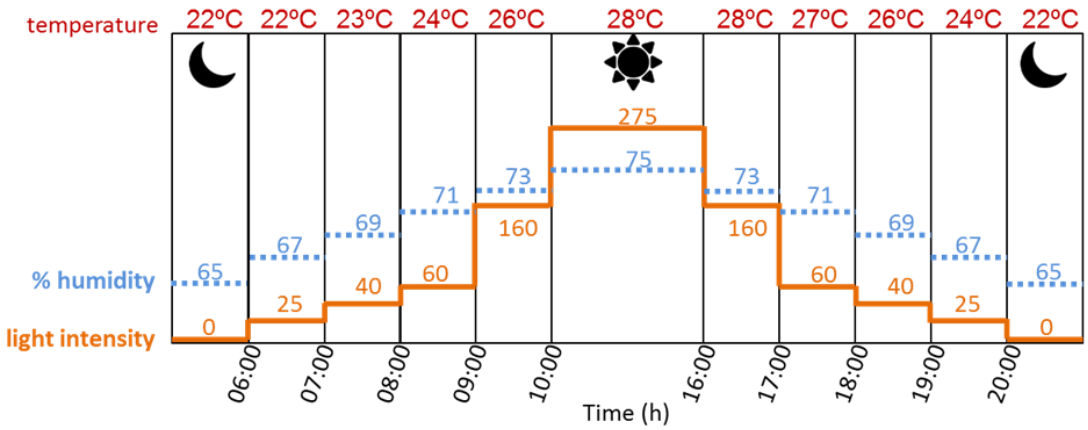


Figure S1. Experimental temperature, relative humidity and light long-day rice cultivation conditions at the growth chamber.

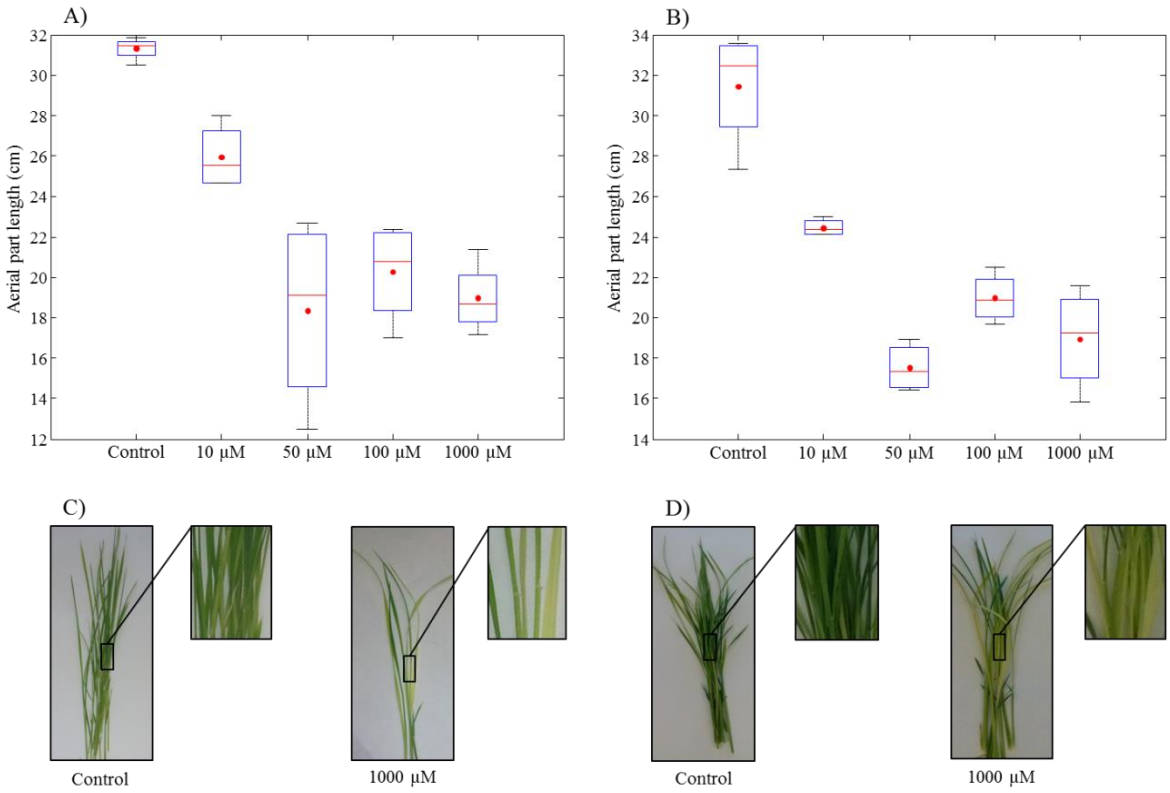


Figure S2. Phenotype information of the studied rice plants. A) Boxplot of the length of aerial parts for Control and Cd(II)-treated plants. B) Boxplot of the length of aerial parts for Control and Cu(II)-treated plants. C) Photographs of Control and 1000 µM Cd(II)-treated plants. D) Photographs of Control and 1000 µM Cu(II)-treated plants. Zooms in C and D show that treated plants were significantly yellowish compared to controls.

Analysis of metal content in root samples

In the case of root samples, 3 mL of HNO3, 1 mL of H2O2 and 0.2 mL of HF were added to 40 mg of sample, and then the mixture was digested in a Teflon reactor for three days. Finally, the samples were diluted with 30 mL of water.

Table S1. Results obtained in the determination of Cd and Cu in roots of the analyzed rice samples.

Treatments	Content ($\mu g \cdot g^{-1}$ dry weight)	
	Cu	Cd
Control 1	47.0	1.18
Control 2	45.14	0.76
10 µM Cu	50.4	3.04
50 µM Cu	87.5	4.05
100 µM Cu	81.7	1.63
1000 µM Cu	176	1.36
10 µM Cd	46.2	74.2
50 µM Cd	43.4	31.4
100 µM Cd	36.9	55.6
1000 µM Cd	38.8	200

* Limits of detection were 0.18 $\mu g \cdot g^{-1}$ for Cd and 1.75 $\mu g \cdot g^{-1}$ for Cu.

Table S2. AIF identification results. Common metabolites obtained by both metal (Cd(II) and Cu(II)) treatments

					Cd-treatment						Cu-treatment					
Exact mass	**Compound Name**	**Ion Assignment**	**Rel. mass error (ppm)**	**AIF (*m/z*)**	**corrected *p*-value**	**Up/Dow-regulated**	**Fold Change Control vs**				**corrected *p*-value**	**Up/Dow-regulated**	**Fold Change Control vs**			
							10	**50**	**100**	**1000**			**10**	**50**	**100**	**1000**
113.0610	2-Hydroxycyclohexan-1-one *	$[M-H]^-$	1.6	99.0452	5.40E-03	Down	0.50	0.46	0.80	0.46	1.089E-01	Down	1.16	0.81	0.40	0.74
157.0508	2-Isopropylmaleate	$[M-H]^-$	0.9	127.0039/112.0531/ 123.0453	5.97E-03	Up	2.40	2.58	1.07	3.34	3.E-03	Up	1.45	1.66	3.18	1.52
223.0611	Synaptic acid	$[M-H]^-$	0.4	208.0379/93.0347/ 121.0297/149.0246/ 164.0481/1193.0146	5.115E-03	Up	1.89	1.84	2.18	1.65	8.E-03	Down	0.79	0.54	0.41	0.38
327.2186	2,3-Dinor-8-iso prostaglandin F1alpha;	$[M-H]^-$	2.7	310.2150/293.2122/ 125.0972	1.49E-04	Down	0.18	0.51	0.32	0.46	3.E-04	Down	0.20	0.13	0.18	0.09
481.2577	Stearyl citrate	$[M-2H+K]^-$	0.7	190.0149/174.0171/ 268.2727	1.430E-03	Down	0.04	0.12	0.07	0.13	9.E-05	Down	0.52	0.29	0.18	0.17
305.0184	UMP	$[M-H_2O-H]^-$	3.1	111.0201/211.0014	5.461E-03	Down	0.24	0.44	0.42	0.67	4.E-05	Down	0.26	0.23	0.19	0.16
328.0457	Cyclic-AMP	$[M-H]^-$	1.5	134.0054/107.0504/ 192.9911	3.407E-03	Down	0.24	0.44	0.35	0.61	3.E-05	Down	0.25	0.20	0.15	0.13
343.0532	Indole-3-acetyl-methionine *	$[M-2H+K]^-$	2.3	142.0512	5.443E-02	Down	0.78	0.91	0.70	0.55	6.E-01	Down	1.05	1.88	0.92	0.94
145.0619	Glutamine	$[M-H]^-$	0.2	127.0513/128.0355/ 109.0409	3.603E-03	Up	3.77	3.02	1.64	4.33	1.E-01	Up	1.40	1.53	1.72	1.44
342.1110	(6S)-Hydroxyhyoscyamine	$[M-2H+K]^-$	1.0	93.0347/139.0879	2.133E-01	Up	1.33	0.87	1.12	1.17	3.E-01	Up	1.05	1.15	2.74	1.22
401.1303	Cellobiose *	$[M-H+HAc]^-$	0.6	341.1093	2.544E-01	Up	1.31	0.86	1.12	1.17	4.E-01	Up	1.04	1.15	1.29	1.23
102.0563	GABA *	$[M-H]^-$	2.4	-	9.492E-02	Down	0.63	0.60	0.61	0.96	2.E-03	Down	0.49	0.50	0.49	0.56

127.0517	5,6-Dihydrothymine *	$[M-H]^-$	2.7	-	4.08E-03	Up	3.36	2.66	1.58	3.69	4.E-01	Up	1.39	1.51	1.71	3.45
439.0865	3,5-Dihydroxyphenyl 1-O-(6-O-galloyl-beta-D-glucopyranoside)	$[M-H]^-$	3.9	125.0245/93.0347/ 179.0565/147.0662/ 149.0457/133.0509	3.367E-01	Down	1.12	0.85	0.89	0.93	3.E-01	Up	1.17	1.20	1.28	1.07
487.1789	Ptelatoside B	$[M-H+HAc]^-$	4.6	179.0563/147.0661/ 149.0456/133.0505/ 93.0347	5.747E-02	Up	2.50	1.97	0.98	2.65	3.E-01	Up	1.43	1.65	2.18	1.87

* Metabolites identified with less than four IPs.

Table S3. AIF identification results. Metabolites identified only for Cd(II).

Exact mass	Compound Name	Ion Assignment	Rel. mass error (ppm)	AIF (*m/z*)	corrected *p*-value	Up/Dow-regulated	Fold Change Control vs			
							10	50	100	1000
111.0453	Acrolein *	$[2M-H]^-$	1.4	-	1.06.E-01	Down	0.76	0.90	1.41	0.88
134.0374	(Medicarpin	$[M-2H]^{2-}$	0.5	269.1031/254.0956	1.18.E-02	Down	0.49	0.45	0.66	0.48
137.0243	4-Hydroxybenzoate *	$[M-H]^-$	0.8	93.0347	2.10.E-01	Up	1.05	1.26	1.59	1.35
145.0659	Eugenol	$[M-H_2O-H]^-$	4.0	163.0402/148.0533/ 130.0875/102.0279	4.67.E-01	Up	0.84	1.01	1.22	1.04
193.0509	Ferulic acid	$[M-H]^-$	1.1	134.0374/149.0609/ 178.0274	9.92.E-04	Down	0.35	0.45	0.54	0.49
205.0356	(R)-Lipoate	$[M-H]^-$	3.3	100.0487/160.0406	1.12.E-01	Up	1.35	1.28	1.50	1.05
213.1498	(-)-Menthone	$[M-H+HAc]^-$	0.9	111.0453/152.9959/ 138.0199	1.72.E-02	Down	0.95	0.51	1.17	0.64
227.1290	6(E)-8-Oxogeraniol	$[M-H+HAc]^-$	0.4	167.1079/152.0845/ 137.0609/151.1009	3.28.E-03	Down	0.44	0.50	0.66	0.42
241.0120	Galactose 1-phosphate	$[M-H+HAc]^-$	2.7	259.0225/96.9697/ 138.9874	1.55.E-01	Up	1.18	1.20	1.58	1.28
319.0466	Dihydromyricetin	$[M-H]^-$	2.1	194.0284/124.0168/ 177.0196/160.0169/ 143.0139/107.0140	1.49.E-03	Down	0.59	0.58	1.10	0.65
337.0932	Columbamine	$[M-H]^-$	0.8	275.0952/289.1108	1.97.E-02	Down	1.09	0.77	1.60	0.87

367.1041	O-Feruloylquinic acid	$[M-H]^-$	1.9	202.1086/122.0375/ 199.0615/174.0561/ 129.0559/192.0432	1.29.E-02	Up	1.82	1.83	1.63	2.67
515.1231	b-D-Glucuronopyranosyl-(1->3)-a-D-galacturonopyranosyl-(1->2)-L-rhamnose	$[M-H]^-$	4.4	131.0352/192.0262/ 500.0988	2.92.E-01	Down	0.60	0.74	1.11	0.72
671.4685	PA(16:0/18:2(9Z,12Z))	$[M-H]^-$	4.2	391.2256/255.2329/ 433.2364/409.2363/ 415.2257/279.2329	1.20.E-02	Down	0.32	0.43	0.67	0.29
719.4904	PG(16:0/16:1(9Z))	$[M-H]^-$	4.9	687.4284/254.2207	2.48.E-01	Down	0.65	0.85	0.81	0.56
721.5055	PG(16:0/16:0)	$[M-H]^-$	4.2	254.2207/706.4765/ 689.4425	9.19.E-02	Down	0.62	1.06	0.91	0.51
741.4741	1-18:3-2-trans-16:1-phosphatidylglycerol	$[M-H]^-$	3.9	488.2498/464.2501/ 276.2049	3.25.E-01	Down	0.69	0.81	0.79	0.54
793.5193	1-18:1-2-16:0-monogalactosyldiacylglycerol	$[M-2H+K]^-$	4.9	255.2329/295.2643	4.63.E-04	Down	0.46	0.67	0.54	0.36
99.0089	4 -Fumaryl-acetoacetate	$[M-2H]^{2-}$	1.7	100.0123/154.0273/ 112.0123/98.0011	1.76.E-02	Down	1.06	1.06	1.59	0.98
105.0195	Glycerate *	$[M-H]^-$	1.5	-	3.31.E-02	Down	0.98	0.62	1.21	0.81
133.0143	Malic acid *	$[M-H]^-$	0.6	115.0038	3.04.E-03	Down	0.77	0.88	0.93	0.47
145.0143	2-Oxoglutarate *	$[M-H]^-$	0.5	101.0246	8.11.E-03	Up	1.17	1.32	1.15	1.05
153.0323	n-Butyl acetate *	$[M-2H+K]^-$	0.2	-	3.59.E-01	Up	1.47	1.22	1.19	1.26
157.0369	Allantoin	$[M-H]^-$	1.1	129.0196/114.0311	1.63.E-03	Up	4.65	4.17	2.56	5.78
171.0302	3-Dehydroshikimate	$[M-H]^-$	1.6	109.0297/127.0403/ 108.0219/91.0191/ 99.0406	8.31.E-02	Down	0.78	0.71	1.03	0.62
179.0565	Glucose *	$[M-H]^-$	1.9	119.03515	3.94.E-01	Down	0.81	0.78	0.69	0.95
243.0620	Uridine	$[M-H]^-$	1.0	110.0249/140.0356/ 152.0356/200.0569	4.90.E-01	Down	0.43	1.11	0.84	0.81
311.1353	4-Hydroxybutanoate *	$[3M-H]^-$	1.7	-	2.86.E-03	Up	2.01	3.41	2.81	3.18
315.0727	2,5-Dihydroxybenzoate 5-O-β-D-glucoside	$[M-H]^-$	1.6	131.0352/152.0116/ 107.0141/161.0459/ 145.0508/239.0561/ 247.0611	9.70.E-02	Up	1.22	1.26	1.71	1.21

321.1559	Butyl (S)-3-hydroxybutyrate glucoside	$[M-H]^-$	1.3	249.0615/247.0822/ 287.1502/253.1446	2.80.E-01	Up	1.18	1.24	1.44	1.05
337.0570	5-Amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide	$[M-H]^-$	4.5	96.96978/125.04167	9.75.E-03	Down	0.73	0.74	1.19	0.68
343.1042	Benzoate β-D-glucose ester	$[M-H+HAc]^-$	2.1	179.056/163.0616/ 149.0456/165.0405/ 121.0297	4.50.E-01	Up	2.28	1.16	1.61	1.15
344.0407	cyclic-GMP	$[M-H]^-$	1.7	133.0151/150.0422	5.01.E-04	Down	0.23	0.42	0.34	0.53
371.0990	Syringin	$[M-H]^-$	1.8	209.0820/433.1508	3.81.E-02	Up	2.13	2.14	1.95	1.93
563.1421	Apigenin 7-O-[beta-D-apiosyl-(1->2)-beta-D-glucoside]	$[M-H]^-$	2.6	269.0666/431.0988	1.04.E-03	Down	0.52	0.49	1.12	0.67
799.2141	6'''-O-Sinapoylsaponarin	$[M-H]^-$	4.8	92.0256/131.0352/ 142.0053/756.1915	4.51.E-03	Down	0.28	0.30	1.68	0.66
130.0876	Leucine*	$[M-H]^-$	1.6	-	3.29.E-01	Up	1.00	0.98	0.66	1.29
164.0720	Phenylalanine	$[M-H]^-$	1.5	147.04529/103.0554	6.09.E-01	Down	0.69	0.81	0.43	0.86
173.0458	Shikimate	$[M-H]^-$	1.2	93.0347/99.0453/ 111.0453/137.0245/ 155.0352	1.82.E-02	Down	0.87	0.61	1.14	0.60
191.0560	L-Quinic acid*	$[M-H]^-$	0.6	96.96971	3.96.E-03	Down	0.76	0.50	0.99	0.51
237.0615	3-Deoxy-D-(manno)-octulosonate	$[M-H]^-$	0.5	221.0665/165.0752	4.67.E-01	Down	0.95	0.94	1.16	0.94
251.1037	2,6-Dihydroxy-N-methylmyosmine	$[M-H+HAc]^-$	0.1	109.01709/157.0316 7/176.67197	3.29.E-03	Down	0.38	0.53	0.49	0.63
301.0537	Cysteine *	$[2M-H+HAc]^-$	1.0	-	1.58.E-01	Down	0.78	1.07	1.11	0.86

311.1353	4-Hydroxybutyric acid *	$[3M-H]^-$	1.7	-	3.84.E-01	Up	0.90	1.55	1.19	1.64
347.0599	Maclurin 3-C-(2"-galloyl-6"-p-hydroxybenzoyl-glucoside)	$[M-2H]^{2-}$	2.5	695.1251/93.0347/109.0296/124.0166/125.0245/136.0166	2.50.E-08	Up	1000.00	1000.00	1000.00	1000.00
363.0337	Xanthylic acid	$[M-H]^-$	3.0	96.9697/108.0219/120.9698/151.0264/211.0013	1.49.E-01	Down	1.24	0.97	1.51	0.70
365.1092	cis-3,4-Leucopelargonidin	$[M-H+HAc]^-$	4.9	109.02960/137.0245/289.07368	1.46.E-04	Down	0.67	0.42	1.00	0.40
385.0547	Shoyuflavone A *	$[M-H]^-$	4.7	133.0144	4.91.E-02	Down	0.72	0.43	0.82	0.39
623.1642	Apigenin 7-O-beta-D-glucoside	$[M-H+HAc]^-$	3.9	432.1024/431.0988/270.0420/269.0454/239.0345/151.0040	9.80.E-04	Down	0.47	0.57	1.28	0.72
785.2200	Kaempferol 3-(2G-apiosylrobinobioside)	$[M-H+HAc]^-$	4.9	93.0347/177.0198/131.0351/147.0303/117.0194/101.0245	3.24.E-05	Down	0.35	0.28	0.85	0.28
815.2290	Cyanidin-3-O-rutinoside-5-O-ß-D-glucoside	$[M-H+HAc]^-$	4.8	93.0347/177.0197/163.0616/179.0565/149.0457/133.0508	1.80.E-06	Down	0.22	0.25	0.86	0.24
96.9698	Phosphate *	$[M-H]^-$	1.5	-	3.61.E-01	Up	1.00	1.01	1.35	1.73
132.0305	Aspartic acid	$[M-H]^-$	2.0	115.0039/114.0562	3.27.E-01	Up	1.79	2.28	3.70	1.68
171.0068	Glycerol 3-phosphate	$[M-H]^-$	2.1	96.9697/152.9597	9.85.E-06	Down	0.39	0.39	0.48	0.51
176.9363	Diphosphate	$[M-H]^-$	2.1	133.0146/114.9489	2.21.E-04	Down	0.49	0.45	0.78	0.35
195.0515	Gluconic acid	$[M-H]^-$	2.2	98.9561/111.0089/128.9598/130.9439/140.9886/149.0536/151.0153	3.25.E-01	Down	1.00	1.12	1.09	0.77
244.0227	3-Phosphoserine	$[M-H+HAc]^-$	0.4	184.0021/96.9697	1.89.E-02	Up	1.95	1.86	1.65	1.51
253.0567	D-Glucuronic acid	$[M-H+HAc]^-$	0.7	193.0358/113.0246/103.0037/101.0246	1.74.E-01	Down	0.71	0.87	1.06	0.82
259.0225	Glucose 1-phosphate	$[M-H]^-$	0.1	96.9155/138.9805/181.0511/240.9522	1.36.E-02	Down	0.65	0.59	0.66	0.59

277.0333	Caffeoylmalic acid	$[M-H_2O-H]^-$	4.9	134.0351/278.0457/ 227.0337/186.0163/ 141.0182/116.0107/ 178.0240	2.68.E-05	Down	0.36	0.43	0.51	0.34
299.0990	D-Ribulose	$[2M-H]^-$	2.1	149.0459/133.0509	6.11.E-03	Down	0.29	0.64	0.46	0.44
313.1145	(2R)-1-O-beta-D-Galactopyranosylglycerol	$[M-H+HAc]^-$	1.7	259.0945/236.0954/ 191.0509/162.0499	1.43.E-05	Down	0.25	0.56	0.37	0.36
331.0554	S-7-Methylthioheptylhydroximoyl-L-cysteine	$[M-2H+K]^-$	1.1	293.0990/130.0876/ 277.0839/278.0763/ 219.0772	3.88.E-03	Up	9.23	7.69	3.23	9.55
423.1797	Sophoraflavanone G	$[M-H]^-$	3.9	93.0347/109.0297/ 179.0387/163.0431/ 147.0456	3.87.E-01	Up	1.57	1.79	1.70	1.73
104.0355	Serine*	$[M-H]^-$	1.9	-	1.90.E-02	Up	2.20	1.96	1.45	2.46
118.0511	Threonine*	$[M-H]^-$	1.0	-	6.53.E-03	Up	2.78	2.46	1.38	3.47
239.0773	Galactose	$[M-H+HAc]^-$	0.1	179.0565/161.0458	1.32.E-02	Down	1.01	0.92	0.99	0.95
274.0120	6-Pyruvoyltetrahydropterin	$[M-2H+K]^-$	1.2	192.0903/162.0489	2.36.E-02	Down	0.66	0.66	0.34	0.60
315.0717	3'-Hydroxy-N-methyl-(S)-coclaurine	$[M-H]^-$	1.4	108.02183/122.0375 /191.0955	5.19.E-03	Up	1.32	0.87	1.12	1.18
341.1092	Trehalose	$[M-H]^-$	2.3	101.0245/113.0246/ 119.0305/143.0349/ 161.0456	3.68.E-03	Down	0.87	0.77	0.81	0.69
377.0857	1-O-Feruloylglucose	$[M-2H+Na]^-$	0.7	179.0565/149.0457/ 133.0509	5.37.E-01	Down	0.87	0.80	0.80	0.68
402.1047	2,4-Dihydroxy-7,8-dimethoxy-2H-1,4-benzoxazin-3(4H)-one 2-glucoside	$[M-H]^-$	1.3	179.0565/170.0438	5.50.E-01	Down	0.59	0.45	0.13	1.89
404.1370	Dimethylsulfoniopropanoate*	$[3M-H]^-$	0.8	133.0338	3.23.E-01	Down	1.19	0.74	1.00	0.97
683.2288	Resveratrol	$[3M-H]^-$	0.3	185.0935/227.0647/ 183.0180/159.9910/ 143.0349/157.0987	1.32.E-01	Up	3.32	2.67	1.59	3.63
128.0356	5-Oxoproline*	$[M-H]^-$	2.1	-	3.89.E-03	Up	3.61	2.97	1.13	4.56
329.0874	1-O-vanilloyl-beta-D-glucose	$[M-H]^-$	1.1	179.0565/149.0457/ 165.0406/93.0347	2.92.E-02	Up	9.52	7.42	0.93	61.69
743.2493	Cellobiose	$[2M-H+HAc]^-$	4.1	341.1093/179.0565/ 149.0457/133.0509	1.75.E-01	Down	1.20	0.74	1.01	0.98

* Metabolites identified with less than four IPs.

Table S4. AIF identification results. Metabolites identified only for Cu(II).

Exact mass	Compound Name	Ion Assignment	Rel. mass error (ppm)	AIF (*m/z*)	corrected *p*-value	Up/Dow-regulated	Fold Change Control vs			
							10	50	100	1000
409.2365	1-Palmitoylglycerol 3-phosphate	$[M-H]^-$	1.0	152.9959/255.2329/171.0065	1.35E-04	Down	0.27	0.25	0.19	0.17
431.2209	6-Aminopenicillanate	$[2M-H]^-$	4.4	215.1078/171.0066	4.893E-05	Down	0.45	0.31	0.14	0.14
433.2365	LPA(0:0/18:2(9Z,12Z))	$[M-H]^-$	1.1	278.2207/153.9993/169.9923/416.2294	2.20E-05	Down	0.43	0.33	0.14	0.14
459.2337	Fusicoplagin A*	$[M-2H+Na]^-$	4.8	347.2593	4.90E-05	Down	0.11	0.03	0.07	0.03
503.2417	PG(18:4(6Z,9Z,12Z,15Z)/0:0)	$[M-H]^-$	1.4	152.9959/90.0350	1.990E-05	Down	0.70	0.41	0.24	0.24
134.0474	Adenine	$[M-H]^-$	1.5	106.0664/107.0365	3.97E-03	Down	0.52	0.52	0.43	0.42
179.0387	Methyl-5-thio-D-ribose	$[M-H]^-$	2.2	161.0403/144.0305/130.0229/116.0072/102.0279	7.104E-03	Up	2.03	1.85	2.21	1.15
226.9966	3-Dehydroquinate	$[M-2H+K]^-$	1.0	127.0403/171.0302/109.0295/153.0194/125.0245/189.0199/117.0348	1.260E-04	Up	1.80	1.65	1.62	2.51
266.0895	Adenosine	$[M-H]^-$	0.0	134.0474/107.0365	1.37E-03	Down	0.52	0.45	0.41	0.40
267.1084	2'-Deoxyribose	$[2M-H]^-$	0.5	133.0509/103.0403	5.405E-04	Up	24.43	22.25	30.03	14.04

299.0776	1R,6R)-6-Hydroxy-2-succinylcyclohexa-2,4-diene-1-carboxylate	$[M-H+HAc]^-$	1.1	239.0561/137.0246/101.0246/222.0532/194.0588	3.914E-07	Up	0.80	1.36	5.65	5.64
312.0954	5-(3'-Carboxy-3'-oxopropenyl)-4,6-dihydroxypicolinate	$[M-H+HAc]^-$	1.3	99.0088/151.0262/219.0161	1.936E-03	Down	0.48	0.42	0.34	0.34
430.1568	Oxynarcotine	$[M-H]^-$	4.6	137.0246/196.0986/193.0511	9.E-01	Up	1.01	1.20	1.14	1.29
110.0251	3,4-Dihydroxypyridine *	$[M-H]^-$	3.5	-	5.E-01	Down	0.68	2.09	0.64	0.69
113.0245	Prop-2-ynal *	$[M-H+HAc]^-$	1.0	-	1.E-01	Down	0.63	0.86	0.72	0.74
242.0800	Pantothenol	$[M-2H+K]^-$	0.2	102.0562/126.0938	1.60E-03	Up	1.44	1.45	1.34	1.62
316.1173	Glycerophosphocholine *	$[M-H+HAc]^-$	2.0	199.0341	4.36E-03	Up	1.46	1.47	1.37	1.62
489.1643	Phloroacetophenone 6'-[xylosyl-(1->6)-glucoside]	$[M-H]^-$	4.9	195.0615/179.0561/149.0457/165.0406	6.E-01	Up	1.30	1.28	0.90	1.05
175.0251	Ascorbic acid	$[M-H]^-$	1.8	113.02462/157.03676/115.07661/139.08801/130.28988	2.686E-05	Up	5.62	7.40	36.71	22.14
336.0875	S-(Hydroxymethyl)glutathione	$[M-H]^-$	1.3	320.0678/319.0844/305.0689/277.0738/185.0571	3.201E-07	Up	3.65	5.67	13.53	9.87

* Metabolites identified with less than four IPs.

4.3. Publicació 4

Untargeted lipidomic evaluation of hydric and heat stresses on rice growth.

Navarro-Reig, R. Tauler, G. Iriondo-Frias, J. Jaumot.

Enviat

Untargeted lipidomic evaluation of hydric and heat stresses on rice growth.

Meritxell Navarro-Reig*, Romà Tauler, Guillermo Iriondo-Frias and Joaquim Jaumot

Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

Abstract

Environmental stresses are the major factors that limit the geographical distribution of plants. As a consequence, plants have developed different strategies to adapt to these environmental changes among which can be outlined the maintenance of membranes' integrity and fluidity. Lipids are key molecules for this environmental adaptation and a comprehensive understand of the molecular mechanisms underlying is still required. Here, lipidome changes in Japanese rice (*Oryza sativa var. Japonica*) upon heat and hydric stresses are assessed using an untargeted approach based on liquid chromatography coupled with mass spectrometry (LC-MS). The obtained data were analyzed using different multivariate data analysis tools. A total number of 298 lipids responded to these abiotic stresses, and 148 of them were tentatively identified. Diacylglycerols (DAG), triacylglycerols (TAG), phosphatidylcholines (PC) and phosphatidylethanolamines (PE) were the most altered lipid families heat and hydric stress. Interpretation of the obtained results showed relevant changes related to the unsaturation degree in the identified lipids. In the case of heat stress, a decrease in the unsaturation degree of lipids can be linked to an increase in the cell membranes' rigidity. In contrast, the hydric stress produced an increase in the lipids unsaturation degree causing an increase in the cell membranes' fluidity, in an attempt to adapt to these non-optimal conditions.

Keywords: Heat stress, hydric stress, environmental changes, lipidomics, rice, multivariate data analysis.

1. Introduction

Most plants are subjected to a range of stresses in their natural environment that affect their growth and development[1]. Environmental stresses, including heat and hydric stresses, are the primary limiting factors of the geographical distribution of plants[2-4]. Heat stressing conditions are expected to increase in the near future because of the climate change. For instance, mean surface air global temperatures have been predicted to increase from 1.3 to 3.1 ºC by the end of the 21st century[5]. Therefore, global warming is expected to produce increasing day-maximum and night-minimum temperatures[2,6,7]. On the other side, water deficit is one of the major environmental stresses constraining plant growth, and it is predicted to be the principal abiotic factor affecting global crop yields[1,8,9]. Climatic change models predict that drought episodes could become more frequent and severe in the near future. Linking both stresses, global warming and water deficit effects, the increment of evapotranspiration will be intensified, increasing the frequency and intensity of droughts from 1% to 30% in the extreme drought land areas by the end of this century[10]. Therefore, understanding the responses of edible plants to heat and hydric stresses is essential to anticipate the impacts of climate change on crops. Rice (*Oryza sativa L.)* is a plant of remarkable alimentary and economic importance, being one of

the most consumed cereals in the world[11]. Together with wheat and maize, they supply more than 50% of calories consumed by the human population. A total of 154 million hectares of rice are harvested each year, and human consumption accounts for 85% of total production of rice[11-13]. The cultivar *Japonica Nipponbare* was selected as a model plant specimen for this work, since it is one of the best-known rice varieties, with a relatively short period of growth and easy genetic modification[12-14]. The effects of heat and hydric stresses have been previously studied in some plant species, such as wheat (*Triticum aestivum L.*)[2], *Arabidopsis thaliana*[1,4,9,15] or soybean (*Glycine max L.*)[8]. However, to the authors' knowledge, this is the first simultaneous study of the effects of these two environmental stresses on the rice lipidome.

Plants have developed multiple strategies to adapt to these environmental stresses. The maintenance of the integrity and fluidity of membranes is of fundamental importance from physiological, biochemical and molecular levels[4]. Taking into account the major molecules present in membranes, lipids have a crucial role in cell, tissue and organ physiology. Stress-induced lipid peroxidation and other changes in membrane lipid profile can lead to membrane damage, electrolyte leakage and cell death[1,2,4]. In this context, lipidomics appears as a powerful tool to evaluate the effects of heat and hydric stresses on plants. Lipidomics is a branch of metabolomics consisting in the study of lipids, molecules with which lipids interact, and their function within the cell[16-19]. Therefore, lipidomics consists of the comprehensive analysis of all lipids on a biological system and the assessment of alterations in lipid content and composition in response to external perturbations. There are two principal lipidomics type of approaches: targeted and untargeted[16-19]. The targeted approach is only focused on the study of a specific list of lipids, typically of the same class or subclass, in an attempt to validate a previous hypothesis. In contrast, untargeted lipidomics aims to screen the entire lipidome content of an organism without any preliminary information[16,18,19]. In this work, the untargeted approach is preferred because it enables the simultaneous profiling of a very large number of lipids present in a particular biological system without prior assumptions, providing the possibility of elucidating lipid species associated with previously unexplored biological pathways.

Considering the vast lipids structural diversity and the high complexity of samples analyzed in plant lipidomics, analytical techniques used in this field should have high separation power[20]. Reverse phase liquid chromatography coupled with high resolution mass spectrometry (RPLC-HRMS) is the most currently used analytical approach in lipidomics. RPLC allows the analysis of a wide range of lipids with high resolution and good reproducibility[16-20]. HRMS-based techniques offer high sensitivity and resolution for the characterization of lipids. Moreover, HRMS has very powerful identification ability[16,17,19]. However, when this analytical methodology is applied to complex lipid plant extracts, massive amounts of MS data are generated. Processing of these data is a challenging task and a crucial step of the whole analytical process[21]. The application of multivariate data analysis tools is opening new ways in omics sciences, allowing for the reliable evaluation of a large number of lipids concentration changes among samples and establishing their biological relationships[22].

The main aim of this work was the investigation of the effects of heat and hydric stresses on the lipidome of Japanese rice (*Oryza sativa var. Japonica*). With this aim, an untargeted LC-MS

lipidomic analysis has been performed to extract maximum information about the investigated stressing effects and found a biochemical interpretation.

2. Materials and Methods

2.1. Chemicals and Reagents

LC-MS grade water, methanol (MeOH, LC-MS grade), methyl tert-butyl ether (MTBE), ammonium formate (≥99.0%) and formic acid (≥95.0%) were supplied by Sigma-Aldrich (Steinheim, Germany).

Eight lipid standards from different families were used as surrogates: 17:0 monoacylglycerol, 17:1 lyso phosphatidylethanolamine, 17:0 lyso phosphatidylcholine, 1,3-17:0 D5 diacylglyceride, 17:0 cholesteryl ester, 1,2,3-17:0 triglyceride, 16:0 D31-18:1 phosphatidylcholine, 16:0 D31-18:1 phosphatidylserine. Three sphingolipids were used as internal standards: N-dodecanoylsphingosine, N-dodecanoylglucosyl-sphingosine and N-dodecanoylsphingosylphosphorylcholine. All these lipid standards were obtained from Avanti Polar Lipids (Alabaster, AL, USA).

Water used for plant watering was purified using an Elix 3 coupled to a Milli-Q system (Millipore, Belford, MA, USA), and filtered through a 0.22 µm nylon filter integrated into the Milli-Q system.

2.2. Plant growth and sample preparation

Rice seeds, obtained from the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain), were incubated for two days at 30 ºC in a wet environment. After this period, plants were grown on an Environmental Test Chamber MLE-352H (Panasonic®) for 22 days simulating cyclic environmental changes of temperature, relative humidity, and light intensity, as shown in Figure 1A. In this work, effects of heat and hydric stresses on rice lipidome were assessed. With this aim, two rice crops were performed, one at standard temperature conditions (daily range from 22 to 28 ºC) and the other at higher temperature conditions (daily range from 25 to 31 ºC). Each crop consisted of 20 trays of samples, and each tray contains nine pots of a single rice plant. The first ten days of growth, all trays were watered three times per week with 150 mL of Milli-Q water. After this period, to study the effects of the watering factor, the next ten days of growth samples were watered with four different levels of Milli-Q water (five trays per level): 150 (control), 100, 50 and 5 mL. Finally, the capacity of rice plants to recover from watering stress was also evaluated. With this purpose, two trays of rice plants of each watering treatment level (100, 50 and 5 mL) were watered with 150 mL of Milli-Q water (control level) during the last two days of growth. The other three trays of rice plants were watered according to the corresponding watering treatment (100, 50 and 5 mL). Figure 1B shows this experimental design schematically. After harvest, aerial parts and roots were separated and frozen at liquid nitrogen temperature for metabolism quenching. Then, samples were stored at -80 ºC until extraction. Four biological replicates were made for each sample condition, and four quality control samples were also prepared for aerial part and root samples. Therefore, a total number of 136 samples (68 aerial part samples and 68 root samples) were analyzed.

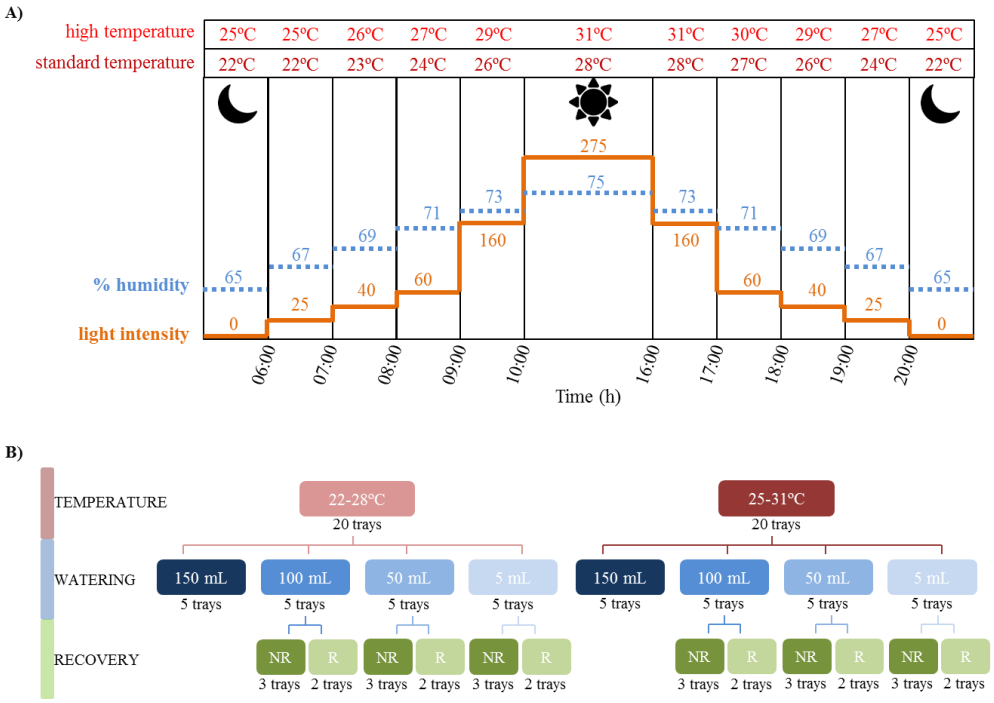


Figure 1. A) Experimental temperature, relative humidity and light long-day rice cultivation conditions at the growth chamber. B) Experimental design. The two levels of the temperature factor (standard and high) appear in red boxes. Blue boxes contain the four levels of watering factor (150, 100, 50 and 5 mL). Finally, the two levels of recovery factor are represented in green boxes: samples watered at control level (150 mL) during the two last days of growth (recovered, R) and samples watered at treatment levels (100, 50 and 5 mL) until harvest (not recovered, NR).

Before extraction, rice samples were ground to a fine powder using a liquid nitrogen mortar and lyophilized for 24 hours to dryness. Lipid extraction was performed using a procedure described elsewhere[23]. Briefly, 5 mg of sample was dispersed in 1 mL of MTBE:MeOH (3:1) in a 2.0 mL Eppendorf tube. The mixture was fortified with 10 µL of the surrogates mix, and then, it was vortexed for 1 min and sonicated for 10 min. Then, 0.5 mL of H_2O:MeOH (3:1) were added, and the mixture was again vortexed for 1 min. After centrifuging for 5 min at 2000 x g, the organic fraction (upper) was collected and transferred to a 1.5 mL Eppendorf tube. The aqueous phase (lower) was re-extracted with 0.65 mL of MTBE and 0.35 mL of MeOH:H_2O (1:0.85). Next, the mixture was vortexed for 1 min and centrifuged for 5 min at 2000 x g. After that, combined organic phases were evaporated to dryness under nitrogen gas. All of the extracts were stored at -80ºC until analysed. Before injection, extracts were reconstituted with 250 µL of MeOH:H_2O (4:1) and 10 µL of the internal standards mix was added.

2.3.LC-MS analysis

Chromatographic separation was performed on an Acquity UHPLC system (Waters, Milford, MA, USA), using a procedure described elsewhere[24]. LC column was a Kinetex EVO C8 (100 x 2.1 mm i.d.; 1.7 µm) provided by Phenomenex (Torrance, CA, US). Elution gradient was performed using solvent A (MeOH 1 mM ammonium formate and 2% formic acid) and solvent B (H_2O 2 mM ammonium formate and 2% formic acid) as follows: 0-3 min, linear gradient from 80 to 90% A; 3-6 min, isocratic gradient at 90% A; 6-15 min, linear gradient from 90 to 99% A; 15-18 min, isocratic gradient at 99% A, and 18-20 min back to initial conditions at 80% A. The mobile phase flow rate was 0.3 mL·min^{-1}, the column temperature

was set at 30ºC and the injection volume was 10 µL.

The mass spectrometer was an LCT Premier XE-time-of-flight (TOF) analyzer (Waters, Milford, MA, USA) equipped with an electrospray (ESI) as ionization source working in both negative and positives modes. Nitrogen (purity >99.98 %) was used as cone and desolvation gas at flow rates of 50 and 600 $L \cdot h^{-1}$, respectively. Desolvation temperature was set at 350 ºC, and electrospray voltages were set at 3.0 kV (positive mode) and at 2.2 kV (negative mode). The mass acquisition range was 50-1500 Da.

2.4. Data analysis

2.4.1. Lipid resolution by ROIMCR procedure

The pure elution profiles and mass spectra of the lipids contained in the analyzed rice samples were resolved using the ROIMCR procedure. This procedure uses the regions of interest (ROI) strategy to compress the LC-MS data and the multivariate curve resolution alternating least squares (MCR-ALS) method for the resolution of pure component contributions present in unresolved complex mixtures. ROIMCR procedure has been already described in the literature[21] and it is only briefly explained here focusing on the particular case of untargeted LC-MS lipidomics study.

First, the ROI approach allowed for the selection of the most interesting mass traces, which meant those *m/z* values whose intensity signals were higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and appeared a minimum number of times consecutively in the time direction[21]. The required parameters for the implementation of ROI approach were the SNR_{Thr} (set at 0.1% of the maximum MS signal intensity of each sample), the mass accuracy of the spectrometer (set at 0.05 Da/e for the TOF MS analyzer used in this work) and the minimum number of consecutive retention times (data points) to be considered as a chromatographic peak (set at 20). Using this approach, a new data matrix containing the intensities at all retention times (rows) only for a reduced number of ROI *m/z* values (approximately 500 columns) was finally stored for each sample. Every ROI data matrix was normalized to correct for instrumental intensity drifts among different sample injections. This normalization was done by dividing all the measured MS intensity values of every data matrix by the mean of the chromatographic area of the seven surrogates and of the three internal standards added to the metabolite extract of that sample.

Then, MCR-ALS allowed the resolution of pure elution profiles and pure mass spectra of the lipids present in rice samples[25-27]. The successful application of MCR-ALS on LC-MS in omic studies for the resolution of coeluted peaks has already been reported[24,28-31]. Relative quantitative information about the different lipids in the 68 analyzed samples can be obtained from the chromatographic peak areas of the resolved elution profiles. On the other hand, the resolved mass spectra were used for tentative identification of lipids.

More details regarding the ROIMCR procedure can be found in the Supporting Information and the reference therein.

2.4.2 Statistical evaluation of the changes in the lipid peak areas

Chromatographic peak areas of the resolved lipids were directly obtained from the MCR-ALS profiles. These areas were arranged in a new data matrix (**A**), where the areas of every component are in the columns and the different samples in the rows. This peak areas data matrix was further analyzed to assess the effects of experimental factors on the concentration changes of the lipids of the investigated rice (aerial and root) samples

using ANOVA-simultaneous component analysis (ASCA)[32], principal component analysis (PCA)[33] and partial least squares – discriminant analysis (PLS-DA)[34].

ASCA was applied to statistically assess the significance of temperature, watering and recovery factors. ASCA was performed on a well-balanced experimental design, and the permutation test consisted on 1000 permutations for each model. PCA was used to explore the behaviour of rice samples under the effects of the experimental factors studied in this work (temperature, watering and recovery).

Finally, PLS-DA was used to evaluate the effects of experimental factors (temperature, watering and recovery). In this work, PLS-DA had been used to discriminate between: samples grown up at standard and high temperature (temperature factor); samples watered with 150, 100, 50 or 5 mL of Milli-Q water (watering factor); and, finally, samples grown up with or without the recovery step (recovery factor). In addition to sample class classification, PLS-DA provides information about which are the most relevant variables (resolved lipids) for the samples discrimination using methods such as the variable importance on projection (VIP) scores method[35]. Finally, the discrimination power of PLS-DA models can be assessed by using the Mathews correlation coefficient (MCC), which measures the quality of binary classifications of belonging or not to a particular sample class (environmental conditions or treatments). MCC can vary from -1 to 1, with values closer to 1 indicating good predictions and values closer to -1 showing bad predictions[36].

An extended description of ASCA, PCA and PLS-DA can be found in SI. Lipid peak areas were mean centered before applying ASCA and autoscaled (mean-centered and scaled) prior to the application of PCA and PLS-DA. ASCA, PCA and PLS-DA methods All these procedures were performed using the PLS Toolbox 8.0.2 (Eigenvector Research Inc, Wenatchee, WA, USA) working under MATLAB 2015b. MCR-ALS analyses were carried out using the MCR-ALS 2.0 toolbox available at www.mcrals.info.

3. Results and Discussion

3.1. Chemometric analysis of rice LC-HRMS lipidomics data

Untargeted LC-HRMS lipidomics provides highly complex data, with large amounts of both useful and meaningless information. Here, the ROI approach was used to compress the amount of information present in the experimental data and to select the more relevant one to simplify the analysis. Using this method, a relatively low number of high resolution *m/z* values was finally taken into account (approximately 500), with more than 100-fold computer storage reduction without loss of spectral accuracy. After this ROI compression, four column-wise augmented data matrices were obtained: 1) augmented data matrix of the aerial part samples analyzed in positive mode ($\mathbf{D}_{\mathbf{augAP}}$); 2) augmented data matrix of the aerial part samples analyzed in negative mode ($\mathbf{D}_{\mathbf{augAN}}$); 3) augmented data matrix of the root samples analyzed in positive mode ($\mathbf{D}_{\mathbf{augRP}}$); and 4) augmented data matrix of the root samples analyzed in negative mode ($\mathbf{D}_{\mathbf{augRN}}$). Each one of these four column-wise augmented data matrices had 42679 rows (total number of retention times) and 528, 622, 466 and 401 columns (number of *m/z* ROI values) respectively. As an example, Figure 2 shows LC-HRMS chromatograms after ROI compression for aerial part samples analyzed in negative mode ($\mathbf{D}_{\mathbf{augAN}}$ matrix). The visual inspection of the obtained chromatograms showed some signals that already allowed the preliminary differentiation of samples according to the

temperature conditions of each crop. However, these LC-HRMS chromatograms presented complex profiles with multiple coeluted compounds (see the zoomed sample in Figure 2). Therefore, the direct detection and identification of the rice lipids changing their concentration at different environmental conditions from ROI compressed data were not straightforward. For this reason, advanced data analysis methods, such as MCR-ALS, were needed to get a more in-depth insight into this experimental lipidomic data.

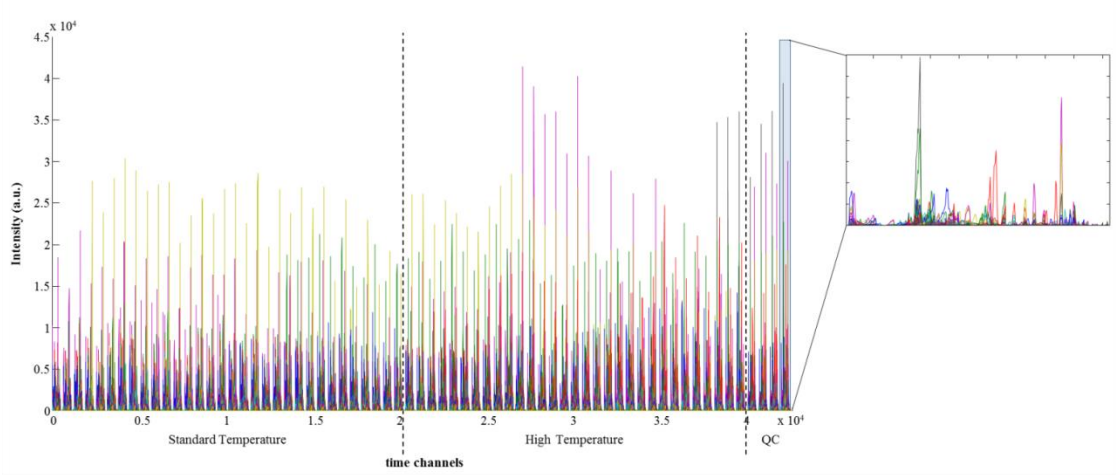


Figure 2. LC-HRMS ROI chromatograms for 68 multiple rice samples under different environmental conditions giving the D_{augAN} column-wise augmented data matrix. On the left side of the plot, the 32 samples grown up at standard temperature conditions, on the right side the 32 samples grown up at high-temperature conditions and on the extreme right side are the four additional quality control (QC) samples. The zoomed view of one QC sample is depicted.

3.1.1. MCR-ALS resolution of rice LC-HRMS lipidomics data

MCR-ALS was applied to the four augmented matrices corresponding to aerial part or root samples and MS ionization modes ($\mathbf{D_{augAP}}$, $\mathbf{D_{augAN}}$, $\mathbf{D_{augRP}}$ and $\mathbf{D_{augRN}}$). These four matrices were resolved by an MCR-ALS model using approximately 200 components. This large number of components included all detected lipid contributions as well as other noisy chromatographic signals, such as instrumental background and solvent contributions. In the case of $\mathbf{D_{augAP}}$, the percentage of explained variance (R^2) was 99.8%, and the lack of fit (LOF) was 4.2%. $\mathbf{D_{augAN}}$ was resolved with R^2 of 98.4% and LOF equal to 12.5%. For $\mathbf{D_{augRP}}$ matrix, the obtained MCR-ALS model had R^2 equal to 99.9% and LOF equal to 3.3%. Finally, in the case of $\mathbf{D_{augRN}}$ matrix, R^2 was 99.7% and LOF was equal to 5.2%.

As an example, Figure 3 shows the MCR-ALS resolution of the pure elution and spectral profiles of a reduced group of four coeluted lipids in aerial part samples analyzed in positive mode ($\mathbf{D_{augAP}}$ matrix). Figure 3A depicts the elution profiles of the four metabolites in the 68 aerial part samples, the zoomed view of one sample of the crop at high-temperature conditions shows the strong coleution of four different lipids. The pure mass spectra of each lipid are displayed in Figure 3B.

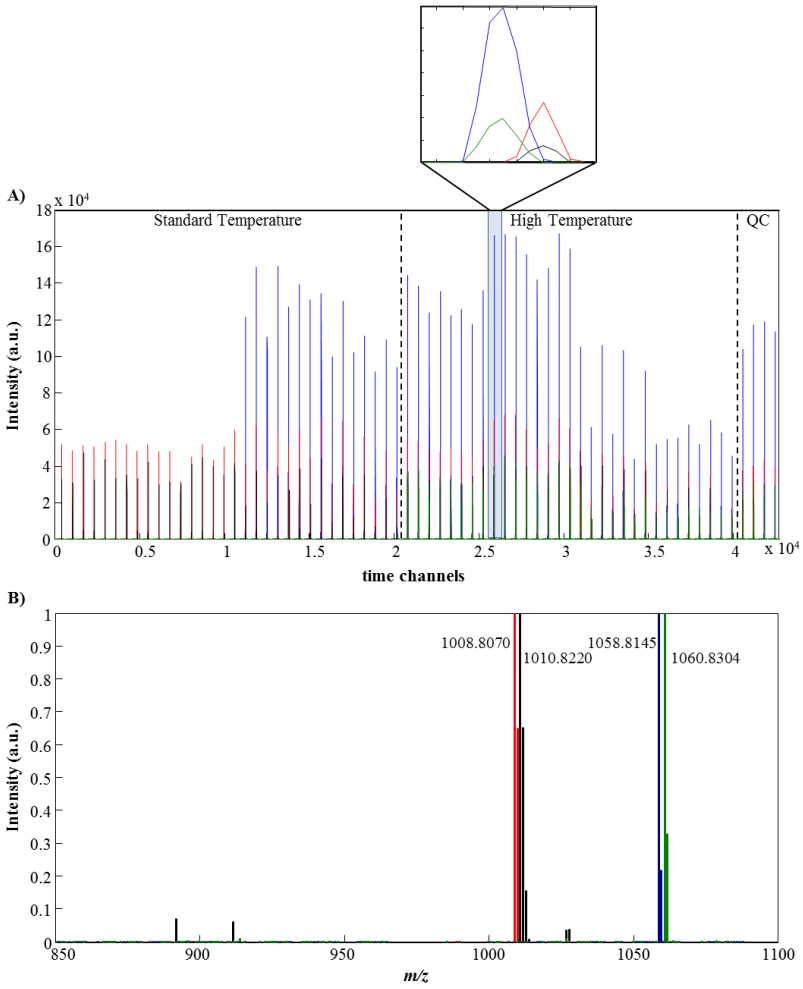


Figure 3. Example of MCR-ALS results in the analysis of LC-HRMS lipidomics data. A) MCR-ALS resolved elution profiles for four coeluted lipids. An example of one rice sample under high-temperature conditions is depicted in the zoomed view B) MCR-ALS resolved mass spectra for the four lipids.

3.1.1. Statistical evaluation of the changes in the lipid peak areas

The effects of the experimental factors (temperature, watering and recovery) were assessed using the chromatographic peak areas of the resolved lipid elution profiles. Four data matrices with the chromatographic peak areas of the lipids in the different type of samples and treatments were considered: aerial part samples analyzed in positive mode ($\mathbf{A_{AP}}$); aerial part samples analyzed in negative mode ($\mathbf{A_{AN}}$); root samples analyzed in positive mode ($\mathbf{A_{RP}}$); and root samples analyzed in negative mode ($\mathbf{A_{RN}}$).

A three-way ASCA model with interactions was applied to every peak areas data matrix to study the statistical significance of the three factors. The results obtained were the same for the four cases (aerial parts or roots samples and negative or positive mass ionization modes), both temperature and watering factors were significant with a *p*-value lower than 0.05 (see Table S1 in SI). On the contrary, the recovery factor and the three binary interactions (watering × temperature, temperature × recovery and watering × recovery) were not significant (*p*-value > 0.05) (see Table S1 in SI).

PCA was also applied to each peak area data matrix to explore the adaptation of rice samples under the effects of the studied experimental factors. In all the cases, effects of the temperature factor were visible. In the four PCA models, using

first and second principal components (PCs), samples of the crop at standard temperature conditions were distinguished from samples of the crop at high-temperature conditions. For instance, in in the case of aerial part samples analysed in negative mode, PCA scores plot shows this sample distinction in the two first PCs (Figure 4A), which already explained the 62% of all data variance. After that, the effects of hydric stress and recovery factors were independently investigated for low and high-temperature samples. Differentiation of samples according to the level of watering treatment was only achieved for aerial part samples, but not for root samples. For example, Figure 4B depicts the PCA scores plot for aerial part samples at high-temperature conditions. A clear arc-shaped continuous trend combining PC1 and PC2 can be distinguished from 5 mL watered samples without recovery to control samples (150 mL). It should be highlighted that in aerial part samples the effect of recovery is observed at the higher level treatments (5 and 50 mL), but not at the lowest level treatment (100 mL). This can be observed in the scores plot of Figure 4B. Samples watered with 5 and 50 mL and recovery (150 mL) during the last two days of treatment (coded 5R and 50R) are more similar to the samples treated at a lower level. For instance, 5R samples can be distinguished from 5 mL watered samples and they are similar to samples watered with 50 mL without the recovery step. The same tendency is observed considering 50R samples (50 mL with the recovery step) which are in an intermediate position between 50 mL watered samples and 100 mL watered samples. In contrast, almost no differences can be observed when 100 mL watered samples with and without recovery are considered. Considering this trend, the final recovery step applied to some samples allowed the plant lipidome to be partially recovered from the hydric stress. However, this recovery ability was not complete as recovered samples did not arrive to behave as 150 mL watered samples.

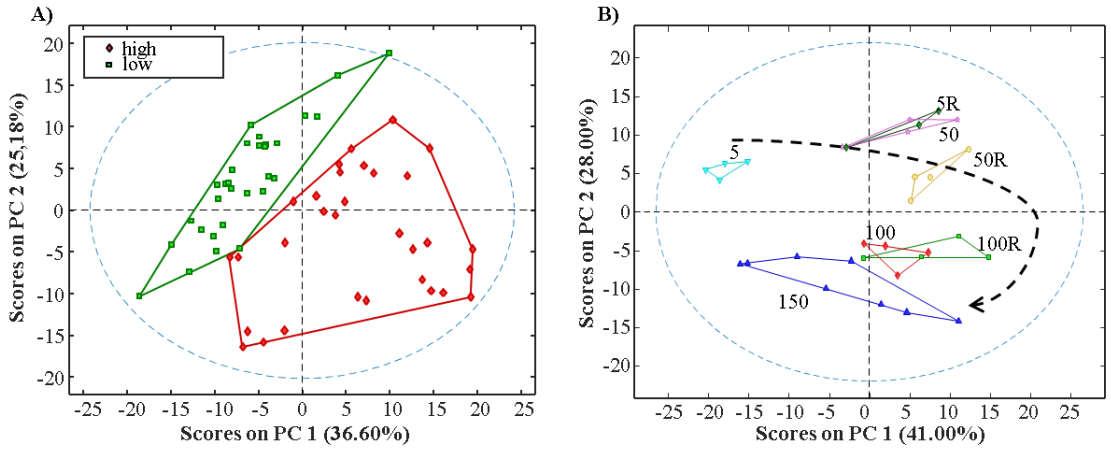


Figure 4. PCA results for aerial part samples analyzed in negative mode ($\mathbf{A}_{AN}$ matrix). A) PCA scores plot for the 64 aerial part samples. Green squares are those samples from the rice crops at standard temperature conditions and red diamonds represent those samples from the rice crops at high-temperature conditions. B) PCA scores plot for 32 aerial part samples from the rice crops at high-temperature conditions. Cyan triangles are 5 mL samples (5), green diamonds are 5 mL with recovery (5R), purple stars are 50 mL samples (50), yellow rounds are 50mL with recovery (50R), red diamonds are 100 mL samples (100), green squares are 100 mL with recovery (100R) and blue triangles are 150 mL samples (control,150).

PLS-DA was applied to complement these results and identify the most significant lipids changing their concentrations because of the experimental conditions. Table S2 gives the values of the Mathews correlation coefficient (MCC) for each of the PLS-DA models obtained in the analysis of each type of samples. Models differentiating samples of the crop at standard temperature conditions from the ones grown up at high-temperature conditions had the best MCC values, between 0.93 and 1.00, which indicated a clear differentiation between these two groups of samples. On the contrary, models discriminating samples with and without the recovery step showed the worst MCC values, between 0.73 and 0.76, indicating that these samples showed minor differences in their lipidic composition and cannot be perfectly distinguished. Considering these MCC values, the recovery step could be disregarded as a factor to identify potential biomarkers as the discrimination models were not good enough. These results are consistent with those previously obtained in ASCA and PCA data analysis. In the case of the watering factor, PLS-DA models were independently built for high and standard temperature samples. Then, the comparison of control samples with each level of water treatment (100, 50 and 5 mL) was performed. As observed in PCA, the models obtained for aerial part samples discriminated better control and treated samples than the models obtained for root samples. This fact indicated that concentration changes in lipids in the aerial parts were larger than in roots. The MCC values in the case of the aerial part sample were all equal to 1.00 (perfect discrimination) except the one differentiating controls and the lowest level of treatment (100 mL) samples analyzed in positive mode. Whereas, the MCC values obtained for the PLS-DA models of root samples were only equal to 1.00 for the models discriminating controls and the highest level of treatment (5 mL) samples. These results can be associated with major effects of hydric stress in the aerial parts than in roots as samples showed differences in the ability of models to discriminate between them.

Finally, VIP scores of the PLS-DA models were used as feature selection tool in order to select the most relevant resolved lipids for the two significant factors, temperature, and watering. Selected lipids where those with a VIP value higher than 2. In the case of temperature factors, the number of selected lipids for RP, RN, AP and AN were 32, 37, 40 and 27. For the watering factor at standard temperature conditions, the number of selected lipids for RP, RN, AP and AN were 40, 43, 35 and 42. Finally, at high-temperature conditions, the number of selected lipids for RP, RN, AP and AN were 60, 49, 44 and 35. The selected lipids were tentatively identified by comparing their resolved accurate mass value with theoretical exact mass values included in public databases.

From the 298 lipids selected using the VIP scores, 148 were tentatively identified with an error lower than 10 ppm, as recommended for this type of TOF analyzer. Moreover, the chromatographic retention times were also used to identify the lipids by comparison with home-made databases[24,37-39]. Results of this identification are in Tables S3-5 in Supporting Information. The most common families of the identified lipids were: diacylglycerols (DAG), triacylglycerols (TAG), phosphatidylcholines (PC) and phosphatidylethanolamines (PE).

3.2. Biological interpretation of environmental stresses

All these results allowed to identify the main effects caused by these environmental stresses (heat and hydric) on lipid concentrations of rice.

First, in the case of the heat stress, the major part of the lipids selected as VIPs for the aerial part and root samples were upregulated. In both types of samples, glycerolipids (DAG and TAG) were one of the most altered families among the identified lipids. Moreover, and most especially in the case of root samples, the fold changes of these lipids were high (see Table S3). This result can be explained because of cell membranes sensitivity to environmental changes and, as a consequence, adjustments of the glycerolipid composition to maintain the optimal fluidity of membranes (particularly in plasma and chloroplast membranes)[2-4]. At high-temperature, the degree of unsaturation of fatty acids has been shown to decrease[2-4]. In this study, fatty acid 9,10-dihydroxyoctadecanoic (18:0) acid was accumulated in roots with a extremely high fold change (see Table S3). Another evidence of the change in fatty acids under heat stress was that acylcarnitines, for instance the 3-hydroxynonanoyl carnitine, which also appeared with a high fold change in roots (see Table S3). Acylcarnitines are essential compounds for the metabolism of fatty acids. The decrease in the unsaturation degree in lipids was also observed for TAGs and DAGs, in both roots and aerial parts of rice samples. In general, glycolipids with a low number of unsaturations (from 0 to 3), for instance TAG (48:1), were upregulated (fold change equal from 2 to 15). In contrast, glycerolipids with a high number of unsaturations (from, 4 to 10), for example, TAG (52:6), were downregulated (see Table S3). Additionally, glycerophospholipids, in particular PC, PE and phosphatidylinositols (PI), were other upregulated lipid families in both aerial part and roots samples. These glycerophospholipids have been reported to accumulate in *Arabidopsis thaliana* plants under heat stress[3]. This accumulation may be related to changes in glycerolipid synthesis pathway and phospholipid synthesis in the endoplasmic reticulum[3]. Additionally, because of PC, PE and PI are typical lipid membrane constituents, they can also contribute to the modification of membrane fluidity[15].

In the case of the hydric stress, most of the lipids selected with high VIPs values in root samples were downregulated. In contrast, in the case of aerial part samples, there was a different behavior for samples grown up at standard temperature (selected lipids were downregulated) compared to those grown at high-temperature (upregulated). As in the case of heat stress, hydric stress also affects the integrity and fluidity of membranes[1,15]. Plastidic lipids, which mainly include monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG), are the main components of chloroplast membranes[1,15]. This explained the appearance of DGDG (34:2) as a highly altered lipid (large fold change) in root samples raised at standard temperature (see Table S4). PC and PE glycerolipids were also found as altered families in both aerial and roots parts at standard and high-temperatures. These two glycerolipid families are also membrane constituents. Therefore, the observed changes can also be explained as an attempt to control the membrane fluidity[15]. TAGs also appeared as a relevant family of lipids, in agreement with the fact that TAGs have been already reported to accumulate in plants under drought stress[15]. In contrast to the situation under heat stress, the degree of unsaturation has been reported to increase under hydric stress effects [1,15]. In this work, TAGs and DAGs found in samples grown at a standard temperature and with a high number of unsaturations (from 4 to 7 double bonds) were upregulated (see Table S4). In contrast, the highly unsaturated TAGs and DAGs (4 double bonds)

found at roots grown at high-temperature were downregulated (see Table S4) whereas the low unsaturated TAGs and DAGs (0 to 3 double bonds) were upregulated (see Table S5). These results suggest that the effects caused by changes in temperature have larger consequences than those caused by hydric stress. This statement is in agreement with the ASCA results in which the heat stress showed a much more significant effects.

Conclusions

The untargeted lipidomics approach proposed in this work allowed the selection of a large number (298) of rice lipids to be good indicators of heat and hydric stresses. DAG, TAG, PC and PE were the most altered lipid families by heat and hydric stresses. The identified lipids in response to these stresses confirmed that the maintenance of membranes' fluidity is one of the most significant mechanisms for plant adaptation under these environmental stresses. In the case of heat stress, the degree of unsaturation of fatty acids, TAGs and DAGs decreased significantly. This fact could also be directly related to an augment in the rigidity of membranes. On the contrary, under hydric stress, the degree of unsaturation in lipidic molecules increased, especially in the case of fatty acids, TAGs and DAGs. This corroborates an opposite effect with an increment of the membranes' fluidity. Finally, from the simultaneous evaluation of the effects of the two stressing conditions, heat stress seemed to cause a stronger effect because a decrease in the unsaturation degree in lipids was observed.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 320737. The authors would like to thank CRAG for kindly supplying Japanese rice seeds.

Conflict of interest statement

The authors declare that they have no competing interests.

References

(1) Yu, B.; Li, W. *Phytochemistry* **2014**, *108*, 77-86.
(2) Narayanan, S.; Tamura, P. J.; Roth, M. R.; Prasad, P. V. V.; Welti, R. *Plant, Cell and Environment* **2016**, *39*, 787-803.
(3) Higashi, Y.; Okazaki, Y.; Myouga, F.; Shinozaki, K.; Saito, K. *Scientific Reports* **2015**, *5*.
(4) Zheng, G.; Tian, B.; Zhang, F.; Tao, F.; Li, W. *Plant, Cell and Environment* **2011**, *34*, 1431-1442.
(5) Stocker, T. F.; Qin, D.; Plattner, G. K.; Tignor, M. M. B.; Allen, S. K.; Boschung, J.; Nauels, A.; Xia, Y.; Bex, V.; Midgley, P. M. *Climate change 2013 the physical science basis: Working Group I contribution to the fifth assessment report of the intergovernmental panel on climate change*, 2013, p 1-1535.
(6) Easterling, D. R.; Meehl, G. A.; Parmesan, C.; Changnon, S. A.; Karl, T. R.; Mearns, L. O. *Science* **2000**, *289*, 2068-2074.
(7) Harvey, L. D. D. *Nature* **1995**, *377*, 15-16.
(8) Manavalan, L. P.; Guttikonda, S. K.; Phan Tran, L. S.; Nguyen, H. T. *Plant and Cell Physiology* **2009**, *50*, 1260-1276.
(9) Gigon, A.; Matos, A. R.; Laffray, D.; Zuily-Fodil, Y.; Pham-Thi, A. T. *Annals of Botany* **2004**, *94*, 345-351.
(10) Xu, Z.; Zhou, G.; Shimizu, H. *Plant Signaling & Behavior* **2010**, *5*, 649-654.
(11) Gnanamanickam, S. S. In *Biological Control of Rice Diseases*, Gnanamanickam, S. S., Ed.; Springer Netherlands: Dordrecht, 2009, pp 1-11.
(12) Kim, J. K.; Park, S.-Y.; Lim, S.-H.; Yeo, Y.; Cho, H. S.; Ha, S.-H. *Journal of Cereal Science* **2013**, *57*, 14-20.
(13) Aina, R.; Labra, M.; Fumagalli, P.; Vannini, C.; Marsoni, M.; Cucchi, U.; Bracale, M.; Sgorbati, S.; Citterio, S. *Environ. Exp. Bot.* **2007**, *59*, 381-392.

(14) Fukusaki, E.; Kobayashi, A. *J. Biosci. Bioeng.* **2005**, *100*, 347-354.
(15) Tarazona, P.; Feussner, K.; Feussner, I. *Plant Journal* **2015**, *84*, 621-633.
(16) Cajka, T.; Fiehn, O. *TrAC - Trends in Analytical Chemistry* **2014**, *61*, 192-206.
(17) Wolf, C.; Quinn, P. J. *Progress in Lipid Research* **2008**, *47*, 15-36.
(18) Navas-Iglesias, N.; Carrasco-Pancorbo, A.; Cuadros-Rodríguez, L. *TrAC - Trends in Analytical Chemistry* **2009**, *28*, 393-403.
(19) Li, M.; Zhou, Z.; Nie, H.; Bai, Y.; Liu, H. *Analytical and Bioanalytical Chemistry* **2011**, *399*, 243-249.
(20) Bou Khalil, M.; Hou, W.; Zhou, H.; Elisma, F.; Swayne, L. A.; Blanchard, A. P.; Yao, Z.; Bennett, S. A. L.; Figeys, D. *Mass Spectrometry Reviews* **2010**, *29*, 877-929.
(21) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC - Trends in Analytical Chemistry* **2016**, *82*, 425-442.
(22) Checa, A.; Bedia, C.; Jaumot, J. *Analytica Chimica Acta* **2015**, *885*, 1-16.
(23) Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. *Journal of Lipid Research* **2008**, *49*, 1137-1146.
(24) Dalmau, N.; Jaumot, J.; Tauler, R.; Bedia, C. *Molecular BioSystems* **2015**, *11*, 3397-3406.
(25) Jaumot, J.; de Juan, A.; Tauler, R. *Chemometrics Intell. Lab. Syst.* **2015**, *140*, 1-12.
(26) De Juan, A.; Jaumot, J.; Tauler, R. *Anal. Chim. Acta* **2014**, *6*, 4964-4976.
(27) Ruckebusch, C.; Blanchet, L. *Anal. Chim. Acta* **2013**, *765*, 28-36.
(28) Navarro-Reig, M.; Jaumot, J.; García-Reiriz, A.; Tauler, R. *Analytical and Bioanalytical Chemistry* **2015**, *407*, 8835-8847.
(29) Navarro-Reig, M.; Jaumot, J.; Piña, B.; Moyano, E.; Galceran, M. T.; Tauler, R. *Metallomics* **2017**, *9*, 660-675.
(30) Ortiz-Villanueva, E.; Navarro-Martín, L.; Jaumot, J.; Benavente, F.; Sanz-Nebot, V.; Piña, B.; Tauler, R. *Environmental Pollution* **2017**, *231*, 22-36.
(31) Pérez, I. S.; Culzoni, M. J.; Siano, G. G.; García, M. D. G.; Goicoechea, H. C.; Galera, M. M. *Analytical Chemistry* **2009**, *81*, 8335-8346.
(32) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Smilde, A. K. *Anal. Chim. Acta* **2005**, *530*, 173-183.
(33) Wold, S.; Esbensen, K.; Geladi, P. *Chemometrics Intell. Lab. Syst.* **1987**, *2*, 37-52.
(34) Barker, M.; Rayens, W. *J. Chemometr.* **2003**, *17*, 166-173.
(35) Wold, S.; Johansson, A.; Cocchi, M. In *3D QSAR in Drug Design*, Kubiny, H., Ed.; ESCOM Science Publishers: Leiden, 1993, pp 583-618.
(36) Matthews, B. W. *BBA - Protein Structure* **1975**, *405*, 442-451.
(37) Gorrochategui, E.; Casas, J.; Pérez-Albaladejo, E.; Jáuregui, O.; Porte, C.; Lacorte, S. *Environmental Science and Pollution Research* **2014**, *21*, 11907-11916.
(38) Gorrochategui, E.; Casas, J.; Porte, C.; Lacorte, S.; Tauler, R. *Analytica Chimica Acta* **2015**, *854*, 20-33.
(39) Bedia, C.; Dalmau, N.; Jaumot, J.; Tauler, R. *Environmental Research* **2015**, *140*, 18-31.

Informació Suplementària a la Publicació 4

Untargeted lipidomic evaluation of hydric and heat stresses on rice growth.

Navarro-Reig, R. Tauler, G. Iriondo-Frias, J. Jaumot.

Enviat

Description of chemometric tools

LC—MS data preprocessing

Waters raw chromatographic data files (.raw format) were converted to the standard CDF format by the Databridge function of MassLynx^{TM} 4.1 software (Waters, Milford, MA, USA). Then, these data files were imported into MATLAB environment (Release 2015b, The Mathworks Inc, Natick, MA, USA) by using the appropriate functions of the MATLAB Bioinformatics Toolbox (4.3.1 version).

LC-MS data were arranged and aligned according to their *m/z* in a data matrix, containing retention times in the rows and selected *m/z* values in the columns. Here this data matrix was build up using the previously proposed regions of interest (ROI) strategy[1]. This strategy allowed for the selection of the most interesting mass traces, which meant those *m/z* values whose intensity signals were higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and appeared a minimum number of times consecutively in the time direction. The main advantage of this ROI strategy is that it performs a significant compression of the MS data without any loss of chromatographic relevant information nor of MS spectral accuracy. The required parameters for the implementation of ROI approach were the SNR_{Thr} (set at 0.1% of the maximum MS signal intensity of each sample), the mass accuracy of the spectrometer (set at 0.05 Da/e for the TOF MS analyzer used in this work) and the minimum number of consecutive retention times (data points) to be considered as a chromatographic peak (set at 20). Using this approach, a new data matrix containing the intensities at all retention times (rows) only for a reduced number of ROI *m/z* values (approximately 500 columns) was finally stored for each sample. More details about ROI strategy can be found at the work of Gorrochategui *et. al* [1].

An ROI data matrix was obtained for every analyzed lipids sample with a total number of 136 samples analyzed in both MS ionization modes (positive and negative). Every ROI data matrix was normalized to correct for instrumental intensity drifts among different sample injections. This normalization was done by dividing all the measured MS intensity values of every data

matrix by the mean of the chromatographic area of the seven surrogates and of the three internal standards added to the metabolite extract of that sample. After data normalization, column-wise augmented ROI data matrices were built with the individual data matrices corresponding to the aerial and root parts of the rice plant under the different stressing treatments, analyzed in both positive and negative acquisition modes. Since ROI individual data matrices may have a different number of ROI *m/z* values (columns), a preliminary search for common and uncommon ROI *m/z* values among the different samples under the different treatments (68) was performed. When an individual compressed matrix had not a significant intensity of a particular ROI *m/z* value, low random intensity values at the noise level (below SNR_{Thr}) were assigned. In total, four column-wise augmented data matrices containing 68 samples (two for aerial part samples and two for root samples in both positive and negative acquisition modes) were obtained.

MCR-ALS resolution for feature detection

Multivariate curve resolution alternating least squares (MCR-ALS) is a chemometric method used for the investigation and resolution of pure component contributions present in unresolved complex mixtures. In this work, MCR-ALS allowed the resolution of pure elution profiles and pure mass spectra of the lipids present in rice samples. The MCR-ALS method has been already described in the literature[2-4] and it is only briefly explained here focusing on the particular case of untargeted LC-MS lipidomics study.

MCR-ALS decomposes the experimental data sets according to a bilinear model that can be extended to the simultaneous analysis of several samples. For instance, in the case of this work, the four column-wise augmented data matrices (two for aerial part samples and two for root samples in both positive and negative acquisition modes) were analyzed. Each one of these column-wise augmented data matrices ($\mathbf{D_{aug}}$) contained 68 samples and were decomposed by MCR-ALS method using a bilinear model as shown in Equation 1:

$$\mathbf{D_{aug}} = \begin{bmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \mathbf{D_3} \\ \vdots \\ \mathbf{D_{68}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_1} \\ \mathbf{C_2} \\ \mathbf{C_3} \\ \vdots \\ \mathbf{C_{68}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_1} \\ \mathbf{E_2} \\ \mathbf{E_3} \\ \vdots \\ \mathbf{E_{68}} \end{bmatrix} = \mathbf{C_{aug}} \mathbf{S^T} + \mathbf{E_{aug}} \qquad \text{Equation (1)}$$

MCR-ALS decomposition of the matrix $\mathbf{D_{aug}}$ gives $\mathbf{C_{aug}}$ and $\mathbf{S^T}$. $\mathbf{C_{aug}}$ is the matrix containing the resolved elution profiles for each resolved lipid at the 68 samples. From this $\mathbf{C_{aug}}$ matrix, relative quantitative information about the different lipids in the 68 samples can be obtained from the chromatographic peak areas of the resolved profiles. On the other hand, $\mathbf{S^T}$ contains the pure mass spectra of the resolved lipids, which can be used for tentative identification of lipids, by comparing their resolved accurate mass values with theoretical exact mass values included in public databases such as HMDB[5], LipidMaps (http://www.lipidmaps.org) and MetaCyc[6].

More details regarding the MCR-ALS approach and the initialization and constraints for the MCR-ALS optimization can be found in the literature[7-12].

The procedure that combines the ROI data compression strategy described above and the MCR-ALS analysis of the ROI data is called ROIMCR and has been described elsewhere[1].

ASCA

ASCA is a multivariate analysis of variance method that combines the power of ANOVA to separate variance sources with the advantages of simultaneous component analysis (SCA) to the modelling of the individual separate effect matrices. ASCA is especially useful for the analysis of complex multivariate datasets containing an underlying experimental design, and it allows for a natural interpretation of the variance induced by the different factors in the design[13]. For a more detailed description of ASCA procedure and statistical significance of the results based on a permutation test see the works of Smilde[13], Jansen[14] and Vis[15].

PCA

PCA compresses the information of the original variables into a smaller number of uncorrelated variables known as principal components[16]. The representation of these first components both in

samples (scores maps) and variables (loadings) modes are usually useful to explore the main sources of variance in the analysed data.

PLS-DA

PLS-DA is a supervised multivariate regression method oriented to discriminate among different groups of samples. In this method, the lipid peak areas (data matrix **A**, predictor variable) were correlated with the vector describing the sample type class (different temperature and watering environmental conditions) membership (**y**, predicted variable)[17]. In addition to sample class classification, PLS-DA provides information about which are the most relevant variables (resolved lipids) for the samples discrimination using methods such as the variable importance on projection (VIP) scores method[18]. VIP scores measure the importance of each predictor variable into the final PLS model and they are calculated as the weighted sum of the squared variable weights of the PLS model[18]. Since the average of the squared VIP scores is equal to 1, the "greater than one" rule is commonly used as the criterion to identify the most important variables for a given model[19]. Finally, the discrimination power of PLS-DA models can be assessed by using the Mathews correlation coefficient (MCC), which measures the quality of binary classifications of belonging or not to a particular sample class (environmental conditions or treatments). MCC can vary from -1 to 1, with values closer to 1 indicating good predictions and values closer to -1 showing bad predictions[20].

ASCA Results

Table S1. Obtained *p*-values for temperature, watering and recovery factors and their interaction.

	p-value			
	Aerial Part		Root	
	Negative	Positive	Negative	Positive
temperature	0.001	0.002	0.001	0.001
tecovery	0.322	0.232	0.618	0.431
watering	0.002	0.038	0.007	0.001
temperature × recovery	0.240	0.663	0.773	0.917
watering × temperature	0.078	0.310	0.140	0.086
watering × recovery	0.422	0.192	0.832	0.609

Permutation test done using 1000 permutations.

PLS-DA Results

Table S2. Mathews correlation coefficients (MCC) for the PLS-DA model evaluation of watering effects on the concentration changed of lipids in aerial parts and roots of rice

			AP	**AN**	**RP**	**RN**
TEMPERATURE			1.00	1.00	1.00	0.93
WATERING	Low T	CONTROL vs 100 mL	0.88	1.00	0.73	0.73
		CONTROL vs 50 mL	1.00	1.00	0.87	0.96
		CONTROL vs 5 mL	1.00	1.00	1.00	1.00
	High T	CONTROL vs 100 mL	1.00	1.00	0.75	0.88
		CONTROL vs 50 mL	1.00	1.00	0.88	0.98
		CONTROL vs 5 mL	1.00	1.00	1.00	1.00
RECOVERY			0.73	0.74	0.73	0.76

Table S3. Identification of VIP lipids for the temperature factor.

		Exact Mass	Compound name	Family	Ion Assigment	Rel. Mass error (ppm)	Fold Change	Up/Down-regulated
ROOT	ESI +	338.3415	Docosenamide	Carboxylic acid	[M+H]+	1.0	0.31	Down
		413.3763	Stigmasterol	Phytosterol	[M+H]+	3.6	1.22	Up
		466.2931	LysoPC(14:1)	Lyso PC	[M+H]+	0.7	30.49	Up
		483.4600	n. id.	-	-	-	2.36	Up
		496.3398	LysoPC(16:0)	Lyso PC	[M+H]+	0.0	0.38	Down
		629.3144	Gibberellin	Diterpenoid	[2M+H]+	5.5	35.89	Up
		634.5365	DAG(36:3)	DAG	[M+NH4]+	6.5	150.58	Up
		673.4993	PA(34:2)	PA	[M+H]+	11.0	32.25	Up
		675.6760	n. id.	-	-	-	0.11	Down
		690.5043	PA(34:2)	PA	[M+NH4]+	3.7	2.33	Up
		696.4825	PE(32:1)	PE	[M+Na]+	16.2	2.93	Up
		706.6182	PC(O-32:0)	PC	[M+NH4]+	10.0	2.77	Up
		717.5223	CL(70:3)	CL	[M+2H]2+	3.1	0.55	Down
		736.6771	n. id.	-	-	-	0.48	Down
		761.5852	PE(36:2)	PE	[M+NH4]+	6.4	192.11	Up
		771.6180	n. id.	-	-	-	7.37	Up
		785.6150	PC(P-36:3)	PC	[M+NH4]+	2.3	131.83	Up
		786.5970	PC(36:2)	PC	[M+H]+	4.8	5.02	Up
		790.7683	Glycinoprenol 11	Polyterpenoid	[M+NH4]+	14.6	18.71	Up
		791.7739	n. id.	-	-	-	1.41	Up
		797.7441	n. id.	-	-	-	6.64	Up
		826.6736	PC(40:3)	PC	[M+H]+	6.2	0.84	Down
		827.6775	2-Decaprenyl-6-methoxyphenol	Polyterpenoid	[M+Na]+	11.9	66.26	Up
		848.7610	TAG(50:2)	TAG	[M+Na]+	10.8	2.29	Up
		850.5112	CPA(18:2)	PA	[2M+N4H]+	10.0	0.35	Down
		868.7335	TAG(52:6)	TAG	[M+NH4]+	6.1	0.65	Down
		876.7903	TAG(52:2)	TAG	[M+NH4]+	12.7	207.62	Up
		900.7868	TAG(54:4)	TAG	[M+NH4]+	16.2	587.79	Up
		902.8040	TAG(54:3)	TAG	[M+NH4]+	14.5	112.74	Up

		943.8087	TAG(58:6)	TAG	[M+Na]+	0.2	5.42	Up
		1008.8051	n. id.	-	-	-	0.63	Down
		1368.2885	n. id.	-	-	-	33.16	Up
	ESI -	187.0965	cis-3-Hexenyl acetate	Carboxylic acid ester	[M-H+FA]-	5.8	1.22	Up
		297.2425	9.10-dihydroxyoctadecanoic acid	FA	[M-H-H2O]-	1.0	361.78	Up
		367.2304	Chondrillasterol 3-[glucosyl-(1->4)-glucoside]	Steroidal glycoside	[M-2H]-2	1.0	74.47	Up
		430.3617	n. id.	-	-	-	143.23	Up
		452.2773	LysoPE(16:0)	Lyso PE	[M-H]-	2.1	0.31	Down
		515.0954	n. id.	-	-	-	30.85	Up
		516.4182	n. id.	-	-	-	135.34	Up
		540.3295	LysoPC(16:0)	Lyso PC	[M-H+FA]-	2.1	0.32	Down
		607.4175	Peltatol B	Sesquiterpenoid	[M-H-H2O]-	6.1	20.05	Up
		633.4326	3-hydroxynonanoyl carnitine	Acyl Carnitine	[2M-H]-	0.9	100.88	Up
		672.5747	PE(O-32:0)	PE	[M-H-H2O]-	7.0	789.56	Up
		688.4946	PE(32:1)	PE	[M-H]-	3.4	626.05	Up
		690.5003	PE(32:0)	PE	[M-H]-	11.0	4.96	Up
		691.5823	Glycerol 1-(9Z-octadecenoate) 2-octanoate 3-tetradecanoate	Fatty acyl	[M-H]-	8.6	241.75	Up
		698.6204	n. id.	-	-	-	177.06	Up
		699.6246	n. id.	-	-	-	0.30	Down
		710.9396	n. id.	-	-	-	0.20	Down
		714.6505	n. id.	-	-	-	636.70	Up
		716.5106	PE(34:1)	PE	[M-H]-	18.1	35.20	Up
		718.4708	PE(P-34:3)	PE	[M-2H+Na]-	10.0	517.65	Up
		723.5927	22:3 Cholesteryl ester	Steroid ester	[M-2H+Na]-	18.0	29.45	Up
		731.6463	n. id.	-	-	-	79.63	Up
		742.6862	n. id.	-	-	-	4.52	Up
		777.6170	n. id.	-	-	-	0.06	Down
		796.6225	PC(O-36:0)	PC	[M-2H+Na]-	3.0	3.11	Up
		803.7277	TAG(48:1)	TAG	[M-H]-	17.0	2.02	Up
		831.4962	PI(34:3)	PI	[M-H]-	8.0	0.46	Down
		834.7611	n. id.	-	-	-	1.45	Up
		851.6269	TAG(54:10)	TAG	[M-H+-H2O]-	9.0	0.42	Down

		878.6458	n. id.	-	-	3.0	199.54	Up
		889.6943	n. id.	-	-	-	9.11	Up
		971.5462	n. id.	-	-	-	0.22	Down
		1031.7710	22-Acetylpriverogenin B	Triterpenoid	[2M-H]-	14.8	289.51	Up
		1082.7592	n. id.	-	-	-	177.37	Up
		1084.7785	n. id.	-	-	-	3.86	Up
		1371.2709	n. id.	-	-	-	4.58	Up
		1402.2731	n. id.	-	-	-	1.83	Up
AERIAL PART	ESI +	549.4872	Cohibin A	Fatty Alcohols	[M+H]+	0.0	22.04	Up
		571.4721	Myristoylglycine	Amino Acid	[2M+H]+	6.9	73.35	Up
		609.5297	DAG(34:0)	DAG	[M+NH4]+	5.0	657.43	Up
		622.5483	DAG(34:3)	DAG	[M+NH4]+	12.5	3.74	Up
		636.5551	DAG(36:3)	DAG	[M+NH4]+	1.6	78.74	Up
		647.5101	n. id.	-	-	-	5.68	Up
		678.6167	CE (18:3)	CE	[M+H]+	2.0	1.63	Up
		680.6339	Campesteryl linoleate	Steroid ester	[M+NH4]+	0.0	2.16	Up
		690.6216	PG(34:4)	PG	[M+NH4]+	5.7	2.04	Up
		692.6331	CE(20:3)	CE	[M+NH4]+	1.4	1.44	Up
		693.6364	20:1 Campesteryl ester	Steroid ester	[M+H]+	17.0	68.82	Up
		740.5345	PE(36:4)	PE	[M+H]+	12.3	26.85	Up
		747.6027	PE(P-36:1)	PE	[M+NH4]+	2.0	1.65	Up
		756.5580	PC(34:3)	PC	[M+H]+	5.6	5.68	Up
		771.6180	n. id.	-	-	-	529.53	Up
		782.4926	PS(34:2)	PS	[M+Na]+	2.1	0.40	Down
		786.5970	PC(36:2)	PC	[M+H]+	4.8	251.92	Up
		790.7683	Glycinoprenol 11	Polyprenol	[M+NH4]+	14.6	1.31	Up
		808.5560	PC(P-36:2)	PC	[M+K]+	7.0	30.33	Up
		816.7031	TAG(48:4)	TAG	[M+NH4]+	5.5	0.22	Down
		840.5408	PE(42:7)	PE	[M+Na]+	12.6	5.28	Up
		858.7457	TAG(50:4)	TAG	[M+NH4]+	10.3	1.30	Up
		887.5602	PI(36:1)	PI	[M+Na]+	2.0	1.32	Up
		903.5592	LysoPE(16:0)	Lyso PE	[2M+H]+	8.1	944.34	Up
		930.8364	TAG(56:3)	TAG	[M+NH4]+	12.9	3.46	Up
		936.6546	n. id.	-	-	-	9.23	Up

		937.5813	1.2-Di-(9Z.12Z.15Z-octadecatrienoyl)-3-(Galactosyl-alpha-1-6-Galactosyl-beta-1)-glycerol	Glycosyldiacylglycerol	[M+Na]+	7.0	941.57	Up
		938.5826	n. id.	-	-	-	73.56	Up
		941.5498	n. id.	-	-	-	3.62	Up
		944.5410	Oleanolic acid 3-[rhamnosyl-(1->4)-glucosyl-(1->6)-glucoside]	Triterpene saponin	[M+NH4]+	17.7	1.60	Up
		957.5495	n. id.	-	-	-	126.03	Up
		1023.8023	n. id.	-	-	-	344.99	Up
		1049.8093	n. id.	-	-	-	63.90	Up
		1050.8127	22-Acetylpriverogenin B	Triterpenoid	[2M+NH4]+	15.2	121.07	Up
		1051.7188	Oleanderolide 3-acetate	Triterpenoid	[2M+Na]+	2.0	13.11	Up
		1075.7183	3-α-Acetomethoxy-11-α-oxo-12-ursen-24-oic acid	Triterpenoid	[2M+Na]+	2.4	6.95	Up
		1342.9094	Lipid A -disaccharide-1-P	SL	[M+Na]+	8.0	1.27	Up
		1344.9320	n. id.	-	-	-	319.40	Up
		1345.9290	CL(64:4)	CL	[M+H]+	9.0	48.14	Up
		1691.0986	PI(34:1)	PI	[2M+NH4]+	10.7	7.17	Up
	ESI -	309.2050	9.12.13.TriHODE	FA	[M-H-H2O]-	5.0	745.26	Up
		503.2403	PG(18:4)	PG	[M-H+FA]-	2.0	0.23	Down
		607.4175	Peltatol B	Sesquiterpenoid	[M-H-H2O]-	2.0	7.29	Up
		661.5752	DAG(38:4)	DAG	[M-H]-	5.5	41.88	Up
		689.4956	PG(O-32:0)	PG	[M-H-H2O]-	17.0	2.13	Up
		712.4888	PE(34:3)	PE	[M-H]-	4.8	6.51	Up
		719.4817	PG(32:1)	PG	[M-H]-	7.2	3.55	Up
		816.5307	PC(36:5)	PC	[M-2H+K]-	5.0	6.48	Up
		832.5032	PS(36:2)	PS	[M-H+FA]-	12.3	204.50	Up
		890.6883	n. id.	-	-	-	15.14	Up
		949.5650	n. id.	-	-	-	7.01	Up
		968.5274	n. id.	-	-	-	5.66	Up
		982.5260	n. id.	-	-	-	12.01	Up
		1007.7705	n. id.	-	-	-	8.54	Up
		1015.6660	n. id.	-	-	-	1.07	Up
		1033.7822	n. id.	-	-	-	2.90	Up

		1038.7878	n. id.	-	-	-	5.64	Up
		1043.7414	n. id.	-	-	-	5.25	Up
		1072.6284	n. id.	-	-	-	9.41	Up
		1082.7592	n. id.	-	-	-	767.45	Up
		1088.8068	n. id.	-	-	-	15.28	Up
		1392.0214	PE(P-32:2)	PE	[2M-H+FA]-	10.0	19.34	Up
		1469.0065	n. id.	-	-	-	7.63	Up
		1584.0445	n. id.	-	-	-	3.34	Up
		1681.0119	n. id.	-	-	-	15.31	Up
		1681.0884	n. id.	-	-	-	16.70	Up
		1754.0924	PGP(36:2)	PGP	[2M-H+FA]-	14.1	3.20	Up

Table S4. Identification of VIP lipids for the watering factor at standard temperature.

		Exact Mass	Compound name	Family	Ion Assigment	Rel. Mass error (ppm)	Fold Change 150/100	Fold Change 150/50	Fold Change 150/5	Up/Down-regulated
ROOT	ESI +	381.0802	N1-(5-phospho-a-D-ribosyl)-5.6-dimethylbenzimidazole	Ribonucleoside	[M+Na]+	5.3	0.86	3.71	3.58	Up
		413.1960	n. id.	-	-	-	1.06	0.51	0.45	Down
		497.4570	n. id.	-	-	-	1.33	0.81	0.49	Down
		500.4810	n. id.	-	-	-	1.47	2.20	0.84	Down
		611.5442	1-(14-Methyl-pentadecanyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerol	Dialkylglycerols	[M+Na]+	11.0	0.96	0.94	1.00	Down
		637.5577	1-(8-[3]-Ladderane-octanyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerol	Dialkylglycerols	[M+H]+	3.0	0.95	1.25	1.31	Up
		663.4760	3-Hydroxydecanoyl carnitine	FA	[2M+H]+	9.2	0.85	0.83	0.70	Down
		668.6400	CE(18:1)	CE	[M+NH4]+	9.0	1.00	0.94	1.01	Up

689.4932	PA(36:1)	PA	[M+H]+	18.0	1.31	1.03	0.63	Down
690.5043	PA(34:2)	PA	[M+NH4]+	3.7	1.23	0.93	0.65	Down
691.5081	PG(32:1)	PG	[M+H]+	19.0	1.35	0.89	0.57	Down
696.4825	PE(P-32:1)	PE	[M+Na]+	16.2	1.14	0.96	1.02	Up
706.6382	n. id.	-	-	-	0.89	0.91	0.96	Down
715.5083	PC(32:4)	PC	[M+NH4]+	8.0	1.23	0.97	0.62	Down
720.6509	DAG(42:3)	DAG	[M+NH4]+	1.2	0.78	0.95	0.90	Down
722.6627	DAG(42:2)	DAG	[M+NH4]+	4.2	0.83	0.85	0.92	Down
736.6771	n. id.	-	-	-	0.98	0.98	0.94	Down
770.6095	PC(P-36:2)	PC	[M+H]+	4.8	1.08	0.83	0.70	Down
826.6736	PC(40:3)	PC	[M+H]+	6.2	1.04	0.78	0.64	Down
848.7610	TAG(50:2)	TAG	[M+Na]+	10.8	1.41	0.99	0.61	Down
858.7457	TAG(50:4)	TAG	[M+NH4]+	10.3	0.85	1.17	1.58	Up
859.7469	n. id.	-	-	-	0.87	1.26	1.48	Up
868.7335	TAG(52:6)	TAG	[M+NH4]+	6.1	1.01	1.02	1.16	Up
872.7578	TAG(52:4)	TAG	[M+NH4]+	14.2	1.07	0.96	1.14	Up
884.7630	n. id.	-	-	-	1.05	1.26	1.68	Up
898.7696	14-Oxolanosterol	Triterpenoid	[2M+NH4]+	5.4	1.04	1.25	688.83	Up
937.5813	1.2-Di-(9Z.12Z.15Z-octadecatrienoyl)-3-(Galactosyl-alpha-1-6-Galactosyl-beta-1)-glycerol	Glycerol	[M+H]+	7.0	1.01	1.10	0.52	Down
944.5410	Oleanolic acid 3-[rhamnosyl-(1->4)-glucosyl-(1->6)-glucoside]	Triterpene saponins	[M+NH4]+	17.7	1.87	1.71	1.76	Up
960.6463	n. id.	-	-	-	1.09	1.01	0.44	Down
986.8951	n. id.	-	-	-	3.87	5.27	2.06	Up
1008.8051	n. id.	-	-	-	1.09	1.67	1.26	Up
1011.8169	n. id.	-	-	-	1.04	1.66	1.28	Up
1012.8279	n. id.	-	-	4.8	0.96	0.62	0.25	Down
1034.8121	DAT(18:0/24:0(2Me[S].3OH[S].4Me[S].6Me[S]))	Carbohydrate	[M+NH4]+	4.0	1.06	0.85	0.55	Down
1036.8285	n. id.	-	-	-	0.94	0.74	0.39	Down
1038.8403	n. id.	-	-	5.0	0.98	0.79	0.25	Down
1047.7751	Methyl 3b-hydroxy-13(18)-oleanen-28-oate	Triterpenoid	[2M+Na]+	12.2	1.16	0.70	0.28	Down
1048.7927	n. id.	-	-	-	1.03	0.82	0.53	Down

		1122.9240	n. id.	-	-	-	0.98	0.85	0.39	Down
		1172.8568	n. id.	-	-	-	1.19	0.58	0.36	Down
	ESI -	187.0965	cis-3-Hexenyl acetate	Carboxylic acid ester	[M-H+FA]-	5.8	1.62	1.00	0.36	Down
		257.2113	Myristic acid	Carboxylic acid	[M-H+FA]-	3.0	1.73	2.45	1.15	Up
		296.2315	n. id.	-	-	-	1.31	1.32	0.56	Down
		298.2467	n. id.	-	-	-	1.83	1.45	0.55	Down
		314.2427	Methyl glucosinolate	Alkylglucosinolate	[M-H-H2O]-	1.4	1.49	0.60	0.06	Down
		317.2314	2-Hydroxy palmitic acid	FA	[M-H+FA]-	5.0	1.32	1.12	0.58	Down
		351.2518	8.11.14-Eicosatrienoic acid	FA	[M-H+FA]-	6.0	0.11	5.17	5.81	Up
		355.2121	12.13-Epoxy-11-hydroxy-9.15-octadecadienoic acid	FA	[M-H+FA]-	1.5	1.07	0.38	0.38	Down
		390.2727	Eicosapentaenoyl Ethanolamide	Acylethanolamine	[M-H+FA]-	19.7	0.12	5.65	6.85	Up
		392.2276	N-arachidonoyl taurine	FA	[M-H-H2O]-	4.0	1.00	0.33	0.08	Down
		404.1039	Avenanthramide 2	Avenanthramide	[M-H+FA]-	12.9	0.93	3.88	2.94	Up
		429.3569	n. id.	-	-	-	1.40	1.05	1.20	Up
		457.2979	DGDG (34:2)	DGDG	[M-2H]-2	2.1	1.43	89.47	264.71	Up
		458.2992	n. id.	-	-	-	0.97	0.30	1.65	Up
		540.3295	LysoPC(16:0)	Lyso PC	[M-H+FA]-	2.1	1.24	1.61	1.22	Up
		587.4368	12.13-Epoxy-9.15-octadecadienoic acid	FA	[2M-H]-	8.7	0.29	28.14	200.74	Up
		634.4367	PE(28:0)	PE	[M-H]-	13.6	1.23	1.61	1.74	Up
		641.2812	n. id.	-	-	-	1.48	0.98	1.02	Up
		645.4695	DAG(36:4)	DAG	[M-H+FA]-	6.4	1.59	0.96	0.46	Down
		690.5003	PE(32:0)	PE	[M-H]-	11.0	1.53	1.07	0.64	Down
		710.9396	n. id.	-	-	-	1.43	0.89	1.49	Up
		716.4556	PE(36:6)	PE	[M-H-H2O]-	13.8	1.25	94.15	465.07	Up
		716.5106	PE(34:1)	PE	[M-H]-	18.1	0.44	1.31	0.72	Down
		742.4720	PE(38:7)	PE	[M-H-H2O]-	12.3	1.40	1.07	0.70	Down
		777.5461	1.28-Octacosanediol diferulate	Fatty alcohol	[M-H]-	19.3	1.20	17.96	15.66	Up
		796.6225	PC(O-36:2)	PC	[M-2H+Na]-	3.0	1.18	0.95	0.86	Down
		816.4290	β-D-Glucosyloxydestruxin B	Glycodepsipeptides	[M-H+FA]-	5.2	0.23	3.87	4.28	Up
		826.5587	PC(36:4)	PC	[M-H+FA]-	2.0	1.11	1.35	1.14	Up
		849.6251	Oryzarol		[2M-H+FA]-	0.1	1.01	0.84	0.69	Down
		850.6225	n. id.	-	-	-	0.08	18.54	12.12	Up
		851.6269	TAG(54:10)	TAG	[M-H+-H2O]-	9.0	0.02	1.00	1.48	Down

		864.7983	Bisdiphosphoinositol tetrakisphosphate	Pyrophosphate	[M-H+FA]-	7.1	1.18	0.91	0.78	Down
		866.8098	Glycerol triheptadecanoate	TAG	[M+NH4]+	8.4	1.54	0.91	0.11	Down
		889.6943	n. id.	-	-	-	0.49	1.22	0.97	Down
		949.5650	n. id.	-	-	-	1.21	1.12	0.36	Down
		960.5944	n. id.	-	-	-	1.56	1.18	0.47	Down
		982.5260	n. id.	-	-	-	1.46	1.43	1.25	Up
		1402.2731	n. id.	-	-	-	1.12	0.96	0.87	Down
		1403.2768	n. id.	-	-	-	1.69	1.68	0.68	Down
		1404.2756	n. id.	-	-	-	1.13	0.93	0.84	Down
		1414.3021	n. id.	-	-	-	0.97	0.95	0.38	Down
		1432.2946	n. id.	-	-	-	1.58	1.36	0.78	Down
		1575.1938	n. id.	-	-	-	0.64	0.24	0.06	Down
AERIAL PART	ESI +	379.0827	2-Hydroxy-2-(2-oxopropyl)butanedioic acid	Keto acid	[2M-H]-	14.6	2.02	3.47	3.93	Up
		518.3234	LysoPC(18:3)	Lyso PC	[M+H]+	1.3	1.25	1.35	0.92	Down
		521.3440	LysoPE(20:3)	Lyso PE	[M+NH4]+	17.0	2.09	1.76	1.14	Up
		566.4420	Cholesteryl β-D-glucoside	Sterols	[M+NH4]+	1.0	0.93	0.80	0.86	Down
		606.5092	DAG(34:4)	DAG	[M+NH4]+	0.1	0.70	0.85	3.59	Up
		609.5297	DAG(34:0)	DAG	[M+NH4]+	5.0	0.00	1.45	2.12	Up
		630.5105	DAG(36:6)	DAG	[M+NH4]+	2.0	0.59	0.41	0.68	Down
		637.5577	1-(8-[3]-ladderane-octanyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerol	Dialkylglycerol	[M+H]+	3.0	0.91	0.84	0.90	Down
		680.6339	Campesteryl linoleate	Steroid ester	[M+NH4]+	0.0	1.25	1.55	1.74	Up
		684.2034	7-(4-Carboxy-3-hydroxy-3-methylbutanoyl)sudachitin 4'-glucoside	Flavonoide	[M+NH4]+	14.6	1.87	1.65	1.46	Up
		690.5043	PA(34:2)	PA	[M+NH4]+	3.7	0.91	0.85	0.66	Down
		690.6216	PG(34:4)	PG	[M+NH4]+	5.7	1.26	1.58	1.23	Up
		712.4898	PE(32:1)	PE	[M+Na]+	1.5	0.53	0.37	0.24	Down
		714.5089	PE(34:3)	PE	[M+H]+	3.0	0.91	0.83	0.59	Down
		716.5220	PE(34:2)	PE	[M+H]+	0.6	1.14	1.14	0.71	Down
		747.6027	PC(P-36:1)	PC	[M+Na]+	2.0	1.33	1.52	0.87	Down
		756.5580	PC(34:3)	PC	[M+H]+	5.6	1.18	1.33	0.89	Down
		758.2234	Pelargonidin-3.5-diglucoside-5-O-p-coumaroylglucoside	Glucoside	[M+NH4]+	7.5	1.87	1.88	1.55	Up
		760.5107	PS(34:2)	PS	[M+H]+	2.2	1.30	1.32	0.68	Down

		771.6180	n. id.	-	-	-	0.00	1.90	1.43	Up
		782.4926	PS(34:2)	PS	[M+Na]+	2.1	1.55	1.66	0.62	Down
		816.7031	TAG(48:4)	TAG	[M+NH4]+	5.5	2.49	4.32	5.93	Up
		850.5572	PI(34:3)	PI	[M+NH4]+	10.0	0.79	1.01	1.00	Down
		864.6349	n. id.	-	-	11.5	1.28	1.53	1.07	Up
		912.7747	22:1-Glc-Stigmasterol	Phytosterol	[M+NH4]+	10.0	0.41	0.67	1.35	Up
		922.7175	n. id.	-	-	6.5	1.19	1.84	6.26	Up
		931.5930	PS(44:4)	PS	[M+NH4]+	2.0	0.81	3.33	2.21	Up
		942.5501	n. id.	-	-	-	1.11	0.93	0.87	Down
		944.5410	Oleanolic acid 3-[rhamnosyl-(1->4)-glucosyl-(1->6)-glucoside]	Triterpene saponin	[M+NH4]+	17.7	1.28	1.18	0.92	Down
		988.7026	PA(34:3)	PA	[2M+H]+	16.0	0.34	0.06	0.37	Down
		1058.8145	n. id.	-	-	-	0.93	0.52	0.33	Down
		1342.9094	Lipid A -disaccharide-1-P	SL	[M+Na]+	8.0	0.94	0.82	0.69	Down
		1568.0811	PG(36:3)	PG	[2M+Na]+	14.8	1.56	1.81	1.29	Up
		1691.0986	PI(34:1)	PI	[2M+NH4]+	10.7	1.33	1.48	0.89	Down
		1786.0679	Chlorophyll a		[2M+H]+	5.6	0.00	5.07	4.15	Up
	ESI-	213.1483	Chrycolide		[M-H-H2O]-	4.8	2.34	2.32	2.29	Up
		265.1473	6.7-Dihydro-4-(hydroxymethyl)-2-(p-hydroxyphenethyl)-7-methyl-5H-2-pyrindinium	Phenol	[M-H-H2O]-	2.0	1.37	1.11	0.73	Down
		309.2050	9.12.13.TriHODE	FA	[M-H-H2O]-	5.0	0.39	0.40	0.35	Down
		341.2312	Dihydroxyfumaric acid	FA	[2M-H+FA]-	0.7	0.79	0.36	0.25	Down
		353.1997	n. id.	-	-	-	0.97	0.76	0.53	Down
		377.0857	Bisdemalonylsalvianin	Flavonoide	[M-2H]-2	5.6	2.21	3.79	4.59	Up
		387.1144	Trehalose	Carbohydrate	[M-H+FA]-	0.1	2.51	4.17	3.91	Up
		439.0844	3.5-Dihydroxyphenyl 1-O-(6-O-galloyl-beta-D-glucopyranoside)	Phenolic glycoside	[M-H]-	8.7	2.03	3.53	2.51	Up
		471.0738	n. id.	-	-	-	0.50	0.26	0.17	Down
		482.2607	LysoPE(20:4)	Lyso PE	[M-H-H2O]-	13.2	1.60	1.08	0.17	Down
		483.2713	LPA(18:0)	GP Monoacylglycerophosphates	[M-H+FA]-	3.1	1.65	1.40	0.79	Down

		503.2403	PG(18:4)	PG	[M-H]-	2.0	1.92	1.53	0.42	Down
		566.3418	LysoPC(18:1)	Lyso PC	[M-H+FA]-	8.0	1.28	0.98	0.63	Down
		636.4214	n. id.	-	-	-	2.71	4.16	2.45	Up
		655.4936	n. id.	-	-	-	0.97	1.56	5.56	Up
		701.4770	PG(32:1)	PG	[M-H-H2O]-	1.8	0.59	0.40	0.71	Down
		712.4888	PE(34:3)	PE	[M-H]-	4.8	1.38	1.67	1.10	Up
		796.6225	n. id.	-	-	-	2.01	2.33	2.06	Up
		804.5716	PC(34:1)	PC	[M-H+FA]-	5.4	1.28	1.40	0.96	Down
		832.6023	PC(36:1)	PC	[M-H+FA]-	6.4	1.20	1.25	0.84	Down
		842.5117	n. id.	-	-	-	1.40	1.53	1.21	Up
		843.4674	n. id.	-	-	-	1.23	1.20	0.94	Down
		918.5147	n. id.	-	-	-	0.90	1.15	0.93	Down
		930.8364	n. id.	-	-	12.9	1.42	1.35	0.97	Down
		950.5666	n. id.	-	-	-	1.28	1.21	1.17	Up
		951.5573	LysoPE(16:0)	Lyso PE	[2M-H+FA]-	12.6	1.17	1.24	1.11	Up
		963.6108	n. id.	-	-	-	0.15	2.99	2.18	Up
		968.5274	n. id.	-	-	-	0.81	1.85	1.42	Up
		973.5557	n. id.	-	-	-	0.86	1.20	1.56	Up
		983.7293	n. id.	-	-	-	0.74	0.64	0.45	Down
		989.5393	PIP(36:1)	PIP	[M-H+FA]-	2.1	1.33	1.27	0.92	Down
		1002.6868	n. id.	-	-	-	0.50	0.13	0.13	Down
		1035.7844	n. id.	-	-	-	0.03	1.27	0.76	Down
		1038.7878	n. id.	-	-	-	1.03	0.73	0.39	Down
		1051.7429	3α-Acetomethoxy-11alpha-oxo-12-ursen-24-oic acid	Triterpenoid	[2M-H]-	17.6	1.07	0.77	0.51	Down
		1084.7785	n. id.	-	-	-	0.97	0.52	0.35	Down
		1085.7859	n. id.	-	-	-	0.88	0.44	0.27	Down
		1557.9919	n. id.	-	-	-	1.95	2.35	1.07	Up
		1584.0445	n. id.	-	-	-	1.42	1.54	1.00	Up
		1592.1385	n. id.	-	-	-	0.94	0.71	0.47	Down
		1645.0338	n. id.	-	-	-	1.56	1.81	1.21	Up
		1754.0924	PGP(36:2)	PGP	[2M-H+FA]-	14.1	1.49	1.38	1.13	Up

Table S5. Identification of VIP lipids for the watering factor at high temperature.

		Exact Mass	Compound name	Family	Ion Assigment	Rel. Mass error (ppm)	Fold Change 150/100	150/50	150/5	Up/Down-regulated
ROOT	ESI +	315.1954	Momilactone A	Diterpene lactones	[M+H]+	0.2	5.98	6.67	16.70	Up
		367.1890	11-Dehydrocorticosterone	Steroids	[M+Na]+	2.8	0.52	0.33	0.22	Down
		381.0802	N1-(5-Phospho-a-D-ribosyl)-5.6-dimethylbenzimidazole	Ribonucleotides	[M+Na]+	5.3	11.22	39.70	40.62	Up
		469.4249	n. id.	-	-	-	1.19	1.25	1.40	Up
		483.4600	n. id.	-	-	-	0.53	0.35	0.36	Down
		496.3398	LysoPC(16:0)	Lyso PC	[M+H]+	0.0	0.70	0.80	0.51	Down
		497.4570	n. id.	-	-	-	0.61	0.47	0.34	Down
		498.4777	n. id.	-	-	-	1.09	0.87	16.44	Up
		500.4810	n. id.	-	-	-	0.65	1.29	0.86	Down
		510.3550	n. id.	-	-	-	1.08	1.13	1.20	Up
		608.5550	n. id.	-	-	-	0.45	0.40	0.26	Down
		611.1860	4'-Hydroxyacetophenone 4'-[4-hydroxy-3.5-dimethoxybenzoyl-(->5)-apiosyl-(1->2)-glucoside]	Tanin	[M+H]+	18.0	0.66	1.82	1.75	Up
		619.6018	N-Hexadecanoylpyrrolidine	Acylpyrrolidine	[2M+H]+	19.0	1.08	1.11	1.41	Up
		634.5365	DAG(36:4)	DAG	[M+NH4]+	6.5	0.91	0.91	1.59	Up
		647.5101	n. id.	-	-	-	0.72	0.72	0.82	Down
		673.4993	PA(34:2)	PA	[M+H]+	11.0	0.70	0.55	0.52	Down
		678.6167	CE (18:3)	CE	[M+H]+	2.0	0.63	1.00	1.07	Up
		684.2034	7-(4-Carboxy-3-hydroxy-3-methylbutanoyl)sudachitin 4'-glucoside		[M+NH4]+	14.6	2.32	2.58	3.82	Up
		689.4932	PA(36:1)	PA	[M+H]+	-	0.02	1.60	1.25	Up
		690.5043	PA(34:2)	PA	[M+H]+	3.7	1.03	227.15	274.14	Up
		691.5081	PG(O-32:0)	PG	[M+H-H2O]+	18.0	0.72	0.51	0.37	Down
		692.6331	18:3 Sitosteryl ester	Steroid ester	[M+NH4]+	1.0	0.85	0.78	0.86	Down

696.6705	18:1 Sitosteryl ester	Steroid ester	[M+NH4]+	7.5	1.12	0.86	1.97	Up
712.4898	PE(32:1)	PE	[M+Na]+	1.5	0.66	0.57	0.41	Down
717.5223	CL(70:2)	CL	[M+2H]2+	6.4	1.09	0.89	0.74	Down
758.2234	Pelargonidin-3.5-diglucoside-5-O-p-coumaroylglucoside	Flavonoide	[M+NH4]+	7.5	2.11	2.68	3.11	Up
760.5770	PC(34:1)	PC	[M+H]+	10.7	0.75	0.89	0.79	Down
761.5852	PE(36:2)	PE	[M+NH4]+	6.4	0.43	0.47	0.55	Down
781.5539	PS(34:0)	PS	[M+NH4]+	10.0	0.60	0.80	0.65	Down
782.5640	PC(34:1)	PC	[M+Na]+	3.9	1.03	0.88	1.07	Up
785.5850	PG(O-36:1)	PG	[M+Na]+	10.0	0.58	0.58	0.93	Down
786.5970	PC(36:2)	PC	[M+H]+	4.8	0.54	0.70	142.83	Up
790.7683	Glycinoprenol 11	Polyprenols	[M+NH4]+	14.6	1.09	0.95	369.56	Up
791.7739	n. id.	-	-	-	1.25	1.17	1.26	Up
812.6490	n. id.	-	-	-	2.46	1.55	1.58	Up
827.6775	2-Decaprenyl-6-methoxyphenol	Tetraterpenoids	[M+Na]+	11.9	2.30	1.54	2.81	Up
832.2342	Kaempferol 3-[2"-glucosyl-6"-acetyl-galactoside] 7-glucoside	Flavonoide	[M+NH4]+	19.7	3.31	5.21	1.67	Up
848.7610	TAG(50:2)	TAG	[M+Na]+	10.8	0.65	20.04	22.85	Up
852.5545	PI(34:2)	PI	[M+NH4]+	6.0	0.88	0.91	0.77	Down
858.7457	TAG(50:4)	TAG	[M+NH4]+	10.3	0.63	22.21	0.15	Down
859.7469	n. id.	-	-	-	0.73	1.47	1.83	Up
866.8098	Glycerol triheptadecanoate	TAG	[M+NH4]+	8.4	1.90	4.80	4.22	Up
874.7738	3β-Hydroxy-4β-methyl-5α-cholest-7-ene-4α-carbaldehyde	Sterol	[2M+NH4]+	10.4	0.80	0.94	1.42	Up
876.7903	TAG(52:2)	TAG	[M+NH4]+	12.7	0.69	0.72	1.15	Up
884.7630	n. id.	-	-	-	0.91	1.38	1.95	Up
887.5602	PI(38:4)	PI	[M+H]+	4.7	0.17	0.28	1.01	Up
898.7696	14-Oxolanosterol		[2M+NH4]+	5.4	0.79	1.18	1.47	Up
900.7868	TAG(54:4)	TAG	[M+NH4]+	16.2	0.61	0.67	3.82	Up
913.6615	n. id.	-	-	-	0.84	0.96	1.01	Up
942.7246	n. id.	-	-	-	0.41	1.17	5.06	Up
954.6075	DGDG (36:6)	DGDG	[M+NH4]+	7.7	0.80	0.72	0.67	Down
1008.8051	n. id.	-	-	-	0.55	0.34	0.21	Down
1011.8169	n. id.	-	-	-	0.46	0.29	0.17	Down

		1012.8279	n. id.	-	-	-	0.64	0.35	0.17	Down
		1021.9884	n. id.	-	-	-	3.70	1.75	3.45	Up
		1022.7363	n. id.	-	-	-	0.60	0.54	0.40	Down
		1038.8403	n. id.	-	-	-	0.63	0.40	0.22	Down
		1039.0177	n. id.	-	-	-	0.47	0.67	0.84	Down
		1047.7751	Methyl 3b-hydroxy-13(18)-oleanen-28-oate	Triterpenoid	[2M+Na]+	12.2	0.57	0.40	0.23	Down
		1048.7927	n. id.	-	-	-	0.72	0.53	0.39	Down
	ESI -	133.0122	Malic acid	FA	[M-H]-	15.4	0.36	0.62	0.35	Down
		257.2113	Myristic acid	FA	[M-H+FA]-	3.0	0.26	2.04	1.99	Up
		269.0458	Selenocystathionine	FA	[M-H]-	17.1	2.80	1.92	3.68	Up
		297.2425	9.10-dihydroxyoctadecanoic acid	FA	[M-H-H2O]-	1.0	0.54	1.80	2.69	Up
		351.2518	8.11.14-Eicosatrienoic acid	FA	[M-H+FA]-	6.0	0.07	22.58	40.50	Up
		355.2121	12.13-Epoxy-11-hydroxy-9.15-octadecadienoic acid	FA	[M-H+FA]-	1.5	0.36	0.17	0.17	Down
		367.2304	Chondrillasterol 3-[glucosyl-(1->4)-glucoside]	Steroids	[M-2H]-2	1.0	0.40	0.45	0.39	Down
		377.0857	Bisdemalonylsalvianin	Flavonoide	[M-2H]-2	5.6	3.22	16.79	16.81	Up
		387.1144	Trehalose	Carbohydrates	[M-H+FA]-	0.1	3.61	11.36	14.51	Up
		390.2727	Eicosapentaenoyl Ethanolamide	acylethanolamine	[M-H+FA]-	19.7	0.11	11.72	18.86	Up
		392.2276	N-arachidonoyl taurine	FA	[M-H-H2O]-	4.0	0.60	0.23	0.32	Down
		399.3473	Tricosanoic acid	FA	[M-H+FA]-	1.7	0.99	0.70	0.72	Down
		404.1039	Avenanthramide 2	Avenanthramide	[M-H+FA]-	12.9	6.62	54.21	45.54	Up
		413.1960	n. id.	-	-	-	0.68	0.22	0.29	Down
		415.2125	Dihydrofukinolide	Lactone	[M+Na]+	8.5	1.31	1.00	1.09	Up
		429.3569	n. id.	-	-	-	0.85	0.67	0.69	Down
		430.3617	n. id.	-	-	-	1.06	0.92	1.02	Up
		441.3934	Myristyl laurate	FA	[M-H+FA]-	3.3	0.77	0.58	0.72	Down
		457.2979	DGDG (34:2)	DGDG	[M-2H]-2	2.1	0.76	0.54	0.90	Down
		469.4242	Palmityl laurate	Fatty acyl	[M-H+FA]-	4.0	0.87	0.79	0.91	Down
		516.4182	n. id.	-	-	-	0.52	0.37	0.44	Down
		540.3295	LysoPC(16:0)	Lyso PC	[M-H+FA]-	2.1	0.60	0.70	0.44	Down
		594.5058	n. id.	-	-	-	0.82	1.47	1.55	Up
		633.4326	3-hydroxynonanoyl carnitine	Acyl Carnitine	[2M-H]-	0.9	0.64	0.83	1.12	Up
		634.4367	PE(28:0)	PE	[M-H]-	13.6	0.60	0.94	1.38	Up
		670.4525	n. id.	-	-	-	0.60	1.08	1.09	Up

		672.5747	PE(O-32:0)	PE	[M-H-H2O]-	7.0	0.58	0.48	0.42	Down
		688.4946	PE(32:1)	PE	[M-H]-	3.4	0.89	0.76	0.80	Down
		689.4956	PG(O-32:0)	PG	[M-H+FA]-	10.0	0.59	0.41	0.32	Down
		691.5823	Glycerol 1-(9Z-octadecenoate) 2-octanoate 3-tetradecanoate	TAG	[M-H]-	8.6	0.60	0.62	1.11	Up
		692.5040	n. id.	-	-	-	1.00	0.51	1.19	Up
		698.6204	n. id.	-	-	-	0.91	0.84	1.22	Up
		699.6246	n. id.	-	-	-	0.78	0.70	0.89	Down
		714.6505	n. id.	-	-	-	0.51	0.41	0.38	Down
		715.5006	24.25.26.27-Tetranor-23-oxo-hydroxyvitamin D3	Triterpenoids	[2M-H]-	8.8	0.68	0.67	0.87	Down
		716.4556	PE(36:6)	PE	[M-H-H2O]-	13.8	0.53	0.41	0.68	Down
		716.5106	PE(34:1)	PC	[M-H]-	18.1	0.37	0.40	0.50	Down
		718.4708	PE(36:5)	PE	[M-H-H2O]-	10.0	0.75	0.72	0.67	Down
		735.6172	n. id.	-	-	-	0.60	0.67	0.58	Down
		772.6612	β-Hydroarchaetidylethanolamine	Dialkylglycerophosphoethanolamine	[M-H-H2O]-	3.0	0.56	0.56	0.51	Down
		777.5461	1.28-Octacosanediol diferulate	Fatty alcohol	[M-H]-	19.3	1.31	0.71	0.80	Down
		831.4962	PI(34:3)	PI	[M-H]-	8.0	0.60	0.65	0.50	Down
		850.6225	n. id.	-	-	-	0.41	0.37	0.39	Down
		851.6269	TAG(54:10)	TAG	[M-H+-H2O]-	9.0	0.42	0.50	0.52	Down
		879.6464	3-Decaprenyl-4.5-dihydroxybenzoate	Tetraterpenoid	[M-H+FA]-	5.0	0.70	0.66	0.58	Down
		889.6943	n. id.	-	-	-	0.04	17.94	12.09	Up
		971.5462	n. id.	-	-	-	0.65	0.40	0.28	Down
		982.5260	n. id.	-	-	-	0.16	17.61	17.40	Up
		987.6259	n. id.	-	-	-	0.85	0.64	0.55	Down
AERIAL PART	ESI +	397.3824	n. id.	-	-	-	1.60	2.00	5.23	Up
		549.4872	Cohibin A	Fatty alcohol	[M+H]+	0.0	1.57	1.16	1.95	Up
		551.5012	DAG(P-32:1))	DAG	[M+H]+	4.0	1.84	1.42	2.31	Up
		574.5560	Octatriacontatetraenoic acid	FA	[M+NH4]+	0.2	1.80	2.48	6.30	Up
		591.4972	DAG(32:3)	DAG	[M+H]+	1.9	1.70	1.63	2.71	Up
		606.5092	DAG(34:4)	DAG	[M+NH4]+	0.1	1.75	2.89	8.39	Up
		609.5297	DAG(34:0)	DAG	[M+NH4]+	5.0	1.77	1.13	2.74	Up
		610.5384	DAG(34:2)	DAG	[M+NH4]+	3.5	1.61	1.59	2.95	Up

624.5569	DAG(34:2)	DAG	[M+NH4]+	1.2	1.91	1.64	2.53	Up
636.5551	DAG(36:3)	DAG	[M+NH4]+	1.6	1.68	1.03	1.45	Up
637.5577	1-(8-[3]-Ladderane-octanyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerol	Dialkylglycerol	[M+H]+	3.0	0.03	1.53	2.10	Up
678.6167	CE (18:3)	CE	[M+H]+	2.0	1.50	1.90	3.65	Up
680.6339	Campesteryl linoleate	Steroid ester	[M+NH4]+	0.0	0.34	2.13	4.62	Up
692.6331	18:3 Sitosteryl ester	Steroid ester	[M+NH4]+	1.4	1.44	1.53	2.36	Up
693.6364	20:1 Campesteryl ester	Steroid ester	[M+H]+	17.0	1.85	1.36	1.35	Up
712.4898	PE(32:1)	PE	[M+Na]+	1.5	1.31	0.47	0.46	Down
740.5345	PG(32:0)	PG	[M+NH4]+	12.3	1.88	1.36	1.20	Up
760.5107	PS(34:2)	PS	[M+H]+	2.2	1.53	0.98	0.65	Down
771.6180	n. id.	-	-	-	2.22	1.53	2.45	Up
782.4926	PS(34:2)	PS	[M+Na]+	2.1	1.74	0.89	0.51	Down
816.7031	TAG(48:4)	TAG	[M+NH4]+	5.5	2.60	1.08	0.86	Down
850.5572	PI(34:3)	PI	[M+NH4]+	10.0	0.49	1.49	1.42	Up
858.7457	TAG(50:4)	TAG	[M+NH4]+	10.3	2.45	2.16	0.27	Down
912.7747	22:1-Glc-Stigmasterol	Phytosterol	[M+NH4]+	10.0	0.88	2.76	10.22	Up
931.5930	PS(44:4)	PS	[M+NH4]+	2.0	2.92	2.10	1.47	Up
936.6546	n. id.	-	-	-	1.73	1.36	29.33	Up
937.5813	n. id.	-	-	-	2.13	2.22	2.34	Up
942.5501	n. id.	-	-	-	0.35	1.70	1.07	Up
957.5495	n. id.	-	-	-	1.86	1.29	1.62	Up
988.7026	PA(34:3)	PA	[2M+H]+	16.0	10.00	1.99	1.93	Up
999.7412	n. id.	-	-	-	1.49	0.74	0.43	Down
1020.7228	FMC-5(d18:1/24:1)	Neutral glycosphingolipid	[M+H]+	11.0	1.29	0.63	0.37	Down
1022.7363	n. id.	-	-	-	1.47	0.97	0.63	Down
1023.8023	n. id.	-	-	-	1.44	1.35	1.95	Up
1031.7710	22-Acetylpriverogenin B	Triterpenoids	[2M-H]-	14.8	0.48	0.52	0.53	Down
1049.8093	n. id.	-	-	-	2.22	1.52	1.96	Up
1051.7188	Oleanderolide 3-acetate	Triterpenoids	[2M+Na]+	2.0	1.69	1.87	4.50	Up
1075.7183	3α-Acetomethoxy-11α-oxo-12-ursen-24-oic acid	Triterpenoids	[2M+Na]+	2.4	1.24	0.83	1.71	Up
1157.9246	n. id.	-	-	-	1.82	1.31	1.30	Up
1172.8568	n. id.	-	-	-	1.65	1.58	1.70	Up

		1344.9320	n. id.	-	-	-	2.00	1.38	3.88	Up
		1345.9290	n. id.	-	-	-	1.66	1.66	1.98	Up
		1691.0986	PI(34:1)	PI	[2M+NH4]+	10.7	0.25	1.92	4.36	Up
		1786.0679	Chlorophyll a		[2M+H]+	5.6	2.64	1.27	1.89	Up
	ESI -	213.1483	Chrycolide		[M-H-H2O]-	4.8	1.05	1.98	1.66	Up
		309.2050	9.12.13.TriHODE	FA	[M-H-H2O]-	5.0	2.41	2.12	2.90	Up
		518.1491	n. id.	-	-	-	1.83	1.13	0.81	Down
		674.5793	n. id.	-	-	-	2.34	0.90	1.10	Up
		683.5215	PA(36:1)	PA	[M-H-H2O]-	18.0	1.73	1.68	2.37	Up
		701.4770	PG(32:1)	PG	[M-H-H2O]-	1.8	1.76	2.48	3.43	Up
		701.6051	n. id.	-	-	-	1.90	1.26	1.27	Up
		703.6241	n. id.	-	-	-	1.28	1.67	1.43	Up
		717.4664	PG(32:2)	PG	[M-H]-	6.7	1.45	1.04	1.19	Up
		729.6374	n. id.	-	-	-	1.81	1.21	1.44	Up
		732.6279	PC(O2-34:0)	PC	[M-H]-	0.3	1.46	1.63	1.52	Up
		836.5200	n. id.	-	-	-	1.64	1.54	1.82	Up
		844.5307	n. id.	-	-	-	1.70	1.40	1.96	Up
		851.6269	TAG(54:10)	TAG	[M-H+-H2O]-	9.0	2.29	1.14	1.16	Up
		870.5545	PC(40:6)	PC	[M-2H+K]-	14.0	1.85	1.11	0.96	Down
		883.6601	2.3-Diacetoxypropyl stearate	TAG	[2M-H]-	9.6	1.46	1.02	1.05	Up
		889.6830	n. id.	-	-	-	1.64	1.20	1.65	Up
		963.6108	n. id.	-	-	-	1.38	3.30	2.47	Up
		973.5557	n. id.	-	-	-	1.38	1.49	3.84	Up
		979.6972	n. id.	-	-	-	1.25	0.70	0.49	Down
		983.5278	n. id.	-	-	-	1.23	0.96	1.14	Up
		985.5398	Methyl lucidenate F	Triterpenoids	[2M-H+FA]-	8.1	2.30	1.11	1.03	Up
		988.5419	CDP-DAG(36:1)	DAG	[M-H-H2O]-	1.0	2.11	1.56	1.69	Up
		1002.6868	n. id.	-	-	-	1.26	0.62	0.37	Down
		1025.6057	n. id.	-	-	-	2.46	1.57	1.31	Up
		1035.7844	n. id.	-	-	-	2.71	1.15	1.38	Up
		1057.7632	n. id.	-	-	-	1.64	1.51	1.79	Up
		1060.7883	n. id.	-	-	-	2.09	1.39	1.54	Up
		1064.8043	n. id.	-	-	-	1.64	1.29	1.21	Up
		1072.6284	n. id.	-	-	-	1.50	1.29	1.57	Up

		1087.8006	n. id.	-	-	-	1.67	1.63	1.90	Up
		1088.8068	n. id.	-	-	-	2.65	1.13	0.88	Down
		1391.9622	CL(64:3)	CL	[M-H+FA]-	19.7	1.59	0.91	0.86	Down
		1646.0336	n. id.	-	-	-	1.66	1.01	1.10	Up
		1656.0516	n. id.	-	-	-	1.78	2.01	0.87	Down

References

(1) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC - Trends in Analytical Chemistry* **2016**, *82*, 425-442.
(2) Jaumot, J.; de Juan, A.; Tauler, R. *Chemometrics Intell. Lab. Syst.* **2015**, *140*, 1-12.
(3) De Juan, A.; Jaumot, J.; Tauler, R. *Anal. Chim. Acta* **2014**, *6*, 4964-4976.
(4) Ruckebusch, C.; Blanchet, L. *Anal. Chim. Acta* **2013**, *765*, 28-36.
(5) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; de souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythei, I. *Nucleic Acids Res.* **2009**, *37*, D603-D610.
(6) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Weerasinghe, D.; Zhang, P.; Karp, P. D. *Nucleic Acids Research* **2014**, *42*, D459-D471.
(7) Golub, G. H.; Loan, C. F. V. *Matrix computations*, third ed.; Johns Hopkins University Press: Baltimore, 1996, p 728.
(8) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425-1432.
(9) Windig, W.; Stephenson, D. A. *Anal. Chem.* **1992**, *64*, 2735-2742.
(10) De Juan, A.; Rutan, S. C.; Tauler, R. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Brown, S. D.; Tauler, R.; Walczak, B., Eds.; Elsevier: Oxford, 2009, pp 325-344.
(11) Tauler, R.; Maeder, M.; de Juan, A. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Brown, S. D.; Tauler, R.; Walczak, B., Eds.; Elsevier: Oxford, 2009, pp 473-505.
(12) Tauler, R. *Chemometrics Intell. Lab. Syst.* **1995**, *30*, 133-146.
(13) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R. J. A. N.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *21*, 3043-3048.
(14) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Westerhuis, J. A.; Smilde, A. K. *J. Chemometr.* **2005**, *19*, 469-481.
(15) Vis, D. J.; Westerhuis, J. A.; Smilde, A. K.; van der Greef, J. *BMC Bioinformatics* **2007**, *8*.
(16) Wold, S.; Esbensen, K.; Geladi, P. *Chemometrics Intell. Lab. Syst.* **1987**, *2*, 37-52.
(17) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
(18) Wold, S.; Johansson, A.; Cocchi, M. In *3D QSAR in Drug Design*, Kubiny, H., Ed.; ESCOM Science Publishers: Leiden, 1993, pp 583-618.
(19) Chong, I. G.; Jun, C. H. *Chemometrics Intell. Lab. Syst.* **2005**, *78*, 103-112.
(20) Matthews, B. W. *BBA - Protein Structure* **1975**, *405*, 442-451.

4.4. Discussió conjunta dels resultats

En aquesta secció es mostren i es comparen els resultats obtinguts en les diferents etapes dels treballs inclosos en aquest capítol. En primer lloc, es discuteixen els resultats obtinguts en els dos treballs en la resolució quimiomètrica dels perfils d'elució i espectres de masses dels metabòlits i lípids. Seguidament, es presenten els resultats obtinguts en l'estudi estadístic dels diferents factors ambientals avaluats i es mostren les diferents metodologies emprades en els dos treballs per la identificació dels metabòlits i lípids. Finalment, es discuteix la interpretació biològica dels resultats obtinguts per cada factor estressant.

4.4.1. Resolució dels metabòlits i lípids de l'arròs en diferents condicions ambientals

En el capítol anterior (Capítol 3) s'ha demostrat com es pot emprar una estratègia basada en MCR-ALS en la resolució dels metabòlits presents en mostres d'arròs. En els dos treballs d'aquest capítol es va utilitzar aquesta estratègia de tractament de les dades per a resoldre els perfils d'elució i espectres de masses purs dels metabòlits i lípids extrets de mostres d'arròs en diferents condicions ambientals. Es van utilitzar diferents mètodes de compressió de les dades en la direcció espectral i per la construcció de les matrius a analitzar. En la publicació 3 es va emprar el mètode de *binning*, mentre que en la publicació 4 es va utilitzar el procediment de cerca de ROIs [14]. La combinació del procediment de compressió de dades mitjançant la cerca de ROIs i del mètode de resolució d'aquestes dades comprimides per MCR-ALS ha proporcionat resultats molt satisfactoris i s'anomena ROIMCR [14]. Aquest procediment s'ha aplicat en altres treballs per LC-HRMS [15, 16], CE-HRMS (electroforesis capil·lar acoblada a espectrometria de masses d'alta resolució) [17] o MSI (espectroscòpia d'imatges de MS) [18].

A partir dels resultats obtinguts en els dos treballs d'aquest capítol, es poden descriure els principals avantatges del procediment de cerca de ROIs en comparació al procediment de compressió tradicional de *binning*: més compressió de dades mantenint la resolució espectral de les dades originals.

Les dades enregistrades en la publicació 3 tenien una mida inicial de 250 Mb per mostra (corresponents a un rang de 90 a 1000 *m/z* enregistrat en alta resolució). Considerant que es van analitzar un total de 80 mostres, el conjunt de dades tenia una mida total d'aproximadament 20 Gb. Un cop aplicat el procediment de *binning* es van considerar 18200 valors de *m/z* (de 90 a 1000 *m/z* amb una resolució de 0,05). Per cada mostra es van adquirir 1010 espectres de *m/z* (temps d'anàlisi total igual a 40 minuts) i es van analitzar un total de 80 mostres, això va donar lloc a una matriu augmentada de dimensions 80800×18200, amb necessitats de memòria RAM de l'ordinador d'aproximadament 12 Gb. Per tant, el processament d'aquestes dades amb un ordinador PC de sobretaula (actualment amb 8 Gb de memòria

RAM) no va ser factible. Per aquest motiu, va ser necessari dividir el cromatograma en diferents finestres de temps. Aquestes finestres de temps es van determinar mitjançant l'aplicació prèvia del procediment de PLS-DA a la matriu que contenia els cromatogrames de ions totals (TIC) de totes les mostres analitzades (de mida molt més reduïda, 0,7 Mb). Tal com s'explica a la publicació 3 (apartat *Analysis of chromatographic regions*), es van seleccionar 5 regions cromatogràfiques d'interès per a les mostres tractades amb Cd i 6 per a les tractades amb Cu. Per cadascuna d'aquestes finestres de temps es van construir les corresponents matrius augmentades, les quals ja tenien una mida adequada pel seu processament mitjançant MCR-ALS (aproximadament 0,8 Gb). Per tant, van ser necessaris un total de 11 models diferents de MCR-ALS per a explicar la variància de les dades de la publicació 3. En el cas de la publicació 4 les dades enregistrades tenien una mida inicial de 100 Mb per mostra (corresponents a un rang de 50 a 1500 *m/z* enregistrat en alta resolució) i l'aplicació del procediment de cerca de ROIs va permetre considerar només un nombre baix de senyals de *m/z*, aproximadament 500 per cada tipus de mostra analitzada (arrels i parts aèries analitzades en mode de ionització positiu i negatiu). Tenint en compte que es van analitzar un total de 68 mostres de cada tipus (arrels i parts aèries) i que per cada mostra es van adquirir 627 espectres de *m/z* (temps d'anàlisi total igual a 20 minuts), la matriu augmentada que es va obtenir per cada tipus de mostra tenia unes dimensions de 42636×500, amb necessitats de memòria RAM de l'ordinador d'aproximadament 0,2 Gb. Per tant, gràcies a l'aplicació de la cerca de ROIs en el cas de la publicació 4 es van poder analitzar simultàniament totes les mostres del mateix tipus (parts aèries o arrels analitzades en mode positiu o negatiu) a partir d'un sol model de MCR-ALS sense grans requisits computacionals.

En relació a l'exactitud dels valors de *m/z*, en les dues publicacions d'aquest capítol les dades originals (r*aw*) tenien una exactitud de 0,0001 unitats de *m/z*. En la publicació 3, després d'aplicar el procediment de compressió *binning* les dades tenien una resolució de 0,05 unitats de *m/z*. Conseqüentment, en aquesta publicació per a poder conèixer el valor de la massa exacta dels metabòlits resolts va ser necessària una cerca a les dades originals (*raw*), ja que la mida del *binning* utilitzat només conservava dos decimals en els valors de massa. En canvi, en la publicació 4 després d'aplicar el procediment ROIMCR les dades conservaven la resolució espectral original. Per tant, els valors de massa exacta dels metabòlits resolts es van poder obtenir directament dels espectres resolts per MCR-ALS.

Cal comentar, a més, que en les dues publicacions, l'aplicació del procediment MCR-ALS va permetre resoldre satisfactòriament els perfils d'elució i els espectres de masses purs d'un gran nombre

dels metabòlits i lípids presents en les mostres analitzades. Per exemple, en el cas de la publicació 3 es van resoldre directament un nombre total de 100 components per les mostres tractades amb Cu i 115 per les mostres tractades amb Cd. En el cas de la publicació 4 es van resoldre un total de 200 components per cada una de les quatre matrius de dades analitzades (parts aèries i arrels analitzades en mode positiu i en mode negatiu). No obstant això, cal comentar que no tots aquests components de MCR-ALS resolts van ser assignats a metabòlits o lípids individuals. Alguns d'ells representaven senyals de soroll provinents de les contribucions del solvent o del soroll de fons. Tots els models de MCR-ALS obtinguts van mostrar un percentatge de variància explicada (R^2) superior al 98% i un percentatge de falta d'ajust (LOF) inferior al 12%.

A mode d'exemple, en la Figura 4.1A es mostra la resolució MCR-ALS obtinguda en la publicació 3 de dos metabòlits coeluïts i en la Figura 4.1B l'obtinguda en la publicació 4 per quatre lípids coeluïts. En aquesta figura es pot observar que en els espectres de masses resolts en la publicació 3 els valors de massa obtinguts tenien dos decimals, mentre que els resolts en la publicació 4 els valors de massa mantenien els quatre decimals d'exactitud originals. En tots dos casos, els perfils d'elució obtinguts van ser útils per a la interpretació dels efectes dels factors ambientals en l'arròs, mentre que els espectres de masses van emprar-se en la identificació dels metabòlits i lípids resolts. Aquests dos aspectes s'expliquen amb més detall en les seccions següents.

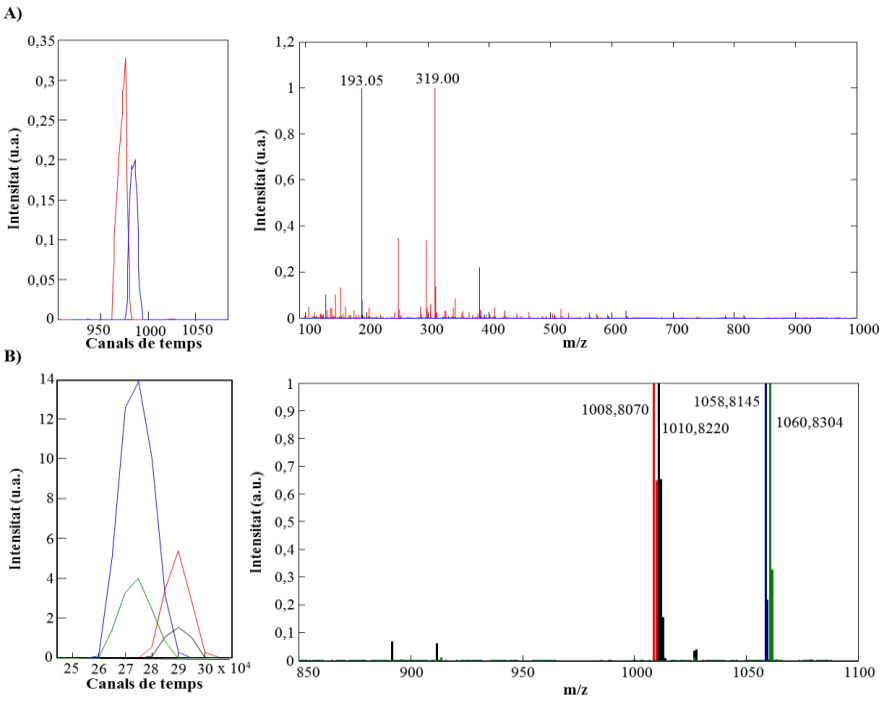


Figura 4.1. Resolució per MCR-ALS dels perfils d'elució i dels espectres de masses de: A) dos dels metabòlits coeluïts en la publicació 3; B) quatre dels lípids coeluïts en la publicació 4.

4.4.2. Estudi estadístic dels efectes de diversos factors ambientals sobre el creixement de l'arròs

Les àrees cromatogràfiques dels perfils d'elució dels metabòlits i lípids resolts pel mètode de MCR-ALS es van analitzar mitjançant diferents mètodes quimiomètrics per a avaluar estadísticament l'efecte dels factors ambientals estressants estudiats sobre el creixement de l'arròs.

Efectes de la contaminació per metalls pesants a l'aigua de regadiu emprada en el creixement de l'arròs

En la publicació 3 es va utilitzar l'ANOVA d'un factor per a detectar quins dels metabòlits resolts presentaven diferències significatives entre els diferents grups de mostres d'arròs analitzades (controls i tractades amb diferents concentracions dels metalls Cu(II) i Cd(II)). En total, es van detectar 48 metabòlits que mostraven un canvi significatiu en les seves concentracions (nivell de significació *p* inferior a 0,05) amb el tractament amb Cd, 17 que el mostraven exclusivament pel tractament amb Cu i 6 que el mostraven pels dos metalls. En les Taules 2, 3 i 4 de la publicació 3 es mostra la identificació d'aquests metabòlits, les principals famílies dels quals es resumeixen en la Taula 4.1. Cal remarcar que la majoria d'aquests metabòlits van presentar un canvi (*fold change*) similar pels quatre nivells de concentració de metalls avaluats (10, 50, 100 i 1000 µM). Aquest resultat indica que l'alteració del metabolisme de l'arròs ja es produïa amb els nivells més baixos de concentració dels metalls (10 µM). No es produïen en canvi alteracions importants en augmentar la concentració dels metalls a nivells superiors (50, 100 i 1000 µM). Aquests metabòlits amb canvis significatius van permetre fer una primera interpretació biològica dels possibles efectes produïts pels metalls sobre el creixement de l'arròs (veure secció següent).

Taula 4.1. Famílies dels metabòlits amb canvis significatius en les mostres tractades amb Cu i Cd.

Tractament	Famílies
Cadmi	Aminoàcids, àcids nucleics, àcids grassos, àcids orgànics, alcaloides, flavonoides, glicerols, glucòsids, lípids, sucres.
Comú	Aminoàcids, àcids nucleics, àcids grassos, àcids orgànics.
Coure	Aminoàcids, Àcids orgànics, alcohols, lípids, nucleòsids, sucres.

Efectes de l'augment de temperatura i de l'estrès hídric sobre el creixement de l'arròs.

En la publicació 4, el procediment d'anàlisi multivariant de la variància ASCA [19] es va aplicar en mostres de diferents parts de la planta de l'arròs (parts aèries i arrels analitzades en mode de ionització positiu o negatiu). Els resultats obtinguts van mostrar que tant l'augment de temperatura com l'estrès

hídric (manca d'aigua) tenien un efecte significatiu en les concentracions dels lípids de l'arròs (nivell de significació *p* inferior a 0,05). En canvi, la recuperació hídrica (després de regar les mostres prèviament sotmeses a estrès de sequera amb 150 mL d'aigua durant els dos últims dies de cultiu) no va mostrar canvis significatius en cap dels casos estudiats (nivell de significació *p* inferior a 0,05).

L'anàlisi mitjançant PCA de les àrees dels lípids resolts també va mostrar que l'augment de temperatura tenia un efecte clar sobre el lipidoma de l'arròs. En tots els casos les mostres d'arròs que van créixer en condicions normals es distingien perfectament de les mostres que havien crescut en condicions d'alta temperatura. Aquesta distinció es va observar en els dos primers PCs, els quals explicaven un 62% de la variància de les dades (veure Figura 4A de la publicació 4). En canvi, la diferenciació de les mostres en funció de la quantitat d'aigua de regadiu disponible (150, 100, 50 i 5 mL) només es va observar per les mostres de les parts aèries de l'arròs, però no per les mostres de les seves arrels. En la Figura 4.2 es mostra el gràfic d'*scores* del model de PCA amb 2 components, PC1 *vs* PC2, obtingut en l'anàlisi en mode negatiu de les parts aèries de l'arròs. En aquesta figura es pot apreciar l'efecte de la recuperació hídrica pels nivells de tractament més alts (5 i 50 mL). En canvi, no s'observen aquests canvis pel nivell de tractament més baix (100 mL). Tal com es pot observar en la Figura 4.2, les mostres s'ordenen seguint una trajectòria corbada en l'espai de dues dimensions PC1 *vs* PC2 en funció del nivell d'aigua de regadiu afegida. Les mostres regades amb 5 mL d'aigua i sotmeses a un pas de recuperació hídrica (codificades 5R) es van distingir de les mostres regades amb 5 mL d'aigua sense el pas de recuperació hídrica (codificades 5), i en canvi aquestes mostres no es distingien de les mostres regades amb 50 mL (codificades 50). El mateix es va observar per les mostres regades amb 50 mL d'aigua i sotmeses a una etapa de recuperació hídrica (codificades 50R), les quals tenien una posició entremig de les mostres regades amb 50 mL i les mostres regades amb 100 mL (codificades 100). En canvi, no es van observar diferències entre les mostres regades amb 100 mL amb o sense recuperació hídrica. Aquesta tendència, va indicar que el procés de recuperació hídrica aplicat a les mostres estudiades en aquest treball (publicació 4) no era suficientment intens, ja que cap de les mostres sotmeses a aquest procés van tenir un comportament semblant al de les mostres regades amb 150 mL (control).

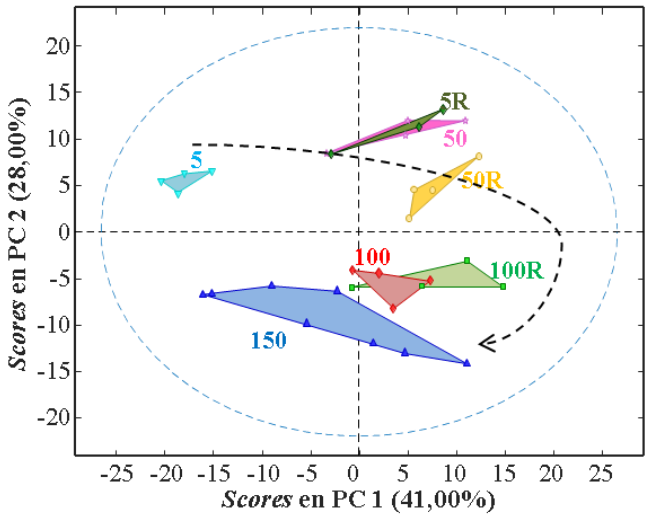


Figura 4.2. Gràfic d'*scores* PC1 *vs* PC2 per les mostres de les parts aèries de l'arròs analitzades en mode negatiu. Les mostres es troben classificades segons el nivell d'aigua emprada en el seu regadiu (5, 50, 100 i 150 ml) i segons si han estat sotmeses o no a un procés de recuperació hídrica (R indica amb recuperació hídrica)

Finalment, es va utilitzar el mètode PLS-DA per identificar quins eren els lípids les concentracions dels quals es trobaven més influenciades pels factors estressants avaluats. Els resultats estaven d'acord amb els obtinguts en les anàlisis per ASCA i PCA. En primer lloc, les mostres cultivades en condicions normals es diferenciaven clarament de les mostres crescudes en condicions d'alta temperatura, amb discriminacions entre les mostres control i tractades que donaven valors de coeficient de correlació de Matthews (MCC, descrit a la introducció de la present Tesi, secció 2.3.4 [20]) entre 0,93 i 1,00. En segon lloc, la diferenciació de les mostres en funció del nivell d'aigua de regadiu va ser millor per les mostres de les parts aèries de l'arròs (MCC entre 0,88 i 1,00) que per les seves arrels (MCC entre 0,73 i 1,00). Els models PLS-DA que discriminaven les mostres sotmeses al procés de recuperació hídrica van mostrar pitjors resultats, amb valors de MCC entre 0,73 i 0,76, la qual cosa estava d'acord amb els resultats previs obtinguts en l'anàlisi de PCA, els quals indicaven que el procés de recuperació hídrica aplicat no havia estat suficient.

Tenint en compte el conjunt dels resultats obtinguts, el procés de recuperació hídrica no es va considerar com a factor a l'hora de seleccionar possibles biomarcadors perquè els models obtinguts en la discriminació de les mostres no van ser prou bons. Es van seleccionar només aquells lípids que van estar més afectats (biomarcadors potencials) per l'alta temperatura i per l'estrès hídric. En la Taula 4.2 es mostren el nombre de lípids seleccionats a partir dels valors dels VIPs (VIP > 2) dels models de PLS-DA obtinguts per aquests dos factors. Aquests lípids van ser identificats tal com s'explica a continuació, la qual cosa va permetre realitzar una interpretació biològica preliminar dels resultats obtinguts.

Taula 4.2. Nombre de lípids seleccionats en la publicació 4 com a biomarcadors potencials dels factors experimentals d'alta temperatura i d'estrès hídric (manca d'aigua). La identificació d'aquests lípids es mostra en les Taules S3, S4 i S5 del material suplementaria de la publicació 4.

		Parts aèries			**Arrels**		
		ESI (+)	**ESI (-)**	**Comú**	**ESI (+)**	**ESI (-)**	**Comú**
ALTA TEMPERATURA		39	41	1	25	30	7
ESTRÈS HÍDRIC	Temperatura òptima	32	59	3	37	40	3
	Temperatura alta	43	34	1	53	42	7

4.4.3. Identificació de metabòlits i lípids

Les dues publicacions incloses en aquest capítol també presenten diferències importants en l'etapa d'identificació dels metabòlits mitjançant espectrometria de masses. En els dos casos es va utilitzar un espectròmetre de masses d'alta resolució, que va permetre identificar temptativament els metabòlits i lípids resolts comparant els valors de massa exacta obtinguts experimentalment amb els valors de masses teòrics disponibles en bases de dades públiques com MassBank [21], METLIN [22] o HMDB [23]. Malauradament, això no era suficient per assegurar la seva assignació completa [24, 25].

Segons la directiva europea 2002/657/CE [26], per poder considerar una identificació completa dels compostos químics, cal aconseguir un mínim de quatre punts d'identificació (IPs). Quan es treballa amb un espectròmetre de masses amb una resolució superior a 20000 FWHM, l'assignació de l'ió precursor a partir d'un metabòlit o d'un lípid concret amb un error de massa inferior a 5 ppm proporciona 2 IPs. En aquest primer aspecte trobem ja la primera diferència important entre els procediments d'identificació emprats en les dues publicacions. En el procediment emprat en la publicació 3 amb l'instrument Q-Exactive Orbitrap es compleix el requisit de resolució de 70000 FWHM a *m/z* 400. En canvi en la publicació 4 amb l'instrument TOF la resolució era de 10000 FWHM a *m/z* 400, i per tant no s'assolia el requisit esmentat. D'altra banda, quan s'aconsegueixen dos ions producte per a confirmar l'assignació de l'ió precursor, s'assoleixen aleshores 2,5 IPs. Aquests ions producte només es van poder enregistrar en el cas de la publicació 3, ja que el Q-Exactive Orbitrap permetia adquirir els espectres de fragmentació de tots els ions detectats (*all ion fragmentation*, AIF). La possibilitat d'adquirir aquests espectres AIF suposa un avantatge molt important en els estudis de metabolòmica no dirigida, ja que permet enregistrar els ions

fills sense tenir cap coneixement previ de quins són els metabòlits (ions precursors) que es troben en les mostres d'arròs analitzades, estalviant a més la necessitat d'analitzar les mostres per duplicat.

En la Figura 4.3 es mostra un exemple de la identificació de la trehalosa feta en el treball de la publicació 3. Tant l'ió precursor (341,1092 *m/z*) com els dos ions productes principals (101,0245 i 179,0564 *m/z*) obtinguts experimentalment es van poder associar als corresponents ions teòrics de la trehalosa amb un error de massa inferior a 2,3 ppm. Conseqüentment, la trehalosa es va poder identificar amb 4,5 IPs (2 IPs guanyats per l'ió precursor i 2,5 IPs guanyats pels ions producte), complint així els criteris de la directiva 2002/657/CE. Seguint aquest mateix procediment d'identificació, en la publicació 3 es van poder assignar completament un total de 97 metabòlits, els quals es mostren en les Taules S2, S3 i S4 del material suplementari de la publicació 3.

.

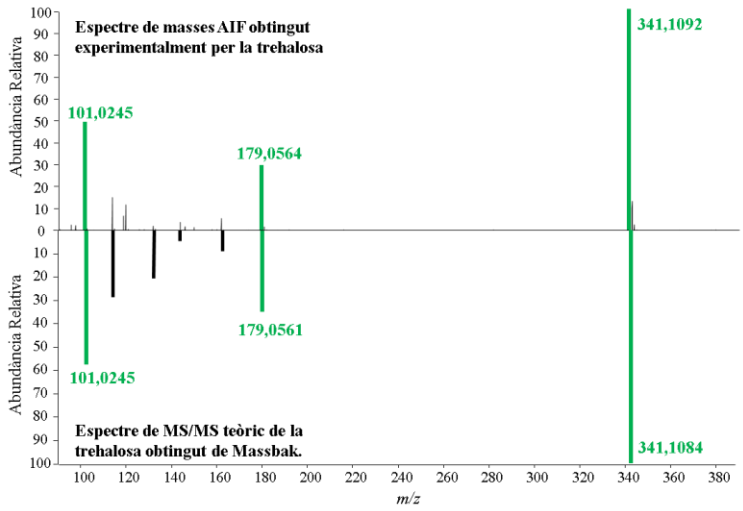


Figura 4.3. Identificació de la trehalosa en la publicació 3. L'espectre de la part de dalt és l'espectre de AIF obtingut experimentalment i l'espectre de baix és el de MS/MS teòric obtingut de la base de dades Massbank [21]. Els senyals marcats amb color verd representen l'ió precursor (341,1092 *m/z*) i els dos ions productes (101,0245 i 179,0561 *m/z*), els quals coincideixen amb els teòrics amb un error de massa inferior a 2,3 ppm.

Per acabar, cal comentar que en la publicació 4 es va utilitzar també el temps de retenció dels lípids resolts com a suport en la seva identificació. El mètode cromatogràfic emprat en aquesta publicació per a l'anàlisi de lípids s'ha utilitzat àmpliament en el grup d'investigació en el que s'ha dut a terme aquesta Tesi, la qual cosa ha permès disposar d'una base de dades amb els temps de retenció de les diferents famílies de lípids [15, 27-34]. Es van identificar temptativament un total de 148 lípids. Aquesta identificació es mostra en les Taules S3, S4 i S5 del material suplementaria de la publicació 4.

4.4.4. Interpretació biològica dels resultats obtinguts

Un cop s'han identificat els metabòlits i lípids representatius dels efectes detectats pels diferents factors ambientals estressants investigats es va procedir a la seva possible interpretació biològica.

Efectes biològics produïts per la contaminació de metalls pesants en l'aigua de regadiu dels cultius d'arròs

En la publicació 3 la interpretació biològica dels resultats assolits es va dur a terme a partir de l'anàlisi de les rutes metabòliques afectades incloses a la base de dades KEGG (*Kyoto Encyclopedia of Gens and Genomes*) [35] per a l'arròs (*O. sativa* L.). En total es van trobar 11 rutes metabòliques amb un mínim de quatre metabòlits significativament afectats (osa01100, osa01110, osa01230, osa00230, osa02010, osa00561, osa01200, osa 00970, osa00970, oisa00630 i osa00260, veure Taula 5 de la publicació 3), entre les quals van destacar la biosíntesi dels metabòlits secundaris, el metabolisme dels glicerolípids, el metabolisme del carboni, el metabolisme de la purina i el metabolisme dels aminoàcids.

Malgrat que el solapament entre els metabòlits afectats pel Cd i el Cu va ser petit (només 6 metabòlits en comú), es va trobar que els dos metalls afectaven les mateixes rutes metabòliques de manera similar (veure Figura 6 de la publicació 3). D'una banda, el contingut d'aminoàcids augmentava. Aquesta resposta es pot explicar com un mecanisme de desintoxicació de les plantes per a protegir els constituents de les cèl·lules. D'altra banda, el metabolisme de la purina i dels glicerolípids va disminuir, la qual cosa es relaciona amb una reducció del creixement i de l'activitat fotosintètica com a conseqüència de l'estrès oxidatiu ocasionat per la contaminació de metalls. Aquestes dues tendències també s'han observat en altres estudis d'organismes vegetals en treballs recents [36-39]. Per exemple, X. Li en el seu treball de 2017 [36] va detectar una acumulació d'aminoàcids en la gespa (*Cynodon dactylon* L.*)* tractada amb Cd. Sobretot destacava un augment de la prolina, la qual també es va trobar significativament afectada en el nostre treball de la publicació 3. D'altra banda, en el treball de A. Alessandro del 2013 [37] es va trobar que el metabolisme de la purina disminuïa en les plantes de mostassa bruna (*Brassica juncea* L.) quan eren tractades amb Cd. Aquestes tendències es van observar també en organismes vegetals tractats amb Cu, per exemple en cogombre, (*Cucumis sativus* L.) en el treball de L. Zhao [39] o en alga bruna (*Ectocarpus siliculosus* D.) en el cas del treball de A. Ritter [38]. A més, alguns treballs basats en proteòmica d'arròs (*O. sativa* L.) i de soja (*Glycine max* L.) també han trobat les mateixes rutes metabòliques afectades per la contaminació de Cd i Cu [40, 41].

Tenint en compte els metabòlits identificats que no es troben anotats al KEGG, es van deduir canvis significatius en la concentració dels glicòsids. Els glicòsids són molt abundants en els organismes vegetals i formen complexos forts amb metalls, la qual cosa explica que es vegin alterats quan l'arròs es troba sotmès a contaminació elevada per Cd i Cu. Aquesta alteració dels glicòsids s'ha observat també en treballs anteriors amb altres plantes [38, 42]. Per exemple, l'any 2010 X. Su ja va observar una acumulació dels glicòsids en mostres d'*A. thaliana* exposades a Cd [43]. També, en el treball de A. Ritter del 2014 es va observar que els nivells dels glicòsids de l'alga bruna (*E. siliculosus* D.) s'alteraven en presència de Cu.

Efectes biològics produïts per l'augment de temperatura i l'estrès hídric

En l'avaluació dels efectes produïts sobre els lípids afectats en la publicació 4, es va observar que la principal resposta de l'arròs a les dues alteracions ambientals estudiades (alta temperatura i estrès hídric) era la d'intentar mantenir la fluïdesa i la consistència de les membranes cel·lulars. Una evidència d'això va ser que tant en les mostres de les parts aèries com de les d'arrels de l'arròs, les concentracions dels glicerolípids (diglicèrids [DAG] i triglicèrids [TAG]) van trobar-se fortament alterades per l'alta temperatura i per l'estrès hídric. D'una banda, la concentració dels lípids amb un grau d'insaturació menor va augmentar quan les mostres d'arròs van créixer en condicions d'alta temperatura. Per exemple, el TAG (48:1) es va acumular en les mostres tractades a temperatura més elevada, i per tant, la rigidesa de la membrana va augmentar. D'altra banda, en condicions de manca d'aigua es va observar l'efecte contrari, la concentració dels lípids amb un grau d'insaturació elevat es va incrementar. Per exemple, la concentració del TAG (48:4) va augmentar en les mostres tractades, de forma que s'augmenta la fluïdesa de la membrana. També cal comentar que en les mostres sotmeses als dos estressos ambientals simultàniament, la concentració dels lípids amb un grau d'insaturació elevat va disminuir. Aquest resultat suggereix que els efectes causats en l'arròs per un augment de temperatura van ser més severs que els causats per l'estrès hídric (manca d'aigua).

A més, en les plantes crescudes en condicions de temperatura més alta es van observar alteracions en les concentracions dels glicerofosfolípids (sobretot en fosfatidiletanolamines [PE], fosfatidilcolines [PC] i fosfatidilinositols [PI]). L'acumulació d'aquests lípids pot estar relacionada amb possibles canvis en la síntesi dels glicerolípids i fosfolípids en el reticle endoplasmàtic. Els glicerofosfolípids són constituents importants de les membranes cel·lulars i, per tant, contribueixen en la modificació de la seva fluïdesa.

Finalment, cal comentar també que en les mostres que van créixer amb estrès hídric de manca d'aigua es van observar canvis en les concentracions dels lípids plastídics: monogalactosildiacilglicerol (MGDG) i digalactosildiacilglicerol (DGDG). Aquests lípids són els principals components de les membranes de cloroplast, i per tant, tenen una influència important en la seva consistència.

Moltes de les alteracions en els lípids observades en la publicació 4 per les mostres que van créixer en condicions d'estrès hídric de manca d'aigua també s'han observat en treballs similars en l'organisme vegetal model *A. thaliana* [5, 44]. Per exemple, en el treball de B. Yu de 2014 [5] es descriu que les famílies de lípids més alterades en l'*A. thaliana* van ser els glicerofosfolípids (PG, PC, PE i PI) i els lípids plastídics (MGDG i DGDG). L'augment del grau d'insaturació dels lípids que incrementa la fluïdesa de la membrana cel·lular va ser observat també en mostres d'*A. thaliana* sotmeses a estrès de sequera en el treball de P. Tarazona del 2015 [44]. S'han trobat altres resultats semblants als de la publicació 4 en el cas de les alteracions causades per l'alta temperatura en diferents organismes vegetals, com l'*A. thaliana* [4, 44], el blat (*T. aestivum* L.) [45] o la *Saussurea medusa* D.C. (lotus de les neus) [46]. En tots aquests treballs es va observar que la resposta de les plantes a un augment de temperatura produïa una modificació de la rigidesa de les membranes cel·lulars associada a una disminució del grau d'insaturació dels glicerolípids de forma anàloga a l'observat aquí.

4.5. Referències

1. Rodziewicz, P.;Swarcewicz, B.;Chmielewska, K.;Wojakowska, A.;Stobiecki, M., Influence of abiotic stresses on plant proteome and metabolome changes, *Acta Physiologiae Plantarum*, 2014, **36**, 1-19.
2. Debnath, M.;Pandey, M.;Bisen, P. S., An omics approach to understand the plant abiotic stress, *Omics : a journal of integrative biology*, 2011, **15**, 739-762.
3. Meena, K. K.;Sorty, A. M.;Bitla, U. M.;Choudhary, K.;Gupta, P.;Pareek, A.;Singh, D. P.;Prabha, R.;Sahu, P. K.;Gupta, V. K.;Singh, H. B.;Krishanani, K. K.;Minhas, P. S., Abiotic Stress Responses and Microbe-Mediated Mitigation in Plants: The Omics Strategies, *Frontiers in Plant Science*, 2017, **8,** doi.org/10.3389/fpls.2017.00172.
4. Higashi, Y.;Okazaki, Y.;Myouga, F.;Shinozaki, K.;Saito, K., Landscape of the lipidome and transcriptome under heat stress in Arabidopsis thaliana, *Scientific reports*, 2015, **5**, 10533.
5. Yu, B.;Li, W., Comparative profiling of membrane lipids during water stress in Thellungiella salsuginea and its relative Arabidopsis thaliana, *Phytochemistry*, 2014, **108**, 77-86.
6. Obata, T.;Fernie, A. R., The use of metabolomics to dissect plant responses to abiotic stresses, *Cellular and Molecular Life Sciences*, 2012, **69**, 3225-3243.
7. Stocker, T., *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 2014.
8. Shanker, A. K.;Maheswari, M.;Yadav, S. K.;Desai, S.;Bhanu, D.;Attal, N. B.;Venkateswarlu, B., Drought stress responses in crops, *Functional & integrative genomics*, 2014, **14**, 11-22.
9. Easterling, D. R.;Meehl, G. A.;Parmesan, C.;Changnon, S. A.;Karl, T. R.;Mearns, L. O., Climate Extremes: Observations, Modeling, and Impacts, *Science*, 2000, **289**, 2068-2074.
10. Singh, S.;Parihar, P.;Singh, R.;Singh, V. P.;Prasad, S. M., Heavy Metal Tolerance in Plants: Role of Transcriptomics, Proteomics, Metabolomics, and Ionomics, *Frontiers in Plant Science*, 2015, **6**, 1143.

11. Agency, U. S. E. P., 2015, **EPA 550-B-15-001**.
12. Ahsan, N.;Nakamura, T.;Komatsu, S., Differential responses of microsomal proteins and metabolites in two contrasting cadmium (Cd)-accumulating soybean cultivars under Cd stress, *Amino acids*, 2012, **42**, 317-327.
13. Sebastiani, L.;Francini, A.;Romeo, S.;Ariani, A.;Minnocci, A., in *Approaches to Plant Stress and their Management*, eds. R. K. Gaur and P. Sharma, Springer India, New Delhi, 2014, DOI: 10.1007/978-81-322-1620-9_15, pp. 267-279.
14. Gorrochategui, E.;Jaumot, J.;Lacorte, S.;Tauler, R., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC Trends in Analytical Chemistry*, 2016, **82**, 425-442.
15. Marques, A. S.;Bedia, C.;Lima, K. M. G.;Tauler, R., Assessment of the effects of As(III) treatment on cyanobacteria lipidomic profiles by LC-MS and MCR-ALS, *Analytical and Bioanalytical Chemistry*, 2016, **408**, 5829-5841.
16. Gómez-Canela, C.;Prats, E.;Piña, B.;Tauler, R., Assessment of chlorpyrifos toxic effects in zebrafish (Danio rerio) metabolism, *Environmental Pollution*, 2017, **220**, 1231-1243.
17. Ortiz-Villanueva, E.;Benavente, F.;Piña, B.;Sanz-Nebot, V.;Tauler, R.;Jaumot, J., Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data, *Analytica Chimica Acta*, 2017, **978**, 10-23.
18. Bedia, C.;Tauler, R.;Jaumot, J., Analysis of multiple mass spectrometry images from different Phaseolus vulgaris samples by multivariate curve resolution, *Talanta*, 2017, **175**, 557-565.
19. Smilde, A. K.;Jansen, J. J.;Hoefsloot, H. C.;Lamers, R. J.;van der Greef, J.;Timmerman, M. E., ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, *Bioinformatics*, 2005, **21**, 3043-3048.
20. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et biophysica acta*, 1975, **405**, 442-451.
21. Horai, H.;Arita, M.;Kanaya, S.;Nihei, Y.;Ikeda, T.;Suwa, K.;Ojima, Y.;Tanaka, K.;Tanaka, S.;Aoshima, K.;Oda, Y.;Kakazu, Y.;Kusano, M.;Tohge, T.;Matsuda, F.;Sawada, Y.;Hirai, M. Y.;Nakanishi, H.;Ikeda, K.;Akimoto, N.;Maoka, T.;Takahashi, H.;Ara, T.;Sakurai, N.;Suzuki, H.;Shibata, D.;Neumann, S.;Iida, T.;Tanaka, K.;Funatsu, K.;Matsuura, F.;Soga, T.;Taguchi, R.;Saito, K.;Nishioka, T., MassBank: A public repository for sharing mass spectral data for life sciences, *Journal of Mass Spectrometry*, 2010, **45**, 703-714.
22. Tautenhahn, R.;Cho, K.;Uritboonthai, W.;Zhu, Z.;Patti, G. J.;Siuzdak, G., An accelerated workflow for untargeted metabolomics using the METLIN database, *Nature Biotechnology*, 2012, **30**, 826-828.
23. Wishart, D. S.;Knox, C.;Guo, A. C.;Eisner, R.;Young, N.;Gautam, B.;Hau, D. D.;Psychogios, N.;Dong, E.;Bouatra, S.;Mandal, R.;Sinelnikov, I.;Xia, J.;Jia, L.;Cruz, J. A.;Lim, E.;Sobsey, C. A.;Shrivastava, S.;Huang, P.;Liu, P.;Fang, L.;Peng, J.;Fradette, R.;Cheng, D.;Tzur, D.;Clements, M.;Lewis, A.;de souza, A.;Zuniga, A.;Dawe, M.;Xiong, Y.;Clive, D.;Greiner, R.;Nazyrova, A.;Shaykhutdinov, R.;Li, L.;Vogel, H. J.;Forsythei, I., HMDB: A knowledgebase for the human metabolome, *Nucleic Acids Research*, 2009, **37**, D603-D610.
24. Creek, D. J.;Jankevics, A.;Breitling, R.;Watson, D. G.;Barrett, M. P.;Burgess, K. E. V., Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction, *Analytical Chemistry*, 2011, **83**, 8703-8710.
25. Falchi, F.;Bertozzi, S. M.;Ottonello, G.;Ruda, G. F.;Colombano, G.;Fiorelli, C.;Martucci, C.;Bertorelli, R.;Scarpelli, R.;Cavalli, A.;Bandiera, T.;Armirotti, A., Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification, *Analytical Chemistry*, 2016, **88**, 9510-9517.
26. Implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, 2002, **Directive 2002/657/CE**.
27. Lima, K. M. G.;Bedia, C.;Tauler, R., A non-target chemometric strategy applied to UPLC-MS sphingolipid analysis of a cell line exposed to chlorpyrifos pesticide: A feasibility study, *Microchemical Journal*, 2014, **117**, 255-261.
28. Bedia, C.;Dalmau, N.;Jaumot, J.;Tauler, R., Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors, *Environmental Research*, 2015, **140**, 18-31.
29. Dalmau, N.;Jaumot, J.;Tauler, R.;Bedia, C., Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells, *Molecular BioSystems*, 2015, **11**, 3397-3406.

30. Gorrochategui, E.;Casas, J.;Pérez-Albaladejo, E.;Jáuregui, O.;Porte, C.;Lacorte, S., Characterization of complex lipid mixtures in contaminant exposed JEG-3 cells using liquid chromatography and high-resolution mass spectrometry, *Environmental Science and Pollution Research*, 2014, **21**, 11907-11916.
31. Gorrochategui, E.;Pérez-Albaladejo, E.;Casas, J.;Lacorte, S.;Porte, C., Perfluorinated chemicals: Differential toxicity, inhibition of aromatase activity and alteration of cellular lipids in human placental cells, *Toxicology and Applied Pharmacology*, 2014, **277**, 124-130.
32. Gorrochategui, E.;Casas, J.;Porte, C.;Lacorte, S.;Tauler, R., Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells, *Analytica Chimica Acta*, 2015, **854**, 20-33.
33. Canals, D.;Mormeneo, D.;Fabriàs, G.;Llebaria, A.;Casas, J.;Delgado, A., Synthesis and biological properties of Pachastrissamine (jaspine B) and diastereoisomeric jaspines, *Bioorganic & Medicinal Chemistry*, 2009, **17**, 235-241.
34. Merrill, A. H., Jr.;Sullards, M. C.;Allegood, J. C.;Kelly, S.;Wang, E., Sphingolipidomics: high-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry, *Methods (San Diego, Calif.)*, 2005, **36**, 207-224.
35. Kanehisa, M.;Goto, S.;Sato, Y.;Furumichi, M.;Tanabe, M., KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Researhc.*, 2012, **40**, D109-D114.
36. Li, X.;Gitau, M. M.;Han, S.;Fu, J.;Xie, Y., Effects of cadmium-resistant fungi Aspergillus aculeatus on metabolic profiles of bermudagrass [Cynodondactylon (L.)Pers.] under Cd stress, *Plant physiology and biochemistry : PPB*, 2017, **114**, 38-50.
37. D'Alessandro, A.;Taamalli, M.;Gevi, F.;Timperio, A. M.;Zolla, L.;Ghnaya, T., Cadmium stress responses in Brassica juncea: hints from proteomics and metabolomics, *Journal of proteome research*, 2013, **12**, 4979-4997.
38. Ritter, A.;Dittami, S. M.;Goulitquer, S.;Correa, J. A.;Boyen, C.;Potin, P.;Tonon, T., Transcriptomic and metabolomic analysis of copper stress acclimation in Ectocarpus siliculosus highlights signaling and tolerance mechanisms in brown algae, *BMC Plant Biology*, 2014, **14**, 116.
39. Zhao, L.;Huang, Y.;Hu, J.;Zhou, H.;Adeleye, A. S.;Keller, A. A., 1H NMR and GC-MS Based Metabolomics Reveal Defense and Detoxification Mechanism of Cucumber Plant under Nano-Cu Stress, *Environmental Science & Technology*, 2016, **50**, 2000-2010.
40. Chen, C.;Song, Y.;Zhuang, K.;Li, L.;Xia, Y.;Shen, Z., Proteomic Analysis of Copper-Binding Proteins in Excess Copper-Stressed Roots of Two Rice (Oryza sativa L.) Varieties with Different Cu Tolerances, *PloS one*, 2015, **10**, e0125367.
41. Hossain, Z.;Hajika, M.;Komatsu, S., Comparative proteome analysis of high and low cadmium accumulating soybeans under cadmium stress, *Amino acids*, 2012, **43**, 2393-2416.
42. Fukusaki, E.;Kobayashi, A., Plant metabolomics: Potential for practical operation, *Journal of Bioscience and Bioengineering.*, 2005, **100**, 347-354.
43. Sun, X.;Zhang, J.;Zhang, H.;Ni, Y.;Zhang, Q.;Chen, J.;Guan, Y., The responses of Arabidopsis thaliana to cadmium exposure explored via metabolite profiling, *Chemosphere*, 2010, **78**, 840-845.
44. Tarazona, P.;Feussner, K.;Feussner, I., An enhanced plant lipidomics method based on multiplexed liquid chromatography-mass spectrometry reveals additional insights into cold- and drought-induced membrane remodeling, *Plant Journal*, 2015, **84**, 621-633.
45. Narayanan, S.;Tamura, P. J.;Roth, M. R.;Prasad, P. V. V.;Welti, R., Wheat leaf lipids during heat stress: I. High day and night temperatures result in major lipid alterations, *Plant, Cell and Environment*, 2016, **39**, 787-803.
46. Zheng, G.;Tian, B.;Zhang, F.;Tao, F.;Li, W., Plant adaptation to frequent alterations between high and low temperatures: Remodelling of membrane lipids and maintenance of unsaturation levels, *Plant, Cell and Environment*, 2011, **34**, 1431-1442.

Capítol 5

Desenvolupament i aplicació de la cromatografia de líquids bidimensional exhaustiva en estudis de metabolòmica no dirigida

5.1. Introducció

En aquest capítol es discuteix l'ús de la LC×LC en estudis de metabolòmica i lipidòmica no dirigides. El nombre de publicacions que utilitzen LC×LC ha crescut en els darrers anys en diferents àmbits científics com, per exemple, l'ambiental [1-4], el clínic [1, 5-8] o l'alimentari [1, 9-11]. Malgrat dècades de desenvolupament de la 1D-LC, incloent avenços tecnològics com les columnes amb una mida de partícula inferior a 2 µm o els instruments capaços de treballar a pressions molt elevades (superiors a 400 bars), aquesta tècnica sovint no és capaç de separar tots els anàlits d'interès de les mostres. Aquest és el principal motiu que porta als cromatografistes a considerar els avantatges d'afegir una dimensió de separació addicional [1]. En la Figura 5.1 es mostra un exemple de la millora en la separació obtinguda per LC×LC en comparació amb la 1D-LC. Es pot observar com pics que coelueixen en la primera dimensió cromatogràfica, es poden separar bé al llarg de la segona columna.

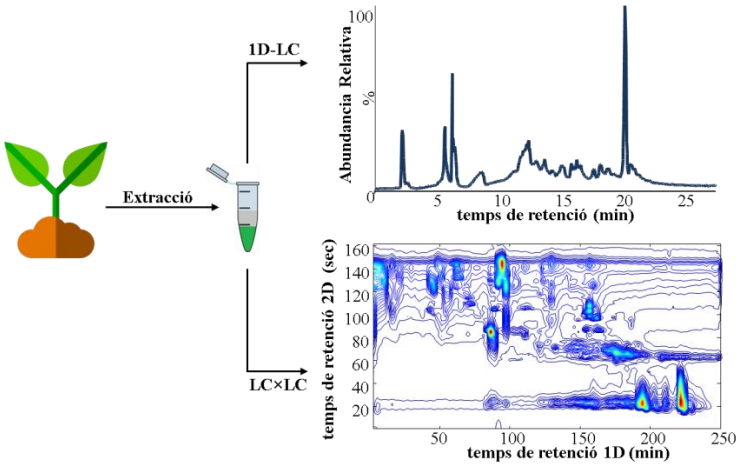


Figura 5.1. Exemple de la millora en la separació obtinguda per LC×LC en comparació amb 1D-LC.

Les limitacions de la 1D-LC a l'hora de separar els anàlits de mostres biològiques en el context de la metabolòmica i la lipidòmica impliquen dos tipus principals de problemes [12, 13]:

A) Les mostres biològiques són extremadament complexes ja que contenen milers de compostos. Aquestes mostres superen la capacitat de la 1D-LC de separar els seus múltiples constituents en pics cromatogràfics únics.

B) Les mostres biològiques contenen alguns compostos molt difícils de resoldre, ja que aquests estan estretament relacionats estructuralment i en les seves propietats cromatogràfiques (per exemple, enantiòmers o isòmers estructurals).

En aquest capítol de la Tesi es proposa l'ús de la LC×LC en estudis de metabolòmica no dirigida. En primer lloc s'estudia l'estructura multidireccional de les dades de LC×LC-MS i s'avalua la possible modelització quimiomètrica del seu comportament. Seguidament es proposa l´ús d'aquesta tècnica per a

l'anàlisi no dirigida en estudis de metabolòmica i lipidòmica. D'una banda, es presenta un mètode de LC×LC-HRMS per a l'anàlisi metabolòmica no dirigida. D'altra banda, es presenta un mètode de LC×LC acoblada a la detecció amb espectrometria de masses en tàndem (LC×LC-MS/MS), pel seu ús en estudis de lipidòmica no dirigida.

Malgrat els avantatges en la separació que ofereix la LC×LC, aquesta presenta un inconvenient important des del punt de vista del tractament de dades, ja que genera dades massives i altament complexes perquè es detecten milions de senyals per cada mostra analitzada. Aquesta complexitat encara és major quan s'utilitza un espectròmetre de masses d'alta resolució en la detecció (LC×LC-HRMS), ja que s'enregistra una quantitat més gran d'informació molt detallada (de l'ordre de Gb per mostra) [12, 14]. Aquest inconvenient fa necessària una estratègia basada en mètodes quimiomètrics per a processar les dades generades, ja que el tractament manual d'aquestes no és pràctic. A més, en els estudis que segueixen una aproximació no dirigida el tractament encara és més complicat, degut a que no es disposa d'informació sobre els anàlits d'interès [12, 15]. Malauradament, a la bibliografia hi ha pocs treballs que utilitzin eines quimiomètriques per analitzar dades de LC×LC i en la majoria es treballa amb dades de LC×LC-DAD. En aquest camp cal destacar el treball realitzat per S. Rutan i els seus col·laboradors, els quals han desenvolupat una estratègia de tractament de dades LC×LC-DAD basada en el mètode MCR-ALS [14, 16, 17]. Per tal d'ajudar a superar el repte que suposa el processament de dades de LC×LC-MS, en aquest capítol es proposa l'ús d'una estratègia de tractament de dades basada en el mètode de MCR-ALS per a resoldre els perfils d'elució i els espectres purs dels components presents a la mostra analitzada. En aquesta estratègia es té en compte l'estructura multidireccional de les dades de LC×LC-MS i s'avalua la seva possible modelització bilineal o trilineal. A més, es proposa una metodologia de compressió d'aquestes dades que faciliti el seu posterior processament.

Aquest treball s'ha realitzat en les següents publicacions:

- **Publicació 5:** *Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil.* M. Navarro-Reig, J. Jaumot, T. A. van Beek, G. Vivó-Truyols, R. Tauler. Talanta 160 (2016), 624-635.

  En aquest article s'avaluen les opcions quimiomètriques existents per a la resolució de les dades de LC×LC-MS. Primer, es discuteix l'estructura multidireccional de les dades de LC×LC-MS i es comparen els resultats obtinguts per diferents mètodes quimiomètrics, basats en models bilineals i trilineals (MCR-ALS, PARAFAC i PARAFAC2). Seguidament, a mode d'exemple,

es descriu pas a pas la resolució i la identificació d'alguns dels compostos coeluïts en una o en ambdues dimensions cromatogràfiques pel cas de l'anàlisi de triacilglicerols (TAGs) en mostres d'oli de blat de moro.

- **Publicació 6:** *Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution*. M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler. Analytical Chemistry 89 (2017), 7675-7683.

  En aquest article es proposa una estratègia de tractament de dades de LC×LC-HRMS. En un primer pas es realitza la compressió de les dades obtingudes en la direcció dels espectres de masses mitjançant la cerca i selecció de les regions d'interès (ROIs), seguida d'una segona compressió en la direcció temporal basada en la transformació d'ondetes (*wavelets*). Posteriorment, el mètode MCR-ALS s'aplica a les dades comprimides per resoldre els perfils d'elució i els espectres de masses purs dels components continguts en les mostres analitzades. La viabilitat de l'estratègia proposada es demostra amb la seva aplicació a un estudi de metabolòmica no dirigida sobre els efectes de diferents condicions ambientals sobre el metabolisme de l'arròs. La part experimental d'aquest article es va dur a terme a la Universitat d'Amsterdam, en el grup de Química Analítica que dirigeix el Prof. Peter Schoenmakers. Aquest grup d'investigació és un dels pioners mundials en el desenvolupament de la LC×LC.

- **Publicació 7:** *An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis*. M. Navarro-Reig, J. Jaumot, R. Tauler. Submited.

  En aquest article es proposa la combinació d'una anàlisi per LC×LC-MS/MS amb l'estratègia de tractament de dades basada en MCR-ALS proposada en la publicació anterior. En aquest cas s'estudien els efectes de la contaminació per arsènic en el creixement de l'arròs. A diferència de les publicacions anteriors la identificació dels lípids d'interès es va realitzar a partir del seu espectre de MS/MS, el qual va fer necessari realitzar l'anàlisi instrumental en dos passos: el primer per a seleccionar els valors de *m/z* de l'ió pare dels lípids d'interès i el segon per a obtenir l'espectre de MS/MS d'aquests lípids.

5.2. Publicació 5

Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oi.

M. Navarro-Reig, J. Jaumot, T. A. van Beek, G. Vivó-Truyols, R. Tauler.

Talanta 160 (2016), 624-635.

Talanta 160 (2016) 624–635

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Chemometric analysis of comprehensive LC × LC-MS data: Resolution of triacylglycerol structural isomers in corn oil

Meritxell Navarro-Reig [a], Joaquim Jaumot [a], Teris A. van Beek [b], Gabriel Vivó-Truyols [c], Romà Tauler [a,*]

[a] IDAEA-CSIC, Department of Environmental Chemistry, Barcelona, Spain
[b] Laboratory of Organic Chemistry, Natural Products Chemistry Group, Wageningen University, Wageningen, The Netherlands
[c] Universiteit van Amsterdam, van't Hoff Institute for Molecular Science, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 May 2016
Received in revised form
28 July 2016
Accepted 1 August 2016
Available online 2 August 2016

Keywords:
LC × LC-MS
Triacylglycerols
Structural isomers
Trilinearity
PARAFAC
MCR-ALS

ABSTRACT

Comprehensive hyphenated two-dimensional liquid chromatography mass spectrometry (LC × LC-MS) is a very powerful analytical tool achieving high throughput resolution of highly complex natural samples. However, even using this approach there is still the possibility of not resolving some of the analytes of interest. For instance, triacylglycerols (TAGs) structural isomers in oil samples are extremely difficult to separate chromatographically due to their very similar structure and chemical properties. Traditional approaches based on current vendor chromatographic software cannot distinguish these isomers from their different mass spectral features. In this work, a chemometric approach is proposed to solve this problem. First, the experimental LC × LC-MS data structure is discussed, and results achieved by different methods based on the fulfilment of the trilinear model are compared. Then, the step-by-step resolution and identification of strongly coeluted compounds from different examples of triacylglycerols (TAGs) structural isomers in corn oil samples are described. As a conclusion, the separation power of two-dimensional chromatography can be significantly improved when it is combined with the multivariate curve resolution method.

1. Introduction

Over the past years, the need for analysis of complex samples has increased considerably in a broad variety of fields, including environmental, clinical or food industry. Liquid chromatography is often the analytical approach chosen for analysing those samples, due to its high ability for resolving complex mixtures. Nevertheless, one-dimension chromatography is not always capable of separating all constituents in complex natural samples. Comprehensive two-dimensional liquid chromatography (LC × LC) appears as a powerful alternative to achieve a better separation of all these constituents [1]. The high resolving power of this technique lies in the fact that under ideal circumstances, when retention mechanisms of the two separation dimensions are uncorrelated, the overall peak capacity is equal to the product of the individual peak capacities of the first and second dimension separations [1,2]. For this reason, during the last few years, major attention has been focused on the development of comprehensive two-dimensional liquid chromatography methodologies coupled to multivariate detectors. For example, molecular absorption diode array or mass spectrometric detectors have been used to deal with complex samples, such as egg yolk, urine, urban aerosols, red wine or polymers, among others [1–9].

Although comprehensive two-dimensional liquid chromatography provides a better resolution than one-dimensional chromatography, there is still the possibility that some of the analytes in complex samples remain unresolved. One example of these limitations is the analysis of triacylglycerols (TAGs) in vegetable oils. Different combinations of three fatty acids connected to the glycerol backbone generate different TAG structural isomers, which can be positional or chain isomers. Positional isomers are generated by variations in the position of the same three fatty acids on the glycerol molecule, whereas chain isomers refer to the different chain lengths, the number of double bonds or position of double bonds of the different fatty acids. Fig. S1 on Supplementary material shows examples of these structural isomers, which have similar chromatographic behaviour. Therefore, their chromatographic separation is highly complex. Moreover, the distinction between structural isomers by their mass spectra is also troublesome. Two TAGs chain isomers can have the same *m/z* value for their protonated molecular ion and some of their diacylglycerols fragments. The case of positional isomers is even more

* Corresponding author.
E-mail address: Roma.Tauler@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.talanta.2016.08.005
0039-9140/

troublesome, because they have the same *m/z* value for their molecular ion and all of their diacylglycerols fragments. Therefore, the achievement of a complete resolution of these compounds is important, especially for those studies which aim at separating multiple compounds with similar structures, such as TAGs in vegetable oils [10].

It is logical to hypothesize that the combination of chemometric methods with comprehensive two-dimensional liquid chromatography could achieve a better resolution of constituents in complex natural samples. There are already multiple examples of the application of chemometric methods to analyse one-dimensional chromatographic data [11–14]. However, in this work, we will focus our study on the analysis of two-dimensional liquid chromatographic data.

Most of the contributions found in the literature dealing with curve resolution of multidimensional separations consider two-dimensional gas chromatography (GC × GC). In this field, it is noteworthy the work done by Synovec [15–18] and by Parastar [19–23]. In contrast to two-dimensional gas chromatography, comprehensive two-dimensional liquid chromatography data analysis is principally done manually using vendor software tools, and the use of chemometric tools for peak detection and resolution is a relatively new concept in this field. For this reason, there are only a few references of using chemometric tools to analyse LC × LC data. In this area, the work of Rutan and co-workers should be highlighted. They have developed a methodology to analyse LC × LC-DAD data based on the application of the multivariate curve resolution–alternating least squares (MCR-ALS) method [24–26]. Other attempts at designing algorithms for detection and quantification of peaks can also be mentioned [27,28]. To the best of our knowledge, the development of a general strategy for the chemometric analysis of LC × LC-MS data for chromatographic peak detection and quantification is still a challenge.

From a chemometric point of view, LC × LC-MS provides three-way data cube and the curve resolution based on trilinear models, such as parallel factor analysis (PARAFAC), might seem appropriate to resolve this type of data. For instance, Synovec has extensively studied the application of the parallel factor analysis (PARAFAC) method to the analysis of GC × GC-MS [15–18]. However, frequent changes in chromatographic peak shapes and in column retention time shifts between consecutive modulations cause the failure in the fulfilment of the trilinear model postulated in these studies. For this reason, bilinear chemometric methods, such as MCR-ALS, emerge as a valid alternative. As stated above, Rutan [9,24–26] and Parastar [19–23] have demonstrated the effectiveness of the MCR-ALS method to deal with LC × LC-DAD and GC × GC-MS data, respectively. Recently, Bortolato and Olivieri [29] compared the application of PARAFAC2 and MCR-ALS for the analysis of chromatographic data considering the effects of changes in peak shapes and shifts. The authors concluded that, in cases where strong coelutions and interferences are present (i.e. in natural samples), successful results could only be obtained using MCR-ALS and they also exposed the limitations of the PARAFAC2 method.

In the presented work, two main goals are pursued. First, the development of a strategy to analyse LC × LC-MS based on the application of curve resolution methods taking as a case of study the complete resolution of TAG structural isomers from vegetable oil samples. Additionally, the study of LC × LC-MS data structure is presented, and consequently the possible use of bilinear and trilinear model based methods for the analysis of LC × LC-MS data is evaluated. Results achieved by bilinear methods are compared with those obtained by methods based on trilinear models.

2. Materials and methods

2.1. LC × LC-MS of triacylglycerols (TAGs) in corn oil samples

In this work, triacylglycerols (TAGs) in corn oil samples were analysed by LC × LC-MS. In the first-chromatographic dimension, an Ag(I)-coated cation exchange (250 × 2.1 mm, 5 μm) column was used, and in the second chromatographic dimension, a C18 (30 × 4.6 mm, 1.8 μm) was employed. Modulations of one minute sections from the first dimension column were introduced into the second dimension column by means of a 10-way valve with two switching loops. Detection was performed by an atmospheric pressure chemical ionization mass spectrometer (APCI-MS) in positive ion mode, and the selected mass spectral range was from *m/z* 250 to 1150 at 1 unit mass resolution.

Fig. 1 describes two possible LC × LC-MS data arrangements obtained in the analysis of a single sample. In the first option, shown in Fig. 1A, the experimental data is arranged in a so-called third-order tensor, i.e. a data cube, **D**. Full scan mass spectra are in the *x*-axis direction of the same data cube, with 901 channels (from 250 to 1150 *m/z* at 1 unit mass resolution). The first-chromatographic dimension is on the *z*-axis direction (from 0 to 179 min) and the second chromatographic dimension is on the *y*-axis direction (from 0 to 60 s). A preliminary data exploration can be done considering different 2D slices of the data cube separately. For instance, a slice of the cube at a particular *m/z* value gives a 2D chromatogram displaying the compounds present at this *m/z* value. A 2D TIC (total ion current) chromatogram can be obtained when all *m/z* intensity values for both chromatographic dimensions are summed. This 2D TIC plot displays the major data features and provides information about retention time regions of more important chromatographic peaks. It is also possible to study individual mass spectra for any combination of the two chromatographic dimensions. Finally, slices at a particular first-dimension elution time show the different chromatographic modulations. Every modulation gives a second dimension chromatographic separation. Therefore, a data matrix can be built for each modulation ($\mathbf{D_K}$ matrix in Fig. 1, where *K* is the number of modulations). These $\mathbf{D_K}$ data matrices are the second dimension LC-MS chromatograms at the different *m/z* values. The rows of this data matrix have the mass spectra at every second dimension retention time, and the columns of this data matrix have the second dimension chromatograms at each *m/z* channel.

Fig. 1B shows a different LC × LC-MS data arrangement in a column-wise augmented data matrix structure. This augmented data matrix ($\mathbf{D_{aug}}$ matrix in Fig. 1B) can be built settling individual $\mathbf{D_K}$ matrices from each modulation one on the top of the other, and keeping *m/z* mode (MS spectra) in common. $\mathbf{D_{aug}}$ matrix contains 901 *m/z* values on its columns (from 250 to 1150 *m/z* at 1 resolution) and 26,492 retention times on its rows, which results from the 179 modulations taken from the first-chromatographic dimension multiplied by the 148 second dimension column retention times in the second chromatographic dimension.

2.2. Data preparation

Acquisition times in the second chromatographic dimension were not an exact multiple of the modulation period, and, therefore, the number of measured mass spectra in each modulation was not constant. In order to construct the data cube **D**, it is necessary that the same number of mass spectra is measured in every modulation. Therefore, the optimal measurement period in the second dimension column was calculated as follows:

$$m = t_m / mean(p) \quad (1)$$

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

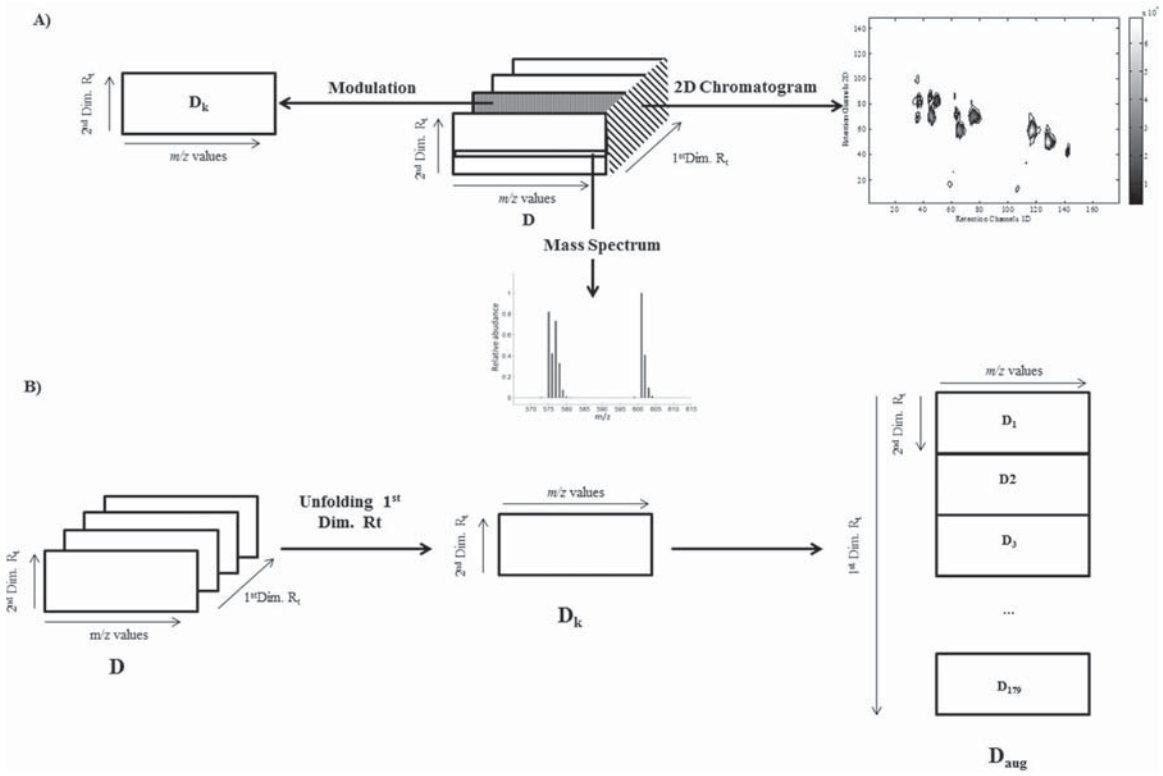


Fig. 1. Graphical description of two possible LC × LC-MS data arrangements. A) LC × LC-MS data arranged in a data cube showing a matrix for each modulation, a mass spectrum for each row of the $\mathbf{D_K}$ matrices and a 2D chromatogram if the cube is considered by the side. B) LC × LC-MS data arranged in a column-wised (in the *m/z* mode) augmented data matrix.

$$m_2 = round(m) \tag{2}$$

$$p_2 = t_m/m_2 \tag{3}$$

Where m is the mean of the number of acquisition times in the second dimension column within each modulation, t_m is the modulation time, mean (p) is the mean of acquisition times in the second dimension column, m_2 is the rounded number of acquisition times within each modulation, and p_2 is the new period (seconds between two consecutive scans) considered. Then, using this new constant period a new acquisition time axis for the second dimension column can be obtained. This axis will be now constant for all modulations. The last step was then the application of a normal kernel smoothing filter with a bandwidth equal to 0.5 to interpolate the signal intensity of every *m/z* value in these newly considered time intervals [30].

2.3. Chemometrics methodology

Two families of chemometric methods have been proposed for the analysis of 2D LC × LC-MS data sets. On the one hand, there are methods that assume that 2D chromatographic data follows a trilinear model such as PARAFAC [31,32] and PARAFAC2 [33,34] methods. On the other hand, there are the chemometric methods based on the basic assumption that 2D chromatographic data fulfils the bilinear model (and not necessarily the trilinear model), such as MCR-ALS [35]. The fulfilment of trilinearity in the case of multidimensional chromatographic (i.e. GC × GC or LC × LC) data is at present a controversial topic, and both types of chemometric methods have been previously used to deal with GC × GC [18,19] and LC × LC data [24,36]. For this reason, in this work, both, bilinear methods (MCR-ALS bilinear), and trilinear methods (PARAFAC, PARAFAC2, MCR-ALS trilinear, MCR-ALS trilinear allowing time shifting) were tested and compared in the analysis of LC × LC-MS data.

2.3.1. MCR-ALS based methods

MCR-ALS is a chemometric method used for the resolution of the contributions of the pure components present in unresolved complex mixtures. MCR-ALS has been used to investigate a wide variety of problems from different fields, in particular multidimensional chromatographic systems [19,24]. MCR-ALS has been extensively described in the literature [37–40] and it is only briefly explained here for the particular case of LC × LC-MS data resolution, with the goal of the resolution of the pure elution profiles in both chromatographic dimensions and the pure mass spectra profiles of the constituents of the analysed sample.

MCR-ALS decomposes experimental data according to a bilinear additive model defined by the multi-sample and multi-wavelength generalization of Lambert-Beer's law for spectroscopic measurements. For a data matrix, such as the one taken from one modulation of the first dimension column in LC × LC-MS, this bilinear model can be written as:

$$\mathbf{D_K} = \mathbf{C_K S^T} + \mathbf{E_K} \tag{4}$$

Where $\mathbf{D_K}$ ($I \times J$) is the experimental data matrix corresponding to one of the second dimension column modulations taken from the first dimension column (where I is the number of rows corresponding to the number of retention times or to the number of mass spectra measured at these retention times, and J is the

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

number of columns corresponding to the *m/z* values of the mass spectra), see Fig. 1B. $\mathbf{C_K}$ ($I \times N$) is the matrix containing the resolved second dimension elution profiles for this modulation (where N is the number of resolved components by MCR-ALS), and $\mathbf{S^T}$ ($N \times J$) has the mass spectra of these N components. Finally, $\mathbf{E_K}$ ($I \times J$) is the matrix of residuals not explained by the MCR model.

This data analysis strategy can be easily extended to the simultaneous analysis of several second dimension column modulations [19]. The same MCR bilinear model described in Eq. (4) can be extended in this case as follows (Fig. 1B):

$$\mathbf{D_{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} = \mathbf{C_{aug}} \mathbf{S^T} + \mathbf{E_{aug}} \tag{5}$$

Where $\mathbf{D_{aug}}$ is the column-wise augmented data matrix containing multiple second dimension modulations. Since K is the number of the second dimension column modulations taken from the first dimension column, the number of rows for $\mathbf{D_{aug}}$ is equal to $I \times K$. Decomposition of matrix $\mathbf{D_{aug}}$ generates $\mathbf{C_{aug}}$ ($IK \times N$) which has second dimension resolved elution profiles at each modulation for the N resolved components. In addition, $\mathbf{S^T}$ ($N \times J$) represents the mass spectra resolved for the N components, common to all the considered modulations, which can be used to identify them. $\mathbf{E_{aug}}$ ($IK \times J$) has the residuals not explained by the model. First dimension elution profiles for every component can be obtained by refolding appropriately every column in $\mathbf{C_{aug}}$ to give a matrix of dimensions ($I \times K$). Every column of the refolded matrix will give the first-dimension elution profile of size (1, K) and, therefore, the matrix of the first-dimension elution profiles of dimensions ($N \times K$) is obtained.

An initial guess of the number of components of $\mathbf{D_{aug}}$ matrix was obtained using singular value decomposition (SVD) algorithm [41]. Initial estimates of pure component spectra ($\mathbf{S^T}$) profiles were computed using a purest spectra detection method based on SIMPLISMA [42,43]. Finally, ALS optimization was carried out applying non-negativity (elution and spectra profiles), spectral normalization (equal height) and spectral equality constraints [44,45].

The described bilinear model for 2D multidimensional chromatographic data can be extended to the trilinear model as follows:

$$\mathbf{D_K} = \mathbf{C T_K S^T} + \mathbf{E_K} \tag{6}$$

Where $\mathbf{D_K}$ is the data matrix for the modulation K, $\mathbf{C}$ is the matrix of the elution profiles of the resolved components, $\mathbf{T_K}$ is a diagonal matrix giving the relative compositions of these components in this particular modulation K, and $\mathbf{S^T}$ has their spectra profiles. In this trilinear model, all $\mathbf{C_K}$ matrices in Eq. (6) are assumed to follow the equation $\mathbf{C_K} = \mathbf{C T_K}$. This implies that profiles in $\mathbf{C_K}$ have exactly the same shape and only can differ in their intensity, defined by a scalar factor in diagonal matrix $\mathbf{T_K}$. From a chromatographic viewpoint, this trilinear model implies that the resolved components have the same elution profile in all the K modulations with exactly the same retention time. MCR-ALS optionally allows imposing the fulfilment of the trilinear model for all or some of the components during the ALS optimization, and also it allows for some between run shift deviations but keeping the same shape. MCR-ALS trilinear forces that all elution profiles of the same component in different modulations have equal shape and appear at the same retention time. MCR-ALS trilinear allowing time shifting, (i.e. partially trilinear), forces also equal shapes in elution profiles but allows variations in retention time among different chromatographic runs.

In all the cases, ALS optimization convergence criterion of the ALS optimization was set to 0.1 (in % of change of standard of deviation of residuals between two consecutive iterations).

2.3.2. PARAFAC based methods

PARAFAC has been extensively used in the resolution of multiway data and claims the uniqueness of obtained solutions as one of its main advantages. PARAFAC models are based on the assumption that a data cube, $\underline{\mathbf{D}}$, can be decomposed into a trilinear combination of pure component responses in each of the three modes. PARAFAC has been proposed for chromatographic data [15–17,32], but the trilinear requirement restricts its use to data where no shift in retention times of the elution profiles of the same component occur among runs and no either peak shape changes are produced. However, these two conditions might be too strong in the current chromatographic practice. PARAFAC has been extensively described in the literature [46–48], and also its application to multidimensional chromatographic data [32,36].

In this work, PARAFAC was initialized using loadings initial estimates obtained by trilinear decomposition (TLD) [49] and the ALS optimization was carried out under non-negativity constraints in the three modes (elution in both chromatographic dimensions and mass spectra profiles) [31].

PARAFAC2 is a variant of PARAFAC method developed by Bro and co-workers [33,34] to deal with three-way data with small shifts in the profiles of one of the data modes. Regarding LC × LC data, PARAFAC2 can supposedly handle small retention time shifts across modulations by allowing a certain freedom in the second dimension elution profiles in matrix $\mathbf{C}$ (see PARAFAC model [31]). Therefore, PARAFAC2 provides as many different $\mathbf{C_K}$ matrices as K number of modulations. To keep uniqueness in the solution, the cross-product of different $\mathbf{C_K}$ has to be constant over all modulations ($\mathbf{C_1^T C_1} = \mathbf{C_2^T C_2} = \ldots = \mathbf{C_K^T C_K}$). In PARAFAC2 constraints cannot be applied in this second dimension chromatographic mode. Consequently, non-negativity constraints can only be applied to mass spectra and first-dimension elution profiles.

The comparison of the capability of PARAFAC2 and MCR-ALS resolving chromatographic data is currently a topic of discussion. Recently, Bortolato and Olivieri [29] compared both methods considering the effects of retention time shifts and changes in peak shapes. The authors concluded that PARAFAC2 would only be able to resolve two overlapped peaks if the time shifts of the two peaks are limited, and no significant changes occur in the profile shapes. These conditions are difficult to accomplish in two-dimensional liquid chromatography. Therefore the study of the structure of LC × LC-MS data is important to know if PARAFAC2 will be able to resolve it.

In all the PARAFAC based analysis the convergence criterion used was the relative change in fit, and was set to 10^{-6}.

2.3.3. Evaluation of the quality of chemometric methods

The quality of the applied methods was assessed using the explained data variance (R^2) and the lack of fit (*LOF*) for both bilinear and trilinear methods. The equations defining these two parameters are:

$$R^2_{2-\text{way}}(\%) = 100 \sqrt{\frac{\sum_{i,j} d_{ij}^2 - \sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \quad R^2_{3-\text{way}}(\%) = 100 \sqrt{\frac{\sum_{i,j,k} d_{ijk}^2 - \sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}} \tag{7}$$

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

$$LOF_{2-way}\,(\%) = 100\sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \qquad LOF_{3-way}\,(\%) = 100\sqrt{\frac{\sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}} \tag{8}$$

Where in the case of two-way arrays $d_{i,j}$ is an element of the experimental data matrix ($\mathbf{D_{aug}}$) and $e_{i,j}$ is the related residual. And, in the case of three-way array d_{ijk} is an element of the experimental data cube ($\underline{\mathbf{D}}$) and e_{ijk} is the related residual.

In order to compare between the mass spectra profiles resolved by the different tested methods in a pairwise mode, angle values between the obtained spectra profiles were calculated. Angle value (α) can be calculated taking into account the cosine of the angle between the two vectors representing the mass spectra profiles obtained for two different methods ($\mathbf{s_1}$ and $\mathbf{s_2}$) [50]:

$$cos\ \alpha = \frac{\mathbf{s}_1\mathbf{s}_2}{\|\mathbf{s}_1\|\|\mathbf{s}_2\|} \tag{9}$$

2.4. Software

PARAFAC and PARAFAC2 were performed using PLS Toolbox 8.0.2 (Eigenvector Research Inc, Wenatche, WA, USA) working under MATLAB R2015b (The Mathworks, Natick, MA, US). MCR-ALS analyses were carried out using the MCR-ALS 2.0 toolbox freely available at www.mcrals.info.

2.5. Abbreviations

TAGs have been abbreviated by means of three letters according to the three fatty acids bound to glycerol. The following abbreviations have been used: (P) palmitic acid (C16:0); (O) oleic acid (C18:1, Δ9); (S) stearic acid (C18:0); (L) linoleic acid (C18:2, Δ9,12); (A) arachidic acid (C20:0); (Po) palmitoleic acid (C16:1, Δ9); (Ln) α-linolenic acid (C18:3, Δ9,12,15). Cx:y indicates x number of carbons and y number of double bonds, and Δ indicates the position of these double bounds.

3. Results and discussion

An example of a 2D TIC LC × LC chromatogram obtained in the analysis of a corn oil sample is shown in Fig. 2A. Seven important chromatographic regions were detected during the elution of TAGs in this LC × LC chromatogram. As an example, one of these seven regions was used to evaluate the trilinear behaviour of LC × LC-MS data. The selected region was number 2, zoomed in Fig. 2B, which corresponded to retention time channels from 41 to 54 in the first-chromatographic dimension, and from 60 to 100 in the second chromatographic dimension (Fig. 2A). Region 2 was the most interesting region of the chromatogram because two pairs of positional isomers were eluted in this region: SLO/SOL and PLO/POL. The SLO and SOL TAG pairs were separated in the first dimension column (peaks 1 and 2 in Fig. 2B) and, therefore, their resolution was straightforward. However, PLO and POL co-eluted in both dimensions (peak 3 in Fig. 2B), and their resolution was more troublesome. The size of the LC × LC-MS data cube, $\underline{\mathbf{D}}$, for this chromatographic region was 41 × 901 × 14 corresponding to 41 retention times of the second chromatographic dimension, 14 modulations from the first-chromatographic dimension and 901 *m/z* values. When the data cube was unfolded into the $\mathbf{D_{aug}}$ matrix (see Fig. 1B), the size of the matrix was 574 rows (41 × 14 retention times) and 901 columns (*m/z* values). Considering $\underline{\mathbf{D}}$ and $\mathbf{D_{aug}}$, the study of the trilinear behaviour of LC × LC-MS data is described below.

3.1. Study of LC × LC-MS data structure

In this section, the three-way structure of LC × LC-MS data is evaluated, in particular if the trilinear model is suitable for its investigation. The first test of this study consisted on the comparison of the SVD of three augmented data matrices containing the same data but arranged in three different ways (Fig. S2A in the Supplementary material). Experimental data was arranged in a column-wise augmented matrix $\mathbf{D_{aug}}$, as well as in a row-wise way (second dimension retention time in the common mode), and in a tube-wise way (each modulation in a row vector). If the trilinear model is accomplished, these three matrices should have the same chemical rank (mathematical rank in the absence of noise) and, therefore, their SVD analysis should give the same number of significant (not-noise) components [51]. Results showed that the number of significant components needed to explain the same amount of variance was lower for the column-wise data matrix, (larger explained variances) than for the other two augmented data matrices, row-wise and tube-wise, (they explain less variance) (Fig. S2B in the Supplementary material). These results indicated that the studied LC × LC-MS experimental data deviated from the trilinear model, and so the application of trilinear methods will be not appropriate for component resolution

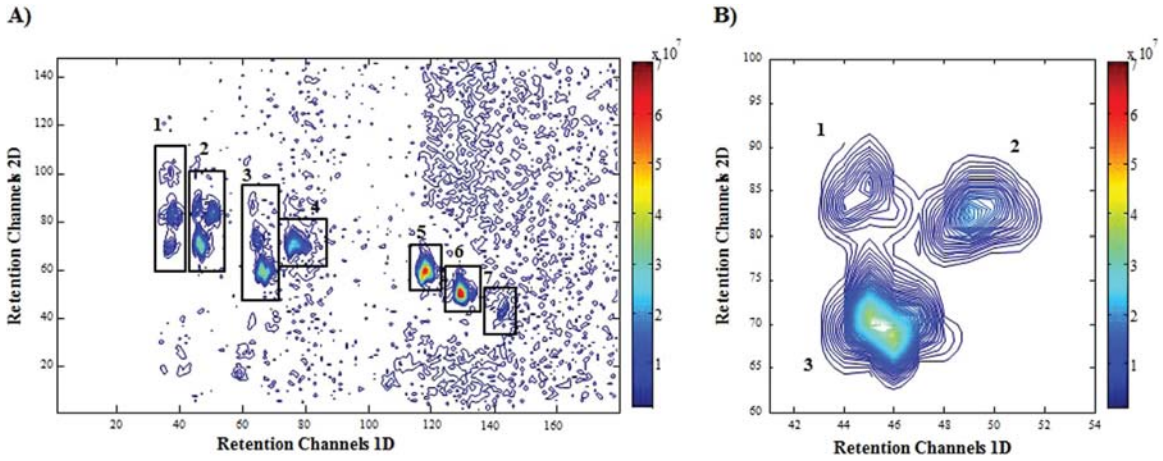


Fig. 2. A) Bidimensional chromatogram of the corn oil sample. Seven important chromatographic regions are marked and numbered. B) Zoomed view of region of interest number 2 with the three important peaks numbered (peak 1 is SOL, peak 2 is SLO and peak 3 corresponds to POL and PLO). Colorbar indicates peak intensity.

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

Table 1
Comparison of results obtained by the application of MCR-ALS, MCR-ALS trilinear allowing time shifting, MCR-ALS trilinear, PARAFAC and PARAFAC2 methods.

Method	Number of components	R^2 (%)[a]	*LOF* (%)[b]	Number of ALS iterations	Core consistency (%)	Selected for comparison[c]
MCR-ALS	3	99.6	6	18	-	
	4	99.8	4	17	-	
	5	99.9	3	23	-	*
MCR-ALS Trilinear allowing time shIfting	3	97.0	18	3	-	
	4	97.0	17	3	-	
	5	97.1	17	4	-	*
MCR-ALS Trilinear	3	89.3	33	2	-	
	4	89.3	32	6	-	
	5	95.3	21	11	-	*
PARAFAC	3	96.2	19	61	67	
	4	97.9	14	51	80	*
	5	98.6	11	23	65	
PARAFAC 2	3	99.6	7	1295	99	
	4	99.8	4	918	99	
	5	99.9	3	11,287	99	*

[a] Calculated according to Eq. (7).
[b] Calculated according to Eq. (8).
[c] Indicates the model chosen to compare the five methods.

purposes.

To further study the effects of this deviation from the trilinear model, MCR-ALS, MCR-ALS trilinear, MCR-ALS trilinear allowing time shifting, PARAFAC and PARAFAC2 were applied and compared in the analysis of the selected chromatographic region (region 2 shown in Fig. 2B). Table 1 shows the information related to the explained variance (R^2) and the lack of fit (*LOF*) obtained by each of the models using three, four and five components. In the case of application of MCR-ALS, MCR-ALS trilinear, MCR-ALS trilinear allowing time shifting and PARAFAC2 methods, the finally selected number of components for further comparison was five. On the contrary, in the case of the PARAFAC method, the number of components was set to four because this was the model that gave the highest core consistency. When a data set does not fulfil the trilinear model requirements (see Chemometrics methodology section) for a particular number of components, methods based on the trilinear model will fit the data less efficiently (with a lower fit) and will give more unreasonable shape component profiles than other methods based on a bilinear model [35,52]. Results presented in Table 1 show a clear gap between the *LOF* values of MCR-ALS and PARAFAC2 and the other three models (MCR-ALS fully or partially trilinear and PARAFAC). Moreover, MCR-ALS trilinear allowing time shifting gave better values of *LOF* and R^2 than MCR-ALS trilinear but worse than bilinear MCR-ALS. This result indicated that not only retention times, but also peak shapes changed among modulations, and consequently the analysed LC × LC-MS data could not be considered as trilinear. An example of the differences in elution profiles resolved by MCR-ALS, MCR-ALS trilinear allowing time shifting and MCR-ALS trilinear is shown in Fig. S3 in Supplementary material.

Furthermore, core consistency diagnostic can also be used to evaluate the trilinear behaviour of a data set [53]. When the evaluated data cannot be described by a trilinear model or too many components are used in the model, core consistency will differ from 100% [53]. In our case, the obtained results for the PARAFAC model gave a core consistency of 80% when 4 components were used, and when the model was performed using three or five components the resulted core consistency decreased significantly (approximately to 65%). All these results indicated that the analysed LC × LC-MS data did not accomplish the trilinear model requirements. Most probably, this deviation from the trilinearity is caused in part by changes in retention times of the same peak on the second dimension column modulation (shifting) and in another part by changes in peak shapes. A short explanation of PARAFAC2 limitations is given below that agrees with those arguments described in reference [29].

In order to complete the comparison between these methods, the resolved elution profiles and mass spectra by each resolution method were contrasted. Fig. 3 shows the obtained elution profiles in both chromatographic dimensions after applying the five methods. Considering the shape of the profiles recovered by the different methods, the obtained profiles may look reasonably good from a chemical point of view. Only the second column profiles resolved by PARAFAC2 (Fig. 3E) were an exception. Since in PARAFAC2 non-negativity constraints could not be applied to second dimension chromatographic mode, the profiles resolved by this method were worse from a physical point of view than those provided by PARAFAC, MCR-ALS, MCR-ALS trilinear methods allowing time shifting or MCR-ALS trilinear. This outcome agreed with the results obtained by Bortolato and Olivieri in the comparison of MCR-ALS and PARAFAC2 for chromatographic analysis [29]. In their work, Bortolato and Olivieri exposed PARAFAC2 limitations and the main reasons for its failure. Briefly, the fact that PARAFAC2 requires that the cross-product of different $\mathbf{X}_k$ should be equal in all modulations implies two important consequences: (1) peak shape should be the same in all modulations for every component *n*, (2) peak shifts are only tolerated in PARAFAC2 for non-coeluting components and in the absence of interferences, otherwise the cross-product constant condition assumed by PARAFAC2 is not fulfilled anymore. To further study retention time shifting and changes in peak shapes, SVD analysis of the resolved elution profiles was performed (Table S1). If the studied LC × LC-MS data were trilinear, then retention time and peak shape of a chemical compound would be equal in all modulations. Consequently, the SVD of the resolved elution profile of an individual compound would give only one significant component. Since MCR-ALS trilinear forced the resolved elution profiles to be constant over all modulations, the SVD of resolved elution profiles gave only one relevant singular component in all cases. Otherwise, the SVD of SLO, POL, SOL and PLO elution profiles resolved by MCR-ALS bilinear gave 6, 5, 10, and 9 significant components respectively, which clearly indicated that these profiles were not trilinear. Moreover, when the SVD was performed on SLO, POL, SOL and PLO elution profiles resolved by MCR-ALS trilinear allowing time shifting, the obtained number of significant components was 3, 4, 4 and 7 respectively. All these results demonstrated that even allowing time shifting, more than one component was necessary to explain the elution profile of the same component over all modulations, indicating the non-fulfilment of the trilinear model for this component. Moreover, the lack of trilinearity of the

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

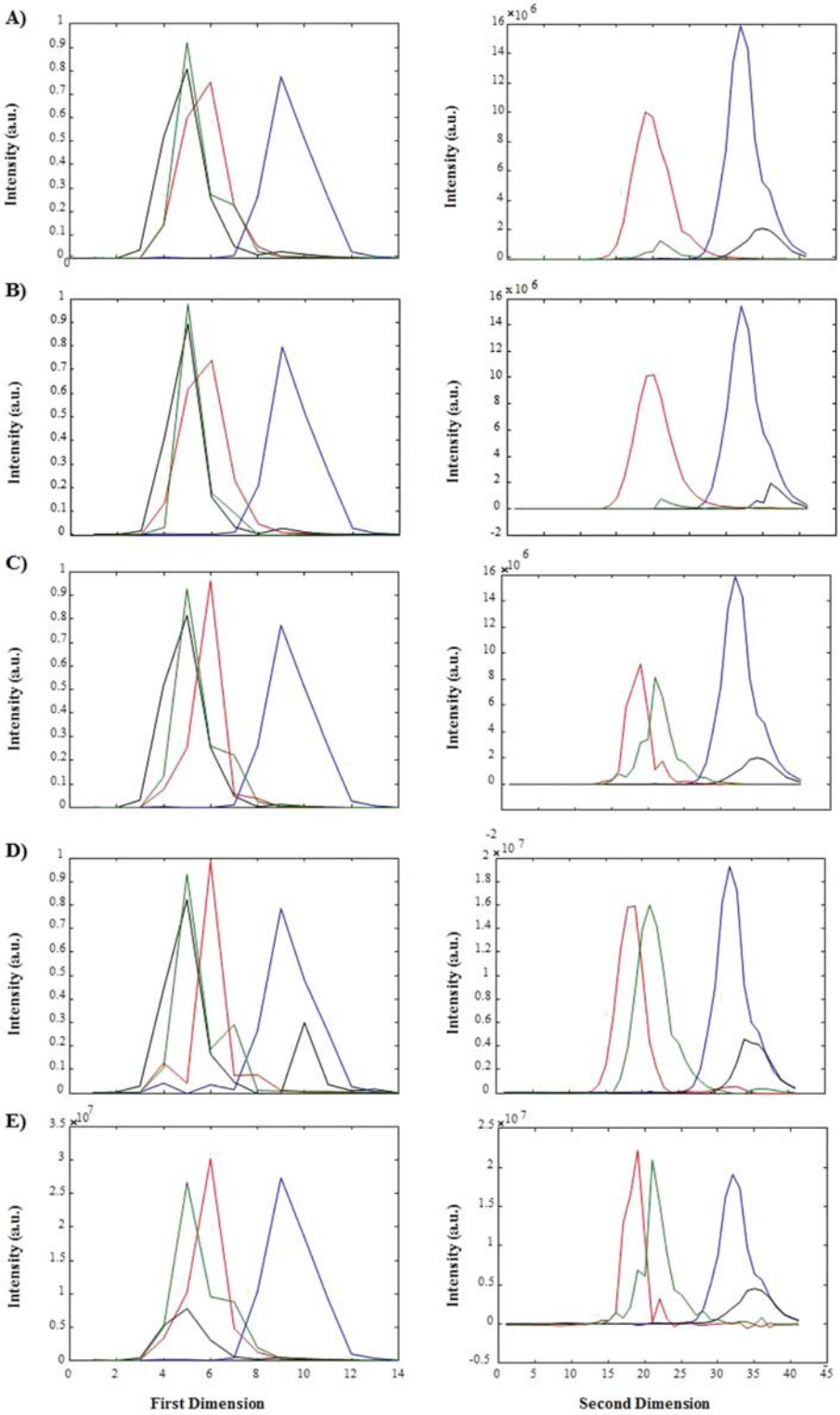


Fig. 3. First- (left) and second- (right) chromatographic dimension elution profiles recovered with: A) MCR-ALS; B) MCR-ALS trilinear allowing time shifting; C) MCR trilinear; D) PARAFAC; and E) PARAFAC2. Colours indicate the four TAGs: Blue is SLO, black is SOL, red is POL and green is PLO.

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

Table 2
Data matching between the resolved mass spectra of SLO, SOL, POL and PLO in region 2 by MCR-ALS, MCR-ALS trilinear allowing time shifting, MCR-ALS trilinear PARAFAC and PARAFAC2.

	Angle (α)			
	SLO	SOL	POL	PLO
MCR-ALS VS MCR-ALS Trilinear	0.6	1.9	0.5	**34**
MCR-ALS VS MCR-ALS Trilinear allowing time shifting	0.2	1.1	1.6	**23**
MCR-ALS VS PARAFAC	0.7	4.5	1.0	**35**
MCR-ALS VS PARAFAC2	0.3	2.3	1.0	**34**
MCR-ALS Trilinear allowing time shifting VS MCR-ALS trilinear	0.4	2.4	1.6	**38**
MCR-ALS Trilinear allowing time shifting VS PARAFAC	0.9	3.9	1.8	**39**
MCR-ALS Trilinear allowing time shifting VS PARAFAC2	0.2	3.1	1.8	**25**
MCR-ALS Trilinear VS PARAFAC	1.2	5.3	0.8	1.3
MCR- Trilinear VS PARAFAC 2	0.4	1.9	0.6	0.6
PARAFAC VS PARAFAC 2	1.0	6.1	0.9	0.6

considered data was not only due to retention time shifting but also to changes in peak shapes between consecutive modulations.

Finally, comparison of the resolved mass spectra for each one of the pure components (SLO, SOL, POL and PLO) obtained by each method is shown in Table 2. Results of this comparison are shown in terms of the angle between the vectors defined by the resolved mass spectra for each pair of methods. Most of these comparisons showed a rather good matching between profiles, with angles lower than 2 degrees. However, the resolution of the PLO isomer showed angles higher than 20 degrees between MCR-ALS and the other methods. This compound was the most difficult to resolve because it was completely co-eluted with POL and only the MCR-ALS bilinear method properly resolved PLO.

Taking into account all the results obtained until now, the analysed LC × LC-MS data were confirmed not to have behaviour consistent with a trilinear model. Consequently, the use of the PARAFAC method would require a preliminary peak alignment procedure as the one suggested by Allen and Rutan to deal with LC × LC-DAD data [36] and, probably, would still have problems with the changes in peak shape changes in case of coelution. On the other hand, MCR-ALS or PARAFAC2 did not require the initial peak alignment step, which can be considered to be an advantage. Finally, PARAFAC2 was not able to obtain reliable second dimension elution profiles because of the application of constraints to this mode were not allowed and in many circumstances. So, PARAFAC2 did not solve data appropriately due to peak shape changes among modulations. On the contrary, MCR-ALS was able to resolve LC × LC-MS data properly and allowed an easy interpretation of the achieved results.

3.2. Application of MCR-ALS to LC × LC-MS data

The potential application of MCR-ALS to LC × LC-MS data is discussed in this section for the resolution of TAGs isomers. With this aim, the analysis of chromatographic regions number 1, 2 and 3 (see Fig. 2A) is described below. Regions 1 and 3 (Zoomed in Fig. S4 of the Supplementary material) were selected because these regions contained a pair of chain isomers that co-eluted in both chromatographic dimensions and could not be resolved by traditional means. In the case of chain isomers, mass spectra show the same protonated molecular ion for both isomers and, also, some diacylglycerol fragments are the same. Region 1 corresponds to retention time channel from 30 to 42 in the first-chromatographic dimension and from 55 to 110 in the second chromatographic dimension. Region 3 corresponds to retention time channel from 61 to 70 in the first-chromatographic dimension and from 50 to 90 in the second chromatographic dimension. The resolution of region number 2 was especially challenging due to the elution of two pairs of positional isomers. In this case both, the protonated molecular ion and all diacylglycerol fragments were the same for both positional isomers, which could be only distinguished by the relative abundance of their fragments. The chromatographic characteristics of this region have already been described in the previous section.

3.2.1. MCR-ALS resolution TAGs positional isomers

The resolution of the two pairs of positional isomers contained in region number 2 shows the advantages of the application of MCR-ALS to LC × LC-MS data. As explained in the previous section, the MCR-ALS model with five components was selected. The percentage of explained data variance (R^2) was 99.9% and the *LOF* was 3.1%. Four of these components (see below) were assigned to TAGs positional isomers (SOL, SLO, POL and PLO) whereas the fifth component was assigned to a baseline contribution. In this last case, both chromatographic profiles and resolved mass spectrum showed profiles that could not be associated with a compound with chemical meaning.

Fig. 4, shows the results obtained after MCR-ALS resolution of region 2. Fig. 4A depicts the first-column elution profiles of the resolved components (baseline component is also shown). As it can be observed, SLO (blue profile) was separated from the other three compounds in the first-column (argentation chromatography) due to the interaction of Ag(I) ions with the double bonds, so its identification was straightforward. Fig. 4B corresponds to the second dimension column (reverse phase C18) elution profiles, where it can be seen that SOL (black) was separated from POL (red) and PLO (green) due to the difference in their partition numbers (48 for SOL and 46 for POL and PLO). Therefore, SOL identification was also possible. On the contrary, PLO (green) was embedded in PLO (red), what made their separation and identification more difficult. Fig. 5A shows the resolved mass spectra of the four components, which were used to identify the TAGs by comparison with their experimental reference mass spectra. TAGs mass spectra were characterised by the mass of the protonated molecular ion ($[M+H]^+$) and the mass of all possible protonated diacylglycerols fragments ($[DG+H]^+$). As it is shown in this Figure, the two TAGs positional isomers have exactly the same *m/z* values for the masses of the $[M+H]^+$ and three $[DG+H]^+$ ions and their mass spectra only differ in the relative abundance of the three different fragments of $[DG+H]^+$. Moreover, since the two positional isomers POL and PLO are eluted at the same retention time, using traditional and commercial chromatographic software based on individual *m/z* signals of various ions their distinction is really difficult and usually fail in their qualitative and quantitative determination. Outstandingly, this difficult problem could be solved by the proposed MCR-ALS method, which was capable of resolving and identifying separately these two isomers. This fact is highly relevant in the study of positional isomers.

As an example of how TAGs identification was carried out, inserts in Fig. 5A show the zoomed diacylglycerols mass region (from 565 to 610 *m/z*) of the resolved MCR-ALS mass spectra for the four TAGs (SOL, SLO, POL and PLO) and Fig. 5B depicts their experimental reference mass spectra obtained from references [10,54]. Fragmentation of the POL and PLO isomers gives three possible diacylglycerols fragments ($[DG+H]^+$): LO (601.6 *m/z*), LP (575.5 *m/z*) and PO (577.5 *m/z*). The relative abundance of these three diacylglycerol fragments depends on the different probability that ester bonds of TAGs break. The fatty acid least likely to split off is the most hindered one, i.e. at position 2 of the glycerol. Consequently, for the POL isomer the signal of LP (breaking O-ester bond, at 575.5 *m/z*) is the least abundant (ratio PO/PL=1.39) whereas for the PLO isomer the PO (breaking L-ester bond, at 577.5 *m/z*) has the lowest abundance (ratio PO/PL=0.71). The main

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

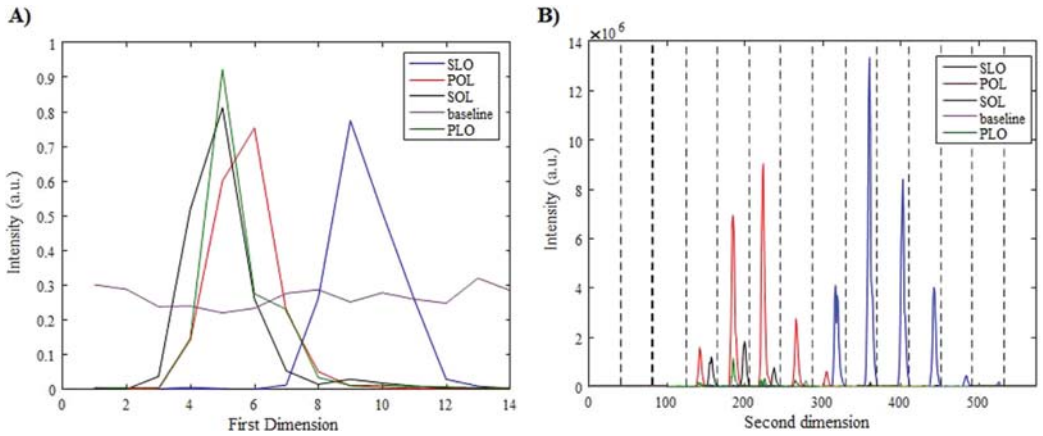


Fig. 4. MCR-ALS resolved elution profiles for chromatographic region 2. A) MCR-ALS resolved first-column elution profiles. B) MCR-ALS resolved second column elution profiles.

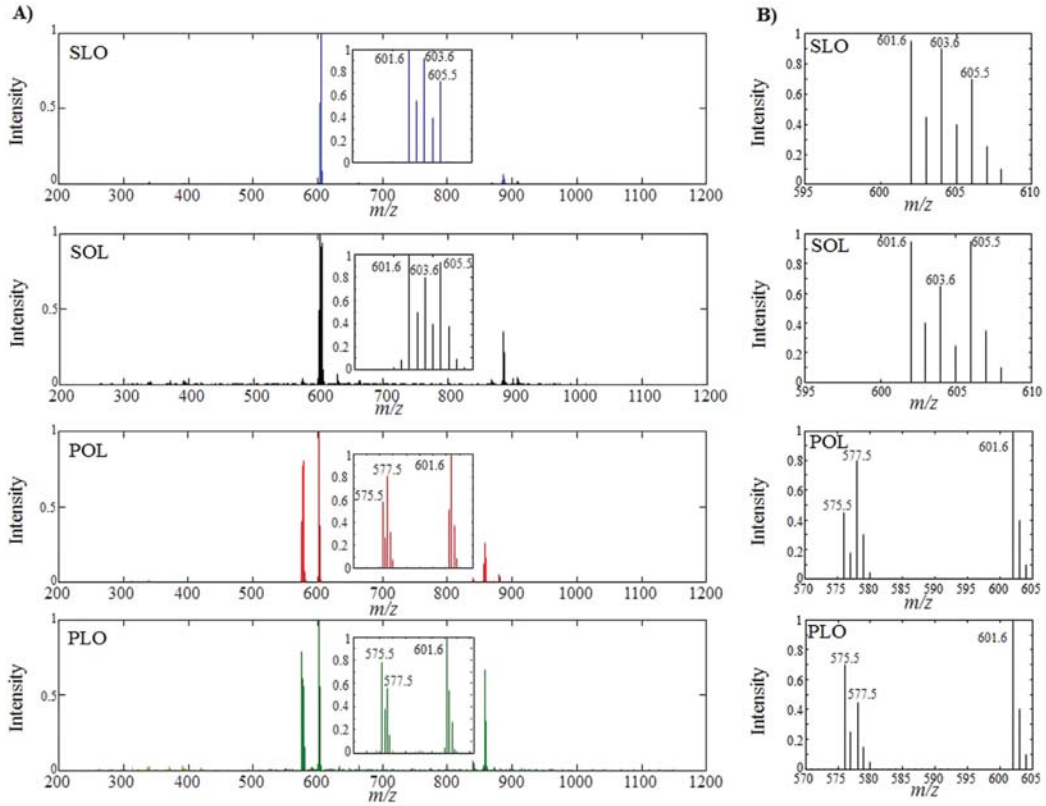


Fig. 5. A) MCR-ALS resolved pure mass spectra profiles for SLO, SOL, POL and PLO. Inserts show the diacylglycerols mass region (from 565 to 610 *m/z*) of each TAG. B) Reference model spectra of the TAGs used for their identification.
SOL and SLO spectra adapted from [54] and POL and PLO adapted from [10].

difference between the mass spectra of PLO and POL isomers lies only in the relative abundance of their signals at 575.5 and 577.5 *m/z*. In Fig. 5A, it can be seen that in the resolved mass spectrum for POL the signal intensity of 577.5 *m/z* (0.8 a.u.) was higher than the signal intensity at 575.5 *m/z* (0.6 a.u.), as happen in its experimental reference spectra (Fig. 5B). In the case of the resolved mass spectrum for PLO (Fig. 5A), the signal intensity at 575.5 *m/z* (0.8 a.u.) was higher than the signal intensity of 577.5 *m/z* (0.55 a.u.), in the same way than in the experimental reference spectra (Fig. 5B). This shows again the great potential of MCR resolution compared to traditional means where these two species would be extremely difficult to be resolved when they are coeluting. SLO and SOL isomers were identified using the same described strategy.

3.2.2. *MCR-ALS resolution of TAGs chain isomers*

Region 1 of Fig. 2A was resolved by an MCR-ALS model with nine components. The percentage of explained variance (R^2) was

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

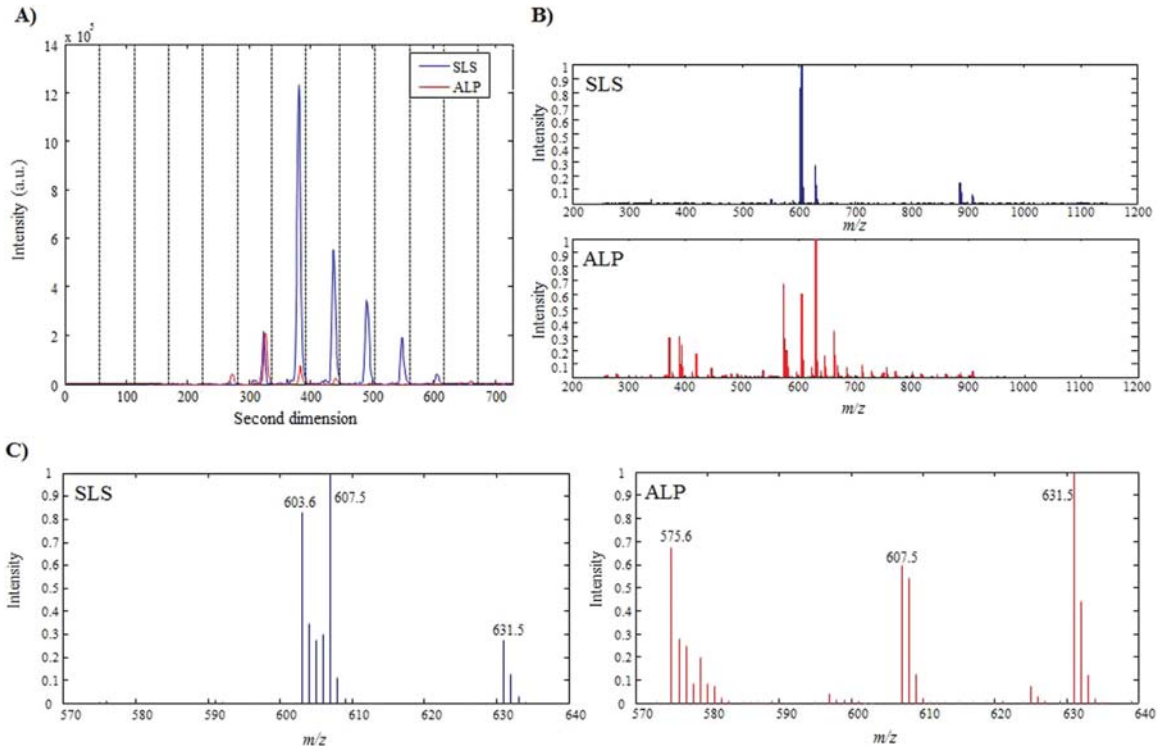


Fig. 6. MCR-ALS resolution of SLS and ALP chain isomers. A) Resolved elution profiles in the fourteen modulations taken from the first dimension column. B) Resolved mass spectra. C) MCR-ALS resolved mass spectra zoomed in the diacylglycerols mass region (from 570 to 640 *m/z*).

99.5% and *LOF* was 6.8%. Eight of these components were assigned to TAGs (SLS, ALP and other TAGs whose resolution was not difficult) and the ninth component was assigned to a baseline contribution. Fig. 6 shows the resolved elution profiles and mass spectra of SLS (MCR-ALS resolved component 3) and ALP (MCR-ALS resolved component 7). Fig. 6A depicts the resolved elution profiles, where SLS and ALP were strongly co-eluting in both chromatographic dimensions, making their resolution challenging. Fig. 6B shows the MCR-ALS resolved mass spectra of SLS and ALP and Fig. 6C depicts a zoom of the diacylglycerols mass region (from 570 to 640 *m/z*) of these two spectra. SLS and ALP chain isomers have exactly the same *m/z* value of the protonated molecular ion $[M+H]^+$ (887.5 *m/z*) and of one of the diacylglycerol fragments, $[DG+H]^+$ (607.5 *m/z* for both AP and SS). However, they also differ in the *m/z* values of other of the two diacylglycerols fragments (575.6 *m/z* for LP, 631.5 *m/z* for AL and 603.6 *m/z* for LS in Fig. 6) [10]. The main difference between the mass spectra of ALP and SLS lies in the presence of the signals at 575.6 and 631.5 *m/z*, which should appear only in ALP mass spectrum and of the signal at 603.6, which should appear only in SLS mass spectrum. In Fig. 6C, it can be seen that signal intensities at 575.6 and 631.6 were low in the MCR-ALS resolved mass spectrum of SLS, whereas in the MCR-ALS resolved mass spectrum of ALP these signals showed a high intensity. Moreover, the signal at 603.6 *m/z* is only present in SLS mass spectrum. Therefore, the two chain isomers could be identified.

Finally, a new MCR-ALS model with five components was employed to resolve region 3 of Fig. 2. The percentage of explained variance (R^2) was 99.8% and the *LOF* was 3.2%. Four of these components were assigned to TAGs (POLn, PoOL and to other TAGs whose resolution was rather simple with no coelution), and the fifth component was again assigned to a noisy background contribution. Fig. S5 in Supplementary material shows the resolved elution profiles and mass spectra of PoOL (MCR-ALS resolved component 1) and POLn (MCR-ALS resolved component 4), two chain isomers. The resolved elution profiles are represented on Fig. S5A, where can be observed that the two chain isomers were heavily coeluting, with POLn totally embedded in PoOL which hindered their identification. Fig. S5B shows the MCR-ALS resolved mass spectra of PoOL and POLn and Fig. S5C gives the zoomed *m/z* region where these two diacylglycerols have their signals (from 555 to 625 *m/z*). PoOL and POLn chain isomers have exactly the same *m/z* value of the masses of $[M+H]^+$ (855.6 *m/z*) and one of the fragments $[DG+H]^+$ (573.5 *m/z* for PoL and PLn) but they differ in the values of the other two diacylglycerols fragments (575.5 *m/z* for PoO, 603.5 *m/z* for LO, 577.5 for PO and 599.5 *m/z* for OLn) [10]. Fig. S5C shows that the main differences observed in the resolved mass spectra of PoOL and POLn were in the presence of the signal at 575.5 *m/z* in PoOL mass spectrum and of the signals at 577.5 *m/z* and 599.5 *m/z* in the POLn one. Therefore, this is again a clear example of the potential of using the MCR-ALS proposed approach to resolve and distinguish these two extremely similar chain isomers.

4. Conclusions

A comprehensive study about the potential of using the MCR-ALS method for the analysis of LC × LC-MS data has been performed. First, it has been proven that LC × LC-MS data does not fulfil the frequently postulated trilinearity requirements since both large retention time shifts and changes in peak shapes between consecutive modulations are occurring in general. Thus, methods based on bilinear models are more suitable to analyse LC × LC-MS data than trilinear model based methods like PARAFAC or PARAFAC2. Using trilinear model based methods, peak retention time

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

shifts and shape changes between modulations and therefore peak alignment procedures are needed. It is proposed therefore the general use of the MCR-ALS method to take full advantage of LC × LC-MS data in the analysis of complex natural samples where strong coelutions and spectra overlap (isomeric species) are frequently encountered. MCR-ALS strategy is shown to be rather simple, and results can be easily interpreted from the directly resolved elution profiles in the two chromatographic dimensions and from the resolved mass spectra of every constituent contained in the mixture sample. Using these profiles, qualitative (identification) information can be easily derived and used for interpretation. Different examples are shown and described in detail in the analysis of different co-eluting triacylglycerols structural isomers in vegetable oils. For instance, POL and PLO positional isomers were resolved and identified although they were coeluting completely in both chromatographic dimensions.

Conflict of interest statement

The authors declare that they have no competing interests.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement no. 320737. Also, recognition from the Catalan government (grant 2014 SGR 1106) is acknowledged.

Appendix A. Supplementary information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.talanta.2016.08.005.

References

[1] D.R. Stoll, X. Li, X. Wang, P.W. Carr, S.E.G. Porter, S.C. Rutan, Fast, comprehensive two-dimensional liquid chromatography, J. Chromatogr. A 1168 (2007) 3–43.
[2] R.A. Shellie, P.R. Haddad, Comprehensive two-dimensional liquid chromatography, Anal. Bioanal. Chem. 386 (2006) 405–415.
[3] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11.
[4] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, Recent advancements in comprehensive two-dimensional separations with chemometrics, J. Chromatogr. A 1184 (2008) 341–352.
[5] J.T.V. Matos, S. Freire, R. Duarte, A.C. Duarte, Profiling water-soluble organic matter from urban aerosols using comprehensive two-dimensional liquid chromatography, Aerosol Sci. Technol. 49 (2015) 381–389.
[6] S. Julka, H. Cortes, R. Harfmann, B. Bell, A. Schweizer-Theobaldt, M. Pursch, L. Mondello, S. Maynard, D. West, Quantitative characterization of solid epoxy resins using comprehensive two dimensional liquid chromatography coupled with electrospray ionization-time of flight mass spectrometry, Anal. Chem. 81 (2009) 4271–4279.
[7] M. Lísa, E. Cífková, M. Holčapek, Lipidomic profiling of biological tissues using off-line two-dimensional high-performance liquid chromatography-mass spectrometry, J. Chromatogr. A 1218 (2011) 5146–5156.
[8] Q. Yang, X. Shi, Y. Wang, W. Wang, H. He, X. Lu, G. Xu, Urinary metabonomic study of lung cancer by a fully automatic hyphenated hydrophilic interaction/RPLC-MS system, J. Sep. Sci. 33 (2010) 1495–1503.
[9] H.P. Bailey, S.C. Rutan, Comparison of chemometric methods for the screening of comprehensive two-dimensional liquid chromatographic analysis of wine, Anal. Chim. Acta 770 (2013) 18–28.
[10] E.J.C. van der Klift, G. Vivó-Truyols, F.W. Claassen, F.L. van Holthoon, T.A. van Beek, Comprehensive two-dimensional liquid chromatography with ultraviolet, evaporative light scattering and mass spectrometric detection of triacylglycerols in corn oil, J. Chromatogr. A 1178 (2008) 43–55.
[11] J.M. Bosque-Sendra, L. Cuadros-Rodríguez, C. Ruiz-Samblás, A.P. de la Mata, Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data-A review, Anal. Chim. Acta 724 (2012) 1–11.
[12] L.W. Hantao, H.G. Aleme, M.P. Pedroso, G.P. Sabin, R.J. Poppi, F. Augusto, Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: a review, Anal. Chim. Acta 731 (2012) 11–23.
[13] S. Mas, A. de Juan, R. Tauler, A.C. Olivieri, G.M. Escandar, Application of chemometric methods to environmental analysis of organic pollutants: a review, Talanta 80 (2010) 1052–1067.
[14] Y.P. Lin, D.Y. Si, C.X. Liu, Advances of liquid chromatography/mass spectrometry combined with chemometric approaches applied to metabonomics, Chin. J. Anal. Chem. 35 (2007) 1535–1540.
[15] A.E. Sinha, C.G. Fraga, B.J. Prazen, R.E. Synovec, Trilinear chemometric analysis of two-dimensional comprehensive gas chromatography-time-of-flight mass spectrometry data, J. Chromatogr. A 1027 (2004) 269–277.
[16] A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC-TOF-MS data: analysis of metabolites in human infant urine, J. Chromatogr. A 1058 (2004) 209–215.
[17] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709.
[18] L.W. Hantao, B.R. Toledo, F.A. De Lima Ribeiro, M. Pizetta, C.G. Pierozzi, E. L. Furtado, F. Augusto, Comprehensive two-dimensional gas chromatography combined to multivariate data analysis for detection of disease-resistant clones of Eucalyptus, Talanta 116 (2013) 1079–1084.
[19] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution, Anal. Chem. 83 (2011) 9289–9297.
[20] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC × GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemom. Intell. Lab. Syst. 117 (2012) 80–91.
[21] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares, Anal. Bioanal. Chem. 405 (2013) 6235–6249.
[22] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297.
[23] J.R. Radović, K.V. Thomas, H. Parastar, S. Díez, R. Tauler, J.M. Bayona, Chemometrics-assisted effect-directed analysis of crude and refined oil using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, Environ. Sci. Technol. 48 (2014) 3074–3083.
[24] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography - a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Anal. Chim. Acta 859 (2015) 87–95.
[25] C. Tistaert, H.P. Bailey, R.C. Allen, Y. Vander Heyden, S.C. Rutan, Resolution of spectrally rank-deficient multivariate curve resolution: alternating least squares components in comprehensive two-dimensional liquid chromatographic analysis, J. Chemom. 26 (2012) 474–486.
[26] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemom. Intell. Lab. Syst. 106 (2011) 131–141.
[27] P.G. Stevenson, M. Mnatsakanyan, G. Guiochon, R.A. Shalliker, Peak picking and the assessment of separation performance in two-dimensional high performance liquid chromatography, Analyst 135 (2010) 1541–1550.
[28] G. Vivó-Truyols, Bayesian approach for peak detection in two-dimensional chromatography, Anal. Chem. 84 (2012) 2622–2630.
[29] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV–vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Anal. Chim. Acta 842 (2014) 11–19.
[30] A.W. Bowman, in: A. Azzalini (Ed.), Applied smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations, Clarendon Press, Oxford, 1997.
[31] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[32] J.C. Hoggard, R.E. Synovec, Automated resolution of nontarget analyte signals in GC × GC-TOFMS data using parallel factor analysis, Anal. Chem. 80 (2008) 6677–6688.
[33] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309.
[34] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (1999) 275–294.
[35] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (2001) 749–772.
[36] R.C. Allen, S.C. Rutan, Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets, Anal. Chim. Acta 723 (2012) 7–17.

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

[37] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12.
[38] A. de Juan, R. Tauler, Factor analysis of hyphenated chromatographic data. Exploration, resolution and quantification of multicomponent systems, J. Chromatogr. A 1158 (2007) 184–195.
[39] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976.
[40] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: progress in concepts and applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176.
[41] G.H. Golub, C.F.V. Loan, Matrix Computations, third ed., Johns Hopkins University Press, Baltimore, 1996.
[42] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432.
[43] W. Windig, D.A. Stephenson, Self-modeling mixture analysis of second-derivative near-infrared spectral data using the simplisma approach, Anal. Chem. 64 (1992) 2735–2742.
[44] A. de Juan, S.C. Rutan, R. Tauler, Two-way data analysis: multivariate curve resolution - iterative resolution methods, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, Oxford, 2009, pp. 325–344.
[45] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: extended multivariate curve resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, Oxford, 2009, pp. 473–505.
[46] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605.
[47] N.M. Faber, R. Bro, P.K. Hopke, Recent developments in CANDECOMP/PARAFAC algorithms: a critical review, Chemom. Intell. Lab. Syst. 65 (2003) 119–137.
[48] K.S. Booksh, Z. Lin, Z. Wang, B.R. Kowalski, Extension of trilinear decomposition method with an application to the flow probe sensor, Anal. Chem. 66 (1994) 2561–2564.
[49] N.M. Faber, Towards a rehabilitation of the generalized rank annihilation method (GRAM), Anal. Bioanal. Chem. 372 (2002) 683–687.
[50] J. Jaumot, J.C. Menezes, R. Tauler, Quality assessment of the results obtained by multivariate curve resolution analysis of multiple runs of gasoline blending processes, J. Chemom. 20 (2006) 54–67.
[51] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[52] A. de Juan, S.C. Rutan, R. Tauler, D. Luc Massart, Comparison between the direct trilinear decomposition and the multivariate curve resolution-alternating least squares methods for the resolution of three-way data sets, Chemom. Intell. Lab. Syst. 40 (1998) 19–32.
[53] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (2003) 274–286.
[54] P. Dugo, O. Favoino, P.Q. Tranchida, G. Dugo, L. Mondello, Off-line coupling of non-aqueous reversed-phase and silver ion high-performance liquid chromatography-mass spectrometry for the characterization of rice oil triacylglycerol positional isomers, J. Chromatogr. A 1041 (2004) 135–142.

Please cite this article as: M. Navarro-Reig, et al., Talanta (2016), http://dx.doi.org/10.1016/j.talanta.2016.08.005

Informació Suplementària a la Publicació 5

Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oi.

M. Navarro-Reig, J. Jaumot, T. A. van Beek, G. Vivó-Truyols, R. Tauler.

Talanta 160 (2016), 624-635.

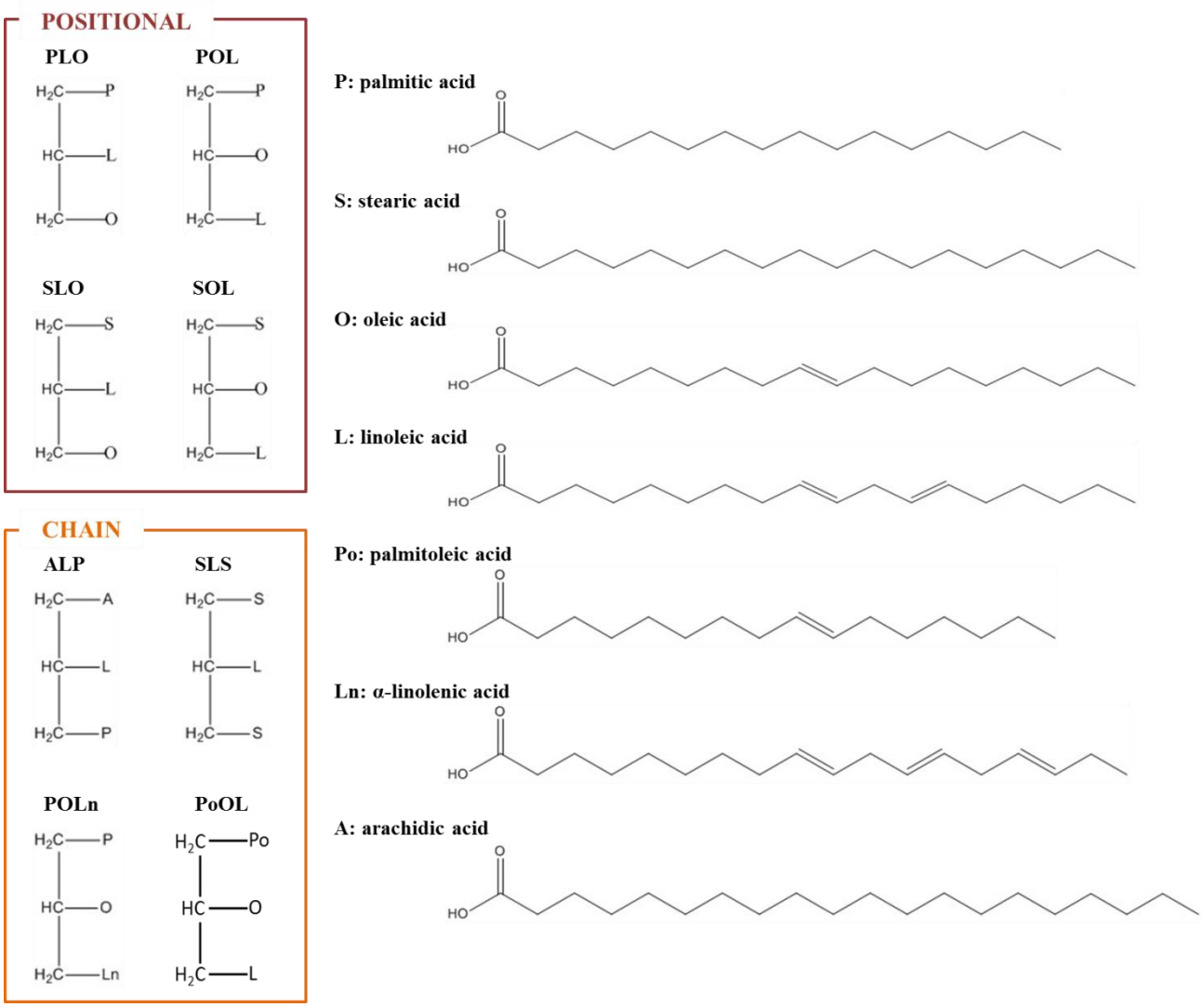


Figure S1. Scheme of the structures of TAGs structural isomers (positional and chain) considered in this work: PLO, POL, SLO, SOL, ALP, SLS, POLn, PoOL.

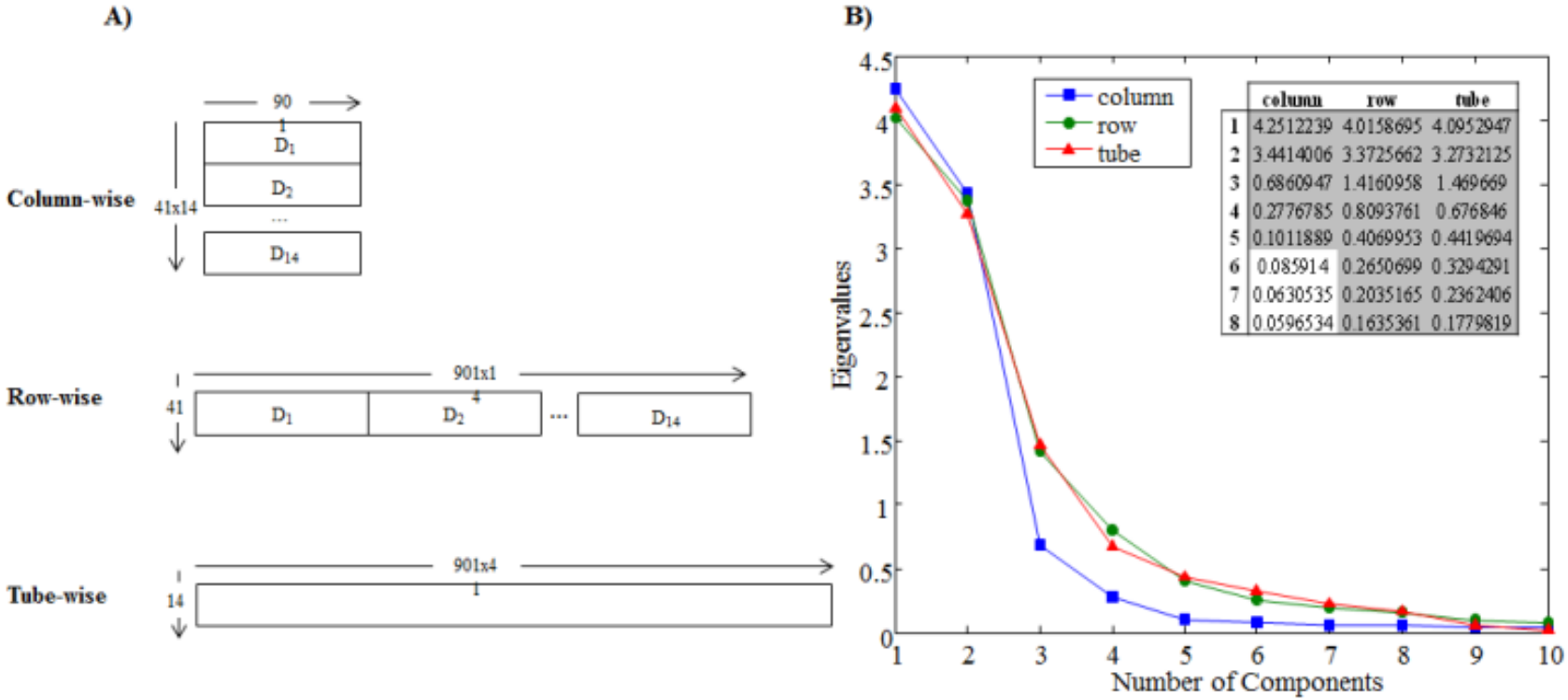


Figure S2. A) Scheme of the different possible data augmentation arrangements of $\mathbf{D}_{aug}$. B) SVD results for column-wise augmented data matrix (blue), row-wise augmented data matrix (green) and tube-wise augmented data matrix (red). The embedded table shows the numeric values for the first eight components.

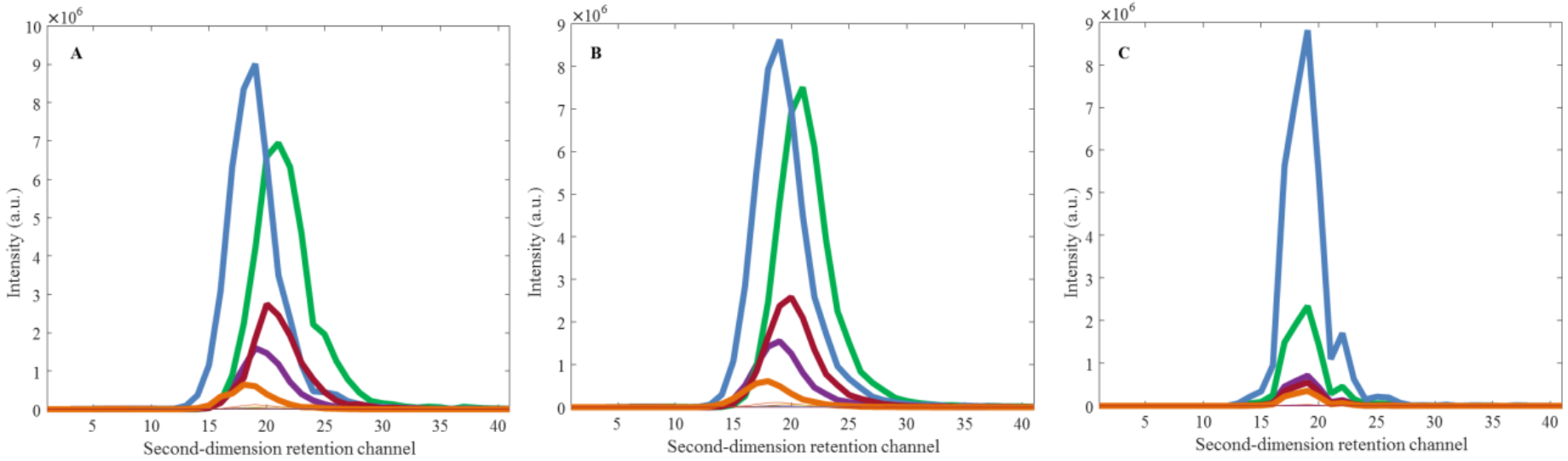


Figure S3.POL second-dimension elution profiles in consecutive modulations resolved as explained in section 3.1. by A) MCR-ALS; B) MCR-ALS trilinear allowing time shifting and C) MCR-ALS trilinear.

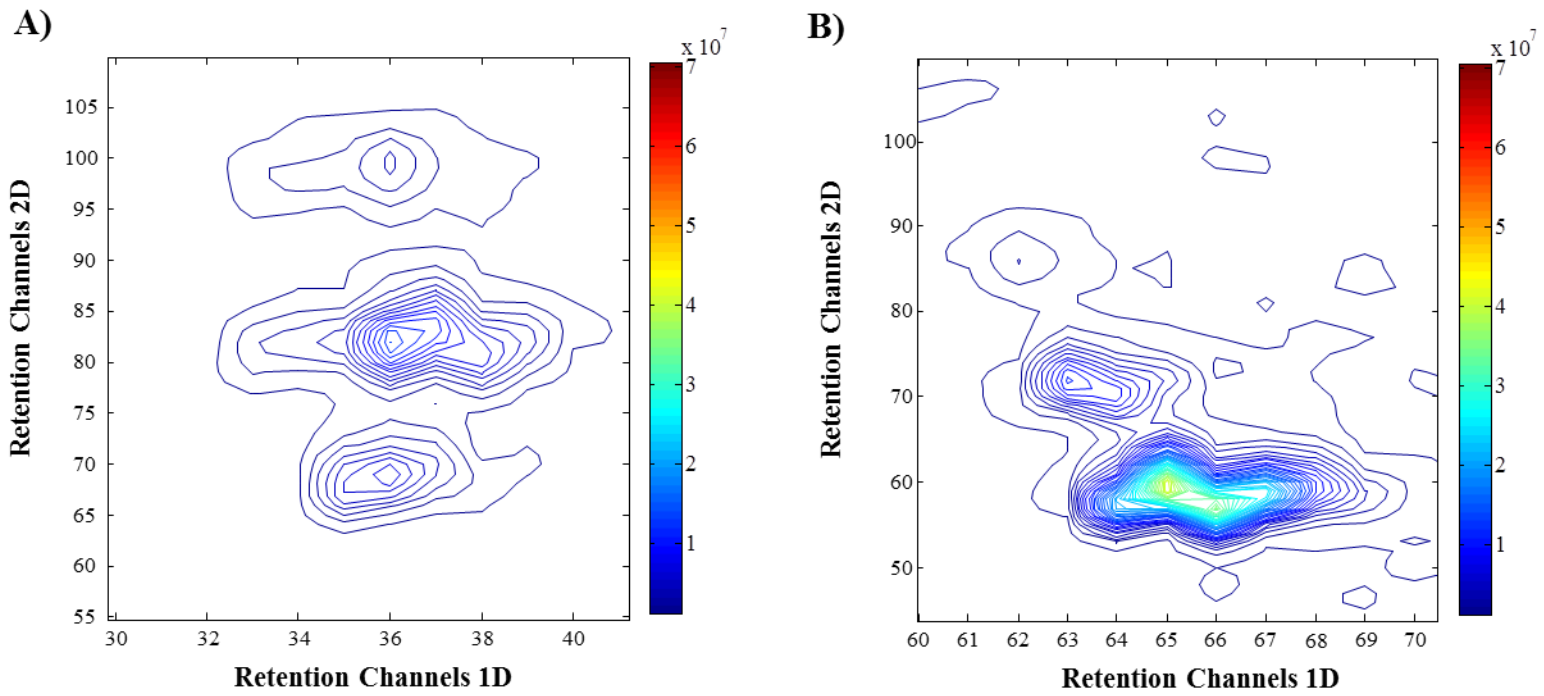


Figure S4. A) Zoomed view of region of interest number 1 (retention time channel from 30 to 42 in the first-chromatographic dimension and from 55 to 110 in the second chromatographic dimension). B) Zoomed view of region of interest number 3 (to retention time channel from 61 to 70 in the first-chromatographic dimension and from 50 to 90 in the second chromatographic dimension). Colorbar indicates peak intensity.

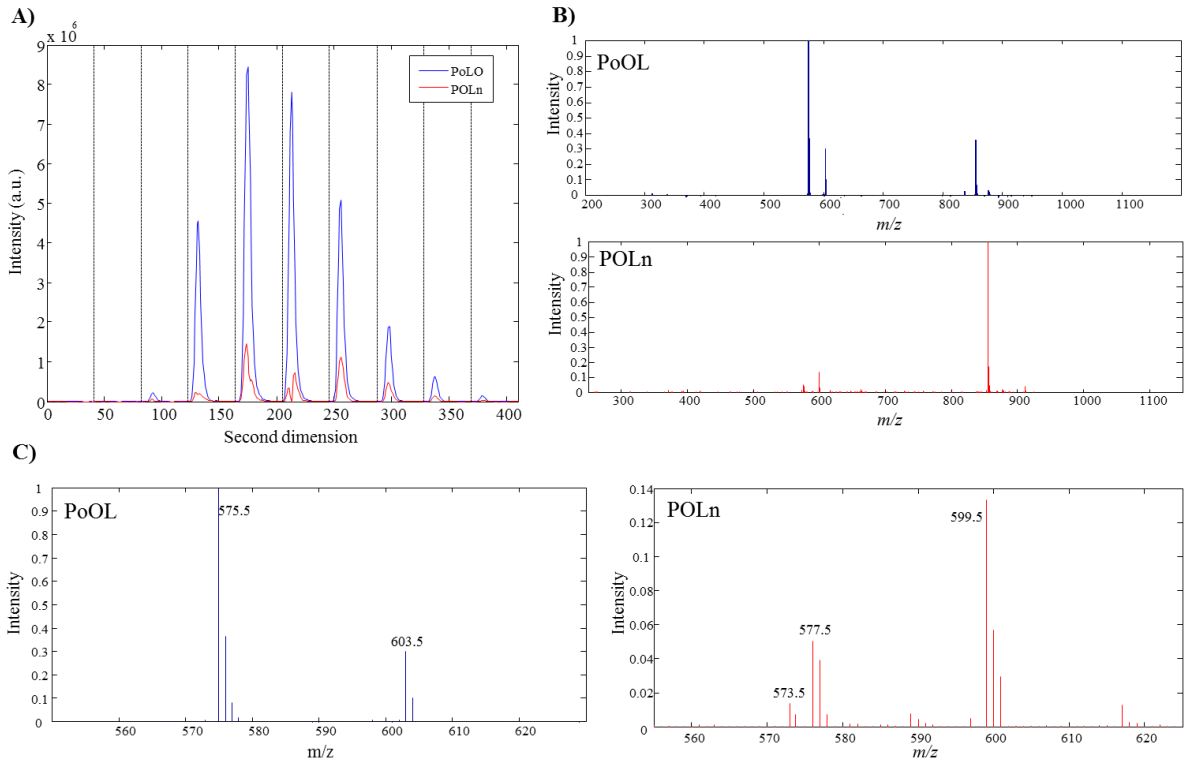


Figure S5. MCR-ALS resolution of PoOL and POLn chain isomers. A) Resolved elution profiles in the fourteen modulations taken from the first dimension column. B) Resolved mass spectra. C) MCR-ALS resolved mass spectra zoomed in the diacylglycerols mass region (from 555 to 620 m/z).

Table S1. SVD values for the elution profiles of SOL, SLO, POL and PLO resolved by MCR-ALS, MCR-ALS trilinear and MCR-ALS trilinear allowing time shifting. Shadowed areas indicate significant values (above computing accuracy).

	MCR-ALS BILINEAR				MCR-ALS TRILINEAR				MCR-ALS TRILINEAR (allowing time shifting)			
	SLO	**POL**	**SOL**	**PLO**	**SLO**	**POL**	**SOL**	**PLO**	**SLO**	**POL**	**SOL**	**PLO**
1	2.903719	2.031219	4.318444	1.860351	2.903774	1.455340	4.296360	1.266495	2.831114	2.092212	2.840144	1.042563
2	4.49E-01	8.47E-01	1.52E+00	4.96E-01	1.96E-16	3.27E-17	3.67E-16	1.11E-16	6.53E-01	7.03E-01	1.661542	2.19E-01
3	3.86E-01	7.31E-02	2.72E-01	3.00E-01	4.41E-17	6.28E-18	1.06E-16	2.00E-17	3.38E-01	5.34E-02	6.43E-01	8.90E-02
4	6.78E-02	2.41E-02	1.16E-01	1.36E-01	2.88E-17	5.25E-18	1.62E-17	8.75E-18	6.69E-03	1.24E-02	8.05E-02	4.58E-02
5	1.47E-02	1.36E-02	6.66E-02	2.78E-02	1.66E-18	1.09E-18	5.06E-18	1.63E-18	4.48E-03	4.56E-03	3.72E-16	2.26E-02
6	1.03E-02	7.25E-03	5.28E-02	2.66E-02	1.23E-18	6.53E-20	1.85E-18	1.09E-19	2.78E-03	6.93E-04	2.72E-17	1.67E-02
7	8.99E-03	2.05E-03	4.19E-02	1.93E-02	1.05E-18	1.96E-20	8.70E-19	8.50E-20	1.47E-04	6.65E-17	9.41E-18	1.31E-02
8	4.84E-03	1.18E-03	1.96E-02	1.83E-02	2.78E-19	3.41E-21	5.22E-19	2.28E-20	6.80E-17	1.11E-17	3.20E-18	2.64E-17
9	3.91E-03	1.07E-03	1.45E-02	1.53E-02	1.17E-19	3.36E-23	1.25E-19	1.42E-20	1.06E-17	1.93E-18	1.90E-18	2.44E-18
10	2.67E-03	8.87E-04	1.42E-02	6.31E-03	6.71E-20	1.34E-33	0	0	3.49E-18	1.07E-18	6.63E-19	0
11	1.61E-03	6.18E-04	4.87E-03	5.17E-03	3.85E-20	2.81E-34	0	0	1.21E-18	3.42E-19	0	0
12	1.14E-03	4.92E-04	3.16E-03	2.42E-03	3.05E-20	0	0	0	7.72E-19	1.07E-19	0	0
13	4.55E-04	4.20E-04	2.84E-03	1.61E-03	7.33E-21	0	0	0	1.17E-19	9.02E-20	0	0
14	3.41E-04	2.67E-04	1.97E-03	1.03E-03	3.61E-21	0	0	0	8.46E-21	9.00E-21	0	0

5.3. Publicació 6

Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution.

M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler.

Analytical Chemistry 89 (2017), 7675-7683.

analytical chemistry

Article

pubs.acs.org/ac

Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution

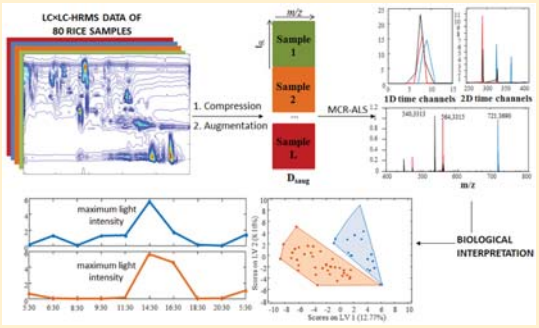
Meritxell Navarro-Reig,[†,‡] Joaquim Jaumot,[†] Anna Baglai,[‡] Gabriel Vivó-Truyols,[‡] Peter J. Schoenmakers,[‡] and Romà Tauler*[,†]

[†]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[‡]Van't Hoff Institute for Molecular Science, University of Amsterdam, 1090 XH Amsterdam, The Netherlands

Supporting Information

ABSTRACT: In this work, a new strategy for the chemometric analysis of two-dimensional liquid chromatography−high-resolution mass spectrometry (LC × LC−HRMS) data is proposed. This approach consists of a preliminary compression step along the mass spectrometry (MS) spectral dimension based on the selection of the regions of interest (ROI), followed by a further data compression along the chromatographic dimension by wavelet transforms. In a secondary step, the multivariate curve resolution alternating least squares (MCR-ALS) method is applied to previously compressed data sets obtained in the simultaneous analysis of multiple LC × LC−HRMS chromatographic runs from multiple samples. The feasibility of the proposed approach is demonstrated by its application to a large experimental data set obtained in the untargeted LC × LC−HRMS study of the effects of different environmental conditions (watering and harvesting time) on the metabolism of multiple rice samples. An untargeted chromatographic setup coupling two different liquid chromatography (LC) columns [hydrophilic interaction liquid chromatography (HILIC) and reversed-phase liquid chromatography (RPLC)] together with an HRMS detector was developed and applied to analyze the metabolites extracted from rice samples at the different experimental conditions. In the case of the metabolomics study taken as example in this work, a total number of 154 metabolites from 15 different families were properly resolved after the application of MCR-ALS. A total of 139 of these metabolites could be identified by their HRMS spectra. Statistical analysis of their concentration changes showed that both watering and harvest time experimental factors had significant effects on rice metabolism. The biochemical insight of the effects of watering and harvesting experimental factors on the changes in concentration of these detected metabolites in the investigated rice samples is attempted.



Over the last several years, comprehensive two-dimensional liquid chromatography (LC × LC) has received much attention as a powerful technique to overcome one-dimensional liquid chromatography (1D-LC) drawbacks related to the separation of highly complex mixtures or the resolution of highly overlapping compounds.[1−3] In LC × LC, the sample is subjected to two independent separation systems, simultaneously connected online using a modulator. Therefore, the higher resolving power of this technique lies in the fact that, under ideal circumstances, when retention mechanisms of the two separation systems are uncorrelated, the overall peak capacity is equal to the product of the individual peak capacities of the first- and second-dimension separations.[2−4] A way to increase this peak capacity significantly is to use the combination of two stationary phases with orthogonal (or at least highly uncorrelated) separation modes as, for instance, hydrophilic interaction liquid chromatography (HILIC) and reversed-phase (RP) separation modes.

Despite the promising abilities of LC × LC, its implementation also has some challenges in practice. First, there are few computational methods described able to analyze the highly complex data sets generated in LC × LC studies. Manual inspection of the thousands of signals detected is not feasible in practice, and therefore, their processing is not straightforward.[5,6] In this regard, a chemometric-based data analysis strategy is strongly needed. There are already some examples of the application of chemometric methods to resolve multidimensional chromatographic data. However, most of these contributions are about two-dimensional gas chromatography (GC × GC).[7−9] LC × LC data is considered to be more complex than GC × GC due to the major retention time shifts and peak shape changes between consecutive modulations and

Received: May 3, 2017
Accepted: June 23, 2017
Published: June 23, 2017

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

between different chromatographic runs (samples). When these changes are important, they prevent the use of chemometric methods based on the fulfillment of the trilinear model like PARAFAC or PARAFAC2.[8,10,11] In fact, only a few works are encountered using chemometric tools to analyze LC × LC data,[6,12,13] mostly related to LC × LC-DAD (diode array detector) data and mainly using the multivariate curve resolution alternating least-squares (MCR-ALS) method. The complexity of LC × LC data analysis is even more challenging when the detector used is a high-resolution mass spectrometer (HRMS). This is caused by the much larger amount of throughput information and larger size of the data sets generated, requiring the application of compression and feature selection approaches to facilitate the data analysis procedures. In a previous work, the effectiveness of MCR-ALS for resolving LC × LC−MS data was initially shown.[10] However, in that preliminary work, only one analytical sample was considered, which simplified the analysis considerably. This is not the case when a large number of samples are simultaneously analyzed and compared. In particular, LC × LC has a doubtless potential in the metabolomics studies, where a large number of complex biological samples with a large number of metabolites with extremely overlapped elution profiles (hundreds of metabolites with similar retention and composition) are simultaneously analyzed.[1]

Metabolites are small molecules that are transformed during metabolism and can provide direct information about the biochemical activity of cells.[5] In this context, the study of metabolites can help to improve the knowledge of perturbations that diseases, drugs, toxins, or environment might cause in organisms.[14] Metabolomics is the field that aims to study how the metabolite content in organisms changes under different scenarios. Metabolomics has become a powerful approach widely used in various areas, such as clinical, environmental, and food sciences.[5,14]

There are two principal metabolomics approaches: targeted and untargeted. The first one is only focused on analyzing a specific list of metabolites, typically related to some known pathways of interest. In contrast, untargeted metabolomics aims to screen the entire (or a large part) metabolite content of biological samples without using any previous knowledge about pathways or a specific list of metabolatites,[5,15] generating highly complex data sets. Until now, 1D-LC coupled with mass spectrometry is one of the most used analytical approaches for analyzing those samples, due to its ability for resolving complex mixtures.[16] Nevertheless, the increasing complexity of samples analyzed by the untargeted metabolomics approach often exceeds the limits of peak capacity achievable by 1D-LC systems.[3] This causes that the analysis has to be focused on particular families of compounds or, alternatively, to the use of techniques that can improve these resolution limits such as LC × LC−MS. Moreover, these complex samples usually contain highly similar compounds and isomers, which coelute in 1D-LC. In contrast, in LC × LC, the second dimension allows the separation of these compounds.[10] There are some examples in the literature of the application of LC × LC to the analysis of metabolites in complex biological samples. However, to the best of our knowledge, its application to untargeted metabolomics studies is still a challenge.[4,6,16]

The main aim of this work is to develop a global methodological strategy for the analysis of untargeted LC × LC−HRMS metabolomics data, based on the application of chemometric tools. The demonstration of the feasibility of the proposed approach to achieve this goal is shown for an untargeted metabolomics study of the changes produced on Japanese rice (*Oryza sativa* var. *Japonica*) metabolism due to environmental factors (watering and harvesting time). In this proof of concept study, an analytical untargeted LC × LC method using the coupling of two orthogonal columns (HILIC × RPLC−HRMS) has been developed and tested in detail. Data generated in the simultaneous processing of a large number of samples is thoroughly analyzed by means of the MCR-ALS method. Results are used to check for the effects produced by watering and harvesting time factors, and to identify the more important metabolites related to these effects.

EXPERIMENTAL SECTION

Chemicals and Reagents. Ammonium acetate (≥99.0%), formic acid (≥95.0%), acetic acid (≥95.0%), and LC−MS water were supplied by Sigma-Aldrich (Darmstadt, Germany). Acetonitrile (ACN, LC−MS grade) was obtained from Avantor Performance Chemicals (Deventer, The Netherlands). Methanol (MeOH, LC−MS grade) and chloroform were obtained from Biosolve (Valkenswaard, The Netherlands). Piperazine-*N*,*N*′-bis(2-ethanesulfonic acid) (PIPES) (≥99.0%) was used as internal standard (Sigma-Aldrich, Steinheim, Germany).

Water used for plant watering was purified using an Elix 3 coupled with a Milli-Q system (Millipore, Bedford, MA, U.S.A.) and filtered through a 0.22 μm nylon filter integrated into the Milli-Q system.

Plant Growth and Metabolite Extraction. Plant growth and metabolite extraction were performed using a procedure described elsewhere.[17] Briefly, rice seeds, obtained from the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain), were incubated for 2 days at 30 °C in a wet environment. After this period, plants were grown on an environmental test chamber MLE-352H (Panasonic) for 22 days simulating cyclic environmental changes of temperature, relative humidity, and light intensity as shown in Supporting Information Figure S1. During this growth period, plants were watered with Milli-Q water three times per week.

Harvest was done at 10 time points, coinciding with the principal changes in environmental conditions: 5:30, 6:30, 8:30, 9:30, 11:30, 14:30, 16:30, 18:30, 20:30 h and 5:30 h of the following day. In order to study the influence of watering on rice metabolism, only half of the samples were watered prior to the first harvest point at 5:30 h. In total, samples of 20 different conditions were analyzed, considering 10 different harvest time points and two different watering conditions. After harvest, aerial parts of rice samples were frozen at liquid nitrogen temperature to quench metabolism. Samples were stored at −80 °C until extraction. Four biological replicates were made for each sample condition. Therefore, a total of 80 samples were analyzed.

Before extraction, aerial parts of rice samples were ground to a fine powder using a liquid nitrogen mortar and lyophilized for 24 h to dryness. Metabolite extraction was carried out by dispersing 40 mg of the dried tissue in 1.4 mL of MeOH/H_2O (70:30). Chlorophyll and hydrophobic compounds were removed using chloroform. Finally, the aqueous fraction was evaporated to dryness under nitrogen gas and reconstituted with 450 μL of acetonitrile/water (1:1 v/v). For internal standard quantification, 50 μL of 50 mg·L^{-1} solution of the internal standard (PIPES) was added to the extract. All of the extracts were stored at −80 °C until analysis.

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

LC × LC−HRMS Analysis. LC × LC analyses were carried out on a Shimadzu 20 AD liquid chromatograph (Shimadzu, Kyoto, Japan) equipped with an autosampler. Second-dimension separation was possible due to the coupling to this instrument of two additional LC pumps (Shimadzu LC-10 AD). The interface between the first and second column was an air-actuated 10-port two-position VICI valve (Valco, Schenkon, Switzerland) equipped with two sample loops with a volume of 100 μL.

In the first chromatographic dimension, an HILIC TSK gel amide-80 column (250 mm × 2.0 mm i.d.; 5 μm) with a guard column (10 mm × 2.0 mm i.d.; 5 μm) of the same material provided by Tosoh Bioscience (Tokyo, Japan) was used. Chromatographic analysis was run using (A) acetonitrile and (B) 5 mM ammonium acetate at pH 5.5, adjusted with acetic acid, as mobile phases, eluted according to the following gradient: 0 min, 5% B; 66.15 min, 15.4% B; 231 min, 65% B; 231−232 min back to the initial conditions at 5% B; and from 232 to 265 min, at 5% B. The mobile phase flow rate was 13 μL·min^{-1}, and the injection volume was 20 μL. The first-dimension elution gradient was optimized using the PIOTR program[18] taking as starting gradient the 1D-LC method used in a previous study of rice metabolomics.[17]

The second chromatographic dimension employed a reversed-phase (RP) Kinetex C18 (50 mm × 2.1 mm i.d.; 1.7 μm) column provided by Phenomenex (Torrance, CA, U.S.A.). During the whole LC × LC separation, 2.7 min repetitive second-dimension gradients were employed, 2.7 min also being the modulation time in the switching valve. Second-dimension elution gradient used water with 0.1% formic acid (A) and acetonitrile with 0.1% formic acid (B) as follows: 0 min, 5% B; 2 min, 90% B; 2−2.1 min back to initial conditions at 5% B; and from 2.1 to 2.7 min, at 5% B. The mobile phase flow rate was 0.5 mL·min^{-1}. In order to avoid breakthrough peaks in the second chromatographic separation, the percentage of acetonitrile in the sample solvent coming from the first dimension should be reduced. For this purpose, an additional LC pump (Shimadzu LC-10 AD) was employed to add water at a flow rate of 16 μL·min^{-1} to the sample solvent using a stainless steel T piece placed between the second-dimension pumps and the second-dimension column.

Mass spectrometry detection was performed using a micrOTOF-Q II (Bruker Daltonics, Billerica, MA, U.S.A.) equipped with an electrospray (ESI) ionization source operated in negative mode at a resolution of 16500 fwhm using the following conditions: dry temperature, 200 °C; mass range, m/z 90−1000 Da; dry gas flow rate, 8.0 L·min; nebulization pressure, 29 psi.

Data Analysis Strategy. *Data Arrangement and Compression.* Bruker raw chromatographic data files (.raw format) were converted to the standard CDF format by the export chromatogram analysis function of Compass DataAnalysis 4.0 SP 1 software (Bruker Daltonics, Billerica, MA, U.S.A.). Then, these data files were imported into the MATLAB environment (release 2015b, The Mathworks Inc., Natick, MA, U.S.A.) by using mzcdfread.m and mzcdf2peak.m functions of the MATLAB Bioinformatics Toolbox (4.3.1. version).

Figure 1A shows the general strategy proposed for LC × LC−HRMS data arrangement and its subsequent chemometric analysis. The figure explains the data arrangement and compression for a single LC × LC−HRMS run. Every full-scan LC × LC−HRMS chromatographic run gives a data

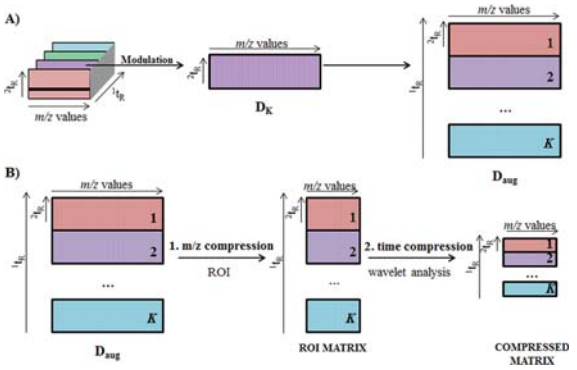


Figure 1. (A) LC × LC−HRMS data arranged in a columnwise (in the m/z mode) augmented data matrix. (B) Size compression scheme: first the spectral dimension (m/z values) was reduced by means of the ROI strategy, and then the time direction was reduced using one-dimensional wavelet analysis strategy.

structure which can be arranged in three orders of measurement. The x-axis corresponds to the full-scan mass spectra, whereas the y-axis and z-axis correspond to the first and second chromatographic dimensions, respectively. The MCR-ALS bilinear model used in this work requires that the data is arranged in a two-way data structure or data matrix. For this reason, the LC × LC−MS data were arranged in a data matrix, where every data slice gives the data matrix $\mathbf{D}_K$, of one modulation with the second chromatographic dimension separation on its rows and mass spectrometry detection on its columns. When K modulations (slices) are considered simultaneously, the columnwise augmented data matrix ($\mathbf{D}_{aug}$) is built up, setting up the individual $\mathbf{D}_K$ matrices from each modulation one on the top of the other, and keeping the mass spectra measured at each second-dimension retention time in the columns of the matrix.

For example, a single LC × LC−HRMS $\mathbf{D}_{aug}$ matrix will have 32 144 rows (328 retention times in the second dimension per 98 time modulations in the first dimension) and, approximately, 5000 m/z values, which resulted in approximately 1 GB of storage for each sample. Since a total of 80 samples were simultaneously analyzed in this work, a total of 80 GB of storage would be needed to store and analyze the complete data set. This big data size hampered its direct analysis and forced the application of a preliminary compression procedure.

Size compression was performed in two steps: first a reduction in the spectral dimension (m/z values), and then, a reduction in the time direction (Figure 1B). MS spectral data compression was performed using the previously proposed regions of interest (ROI) strategy.[14] The ROI approach selects the more interesting mass traces, which are those m/z values whose intensity signals are higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and appear a number of times consecutively in the time dimension. ROI values are searched along the entire chromatogram. A new data matrix containing the intensities at all retention times (rows) only for a reduced number of ROI m/z values (columns) is finally stored. The parameters for the implementation of this ROI approach are the SNR_{Thr}, (usually set at 0.1% of the maximum MS signal intensity), the mass accuracy of the mass spectrometer [set at 0.05 Da/e for the time-of-flight (TOF) MS analyzer used in this work], and the minimum number of consecutive retention times to be considered as a chromatographic peak (set at 25).

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

Using this approach, a relatively low number of m/z values (approximately 700) are usually retrieved. More details on how this strategy works are given in the literature.[14]

On the other hand, data compression in the time dimension was also applied. This was performed using a one-dimensional wavelet analysis strategy to every chromatogram at the previously selected m/z ROI values (see Figure 1B). Wavelet compression allowed reducing the size of the data matrix 2^n times, being n the wavelet compression level. In this work, a four-level compression using the Daubechies wavelet[19] was used. Therefore, the number of rows was reduced eight times without any significant loss of relevant information in the elution time dimension. More details about wavelet compression procedures can be found in the literature.[19−22] In order to accelerate the calculations and to further reduce storage requirements of MCR-ALS analysis, a time windowing approach was also additionally used dividing the whole chromatogram into three separate chromatographic windows: from 0 to 81 min, from 77 to 140 min, and from 136 to 255 min. Considering only one of the time windows, after both compression steps (ROI and wavelets), the compressed data matrix (Figure 1B) for each one of the samples had 4018 retention times and 700 m/z values, which is manageable by standard computers, and it represents more than 50-fold computer storage reduction without loss of information nor accuracy. Every compressed data matrix was then normalized to correct possible instrumental intensity changes among different sample injections and unavoidable differences in sample handling. This normalization was done by dividing all data values of every sample by the chromatographic peak area of the internal standard (PIPES) added to the metabolite extract of every sample.

After normalization, individual compressed matrices for each analyzed sample were arranged in a single superaugmented columnwise data matrix (Figure 2, $\mathbf{D}_{saug}$). Since ROI-compressed individual data matrices may have a different number of columns (ROI m/z values), a preliminary search for common and uncommon ROI m/z values among different samples was done over the 80 samples of this study. This search evaluated common and uncommon ROI values and finally considered both of them. In this way, the final superaugmented data matrix (Figure 2, $\mathbf{D}_{saug}$) had all ROI m/z values from 80 samples (for more details see the Supporting Information and the procedure described ref 14). Finally, every one of these three superaugmented data matrices (one for each of the three selected time windows) having all 80 samples was analyzed by the MCR-ALS method.

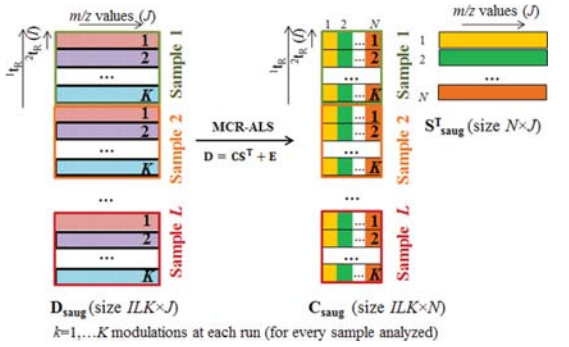


Figure 2. MCR-ALS resolution of LC × LC−HRMS data. The columnwise superaugmented data matrix, $\mathbf{D}_{saug}$, is decomposed into two new data matrices: $\mathbf{C}_{saug}$, which has the resolved pure elution profiles of the components in second column modulations of different samples (chromatographic runs), and $\mathbf{S}^T$, which has the pure mass spectra of the resolved components.

MCR-ALS Resolution of LC × LC−MS Data. MCR-ALS is a chemometric method that allows for the investigation and resolution of pure component contributions present in unresolved complex mixtures. In this work, MCR-ALS allowed the resolution of pure elution profiles in both chromatographic dimensions and of pure mass spectra profiles of the metabolites present in rice samples. The MCR-ALS method has been already described in the literature,[23−25] and it is only briefly explained here focusing on the special requirements for the analysis of multisample LC × LC−HRMS data (see the Supporting Information for an extended description).

MCR-ALS decomposes the experimental data sets according to a bilinear model that can be easily extended to the simultaneous analysis of several chromatographic runs (different modulations and multiple samples). In this case, each compressed columnwise superaugmented data matrix, $\mathbf{D}_{saug}$ (Figure 2), had the information related to each one of the previously described augmented data matrices, $\mathbf{D}_{aug}$, for the L = 1, ..., 80 simultaneously analyzed samples. $\mathbf{D}_{saug}$ was decomposed according to the bilinear model as shown in Figure 2 and eq 1:

$$\mathbf{D}_{saug} = \begin{bmatrix} D_{1,1} \\ D_{1,2} \\ \vdots \\ D_{L,K} \\ \vdots \\ D_{80,98} \end{bmatrix} = \begin{bmatrix} C_{1,1} \\ C_{1,2} \\ \vdots \\ C_{L,K} \\ \vdots \\ C_{80,98} \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} E_{1,1} \\ E_{1,2} \\ \vdots \\ E_{L,K} \\ \vdots \\ E_{80,98} \end{bmatrix} = \mathbf{C}_{saug}\mathbf{S}^T + \mathbf{E}_{saug} \quad (1)$$

MCR-ALS decomposition of the matrix $\mathbf{D}_{saug}$ gives $\mathbf{C}_{saug}$ ($ILK \times N$), the matrix of second-dimension-resolved elution profiles of the N components at retention times (I) for each modulation (K) and sample (L). From this $\mathbf{C}_{saug}$ matrix, relative quantitative information about the different metabolites in the different samples can be obtained from the peak area ratios of the resolved elution profiles in the different samples. On the other hand, the $\mathbf{S}^T$ matrix (size $N \times J$) has the pure mass spectra of the resolved components (Figure 2) which can be used for tentative metabolite identification from their spectral features.

More details regarding the initialization and constraints for the MCR-ALS optimization can be found in the Supporting Information and references therein.

Peak Area Analysis. Chromatographic peak areas of the resolved metabolites were directly obtained from the MCR-ALS profiles resolved in $\mathbf{C}_{saug}$, without needing any additional baseline subtraction (baseline contributions should be resolved in separate MCR-ALS components). These peak areas were arranged in a new data matrix ($\mathbf{A}$), where the areas of every metabolite (M) are in the columns and the samples (L) in the rows of this data matrix. This peak areas data matrix were analyzed to further assess the effects of experimental factors using analysis of variance (ANOVA)-simultaneous component analysis (ASCA),[26] partial least-squares−discriminant analysis (PLS-DA),[27] and MCR-ALS (see the Supporting Information for a complete description of the employed methods).

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

ASCA analysis was applied to statistically assess the significance of watering and time factors. It was performed on a well-balanced experimental design, and allowed the interpretation of the possible sources of experimental variance (watering and time factors). Watering factor effects were also evaluated using PLS-DA, which is a supervised method oriented to discriminate among different groups of samples. In this work, PLS-DA was used to differentiate among watered and nonwatered samples. Apart from sample class discrimination, PLS-DA also provides information about which are the most relevant variables (i.e., resolved metabolites) for achieving this differentiation, which are known as variable important on projection (VIP) scores. Time factor effects were then also studied using the MCR-ALS method, to investigate the evolution of the metabolic concentration profiles over time. In this case, the data matrix of the peak areas, **A** (size $L \times M$), was decomposed by MCR-ALS using a bilinear model into two factor matrices, $\mathbf{A} = \mathbf{T}\,\mathbf{F}^T$, where **T** is the matrix describing the temporal evolution profiles and $\mathbf{F}^T$ provides information about the contribution of the different metabolites to these temporal profiles. Before MCR-ALS analysis, chromatographic peak areas of the metabolites were scaled using a generalized logarithm transformation.[27]

ASCA and PLS-DA methods were applied using PLS Toolbox 8.0.2 (Eigenvector Research Inc., Wenatchee, WA, U.S.A.) working under MATLAB 2015b. MCR-ALS analyses were carried out using the MCR-ALS 2.0 toolbox available at www.mcrals.info.

RESULTS

Resolution of LC × LC−HRMS Data. Figure 3 shows an example of an LC × LC−HRMS chromatogram obtained in the analysis of one of the rice samples investigated in this work. Chromatographic peaks were well-spread over the chromatogram area indicating that their retention times were noncorrelated. This indicates that the two chromatographic separation columns used during the analysis were rather orthogonal and that the separation was satisfactory. Coupling the HILIC-RP C18 columns was quite challenging since the high percentage of acetonitrile (75%) in the solvent from the HILIC column (first dimension) caused breakthrough in the RP C18 column (second dimension). This problem could be satisfactorily solved by the addition of water to the sample solvent before the second-dimension column, which produced a significant reduction of the percentage of acetonitrile down to 33%. More details about the breakthrough reduction are shown in the Supporting Information.

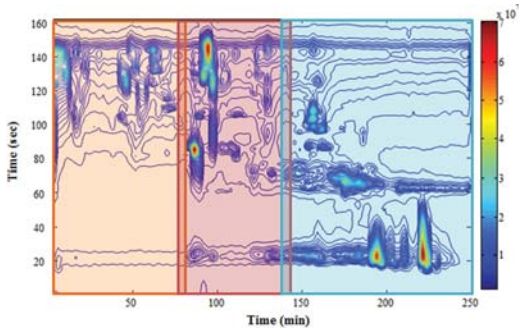


Figure 3. Bidimensional chromatogram obtained in the analysis of a rice sample. The three chromatographic time windows used for MCR-ALS analysis are highlighted. Colorbar indicates signal intensity.

In Figure 3, the three selected chromatographic windows are marked in different colors. To ensure that no chromatographic information was lost (i.e., that no chromatographic peaks were split in between the two areas), these three chromatographic regions were somewhat overlapped. A total number of 80 samples were simultaneously analyzed after the two compression steps (ROI and wavelets) described in the Experimental Section. Details related to the application of the ROI approach and the wavelet transform are shown in the Supporting Information.

Chromatographic region A (from 1 to 1206 retention channels) was resolved by an MCR-ALS model using 50 components. The percentage of explained variance (R^2) was 99.9%, and lack of fit (LOF) was 3.5%. Chromatographic region B (from 1148 to 2091 retention channels) was also resolved with 50 MCR-ALS components with R^2 of 99.9% and LOF equal to 2.7%. Finally, chromatographic region C (from 2031 to 3800 retention channels) was resolved with an MCR-ALS model using 100 components with R^2 equal to 99.9%, and LOF equal to 4.0%. Therefore, a total number of 200 MCR-ALS components were used to explain the variance of the whole data set in the 80 analyzed samples. This large number of selected components includes all detected metabolite contributions as well as other noisy chromatographic signals, such as instrumental background and solvent contributions. In this case, 154 from the 200 resolved components were finally assigned to individual metabolites, whereas the other 46 components were associated with unknown background and noisy signals. Despite the high complexity of this untargeted LC × LC−HRMS data set, MCR-ALS could resolve properly a large number of metabolites in one single analysis of the data from the 80 rice samples. The combination of the data compression and resolution strategies described in this work allowed to overcome most of the encountered challenges, such as the big size of the data, the highly overlapped metabolite elution profiles, and the lack of prior knowledge about the content of the samples (untargeted approach).

Figure 4 shows an example of MCR-ALS resolution of the pure elution and spectral profiles of the four watered samples harvested at 11:30 h. Three of the detected metabolites coeluted strongly in both chromatographic dimensions. Parts A and B of Figure 4 depict the first-dimension and second-dimension elution profiles, respectively, where this strong coelution is displayed. Figure 4C gives the mass spectra of the corresponding metabolites and of their accurate m/z diagnostic ion values, to be used for their subsequent putative identification.

Tentative identification of metabolites was performed by comparing the accurate m/z diagnostic ion values obtained from MCR-ALS-resolved mass spectra with theoretical exact m/z values included in public databases such as MassBank,[28] Metlin,[29] HMDB,[30] and MetaCyc.[31] For instance, following the previous example, the identification of the MCR-ALS-resolved components shown in Figure 4 is given in detail here. The accurate mass of the diagnostic ion observed for metabolite depicted in blue was 721.3690. This mass could be assigned with 8 ppm of relative error to physalolactone B 3-glucoside, whose adduct ion with acetic acid ($[M + CH3COO]^-$) has an exact mass of 721.3748.[29] In the case of the metabolite depicted in red, the accurate mass observed for its diagnostic ion was 564.3315, and it could be associated with 1 ppm of relative

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

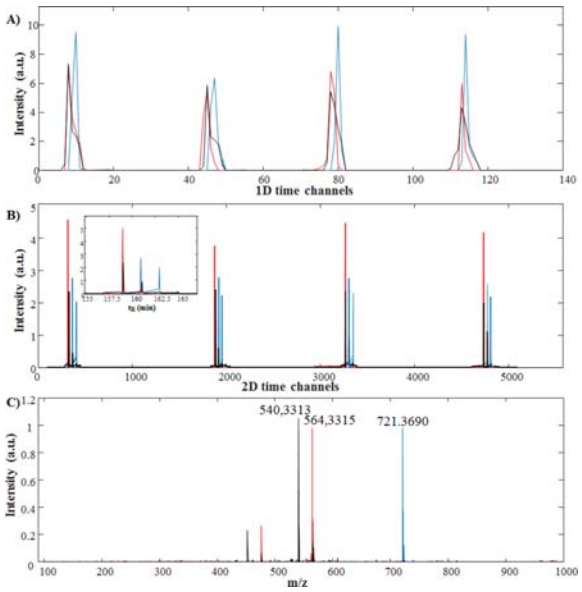


Figure 4. Example of MCR-ALS results obtained in the analysis of the chromatographic region between 155 and 170 min. (A) MCR-ALS first-dimension elution profiles resolved for three coeluted metabolites. (B) MCR-ALS second-dimension elution profiles resolved for three coeluted metabolites. The inset shows a zoomed view of the elution profiles resolved for the first sample. (C) MCR-ALS mass spectra resolved for the three metabolites.

mass error with 18:2-lysophosphatidylcholine, whose adduct ion with formic acid ($[M + FA - H]^-$) has an exact mass of 564.3320.[31] Finally, the accurate mass observed for the metabolite depicted in black was 540.3313, which could be assigned with a relative mass error of 1 ppm to 6:0-2-lysophosphatidylcholine, whose adduct ion with formic acid ($[M + FA - H]^-$) has an exact mass of 540.3320.[31] From the 154 resolved metabolites, 139 were identified with relative errors lower than 10 ppm, as recommended when a TOF analyzer is used.[32] These metabolites are listed in Table S1 (Supporting Information). Identified metabolites were from 15 different metabolite families, including amino acids, carbohydrates, nucleosides, nucleotides, carboxylic acids, hormones, cofactors, aromatic compounds, fatty acids, other lipids, glycosides, flavonoids, alkaloids, terpenes, and other secondary metabolites.

The variety of metabolite families detected in this work, including compounds of different polarity, could be identified in a single chromatographic run because two orthogonal stationary phases, HILIC and RP, were combined in their LC × LC determination. The use of these two noncorrelated separation systems, allowed for the simultaneous profiling of a large number of metabolites with different polarities in the same sample. This is an important benefit of using LC × LC−MS for untargeted metabolomics studies; a wider range of unknown metabolites could be resolved using HILIC × RP-HRMS than when only HILIC-HRMS is used.

Assessment of Experimental Factors. After MCR-ALS resolution of metabolite elution profiles in the simultaneous analysis of the 80 different rice samples, an untargeted metabolomics study was carried out to evaluate the effects of the considered experimental factors (watering and harvesting time) using the chromatographic peak areas of these resolved elution profiles. Finally, 114 from the total 154 resolved metabolites were included in the data table, matrix **A**, to be further analyzed. Metabolites whose peak areas changed very little (with a very low standard deviation) were not considered for further analysis.

The statistical significance of the two factors was assessed using a two-factor design with interaction using an ASCA model which considers peak areas of all metabolites in the same test (multivariate ANOVA). Results showed that both factors were statistically significant, with a p-value of 0.0001 and 0.0032 for watering and harvesting time, respectively. The interaction between factors was also found statistically significant with a p-value of 0.0018.

Assessment of Watering Effects on Rice Metabolism. Watering effects on rice metabolism were examined in more detail by PLS-DA of the peak areas data matrix (**A**). Figure 5 summarizes the results of PLS-DA with a model of two components. Figure 5A shows the scores plot, where a clear differentiation between the two groups of samples (watered and nonwatered) can be observed. The percentage of correctly classified samples after cross validation was a 90% for each class. Figure 5B displays the more important variables in the VIPs plot for this PLS-DA model with two components, which revealed which of the obtained metabolites were more important for the distinction of watered and nonwatered samples. Metabolites with higher VIP values were flavonoids and glycosides. Previous rice epigenetic studies reported that the compounds from these two families are accumulated by plants when they are grown under abnormal conditions, such as those found under drought environments.[33−35] This would agree therefore with the trend observed in this work, since

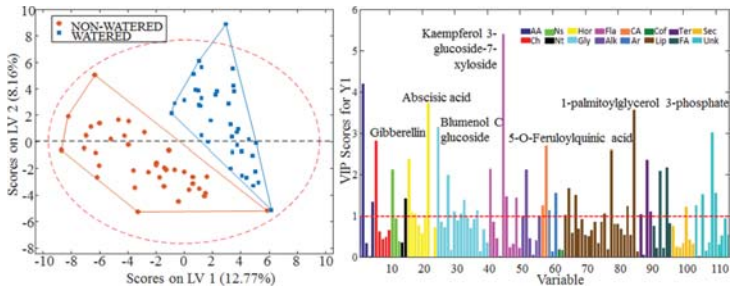


Figure 5. Results of PLS-DA analysis. (A) Scores plot, nonwatered samples are represented with orange rounds, and watered samples with blue rounds. (B) PLS-DA VIP scores plot: variables are colored according to the different metabolites families.

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

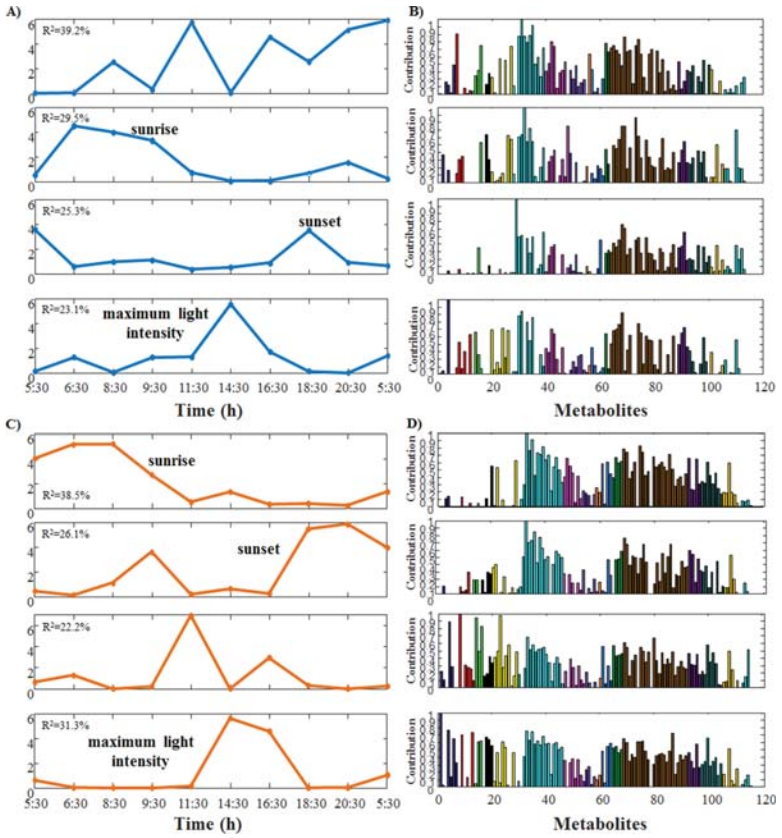


Figure 6. MCR-ALS resolution of peak area matrices. Resolved components are ordered according to their individual explained variance (R^2). (A) Temporal profiles resolved for watered samples. (B) Metabolite profiles resolved for watered samples. (C) Temporal profiles resolved for nonwatered samples. (D) Metabolite profiles resolved for nonwatered samples.

these flavonoids and glycosides metabolites showed a positive fold change (Table S1 in the Supporting Information) when rice was watered with a low amount of water. Moreover, abscisic acid (ABA) and xanthoxin were also detected as significant variables in the VIPs plot. ABA is a plant hormone that plays a major role in plant responses to environmental stress, since it regulates stomatal closure and gas exchange. Xanthoxin is one of the intermediates in ABA biosynthesis.[35] Another plant hormone that appeared as an important variable was gibberellin, which showed in this case a negative fold change (Table S1 in the Supporting Information). This result is also reported in the literature which showed that gibberellic acid content decreased in leaves of rice plants under drought stress.[33] Finally, it should be highlighted that the nucleosides detected in this study appeared also as significant variables. These metabolites were not reported in previously mentioned works. This demonstrates that the proposed methodology can lead to the discovery of additional relevant metabolites in this type of studies.

Assessment of Harvesting Time Effects on Rice Metabolism. In this case, MCR-ALS was applied to the resolved peak areas data matrix (**A**) to investigate the effects of the harvesting time factor on rice metabolites of watered and nonwatered samples. Although MCR-ALS is usually applied for the spectroscopic resolution of unresolved mixtures, its ability to resolve average temporal profiles from environmental and omics studies has also been proven.[36,37] One of the main advantages of using MCR-ALS is its flexibility to implement natural constraints, such as non-negativity, which allows a more straightforward interpretation of the profiles of the resolved components, in contrast with other chemometric methods like PCA.

In the case of the MCR-ALS analysis of the peak areas of watered samples, four components already explained 85.8% of the total peak areas variance. For nonwatered samples, the model with four MCR-ALS components explained 90.1% of the total observed peak areas variance. Figure 6 shows the results obtained after the application of MCR-ALS for watered (Figure 6, parts A and B) and for nonwatered (Figure 6, parts C and D) samples. For both types of samples, the obtained temporal profiles (Figure 6, parts A and C) can be related to metabolic changes due to the increasing light intensity during the daytime. Profiles of components 2 (watered samples) and 1 (nonwatered samples) had their maximum intensity during sunrise. On the other hand, components 4 (for watered and nonwatered samples) describe the changes produced when daylight intensity was the highest. Finally, at sunset, metabolic changes were mostly explained by components 3 (watered samples) and 2 (of nonwatered samples).

The metabolite profiles for the four components resolved for watered and nonwatered samples are shown in Figure 6, parts B and D, respectively. These profiles can be useful to know which metabolites presented higher contributions in each temporal profile of the components described above. For instance,

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

glycosides (depicted in cyan in Figure 6) always showed high contributions in three of the profiles (sunrise, maximum light intensity, and sunset) for both watered and nonwatered samples. This trend agrees with the results from a previous rice transcriptomic study,[38] where no changes of glycosides content were reported during the day time. In contrast, carbohydrates (depicted in red in Figure 6) had higher contributions during sunrise and at maximum daylight intensity, but they did not show significant contributions at sunset. This result was also confirmed in the previous rice transcriptomic study, in which a decrease of carbohydrates concentrations was reported at sunset.[38]

CONCLUSIONS

The experimental untargeted LC × LC−HRMS and data analysis strategies proposed in this work have allowed the investigation of the changes in rice metabolite concentrations due to different experimental (watering and harvesting time) factors. Resolution of a large number of metabolites was achieved satisfactorily using the 2D HILIC × RP proposed method. Peaks were well-spread over the chromatographic area indicating that the separations in both dimensions were noncorrelated. The proposed data compression strategy using ROI and wavelet transform strategies achieved a 50-fold computer storage reduction keeping all the relevant information about the presence and concentration changes of rice metabolites. Additionally, MCR-ALS data analysis allowed the simultaneous analysis of the whole set of 80 samples (chromatographic runs) obtained at the different experimental conditions of watering and harvesting time. A total number of 139 metabolites were identified (out from 154 detected) from their MCR-ALS-resolved mass spectra, and the statistical analysis of the concentration changes of these detected metabolites allowed the investigation of the effects of watering and harvesting time on them. A preliminary biological interpretation of these effects showed that watering altered the concentrations of flavonoids, glycosides, and plant hormones families. In the case of harvesting time factor, the main result was that we could interpret the evolution of some of the resolved metabolite temporal profiles (especially for carbohydrates) with the changes in light intensity during day time. Further work is proposed to further interpret and confirm the biochemical changes induced by the considered environmental factors in rice an also to extend the proposed methodologies to the investigation of other plant's metabolism. Due to the excellent performance of the developed strategy, it is proposed its extension to the general analysis of comprehensive LC × LC−HRMS data sets in omics studies and for other related types of studies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b01648.

Conditions of the environmental test chamber MLE-352H, detailed description of chemometric tools (ROI strategy, MCR-ALS, ASCA, and PLS-DA), Figure S3 showing breakthrough reduction, Figures S4 and S5 showing examples of the application of size compression steps to LC × LC−HRMS data, and Table S1 containing the tentative identification results (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: +34934006140. E-mail: Roma.Tauler@idaea.csic.es.

ORCID

Romà Tauler: 0000-0001-8559-9670

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement No. 320737. The authors would like to thank CRAG for kindly supplying Japanese rice seeds. CTQ2015-66254-C2-1-P project from MINCO (Spain) is also acknowledged.

REFERENCES

(1) Stoll, D. R.; Carr, P. W. *Anal. Chem.* **2017**, *89*, 519−531.
(2) Gargano, A. F. G.; Duffin, M.; Navarro, P.; Schoenmakers, P. J. *Anal. Chem.* **2016**, *88*, 1785−1793.
(3) Vonk, R. J.; Gargano, A. F. G.; Davydova, E.; Dekker, H. L.; Eeltink, S.; De Koning, L. J.; Schoenmakers, P. J. *Anal. Chem.* **2015**, *87*, 5387−5394.
(4) Montero, L.; Ibáñez, E.; Russo, M.; di Sanzo, R.; Rastrelli, L.; Piccinelli, A. L.; Celano, R.; Cifuentes, A.; Herrero, M. *Anal. Chim. Acta* **2016**, *913*, 145−159.
(5) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.
(6) Porter, S. E. G.; Stoll, D. R.; Rutan, S. C.; Carr, P. W.; Cohen, J. D. *Anal. Chem.* **2006**, *78*, 5559−5569.
(7) Mohler, R. E.; Dombek, K. M.; Hoggard, J. C.; Young, E. T.; Synovec, R. E. *Anal. Chem.* **2006**, *78*, 2700−2709.
(8) Parastar, H.; Radović, J. R.; Jalali-Heravi, M.; Diez, S.; Bayona, J. M.; Tauler, R. *Anal. Chem.* **2011**, *83*, 9289−9297.
(9) Parastar, H.; Tauler, R. *Anal. Chem.* **2014**, *86*, 286−297.
(10) Navarro-Reig, M.; Jaumot, J.; van Beek, T. A.; Vivó-Truyols, G.; Tauler, R. *Talanta* **2016**, *160*, 624−635.
(11) Bortolato, S. A.; Olivieri, A. C. *Anal. Chim. Acta* **2014**, *842*, 11−19.
(12) Cook, D. W.; Rutan, S. C.; Stoll, D. R.; Carr, P. W. *Anal. Chim. Acta* **2015**, *859*, 87−95.
(13) Tistaert, C.; Bailey, H. P.; Allen, R. C.; Vander Heyden, Y.; Rutan, S. C. *J. Chemom.* **2012**, *26*, 474−486.
(14) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC, Trends Anal. Chem.* **2016**, *82*, 425−442.
(15) Monteiro, M. S.; Carvalho, M.; Bastos, M. L.; De Pinho, P. G. *Curr. Med. Chem.* **2013**, *20*, 257−271.
(16) Montero, L.; Herrero, M.; Ibáñez, E.; Cifuentes, A. *J. Chromatogr. A* **2013**, *1313*, 275−283.
(17) Navarro-Reig, M.; Jaumot, J.; García-Reiriz, A.; Tauler, R. *Anal. Bioanal. Chem.* **2015**, *407*, 8835−8847.
(18) Pirok, B. W. J.; Pous-Torres, S.; Ortiz-Bolsico, C.; Vivó-Truyols, G.; Schoenmakers, P. J. *J. Chromatogr. A* **2016**, *1450*, 29−37.
(19) Daubechies, I. *Ten Lectures on Wavelets*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1992; Vol. *61*.
(20) Mallat, S. G. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1989**, *11*, 674−693.
(21) Trygg, J.; Kettaneh-Wold, N.; Wallbäcks, L. *J. Chemom.* **2001**, *15*, 299−319.
(22) Izadmanesh, Y.; Garreta-Lara, E.; Ghasemi, J. B.; Lacorte, S.; Matamoros, V.; Tauler, R. *J. Chromatogr. A* **2017**, *1488*, 113−125.
(23) Jaumot, J.; de Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1−12.
(24) De Juan, A.; Jaumot, J.; Tauler, R. *Anal. Methods* **2014**, *6*, 4964−4976.

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

(25) Ruckebusch, C.; Blanchet, L. *Anal. Chim. Acta* **2013**, *765*, 28−36.
(26) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Smilde, A. K. *Anal. Chim. Acta* **2005**, *530*, 173−183.
(27) Parsons, H. M.; Ludwig, C.; Günther, U. L.; Viant, M. R. *BMC Bioinf.* **2007**, *8*, 234.
(28) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(29) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.
(30) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; de souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. *Nucleic Acids Res.* **2009**, *37*, D603−D610.
(31) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Weerasinghe, D.; Zhang, P.; Karp, P. D. *Nucleic Acids Res.* **2014**, *42*, D459−D471.
(32) Martos, P.; McCormick, H. *Compr. Anal. Chem.* **2012**, *58*, 111−167.
(33) Phelix, C. F.; Feltus, F. A. *Plant Biol.* **2015**, *17*, 63−73.
(34) Mohanty, B.; Kitazumi, A.; Cheung, C. Y. M.; Lakshmanan, M.; de los Reyes, B. G.; Jang, I. C.; Lee, D. Y. *Plant Sci.* **2016**, *242*, 224−239.
(35) Chow, B. Y.; Kay, S. A. *Semin. Cell Dev. Biol.* **2013**, *24*, 383−392.
(36) Alier, M.; Felipe-Sotelo, M.; Hernàndez, I.; Tauler, R. *Anal. Chim. Acta* **2009**, *642*, 77−88.
(37) Malik, A.; Jordao, R.; Campos, B.; Casas, J.; Barata, C.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2016**, *159*, 58−68.
(38) Izawa, T.; Mihara, M.; Suzuki, Y.; Gupta, M.; Itoh, H.; Nagano, A. J.; Motoyama, R.; Sawada, Y.; Yano, M.; Hirai, M. Y.; Makino, A.; Nagamura, Y. *Plant Cell* **2011**, *23*, 1741−1755.

DOI: 10.1021/acs.analchem.7b01648
Anal. Chem. 2017, 89, 7675−7683

Informació Suplementària a la Publicació 6

Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution.

M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler.

Analytical Chemistry 89 (2017), 7675-7683.

Conditions of the environmental test chamber MLE-352H.

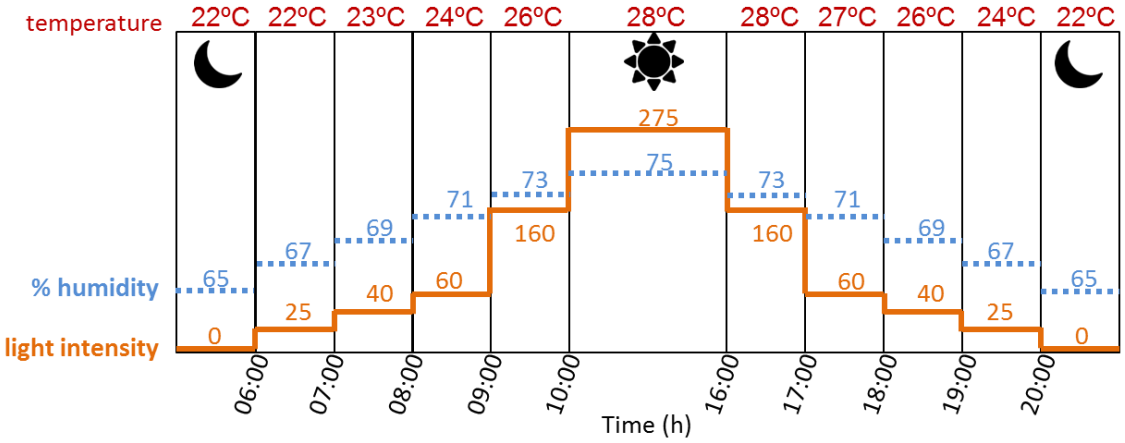


Figure S1. Experimental temperature, relative humidity and light long-day rice cultivation conditions at the growth chamber.

Detailed description of chemometric tools.

1. Regions of interest (ROI) strategy

The regions of interest (ROI) strategy was used to reduce the size of MS spectra. ROI strategy allows for the selection of interesting mass traces, which means *m/z* values whose intensity signals are higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and, also, showing a number of occurrences that allows the proper definition of the chromatographic peak.

Figure S2 shows the general scheme of ROI strategy. First (step 1), for every retention in a chromatogram, the ROI *m/z* values are searched. This search generates a couple of vectors containing the found *m/z* values and their related intensities for every retention time. Logically, as different compounds are detected at different retention times, these found *m/z* values are different among retention times and, therefore, the obtained vectors at different retention times have distinct lengths (depending on the number of ROIs found at each retention time). Finally, these vectors are reorganized into a matrix grouping the common ROIs among all the retention times. The final *m/z* values of each ROI are calculated as the mean of the *m/z* values obtained for that specific ROI (step 2). In the case that a particular ROI is not present at a retention time, the intensity for this ROI at this retention time is set to a low random intensity value at the noise level. This allows preserving all the information detected in both common and uncommon ROIs in the finally obtained matrix for each sample.

For the metabolomic study, the whole set of samples should be simultaneously considered. For this reason, a superaugmented data matrix has to be build up from the previously 80 individual compressed matrices obtained by the ROI search described above. Every compressed matrix was then normalized to correct the instrumental intensity drifts among injections. Normalization was done by dividing each matrix by the chromatographic area of the internal standard (PIPES) added to the metabolite extract of that sample. After normalization, individual compressed matrices were arranged in a column-wised augmented data matrix.

Since compressed individual data matrices have a different number of ROI *m/z* values (each sample has a different number of columns/ROI *m/z* values, left side of step 3), a preliminary pairwise search of ROIs among different samples is performed. Again, common and uncommon ROI *m/z* values are evaluated and finally considering both of them. This strategy allows reducing the number of ROI *m/z* values by grouping those with an *m/z* difference below the mass error tolerance. When an individual compressed matrix has not a significant intensity at a particular ROI *m/z* value, low random intensity values at the noise level were assigned. In this work, the strategy for column-wise augmentation consists of the independent augmentation of watered and not watered samples. For instance, the first step considered ROI matrices for

WATERED_1 and WATERED_2 samples resulting in a matrix containing ROIs found in both samples (WATERED_1_2). Next, this matrix is compared with WATERED_3 generating an augmented matrix with all the relevant information of the three samples (WATERED_1_2_3). These comparisons are repeated until all watered (or not watered) samples have been considered. In a final step, watered (watered_1_2_3_n) and not watered samples (NOT_WATERED_1_2_3_n) to obtain the final column-wise superaugmented matrix. More details about ROI strategy can be found at the works of Gorrochategui *et. al.* [1,2].

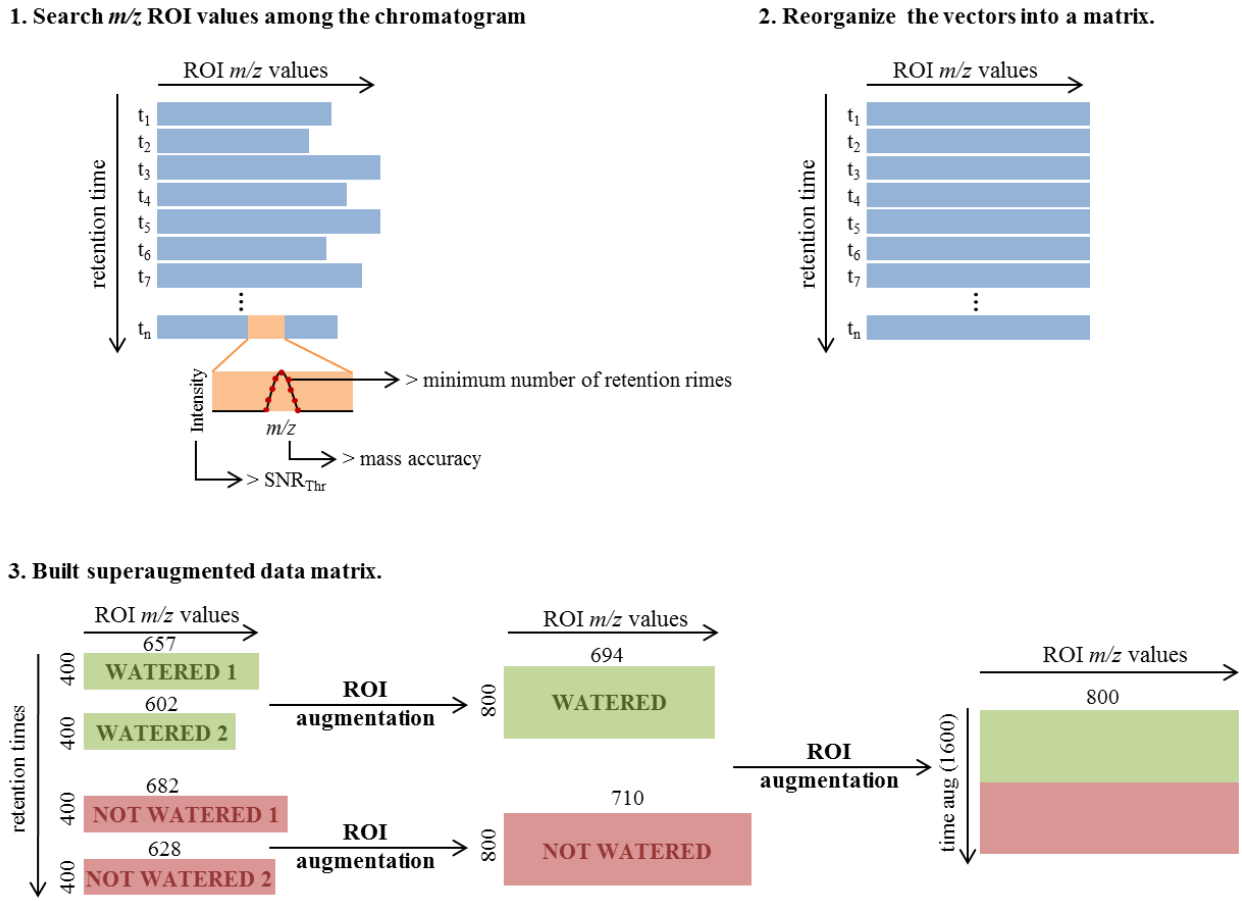


Figure S2. General Scheme of ROI strategy. Step 1: ROI *m/z* values are searched for every retention in a chromatogram. Step 2: Reorganization of the vectors into a matrix grouping the common ROIs among all the retention times. Step 3: Building of the superaugmented data matrix.

2. MCR-ALS for feature detection

MCR-ALS decomposes experimental data sets arranged in a data matrix according to the following bilinear model:

$$\mathbf{D}_K = \mathbf{C}_K\mathbf{S}^\mathbf{T} + \mathbf{E}_K \qquad \text{Equation (1)}$$

Where, in the case of this work, $\mathbf{D_K}$ (size $I\times J$) is the experimental LC-HRMS matrix corresponding to one of the second dimension column modulation taken from the first dimension column. Rows of this matrix are the MS spectra at all second dimension retention times, and columns are the second dimension chromatograms at all *m/z* ROI values. $\mathbf{C_K}$ (size $I\times N$) is the matrix containing the resolved second dimension elution profiles for this modulation and $\mathbf{S^T}$ (size $N\times J$) is the matrix containing their corresponding mass spectra. N represents the number of resolved components using the MCR-ALS method. Finally $\mathbf{E_K}$ (size $I\times J$) is the matrix of the residuals not explained by the MCR model.

This data analysis strategy can be easily extended to the simultaneous analysis of several chromatographic runs (several second dimension column modulations and several samples). The same bilinear model used in Equation 1 can be extended as follows:

$$\mathbf{D_{saug}} = \begin{bmatrix} \mathbf{D_{1,1}} \\ \mathbf{D_{1,2}} \\ \vdots \\ \mathbf{D_{L,K}} \\ \vdots \\ \mathbf{D_{80,98}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{1,1}} \\ \mathbf{C_{1,2}} \\ \vdots \\ \mathbf{C_{L,K}} \\ \vdots \\ \mathbf{C_{80,98}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_{1,1}} \\ \mathbf{E_{1,2}} \\ \vdots \\ \mathbf{E_{L,K}} \\ \vdots \\ \mathbf{E_{80,98}} \end{bmatrix} = \mathbf{C_{saug}}\mathbf{S^T} + \mathbf{E_{saug}} \qquad \text{Equation (2)}$$

Where $\mathbf{D_{saug}}$ is the column-wise superaugmented data matrix build using the previously described ROI augmentation. Since L is the number of analyzed samples and K is the number of the second dimension modulations taken from the first columns, the number of rows for $\mathbf{D_{saug}}$ is equal to $I\times L\times K$ (figure 2B). Decomposition of matrix $\mathbf{D_{saug}}$ generates $\mathbf{C_{saug}}$ (size $ILK\times N$) which has second dimension resolved elution profiles at each modulation (K) and at each sample (L) for the N components (figure 2B). In addition, $\mathbf{S^T}$ (size $N\times J$) contains the mass spectra for the resolved N components (figure 2B). According to Equation 2, resolved mass spectra ($\mathbf{S^T}$) were forced to be the same for the common components (metabolites) in the different modulations from all the analyzed rice samples. However, elution profiles of the same component resolved in the column-wise augmented concentration matrix $\mathbf{C_{saug}}$ were allowed to be different for each one of the chromatographic runs. Fist dimension elution profiles for every component in each sample can be obtained by refolding appropriately every column in $\mathbf{C_{saug}}$ to give a matrix of dimensions ($I\times K$) for each sample. Every column of the refolded matrix will give the first dimension elution profile size ($1\times K$). Therefore, for each sample a matrix of the first dimension elution profiles (size $N\times K$)) is obtained.

An initial guess of the number of components of $\mathbf{D}_{saug}$ matrix was obtained using singular value decomposition (SVD) algorithm [3]. This number is only an initial approximation, as the final number of components is decided by taking into account data fitting results and the reliability of the resolved profiles. Initial estimates of pure spectra ($\mathbf{S}^T$) for the iterative optimization should be provided, which were computed using a purest spectra detection method based on SIMPLISMA approach [4,5]. Finally, ALS optimization was carried out applying non-negativity (for chromatographic elution and spectra profiles of every component) and spectral normalization (equal height). The application of constraints provides chemical meaning to the pure mathematical solution.

3. MCR-ALS for resolution of temporal profiles

In this case, MCR-ALS was used to resolve the metabolic profiles evolving over the time. A matrix **A** (size $L \times M$) containing the chromatographic peak areas of the M previously resolved metabolites in the L analyzed samples, was decompose according the bilinear model:

$$\mathbf{A} = \mathbf{T}\mathbf{F}^T + \mathbf{E} \qquad \text{Equation (3)}$$

Where, decomposition of matrix **A** generates matrices **T** and $\mathbf{F}^T$. **T** (size $L \times N$) is the matrix containing the temporal evolution profiles for the N resolved components, while $\mathbf{F}^T$ (size $N \times M$) contains the contribution of the different metabolites to the temporal profiles. N represents the number of resolved components using the MCR-ALS method. Prior to MCR-ALS analysis, the chromatographic areas of the metabolites were scaled using a generalized logarithm transformation.

4. ANOVA-simultaneous component analysis (ASCA)

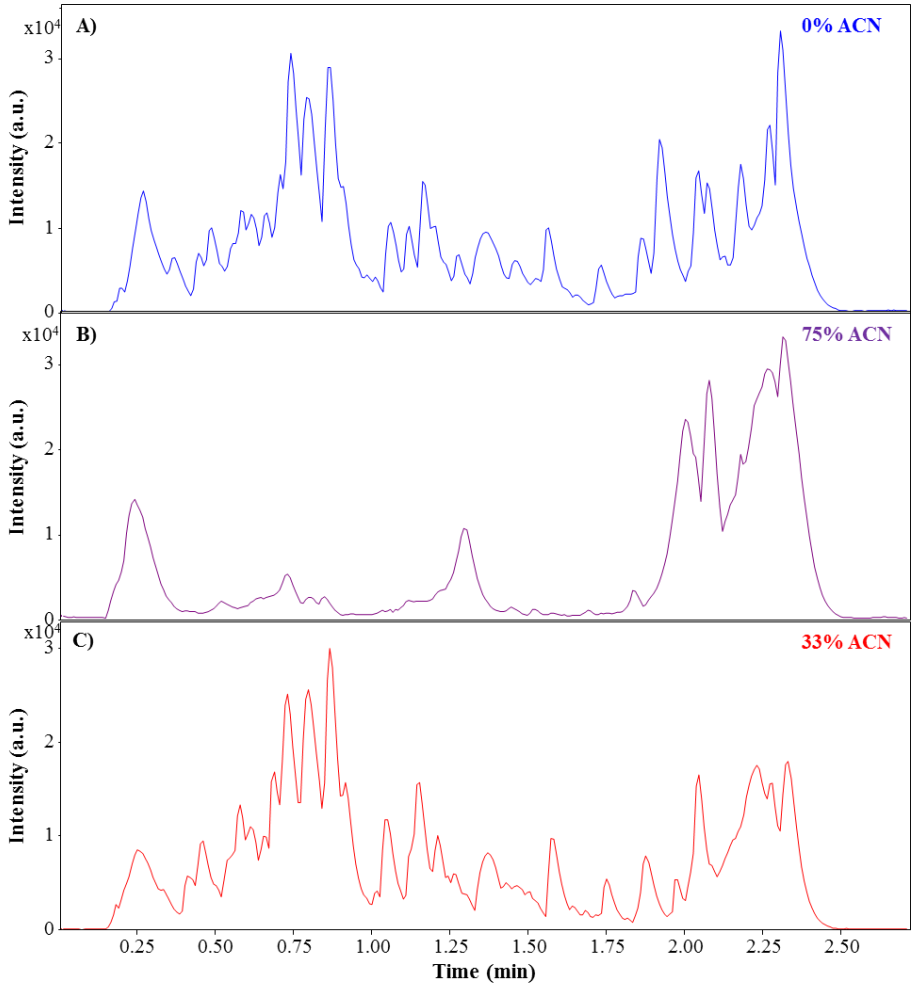
ASCA is a multivariate analysis of variance method that combines the power of ANOVA to separate variance sources with the advantages of simultaneous component analysis (SCA) to the modelling of the individual separate effect matrices. SCA is a generalization of PCA for the situation where the same variables have been measured in multiple conditions [6]. ASCA is especially useful for the analysis of complex multivariate datasets containing an underlying experimental design, and it allows for a natural interpretation of the variance induced by the different factors in the design [7]. In ASCA, ANOVA is firstly performed on the raw data matrix, which is decomposed into the sum of different data matrices characterising the variance caused by each one of the considered factors, plus a residual matrix containing the unexplained variance. Then, SCA is applied to each ANOVA factor matrix individually [7,8]. For a more detailed description of ASCA procedure see the work of Smilde [7] and Jansen [8]. In order to check the statistical significance of the effects of the investigated factors and their interactions, a

permutation test can be performed in which the null hypothesis (H_0) assumes that there is no effect of the considered factor. More details regarding the statistical assessment of ASCA results by using a permutation test can be found at the work of Vis *et. al.* [9].

5. Partial least squares – discriminant analysis (PLS-DA)

Watering factor was studied using PLS-DA[10]. PLS-DA is a multivariate regression method oriented to discriminate among different groups of samples. For instance, in the particular case studied in this work it has been used to discriminate between watered and dried samples. In this method, peak areas of every sample (**X**, predictor variable) were correlated with the vector describing the sample type class membership (**y**, predicted variable) [11]. Apart from sample class classification, PLS-DA provides information about which are the most important variables (i.e. resolved metabolites) for achieving this discrimination. One of the common ways to do this is, for instance, using the variable importance on projection (VIP) scores [12]. VIP scores are a weighted sum of squared PLS variable weights which measure the importance of each predictor variable into the final PLS model. The average of squared VIP score is equal to 1, therefore the "greater than one" rule is commonly used as variable selection criteria: variables with a VIP score greater than 1 can be considered important for a given model [13].

Breakthrough Reduction

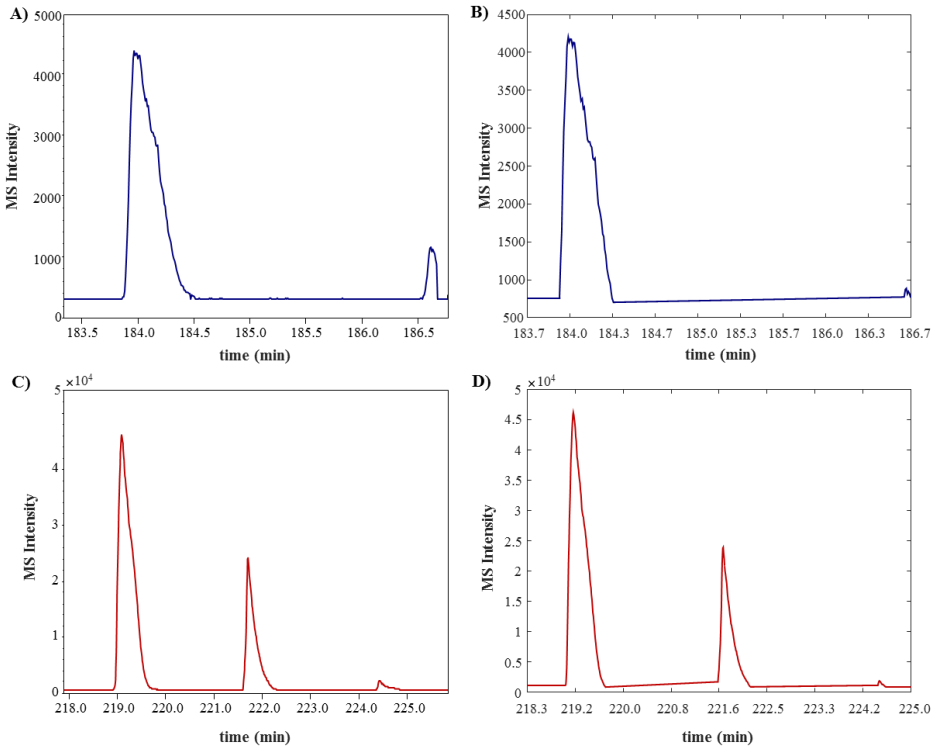


FigureS3. Separation obtained in the second dimension column (C18 RP) when the percentage of acetonitrile in the sample solvent is: 0% (A), 75% (B) and 33% (C).

Figure S3A depicts the total ion current chromatogram obtained in the second dimension column when the rice sample is injected dissolved in 100% of water. However, sample exists from the first dimension column in 75% of acetonitrile, which cause breakthrough, making the achieved separation considerably worse. As can be seen in figure S3B, when the percentage of acetonitrile in the sample solvent is 75%, the peaks obtained between 0.5 and 2.0 minutes are completely lost. With the addition of water at a flow rate of 16 $\mu L \cdot min^{-1}$ to the sample solvent between the second dimension pumps and the second dimension column, the percentage of the acetonitrile in the sample solvent was reduced to 33%. Figure S3C shows the obtained separation on the second dimension column when the percentage of acetonitrile in the sample solvent is 33%. As can be seen, the separation is nearly the same than the one obtained with 100% water as sample solvent, therefore the breakthrough is almost reduced.

Example of the application of size compression steps to LC×LC-HRMS data.

To demonstrate the suitability of ROI strategy in the case of LC×LC-HRMS data, figure S4 shows the extracted ion chromatograms (EIC) of raw data and ROI-compressed data for the internal standard, PIPES (a and b) and one of the most abundant metabolites in rice, trehalose (c and d). In both cases, the elution profiles of raw data and ROI-compressed data were practically identical, they showed the same MS-intensity and retention times in all modulations. In the case of PIPES, two modulations were observed, one at 184.0 min with an MS intensity of 4500 and the other at 186.6 min with an MS Intensify of 1000. In the case of trehalose, three modulations were obtained, the first one appeared at 219.0 min with an MS intensity of 45000, the next one was at 221.9 min with an MS intensity of 25000 and the last one at 225.5 min and an MS intensity equal to 5000. Moreover, the ROI values obtained for PIPES and trehalose (301.0570 and 341.1067 respectively) had an *m/z* error lower than 10 ppm, which is the recommended when a TOF analyzer is used.



FigureS4. Extracted ion chromatogram (EIC). A) Raw data for internal standard PIPES. B) ROI-compressed data for internal standard PIPES. C) Raw data for trehalose. D) ROI-compressed data for trehalose.

The usefulness of wavelet analysis decomposition for compressed the data among the time direction is demonstrated in figure S5, which shows the total current ion chromatogram (TIC) of one rice sample for uncompressed data (figure S5a) and data compressed with a 4-level wavelet analysis (figure S5b). It should be highlighted that the major part of chromatographic peaks observed in figure S5a (uncompressed data) are also observed in figure S5b (compressed data). However, after the compression of time direction shapes of compressed chromatographic peaks were slightly different than the peaks observed in uncompressed data.

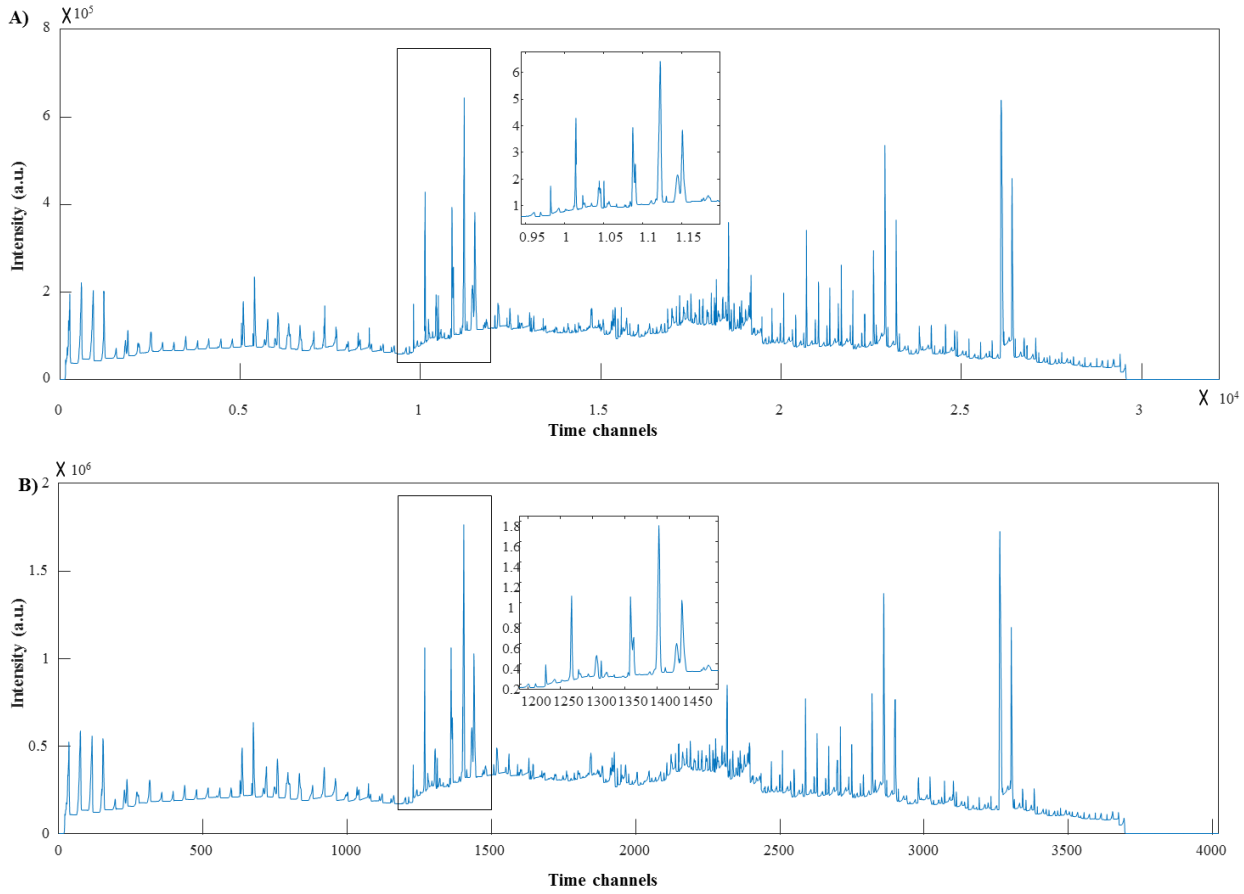


Figure S5. Total current ion chromatogram (TIC) of one rice sample for uncompressed data (A) and data compressed with a 4-level wavelet analysis (B). Inserts show a zoomed view of the region from 81 to 100 min (time channels from 9500 to 12000 in raw data and from 1200 to 1500 in compressed data).

Identified metabolites

Table S1. Tentative identification results.

Exact mass	Compound Name	Ion Assignment	Rel. mass error (ppm)	Identifier			Fold Change	Family
				HMDB	LIPID MAPS	KEGG		
195,0853	L-Arginine	$[M+Na-2H]^-$	5	HMDB00517		C00062	1,00	Amino acids
170,0282	L-Methionine	$[M+Na-2H]^-$	14	HMDB00696		C00073	-24,52	
343,1129	Avenic acid A	$[M+Na-2H]^-$	1	HMDB30416			0,25	
400,2778	Stearidonyl carnitine	$[M-H2O-H]^-$	18	HMDB06463			1,00	
383,1170	S-Adenosyl-L-homocysteine	$[M-H]^-$	7			C00021	-6,00	
191,0484	N-Carbamoyl-2-amino-2-(4-hydroxyphenyl)acetic acid	$[M-H2O-H]^-$	14	HMDB31813			1,00	
405,1023	S-Adenosylhomocysteine	$[M+Na-2H]^-$	15	HMDB00939		C00021	-4,90	
329,2592	Phytal	$[M+Cl]^-$	7	HMDB35654			0,25	Carbohydrates
330,0688	N-Acetylmuramate	$[M+K-2H]^-$	16			C02713	0,85	
683,2281	Maltulose	[M-H]-	4	HMDB29919			0,36	
341,1107	Trehalose	$[M-H]^-$	5	HMDB00975		C01083	-5,86	
377,0874	1-O-Feruloylglucose	$[M+Na-2H]^-$	5	HMDB36938			-0,42	
387,1148	Maltose	$[M+FA-H]^-$	1	HMDB00163		C00208	-0,07	
730,2371	4-Nitrophenyl N,N',N''-triacetyl-b-D-chitotriose	$[M-H2O-H]^-$	15				-1,16	
461,1187	N,N'-diacetylchitobiose	$[M+K-2H]^-$	1	HMDB03556		G10336	0,89	
329,0713	Xanthosine	$[M+FA-H]^-$	7			C01762	1,00	Nucleosides
565,1720	Guanosine	$[2M-H]^-$	7	HMDB00133		C00387	0,64	
326,1196	2'-Deoxyguanosine	$[M+CH3COO]^-$	20	HMDB00085		C00330	1,00	
312,0960	Adenosine	$[M+FA-H]^-$	3	HMDB00050		C00212	-1,59	
215,0310	Orotic acid	$[M+CH3COO]^-$	0	HMDB00226		C00295	-0,17	Nucleotides
305,0164	dUMP	$[M-H]^-$	5	HMDB01409		C00365	0,30	
343,0354	dTMP	$[M+Na-2H]^-$	12	HMDB01227		C00364	0,41	

406,0727	AMP	$[M+CH3COO]^-$	10	HMDB00045		C00020	0,35	
309,1745	Xanthoxin	$[M+CH3COO]^-$	12			C13453	-0,13	Hormones
485,2789	Ecdysone	$[M+Na-2H]^-$	19		LMST01010210	C00477	0,01	
295,1984	Farnesoic acid methyl ester	$[M+FA-H]^-$	23				-0,07	
297,1462	Gibberellin	$[M-H2O-H]^-$	9			C11863	-7,35	
193,0529	5-Hydroxyconiferaldehyde	[M-H]-	11			C12204	0,85	
481,2587	1-Octadecanoyl-sn-glycerol 3-phosphate	$[M+FA-H]^-$	3				-0,27	
467,2633	3,17-Androstanediol glucuronide	$[M-H]^-$	3		LMST05010037		0,75	
425,1878	Abscisic acid (ABA)	$[M-H]^-$	14			C15970	0,98	
387,2862	2α-Methyl-17β-[(tetrahydro-2H-pyran-2-yl)oxy]-5α-androstan-3-one	$[M-H]^-$	10			C15422	0,16	
491,1222	Genistein	$[M+CH3COO]^-$	5			C09126	0,33	
563,3262	Desglucocoroloside	$[M+CH3COO]^-$	6	HMDB33709			0,77	
353,2003	Blumenol C glucoside	$[M-H2O-H]^-$	11	HMDB40668			1,79	Glycosides
837,4869	PG(16:1(9Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z))	$[M+FA-H]^-$	6	HMDB10599			2,39	
505,2577	(3b,9R)-5-Megastigmene-3,9-diol 9-[apiosyl-(1->6)-glucoside]	[M-H]-	15	HMDB38327			4,96	
781,2141	Kaempferol 3-(2"-rhamnosyl-6"-acetylgalactoside) 7-rhamnoside	[M-H]-	7	HMDB40541			3,70	
607,4235	Campesteryl glucoside	$[M+FA-H]^-$	3		LMST01031126		0,33	
623,1642	Pelargonidin-3-O-rutinoside	$[M+FA-H]^-$	3			C12644	0,94	
739,1938	Pelargonidin-3,5-diglucoside-5-O-p-coumaroylglucoside	$[M-H]^-$	7			C16349	0,69	
415,3116	2-Monooleoylglycerol	$[M+CH3COO]^-$	12				0,13	
563,1431	Apigenin 6-C-glucoside 8-C-arabinoside	$[M-H]^-$	4	HMDB29260			0,25	
729,2322	Galabiose	$[M+FA-H]^-$	2	HMDB29902			0,53	
529,0993	Resveratrol 3-glucoside 4'-sulfate	$[M+CH3COO]^-$	5	HMDB37073			0,12	
593,1535	Pelargonidin 3-O-sophoroside	$[M-H]^-$	3	HMDB33679		C16305	0,48	

739,2016	Kaempferol 3-glucuronide	$[M-H]^-$	10		LMPK12111775		0,17	
527,0988	(-)-Epigallocatechin 3'-glucuronide	$[M+FA-H]^-$	10	HMDB41638			0,11	
733,2013	Glucoliquiritin apioside	$[M+Na-2H]^-$	7	HMDB41149			0,68	
439,0901	3,5-Dihydroxyphenyl 1-O-(6-O-galloyl-beta-D-glucopyranoside)	$[M-H]^-$	4	HMDB39307			0,21	
432,1409	Benzyladenine-7-N-glucoside	$[M+FA-H]^-$	16				0,30	
497,3347	Protopanaxadiol	$[M+K-2H]^-$	11		LMPR0106080003		0,26	
579,1353	Kaempferol 3-glucoside-7-xyloside	[M-H]-	0		LMPK12111736		0,26	Flavonoids
409,0324	Quercetin 3'-isobutyrate	$[M+K-2H]^-$	1		LMPK12112302		3,30	
329,0706	2-hydroxyformononetin	$[M+FA-H]^-$	11		LMPK12050081	C02920	0,08	
443,0607	Kaempferol 3-glucuronide	$[M-H2O-H]^-$	1		LMPK12111843		0,40	
659,1434	Kaempferol 7,4'-dimethyl ether 3-neohesperidoside	$[M+Na-2H]^-$	7		LMPK12112593		0,38	
571,1481	Maysin 3'-methyl ether	$[M-H2O-H]^-$	5		HMDB37420		2,44	
337,0704	Ageconyflavone A	$[M-H2O-H]^-$	4		LMPK12111248		0,00	
675,1364	Quercetagetin 7-methyl ether 3-neohesperidoside	$[M+Cl]^-$	4		LMPK12112966		1,47	
455,1043	Epicatechin 3-O-(4-O-methylgallate)	$[M-H]^-$	13		LMPK12020094		-0,66	
321,0336	Scutellarein 7-methyl ether	$[M+Na-2H]^-$	13				1,40	
492,1255	8C-glucosyl-2,5,7-trihydroxyflavanone	$[M+CH3COO]^-$	3				0,10	
265,1481	6,7-Dihydro-4-(hydroxymethyl)-2-(p-hydroxyphenethyl)-7-methyl-5H-2-pyrindinium	$[M-H2O-H]^-$	5	HMDB33483			0,55	
325,1847	Ajmaline	[M-H]-	12			C06542	-10,57	Alkaloids
354,2036	Pipercide	[M-H]-	10	HMDB33449			0,97	
312,1725	Bicyclomahanimbicine	$[M-H2O-H]^-$	8	HMDB30221			0,69	
266,1502	Arenaine	$[M+CH3COO]^-$	3	HMDB30354		C09912	0,16	
523,2683	Hovenine A	$[M+K-2H]^-$	1	HMDB30200			-1,00	

351,1309	Feruloylserotonin	$[M-H]^-$	11	HMDB32759			0,39	
277,0342	PhaselateP	$[M-H2O-H]^-$	2				0,33	Carboxylic acids
367,1042	5-O-Feruloylquinic acid	$[M-H]^-$	2			C02572	0,56	
267,1399	Geranylhydroquinone	[M+Na-2H]-	12				0,58	Aromathic Componds
345,1905	3-(1,1-Dimethyl-2-propenyl)-8-(3-methyl-2-butenyl)xanthyletin	$[M-H2O-H]^-$	14	HMDB30730			0,49	
173,0463	Shikimic acid	$[M-H]^-$	4	HMDB03070		C00493	0,70	
226,9969	3-Dehydroquinate	$[M+K-2H]^-$	2	HMDB12710		C00944	0,62	
431,1259	1-O-sinapoyl-β-D-glucose	$[M+FA-H]^-$	14			C01175	-14,65	
681,2980	Coproporphyrinogen III	$[M+Na-2H]^-$	11	HMDB01261		C03263	-0,11	Cofactors
293,0499	Pyridoxamine 5'-phosphate	$[M+FA-H]^-$	15	HMDB01555		C00647	0,37	
279,0506	2-Succinylbenzoate	$[M+CH3COO]^-$	1			C02730	0,80	
269,1182	4-Hydrocinnamoyl-2,2,5-trimethyl-4-cyclopentene-1,3-dion	[M-H]-	0		LMPK12120611		0,50	Lipids
514,3251	LysoPE(0:0/22:2(13Z,16Z))	$[M-H2O-H]^-$	9	HMDB11492			0,50	
399,2438	Prostaglandin F2a	$[M+FA-H]^-$	12	HMDB01139		C00639	0,83	
425,2564	Prostaglandin PGE2 1-glyceryl ester	[M-H]-	4	HMDB13043			-8,25	
277,2085	2-Hydroxyhexadecanal	[M+Na-2H]-	15				0,30	
389,1698	Prostaglandin E2	$[M+K-2H]^-$	9	HMDB01220		C00584	0,34	
563,5042	DG(15:0/18:0/0:0)	$[M-H2O-H]^-$	1	HMDB07071			0,04	
503,2425	PG(18:4(6Z,9Z,12Z,15Z)/0:0)	[M-H]-	1		LMGP04050021		0,12	
483,2728	PG(16:0/0:0)	[M-H]-	0		LMGP04050008		0,13	
559,3141	PG(22:4(7Z,10Z,13Z,16Z)/0:0)	[M-H]-	17		LMGP04050017		-0,02	
507,4411	DG(14:0/15:0/0:0)	$[M-H2O-H]^-$	1	HMDB07010			-0,05	
655,4449	DG(18:4(6Z,9Z,12Z,15Z)/20:5(5Z,8Z,11Z,14Z,17Z)/0:0)	$[M+Na-2H]^-$	16	HMDB07346			0,18	
461,2689	2-Acetoxy-3-geranylgeranyl-1,4-dihydroxybenzene	$[M+Na-2H]^-$	4	HMDB40748			0,39	
755,4623	PG(16:0/16:1(9Z))	$[M+Cl]^-$	2	HMDB10571			-0,13	
659,4752	3-Heptaprenyl-4-hydroxybenzoate	$[M+FA-H]^-$	10				0,02	

721,3690	Physalolactone B 3-glucoside	$[M+CH3COO]^-$	8	HMDB34201			0,36	
564,3315	1-18:2-Lysophosphatidylcholine	$[M+FA-H]^-$	1	HMDB10386		C04230	0,98	
981,5884	1-18:3-2-18:3-Digalactosyldiacylglycerol	$[M+FA-H]^-$	9				0,41	
540,3313	1-16:0-2-Lysophosphatidylcholine	$[M+FA-H]^-$	1	HMDB10382		C04230	0,21	
476,2793	LysoPE(0:0/18:2(9Z,12Z))	$[M-H]^-$	2	HMDB11477			0,21	
562,3180	LysoPC(18:3(6Z,9Z,12Z))	$[M+FA-H]^-$	5	HMDB10387		C04230	-0,05	
566,3427	LysoPE(0:0/20:1(11Z)	$[M+CH3COO]^-$	6	HMDB11482			-2,36	
433,2302	1-Oleyl-2-lyso-phosphatidate	$[M-H]^-$	13	HMDB07852		C00416	-1,57	
959,6033	1-16:0-2-18:3-Digalactosyldiacylglycerol	$[M+FA-H]^-$	8				-435,12	
229,0117	Glycerol 1-phosphate	$[M+CH3COO]^-$	0			C00661	-0,10	
520,2660	LysoPE(0:0/18:3(6Z,9Z,12Z))	$[M+FA-H]^-$	4	HMDB11478			0,48	
637,1804	Apigenin 7-rutinoside	$[M+CH3COO]^-$	4		LMPK12110355		0,15	
595,2885	PI(18:2(9Z,12Z)/0:0)	$[M-H]^-$	0		LMGP06050010		0,49	
452,2792	LysoPE(0:0/16:0)	$[M-H]^-$	2	HMDB11473			0,16	
431,2127	1-Palmitoylglycerol 3-phosphate	$[M+Na-2H]^-$	12			C04036	0,34	
565,3326	1-18:2-Lysophosphatidylcholine	$[M+FA-H]^-$	10	HMDB10386		C04100	0,32	
365,1090	Crocetin	$[M+K-2H]^-$	19				0,30	
311,1695	4'-Hydroxy-5,5'-diisopropyl-2,2'-dimethyl-3,4-biphenylquinone	[M-H]-	13	HMDB40761			0,53	
293,1796	Germacrone 4,5-epoxide	$[M+CH3COO]^-$	12	HMDB35889		C17489	0,76	
577,2726	Geranylfarnesyl diphosphate	$[M+CH3COO]^-$	4				-0,42	
213,1508	Menthone	$[M+CH3COO]^-$	5	HMDB35783			0,31	Terpenes
227,1299	(3S)-3-Hydroxycyclocitral	$[M+CH3COO]^-$	4			C19731	0,60	

389,1204	16-Hydroxytabersonine	$[M+K-2H]^-$	17			C11643	0,14	
397,2269	Ent-8-D2t-IsoP	$[M+FA-H]^-$	9		LMFA03110055		0,08	Fatty acids
405,3299	24-Hydroxy-tetracosanoic acid	$[M+Na-2H]^-$	12		LMFA01050213		0,20	
221,1471	Lauric acid	$[M+Na-2H]^-$	15		LMFA01010012	C02679	0,09	
390,2023	(+)-Mahanimbine	$[M+CH3COO]^-$	13	HMDB30318		C09220	-0,04	
557,4939	Mycocerosic acid	$[M+Cl]^-$	18		LMFA01020319		0,21	
242,1784	Undecanoylglycine	[M-H]-	9	HMDB13286			0,00	
381,2311	Leukotriene B4	$[M+FA-H]^-$	7	HMDB01085		C02165	0,17	
294,1852	N-Isobutyl-2,4,8,10,12-tetradecapentaenamide	$[M+CH3COO]^-$	9	HMDB31184			0,10	
511,4738	Melissic acid A	$[M+CH3COO]^-$	1	HMDB30925			0,26	
327,2511	Oleic acid	$[M+FA-H]^-$	9	HMDB00207		C00712	0,58	
819,5307	1-18:1-2-18:3-phosphatidylcholine	$[M+K-2H]^-$	14				0,29	
329,2264	16-Oxo-palmitate	$[M+CH3COO]^-$	19				-0,21	
409,0257	2-O-caffeoylglucarat	$[M+K-2H]^-$	19				0,84	Secundary Metabolites
621,4384	Oryzalexin A	$[M+Na-2H]^-$	15				-5,84	
351,0726	Gallocatechin	$[M+FA-H]^-$	1		LMPK12020002	C12127	-0,19	
365,2041	Oryzalide A	$[M+FA-H]^-$	19	HMDB37591			0,42	
285,0438	Kaempferol	$[M-H]^-$	11	HMDB05801		C05903	0,50	
404,1042	6-methylthiohexyldesulfoglucosinolate	$[M+Cl]^-$	15				-0,08	
237,0439	Esculetin	$[M+CH3COO]^-$	14				0,31	
371,1087	Sesamolinol	$[M-H]^-$	13			C10883	-0,20	
403,0727	Indolylmethyl-desulfoglucosinolate	$[M+Cl]^-$	2			C16517	-0,24	

References.

(1) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC, Trends Anal. Chem.* **2016**, *82*, 425-442.
(2) Gorrochategui, E.; Jaumot, J.; Tauler, R. **2015**.
(3) Golub, G. H.; Loan, C. F. V. *Matrix computations*, third ed.; Johns Hopkins University Press: Baltimore, 1996, p 728.
(4) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425-1432.
(5) Windig, W.; Stephenson, D. A. *Anal. Chem.* **1992**, *64*, 2735-2742.
(6) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Smilde, A. K. *Anal. Chim. Acta* **2005**, *530*, 173-183.
(7) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R. J. A. N.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *21*, 3043-3048.
(8) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Westerhuis, J. A.; Smilde, A. K. *J. Chemometr.* **2005**, *19*, 469-481.
(9) Vis, D. J.; Westerhuis, J. A.; Smilde, A. K.; van der Greef, J. *BMC Bioinformatics* **2007**, *8*.
(10) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166-173.
(11) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
(12) Wold, S.; Johansson, A.; Cocchi, M. In *3D QSAR in Drug Design*, Kubiny, H., Ed.; ESCOM Science Publishers: Leiden, 1993, pp 583-618.
(13) Chong, I. G.; Jun, C. H. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103-112.

5.4. Publicació 7

An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis..

M. Navarro-Reig, J. Jaumot, R. Tauler

Enviat

An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis

Meritxell Navarro-Reig, Joaquim Jaumot* and Romà Tauler

Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

Abstract

Untargeted lipidomics sample are extremely complex and often exceed the limits of peak capacity achievable by one-dimensional liquid chromatography (LC). Comprehensive two-dimensional liquid chromatography (LC×LC) appears as a promising alternative to overcome this drawback. Unfortunately, this approach generates highly complex datasets which untargeted analysis is challenging. In this work, a global methodological strategy combining LC×LC-MS with chemometric data analysis is proposed for untargeted lipidomic studies. The feasibility of the proposed methodology is demonstrated by its application to assess the effects of arsenic exposure on the lipidome of growing rice samples. A two-dimensional chromatographic setup coupling reversed phase (RP) and hydrophilic interaction liquid chromatography (HILIC) modes together with a triple quadrupole mass detector (TQD) is proposed to analyze lipid extracts from rice samples at different experimental conditions. Chemometric tools were used for data compression, spectral and elution profiles resolution, feature detection and statistical analysis of the multidimensional LC×LC data. The obtained results revealed that the proposed methodology was useful to gather relevant information from untargeted lipidomic studies and detect potential biomarkers.

Keywords: Untargeted lipidomics, LC×LC-MS, comprehensive, chemometrics, ROIMCR.

1. Introduction

Lipidomics is the branch of metabolomics consisting in the comprehensive study of lipid species and their related networks and metabolomic pathways of a biological system. Lipids are a group of biomolecules involved in a wide range of structural and functional activities in cells. For example, lipids are structural components of cell membranes influencing both, its fluidity and its interactions (such as nutrient transport or waste products expulsion). Moreover, lipids are also involved in energy transport and storage, as well as in cell communication and signaling. For this reason, the comprehensive analysis of lipids is gaining more attention in many research fields, from biomedical to environmental studies[1-5].

There are two major analytical strategies in lipidomic studies: targeted and untargeted approaches. The target approach focuses on analyzing a specific list of lipids, typically related to some known pathway of interest, in an attempt to corroborate a previous hypothesis. In contrast, untargeted lipidomics aims to the global analysis of all measurable lipids present in a sample, without any prior assumption about affected pathways or lipid species[4,6-8]. In this work, the untargeted approach is used, with the aim of discovering which lipid species are associated with previously unexplored biological pathways[9].

The main drawback of untargeted lipidomics is the high complexity of the generated data because of the structural diversity of lipids. The International Lipid Classification and Nomenclature Committee established a "Comprehensive Classification

System for Lipids" based on their chemical and biochemical properties[10]. This classification system grouped lipids into eight different categories: fatty acids, glycerolipids, glycerophospholipids, sphingolipids, sterols, prenol lipids, saccharolipids, and polyketides. Moreover, there are several subcategories within each of the mentioned groups, resulting in a large number of combinations[1]. This great structural diversity and chemical complexity among lipids give a wide range of different physical properties, which causes the profiling of the complete lipidome of biological samples to be still a challenge[1-3,11].

To overcome this hurdle, analytical platforms used in untargeted lipidomics must have a high separation power and at the same time be highly sensitive and selective. Until now, liquid chromatography (LC) coupled to mass spectrometry (MS) is the most frequently used analytical platform in lipidomic studies[1,2,6-8,11]. Nevertheless, biological lipid samples contain thousands of lipids and many of them severely coeluted. The high complexity of these biological lipid extracts often exceeds the limits of peak capacity achievable by LC systems. Consequently, despite its high capacity resolving complex samples, LC-MS may still lose important information in lipidomic studies. Therefore, the use of multidimensional separation systems coupled to MS is proposed to overcome this drawback.

Among multidimensional analytical platforms, comprehensive two-dimensional liquid chromatography (LC×LC) offers two relevant advantages for untargeted lipidomic studies. The first benefit is related to the increase of the resolution capacity in comparison with mono-dimensional systems. The higher resolution power of LC×LC lies in the fact that under ideal circumstances (when the two separations systems are completely orthogonal), the total peak capacity is equal to the product of individual peak capacities of the first- and the second-dimension separations. In practice, the complete orthogonality of both separation systems is difficult to achieve, but a high peak capacity can be reached if two uncorrelated separation modes are used, such as reverse phase (RP) and hydrophilic interaction liquid chromatography (HILIC)[12-14]. The other significant advantage of LC×LC is that this powerful approach increases the identification ability because of the information provided on the different separation patterns of the two columns. In 2D chromatograms, peaks are usually observed along lines or arcs related to specific analyte functionalities (alkanes, aldehydes or degree of unsaturation among others), which provides additional help for identifying unknown lipids[12].

In contrast, the main drawback of LC×LC-MS is that generates complex datasets where the relevant information can remain hidden. Thousands of signals can be detected when analyzing lipid extracts by LC×LC-MS. The manual inspection of these signals is not feasible in practice and, therefore, their processing is not straightforward [15].

In order to achieve a complete resolution of complex lipid extracts and gather as much information as possible from the analyzed biological system, the use of advanced chemometric methods is recommended[15-17]. However, in the analytical literature, only a few number of works have been focused on the chemometric analysis of LC×LC data. Also, up to date, most of these works deal only with LC×LC-DAD (diode array detector) data[14,15,17-19].

The primary goal of this work is to present a new global methodological strategy combining the LC×LC-MS powerful analytical approach with advanced chemometric data analysis tools. This

procedure will provide as much information as possible from untargeted lipidomic studies. With this aim, the combination of the LC×LC-MS/MS method (coupling RPLC with HILIC) with some chemometric tools is proposed for the analysis of the entire lipidome of complex biological samples. The feasibility of the proposed approach is demonstrated by its application to the study of the changes produced on Japanese rice (*Oryza sativa* L.) lipidome under arsenic exposure.

2. Materials and Methods

2.1. Chemicals and Reagents

Sodium arsenate dibasic heptahydrate (≥98.0 %), HPLC grade water, HPLC grade acetonitrile, HPLC grade isopropanol, methanol (MeOH, HPLC grade), methyl tert-butyl ether (MTBE), ammonium acetate (≥99.0%), acetic acid (≥95.0%) and formic acid (≥95.0%) were supplied by Sigma-Aldrich (Steinheim, Germany).

Eight lipid standards from different families were used as surrogates: 17:0 monoacylglycerol, 17:1 lysophosphatidylethanolamine, 17:0 lysophosphatidylcholine, 1,3-17:0 D5 diacylglyceride, 17:0 cholesteryl ester, 1,2,3-17:0 triglyceride, 16:0 D31-18:1 phosphatidylcholine, 16:0 D31-18:1 phosphatidylserine. Three sphingolipids were used as internal standards: N-dodecanoylsphingosine, N-dodecanoylglucosyl-sphingosine and N-dodecanoylsphingosylphosphorylcholine. All these lipid standards were obtained from Avanti Polar Lipids (Alabaster, AL, USA).

Solutions containing 1 and 1000 µM of arsenic (As) were weekly prepared by adequately diluting a 10000 µM stock solution. The stock solution was prepared by dissolution of the appropriate amount of sodium arsenate salt. All solutions were stored at 6 ºC until their use.

Water used for plant watering and for preparing arsenic solutions was purified using an Elix 3 coupled to a Milli-Q system (Millipore, Belford, MA, USA), and filtered through a 0.22 µm nylon filter integrated into the Milli-Q system.

The following abbreviations have been used to describe lipid families: (PA) phosphatidic acid, (PC) phosphatidylcholines, (PG) Phosphatidylglycerol, (PI) phosphatidylinositols, (Cer) Ceramides, (MGDG) monogalactosyldiacylglycerol, (DAG) diacylglycerols, (TAG) triacylglycerols.

2.2. Plant growth and sample preparation

Plant growth and lipid extraction were performed using the procedure described elsewhere[20,21]. Briefly, rice seeds, obtained from the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain), were incubated for two days at 30ºC in a wet environment. After this period, plants were grown on an Environmental Test Chamber MLE-352H (Panasonic®) for 22 days simulating cyclic environmental changes of temperature, relative humidity, and light intensity, as shown in the Supplementary Information (Figure S1). During the first 10 days of growth, rice plants were watered with Milli-Q water three times per week. Since then, the plant treated samples were subjected to irrigation water containing low (1 µM) or high (1000 µM) concentrations of As, whereas the plant control samples were watered with Milli-Q water until harvest. The lower concentration was set to 1 µM because it is the limit of the acceptable As concentration in water by European legislation (Groundwater Directive2006/118/EC)[22]. After harvest, aerial parts and roots were separated and immediately samples were frozen at liquid nitrogen temperature for metabolism quenching. Then, samples were stored at -80ºC until extraction. Five biological replicates were made for each sample condition. Therefore, a total

number of 30 samples (15 aerial part samples and 15 root samples) were analyzed.

Before extraction, rice samples were ground to a fine powder using a liquid nitrogen mortar and lyophilized for 24 hours to dryness. Lipid extraction was carried out by dispersing 10 mg of the dried tissue in 1 mL of MTBE:MeOH (3:1). The mixture was fortified with 20 µL of the surrogates mix, and then, vortexed for 1 min and sonicated for 10 min. Next, 0.5 mL of H_2O:MeOH (3:1) were added, and the mixture was again vortexed for 1 min. After centrifuging for 5 min at 2000 x g, the organic fraction (upper) was collected. The aqueous phase (lower) was re-extracted with 0.65 mL of MTBE and 0.35 mL of MeOH:H_2O (1:0.85). Next, the mixture was vortexed for 1 min and centrifuged for 5 min at 2000 x g. After that, organic phases were combined and evaporated to dryness under nitrogen gas. All of the extracts were stored at -80ºC until analysed. Before injection, extracts were reconstituted with 250 mL of MeOH:H_2O (4:1) and 20 µL of the internal standards mix were added. Quality control (QC) samples were prepared by pooling 15 µL of all studied samples (extracts). QC for aerial part and root samples were separately prepared.

2.3.LC×LC-MS/MS analysis

LC×LC analyses were carried out on an Acquity UHPLC system (Waters, Milford, MA, USA) equipped with a quaternary pump and an autosampler. Second-dimension separation was possible due to the coupling to this instrument of one additional LC pump (Waters 1525 binary HPLC pump). The interface between the first and the second column was an Acquity UPLC Column Manager (Waters, Milford, MA, USA), equipped with two 6-port two-position valves. Figure 1 shows the scheme of the LC×LC system used in this work.

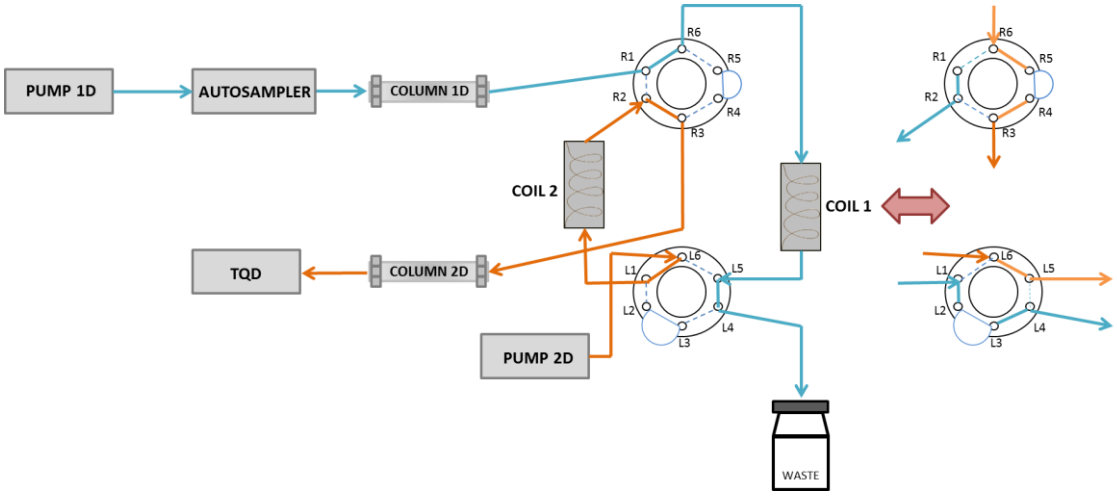


Figure 1. Scheme of the LC×LC-MS system. The red double arrow shows the change in modulator.

In the first chromatographic dimension, an RP ZORBAX Eclipse XDB-C18 (150 mm × 2.1 mm i.d.; 5µm) column provided by Agilent (Santa Clara, CA, USA) was used. Chromatographic analysis was run using (A) acetonitrile:isopropanol (1:2) 0.1% formic acid and (B) water 0.1% formic acid, as mobile phase, eluted according to the following gradient: 0 min, 80% A; 20.5 min 90% A; 78.0-98.5 min, 100% A; 98.5-99.5 min, back to initial conditions at 80% A and from 99.5 to 130 min, at 80% A. The mobile phase flow rate was 39 $\mu L \cdot min^{-1}$, and the injection volume was 20 µL.

The second chromatographic dimension employed a KINETEX HILIC (30 mm × 3 mm i.d.; 2.6 µm) column provided by Phenomenex (Torrance, CA, USA). Second-dimension elution gradient used 5

mM ammonium acetate at pH 5.5, adjusted with acetic acid (A) and acetonitrile (B), in an isocratic elution gradient at 16% A. The modulation time in the switching valve was set at 1.8 min. The mobile phase flow rate was 0.5 $mL \cdot min^{-1}$.

Mass spectroscopic detection was performed in a triple quadrupole detector (TQD, Waters, Milford, MA, USA) equipped with an electrospray (ESI) as ionization source working in both negative and positives modes. Nitrogen (purity >99.98 %) was used as desolvation gas at the flow rate of 800 $L \cdot h^{-1}$. Desolvation temperature was set at 450 ºC, and the con voltage was set at 50 V. First, samples were analysed in full scan mode in an untargeted manner using a mass acquisition range from 90 to 1800 Da. Then, after the application of the chemometric data analysis strategy (see the following section), the most important mass traces were selected. Finally, samples were reinjected in the same chromatographic conditions to obtain the MS/MS spectra of the selected mass traces. All MS/MS spectra were recorded at 10, 20, 30 and 40 eV collision energies (CE).

2.4. Chemometric data analysis strategy

The first step of the proposed data analysis strategy consisted on the compression and the arrangement of the raw LC×LC-MS data using an approach based on the selection of the so-called regions of interest (ROI)[23]. Data matrices generated by this strategy were then analyzed and resolved by means of the multivariate curve resolution by alternating least squares method (MCR-ALS)[24,25], which provides the pure elution and mass spectra profiles of the constituents (lipids) present in the analyzed samples. The procedure that combines the ROI data compression and the MCR-ALS analysis of the ROI compressed data is called ROIMCR and has been described in more detail elsewhere[23]. A further data analysis step was the statistical assessment of the effects of As exposure on rice based on the changes of the peak areas of the elution profiles of the resolved lipidic constituents by principal component analysis (PCA)[26] and ANOVA-simultaneous component analysis (ASCA)[27]. In addition, partial least squares – discriminant analysis (PLS-DA)[28] allowed the differentiation between control and As treated samples and the identification of the main features that allowed this differentiation. Lipids whose concentration changed because of As treatment were tentatively identified using their MS/MS spectra. In the following sections, these different steps are described in more detail.

2.4.1. Compression and data arrangement

Waters raw chromatographic data files (.raw format) were converted to the standard CDF format by the Databridge function of MassLynx™ 4.1 software (Waters, Milford, MA, USA). Then, these data files were imported into MATLAB environment (Release 2016b, The Mathworks Inc, Natick, MA, USA) by using mzcdfread.m and mzcdf2peak.m functions of the MATLAB Bioinformatics Toolbox (4.3.1 version).

ROI strategy[23] was used to compress the MS data without losing mass accuracy and to build the data matrices to be analyzed by the MCR-ALS method. This strategy allowed the selection of the most interesting mass traces, which means those *m/z* values whose intensity signals were higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and appeared a minimum number of times in the time direction. The required parameters for the implementation of the ROI approach were the SNR_{Thr} (set at 0.1% of the maximum MS signal intensity of each sample), the mass accuracy of the spectrometer (set at 0.5 Da/e for the TQD analyzer used in this work) and the minimum number of retention times to be considered as a chromatographic peak (set at 25). These ROI values are searched along the entire

chromatogram. Using this approach, the data matrix containing the intensities at all retention times (rows) for the selected number of ROI *m/z* values (columns) was finally obtained for each sample. More details about ROI strategy can be found at the work of Gorrochategui[23] and in Supplementary Information. A total number of 60 ROI matrices were obtained, one for each of the 30 samples of the presented work acquired at both ionization modes (positive and negative). Every ROI matrix was then normalized to correct for possible instrumental intensity changes among different sample injections and unavoidable differences in sample handling. This normalization was done by diving each matrix by the mean of the chromatographic area of the seven surrogates and the three internal standards added to the lipidomic extract of each sample.

Once ROI compression was performed, every full-scan LC×LC-MS chromatographic run was arranged in a two-way data structure, as shown in Figure 2A. For each second-dimension chromatographic modulation, an individual LC-MS data matrix ($\mathbf{D_k}$, see figure 2A) was built. These $\mathbf{D_k}$ matrices contained the second-dimension retention times on the rows and the *m/z* values on the columns. When *K* modulations were considered simultaneously, an LC×LC-MS column-wise augmented data matrix ($\mathbf{D_{aug}}$, see figure 2A) was built up settling the individual $\mathbf{D_k}$ matrices from each modulation one on the top of each other, and keeping their *m/z* values in common. Thus, $\mathbf{D_{aug}}$ matrix contained the retention times of all modulation (*K=1,…,72*) on the rows and the *m/z* values on the columns.

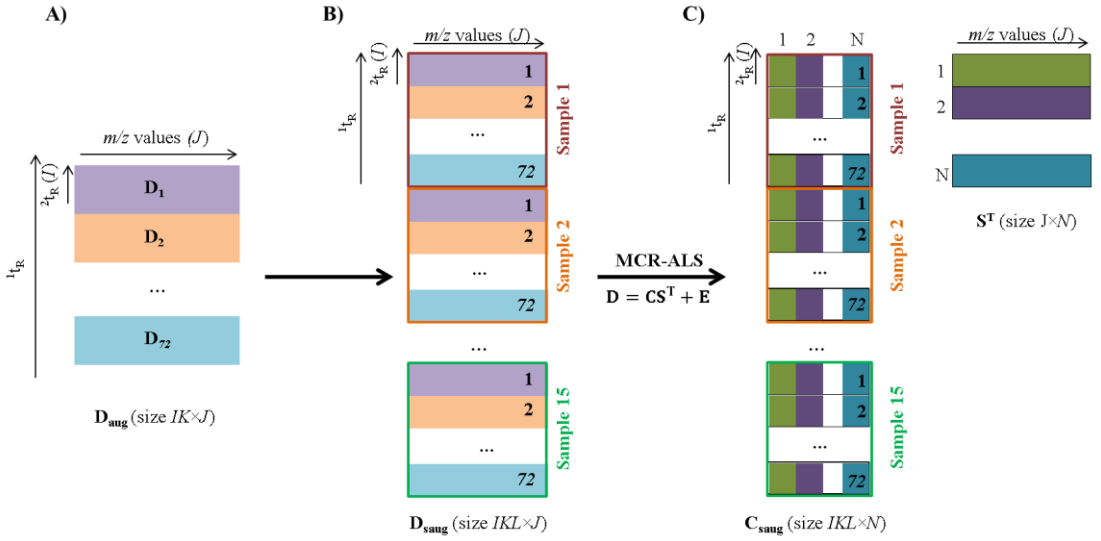


Figure 2. A) LC×LC-MS analysis of a single sample (run) arranged in a column-wise augmented data matrix ($\mathbf{D_{aug}}$). Every single $\mathbf{D_k}$ matrix corresponds to one LC-MS second column modulation. B) LC×LC-MS analysis of multiple samples (runs) arranged in a column-wise superaugmented data matrix ($\mathbf{D_{saug}}$), containing all the analysed samples (chromatographic runs) with all their corresponding second column modulations. C) MCR-ALS resolution of LC×LC-MS data. Matrix $\mathbf{D_{saug}}$ is decomposed into two matrices: $\mathbf{C_{saug}}$ which has the resolved pure elution profiles of the *N* components in all second column modulations of all different samples (chromatographic runs), and $\mathbf{S^T}$, which has the pure mass spectra of the corresponding resolved components.

After normalization, individual ROI matrices for each sample were arranged in four single superaugmented data matrices (Figure 2B, $\mathbf{D_{saug}}$): one for the aerial part samples analyzed in positive mode ($\mathbf{D_{AP}}$); one for the aerial part samples analyzed in negative mode ($\mathbf{D_{AN}}$); one for the root

samples analyzed in positive mode ($\mathbf{D_{RP}}$); and one for the root samples analyzed in negative mode ($\mathbf{D_{RN}}$). As individual data matrices had a different number of ROI *m/z* values; a preliminary search considering common and uncommon ROI *m/z* values among different data samples was performed. When in an individual compressed matrix a particular ROI *m/z* value did not exist, a low random intensity value at the noise level was assigned (below SNR_{Thr}). In this way, the final superaugmented data matrices had the same ROI *m/z* values for the 15 samples simultaneously analyzed. For more details about this ROI matrix augmentation strategy see the Supporting Information and the protocol described in the work of Gorrochategui *et al.*[23]. Next, every one of these four superaugmented data matrices was analyzed by the MCR-ALS method.

2.4.2. MCR-ALS resolution of LC×LC-MS data.

MCR-ALS is a chemometric method that allows the resolution of pure contributions present in unresolved complex mixtures. In this work, MCR-ALS was applied for the resolution of the elution profiles of the lipid constituents of the samples in both chromatographic dimensions and of their pure mass spectra. The MCR-ALS method has been already extensively described in the literature[24,25,29] and is only briefly explained here focusing on the particular case of untargeted multisample LC×LC-MS data.

MCR-ALS decomposes the experimental data sets according to a bilinear model extended to the simultaneous analysis of a large number of chromatographic runs (multiple second column modulations from several samples). For instance, in the case of this work, each of the four column-wise superaugmented LC×LC-MS data matrices ($\mathbf{D_{saug}}$, Figure 2B) had information related to the 15 single sample augmented data matrices ($\mathbf{D_{aug}}$, Figure 2A). $\mathbf{D_{saug}}$ matrices were decomposed by MCR-ALS method using a bilinear model as shown in Equation 1 and Figure 2C.

$$\mathbf{D_{saug}} = \begin{bmatrix} \mathbf{D_{11}} \\ \mathbf{D_{1,2}} \\ \vdots \\ \mathbf{D_{l,k}} \\ \vdots \\ \mathbf{D_{15,72}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{1,1}} \\ \mathbf{C_{1,2}} \\ \vdots \\ \mathbf{C_{l,k}} \\ \vdots \\ \mathbf{C_{15,72}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_{1,1}} \\ \mathbf{E_{1,2}} \\ \vdots \\ \mathbf{E_{l,k}} \\ \vdots \\ \mathbf{E_{15,72}} \end{bmatrix} =$$

$$\mathbf{C_{saug}S^T} + \mathbf{E_{saug}} \qquad \text{Equation (1)}$$

MCR-ALS decomposition of matrix $\mathbf{D_{saug}}$ (*IKL*×*J*) gave $\mathbf{C_{saug}}$ (*IKL*×*N*) and $\mathbf{S^T}$ (*N*×*J*). $\mathbf{C_{saug}}$ (*IKL*×*N*) contained second-dimension resolved elution profiles of the *N* components at all retention times (*I=59*) for each modulation (*K=72*) and sample (*L=15*). From this $\mathbf{C_{saug}}$ matrix, chromatographic peak areas of the resolved profiles of the different lipids and their relative quantitative information in the 15 analyzed samples can be obtained. On the other hand, $\mathbf{S^T}$ (*N*×*J*) contained the pure mass spectra of the resolved components.

More details regarding the initialization and constraints for the MCR-ALS optimization can be found in the Supplementary Information and references therein. MCR-ALS analyses were carried out using the MCR-ALS 2.0 toolbox available at www.mcrals.info.

2.4.3. Statistical assessment of As effects on rice lipidome

Chromatographic peak areas of the resolved lipids (*N*, in the columns) in all samples (*L*, in the rows) were arranged in a new data matrix (**A**). In total, four peak area matrices were obtained: aerial part samples analyzed in positive mode ($\mathbf{A_{AP}}$); aerial part samples analyzed in negative mode ($\mathbf{A_{AN}}$); root samples analyzed in positive mode ($\mathbf{A_{RP}}$); and root samples analyzed in negative mode ($\mathbf{A_{RN}}$). These peak area matrices were analyzed using PCA[26] and ASCA[27] to evaluate the effects of As exposure on rice.

PCA compresses the information of the original variables into a smaller number of uncorrelated variables known as principal components[26]. The

representation of these components both in samples (scores maps) and variables (loadings) modes are useful to explore and interpret the variance sources in the analysed data.

ASCA is an extension of the multivariate analysis of variance method that combines the power of ANOVA to separate variance sources with the advantages of simultaneous component analysis (SCA) for the modelling of the individual separate factor effects matrices. In this work, ASCA was applied to statistically assess the significance of As exposure by using a permutation test (1000 permutations). Experimental design allowed performing ASCA analysis to well-balanced peak area matrices (*i.e.* same number of samples for each analyzed condition). For a more detailed description of the ASCA procedure and permutation tests to assess the significance of factors see the works of Smilde[30], Jansen[27] and Vis[31].

Data was autoscaled before the application of PCA and only mean-centered before applying ASCA. Both methods were applied using PLS Toolbox 8.0.2 (Eigenvector Research Inc, Wenatchee, WA, USA) working under MATLAB 2015b.

2.4.4 Feature detection

After statistical evaluation of As effects on rice lipidome, PLS-DA[28] was used to detect what variables (lipids) were responsible for the observed differences between control and As-treated samples.

PLS-DA is a supervised multivariate regression method oriented to discriminate among different groups of samples. In this work, PLS-DA discriminated between control and As-treated samples. Here, PLS-DA was used to correlate the matrix of peak areas (**A**, predictor variable) with the vector describing the sample type class membership (**y**, predicted variable)[32]. Apart from sample class discrimination, PLS-DA also provides information about which are the most relevant variables for achieving this discrimination. For instance, variable importance on projection (VIP) scores can be used for this purpose[33]. VIP scores measure the importance of each predictor variable into the final PLS model and are calculated as the weighted sum of squared PLS variable weights. The "greater than one" rule is used as the criterion to identify the most important variables for a given model because the average of squared VIP scores is equal to 1[34].

Data were autoscaled prior the application of PLS-DA. This method was applied using PLS Toolbox 8.0.2 (Eigenvector Research Inc, Wenatchee, WA, USA).

2.4.5 Lipid identification

Finally, the selected VIP lipids were identified using their MS/MS spectra. These MS/MS spectra were obtained reinjecting rice samples and using as parent ion the most intense *m/z* signal for each previously detected component. These *m/z* signals were retrieved from the MCR-ALS resolved mass spectra corresponding to the VIP selected components.

In untargeted studies, there is no information about what daughter ions should be selected for the analysis. Therefore, optimization of CE was not possible. All MS/MS spectra were recorded at four different CE (10, 20, 30 and 40 eV).

Finally, experimentally obtained MS/MS spectra were compared with information available in public databases, such as Metlin[35]. The identification was considered satisfactory when at least two daughter ions of the experimental MS/MS spectra coincided with their *m/z* values and their relative abundances with the fragments of *in-silico* MS/MS spectra.

3. Results and Discussion

3.1. Resolution of LC×LC-MS lipidomic data

From the large number of m/z values acquired by the MS instrument, ROI strategy selected a relatively low number of *m/z* values for statistical analysis. After applying ROI augmentation, four LC×LC-MS augmented data matrices were obtained as described before: $\mathbf{D_{AP}}$, $\mathbf{D_{AN}}$, $\mathbf{D_{RP}}$ and $\mathbf{D_{RN}}$. Each one of these four column-wise augmented data matrices contained the LC×LC-MS analysis of the studied samples. These matrices had a final size of 63060 rows (total number of retention time channels) and 1505, 1502, 1482 and 1225 columns (number of finally selected *m/z* ROI values), respectively.

Figure 3 shows an example of an obtained LC×LC-MS ROI chromatogram for the control aerial part of one of the rice samples analyzed in positive mode. This figure reveals that despite the data compression achieved by selection of mass traces by the ROI procedure, LC×LC-MS chromatograms still presented rather complex profiles with multiple coeluted compounds. Consequently, the detection and identification of lipids cannot be performed directly. For this reason, the application of the MCR-ALS procedure was proposed to get a deeper insight into this experimental lipidomic data.

Quality assessment of the chromatographic runs was performed using QC samples (see Methods section). Instrumental conditions during the entire LC×LC-MS sequence were stable, and therefore, no further corrections were required.

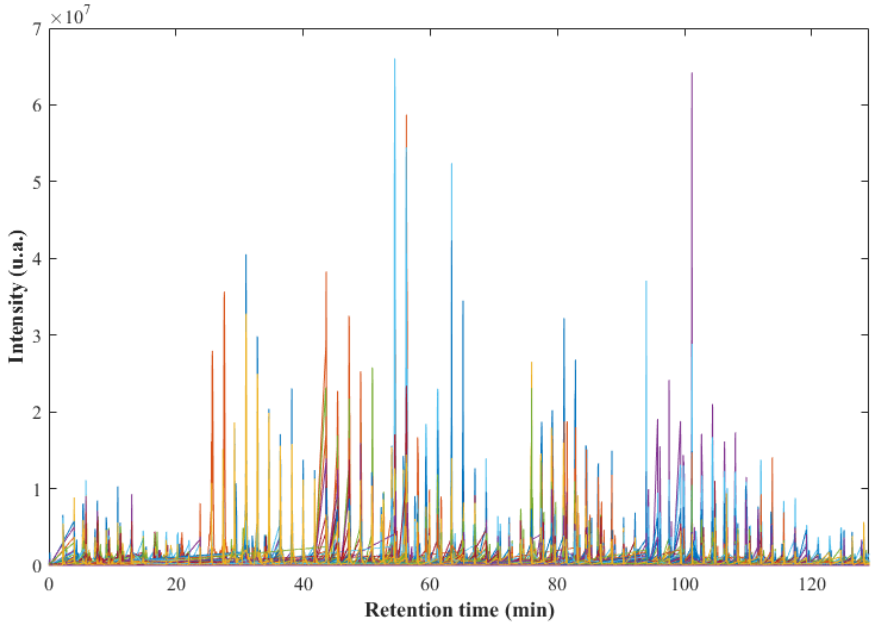


Figure 3. Example of a processed LC×LC-MS ROI chromatogram: control aerial part sample analyzed in positive mode.

In order to resolve the pure elution profiles of the lipids present in the analyzed sample in both chromatographic dimensions and their mass spectra, MCR-ALS was applied separately to the four data matrices obtained in the analysis of the aerial parts and roots of the rice samples, both in positive and negative ionization modes ($\mathbf{D_{AP}}$, $\mathbf{D_{AN}}$, $\mathbf{D_{RP}}$ and $\mathbf{D_{RN}}$). These four matrices were resolved by an MCR-ALS model using a total number of 250 components. This large number of resolved components included all detected lipid contributions as well as other noisy chromatographic signals, such as instrumental background and solvent contributions. In this case,

approximately 200 resolved components were finally assigned to individual lipids for each one of the analyzed matrices. The percentage of explained variance (R^2) and the lack of fit (LOF) were considered satisfactory. For instance, matrix $\mathbf{D_{AP}}$ was resolved with a total variance explained (R^2) of 97.3% and with a lack of fit (LOF) equal to 16.5%. Similar results were obtained for the other three matrices, which are given in Supplementary Information.

Figure 4 shows an example of MCR-ALS resolution of the pure elution and mass spectra profiles of five lipids in the aerial part of the rice samples analyzed in positive mode ($\mathbf{D_{AP}}$ matrix). Figure 4A shows the resolved elution profiles. The dashed lines represent the elution profiles in the first chromatographic dimension, which can be recovered by properly refolding the modulations in the second-dimension column. Each first-dimension peak was divided at least in four modulations, which are represented in Figure 4A with the solid lines. These modulations show the separation achieved at the second-dimension column. Inserts in Figure 4A depict zoomed views of the second-column modulations. Finally, Figure 4 B displays the mass spectra of each component.

In Figure 4, the three typical LC×LC-MS elution cases can be clearly appreciated: i) no coelution (yellow signals in Figure 4A); ii) coelution only in the first chromatographic dimension (red and blue signals in Figure 4A); iii) total coelution with embedded peaks (purple and green signals in Figure 4A). These three situations are detailed below.

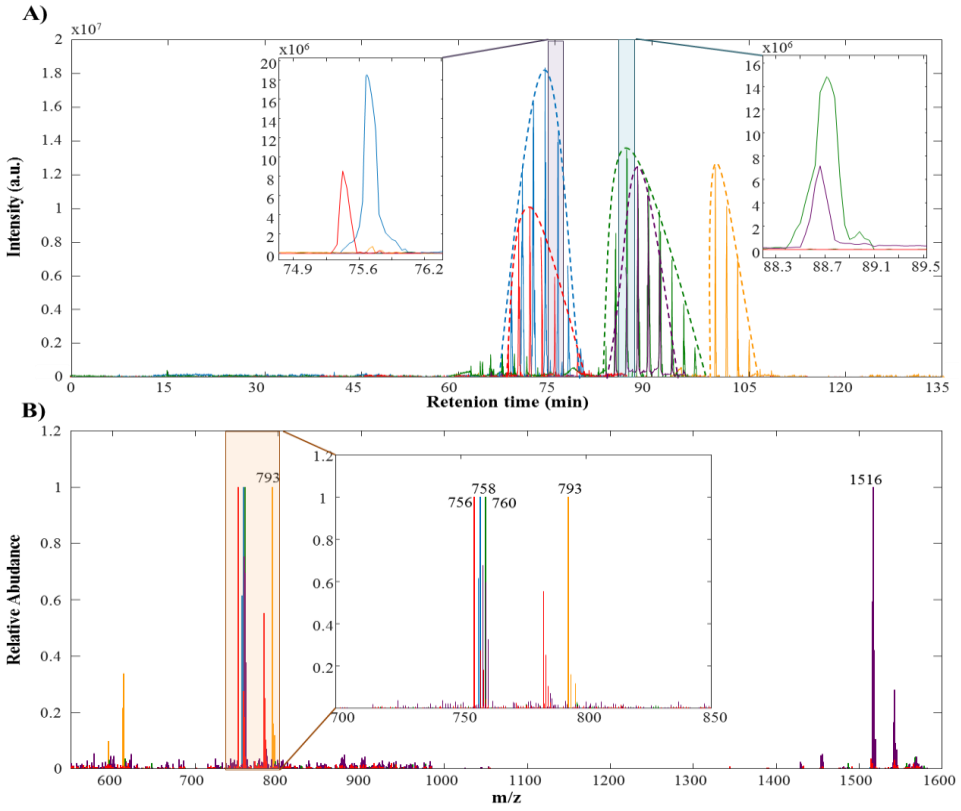


Figure 4. Example of MCR-ALS results of five lipids in aerial part samples analyzed in positive mode. A) Pure resolved elution profiles. Inserts show zoomed views of elution profiles between 74.9 and 76.2 minutes and between 93.3 and 94.1 minutes. B) Pure resolved mass spectra. Insert shows zoomed view of mass spectra between 700 and 850 *m/z*.

The no coelution case was the simplest one and could be easily resolved by traditional means. Signals colored in yellow in Figure 4A are an example of this case of no coelution. The four yellow chromatographic peaks are the four modulations in the second-dimension column of a lipid that eluted between 100 and 105 minutes of the first-dimension column. The resolved mass spectrum of this lipid is colored in yellow in Figure 4B. This spectrum allowed the selection of the ion at 793 *m/z* as the parent ion to be used for MS/MS analysis.

The two chromatographic peaks eluted at 75 minutes (blue and red signals in Figure 4A) were an example of coelution only in the first chromatographic dimension (dashed lines). Figure 4A shows that these two lipids coeluted in the first chromatographic column, both eluted between 70 and 80 minutes. However, as shown in the zoomed view, in the second chromatographic dimension the two lipids were separated. This example revealed one benefit of using LC×LC instead of LC, because their chromatographic separation was achieved on the second dimension. Resolved mass spectra of these lipids are colored in blue and red in Figure 4B. In this case, the selected parent ions were 756 and 758 *m/z*.

Finally, the two chromatographic peaks present at 90 minutes (purple and green signals in Figure 4A) were an example of the total coelution case. In Figure 4A it can be seen that these two lipids eluted between 85 and 95 minutes from the first chromatographic column. Moreover, the zoomed view shows that they were also coeluted in the second chromatographic column. This example exposed the large complexity of samples analyzed in lipidomic studies. Despite the high separation power of LC×LC, a total separation of all the lipids present in rice samples could not be achieved. In these cases, the use of chemometric tools such as MCR-ALS is mandatory. As shown in Figure 4, MCR-ALS was capable of resolving the pure elution and mass spectra profiles of these totally coeluted lipids. The resolved mass spectra of these lipids are colored in green and purple in Figure 4B. In this last case, the selected parent ions were 760 and 1516 *m/z*.

Figure 5 shows the comparison of the results obtained using the ROIMCR approach and by manual inspection of the detected signals considering the hardest case of total coelution. Figure 5A shows the raw extracted chromatograms for *m/z* values 760 (colored in green) and 1516 (colored in purple). In this figure, the chromatographic modulations for both *m/z* values appeared at two different retentions times, one at 70 minutes and the other at 90 minutes. On the contrary, the MCR-ALS resolved profiles for these mass traces (green and purple profiles in Figure 4A) only showed chromatographic modulations at 90 minutes. This result indicated that probably at 70 minutes eluted other lipids that have 760 and 1516 *m/z* values as minor signals on their mass spectra. One of the benefits of using MCR-ALS was that it gives the chromatographic signals for each resolved component (lipids). Moreover, Figure 5B shows the raw mass spectrum obtained between 85 and 95 minutes, which contains five intense mass signals: 663, 760, 785, 872 and 1516 *m/z*. This MS spectrum demonstrated that when the manual inspection is used, it was difficult to determine which of these mass traces belong to the same lipid. On the contrary, when ROIMCR approach was used it gave the pure mass spectra for each of the resolved lipids. In addition, it should be highlighted that a manual inspection of the detected signals is extremely time-consuming, because all intense mass traces detected at every retention time should be checked individually. On

the contrary, ROIMCR approach allowed a rapid resolution of all of them in the entire dataset.

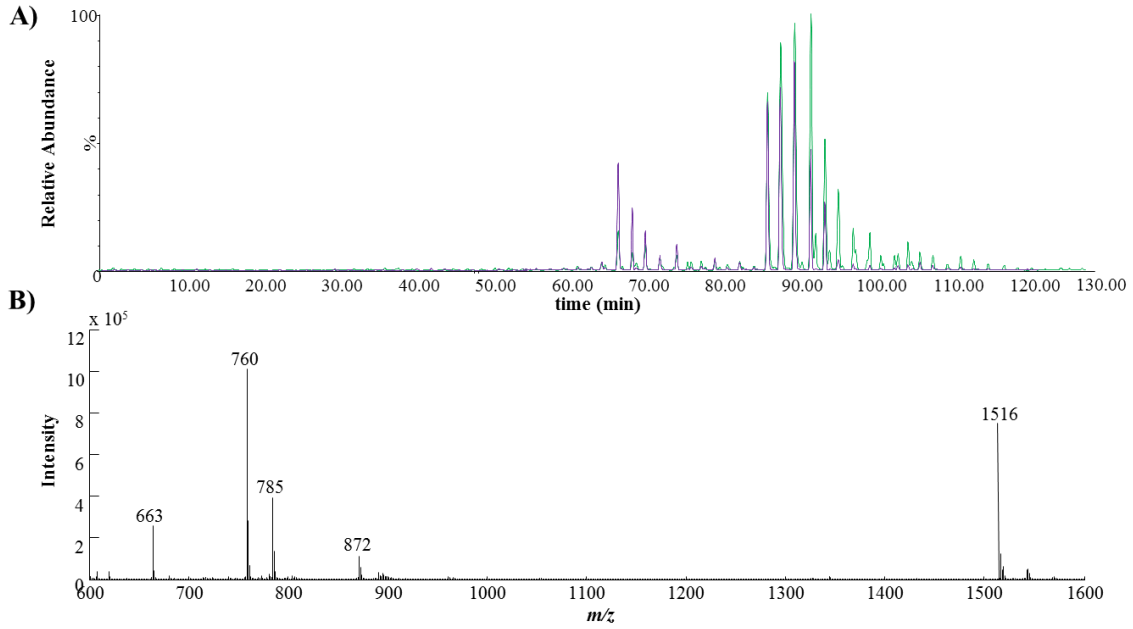


Figure 5. A) Raw extracted chromatograms for *m/z* values 760 (colored in green) and 1516 (colored in purple). B) Raw mass spectrum obtained between 85 and 95 minutes.

3.2. Statistical assessment of As effects on rice lipidome

The application of one-way ASCA model to every peak area data matrix ($\mathbf{A_{AP}}$, $\mathbf{A_{AN}}$, $\mathbf{A_{RP}}$, and $\mathbf{A_{RN}}$) revealed that in the four cases the effects produced by As exposure were significant with *p*-values between 0.003 and 0.001.

The application of PCA in the four cases also showed the effects of As exposure. In all cases, control samples were distinguished from samples exposed to 1000 µM As treatment using first and second principal components (PCs), which already explained more than the 30% of all data variance. For instance, Figure 6 shows PCA scores plots for aerial part (Figure 6A) and root (Figure 6B) samples analyzed in positive ionization mode. In the case of aerial part samples (Figure 6A) PC1 slightly separated control samples from 1µM As watered samples. In contrast, PC2 separates samples watered at 1000 µM. PCA scores plot for root samples (Figure 6B) differentiated samples exposed to 1000 µM As from the others also along PC2. However, in this case, control samples were not distinguished from samples exposed to 1 µM As exposure. These results indicated that As exposure at high doses (1000 µM) affected the rice lipidome, but these effects were low when considering a concentration under the limit accepted by European legislation (1 µM)[22].

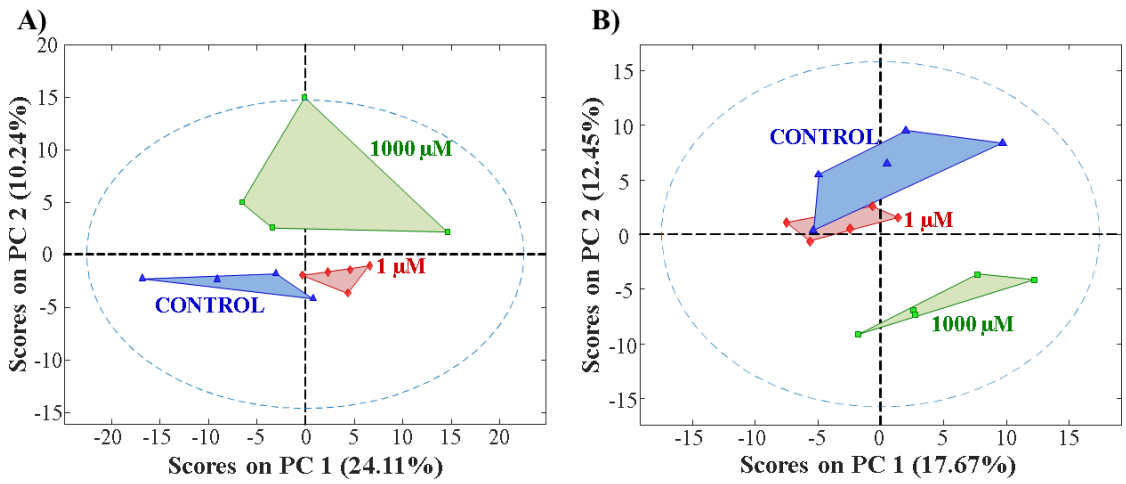


Figure 6. PCA scores plot obtained for A) aerial part samples and B) root samples analyzed in positive mode.

Results obtained for samples analyzed in negative mode are shown in Supplementary Information (Figure S3). From these results, it can be mentioned that samples treated with the lower concentration of As (1 µM) were only clearly distinguished from control samples in the PCA model of roots analyzed in negative mode (Figure S3B).

3.3. Feature Selection

The last step of the chemometric based-strategy for LC×LC-MS data analysis was the selection of the relevant lipids for sample discrimination. These important lipids were those that suffered a significant change under As exposure and, therefore, allowed the differentiation between control and treated samples. With this purpose, PLS-DA was applied to the four peak area matrices ($\mathbf{A_{AP}}$, $\mathbf{A_{AN}}$, $\mathbf{A_{RP}}$, and $\mathbf{A_{RN}}$), but only taking into account control and high dose As (1000 µM) treated samples.

The four obtained PLS-DA models distinguished control from treated samples. As an example, Figure S4 in Supplementary Information shows the PLS-DA results for root samples analyzed in negative mode. Figure S4A represents the cross-validation (CV) class predictions and shows that all samples from both classes (control and treated) were perfectly discriminated. Figure S4B shows the VIP scores plot for the mentioned PLS-DA model. The variables (resolved lipids) with a VIP value greater than one were considered important to discriminate among As exposure factor levels. In the case of the example in Figure S4B (roots analyzed in negative mode), a total of 74 lipids were selected. For roots analyzed in the positive mode the number of selected variables was 54. Finally, in the case of aerial part samples, the number of selected lipids was 70 and 77 for samples analyzed in positive and negative modes, respectively.

3.3. Lipids identification

In order to identify the lipids whose concentration changed under As exposure, samples were reanalyzed by MS/MS. As mentioned above, lipids were identified by comparison of their experimental MS/MS spectra, (recorded at 10, 20, 30 and 40 eV CE) with *in-silico* theoretical spectra available in the METLIN database[35].

Figure 7 shows an example of this MS/MS identification. The upper part of this figure represents the experimental MS/MS spectra at 20 eV CE of the lipid eluted at 100 minutes in Figure 4A (yellow signals in Figure 4). The *m/z* value of the parent ion obtained from the MCR-ALS resolved mass spectrum was 793 (see Figure 4B). This experimental MS/MS spectrum showed three

major product ions at 261, 335 and 613 *m/z*, which were well correlated with the theoretical product ions of MGDG (36:6) (METLIN ID 75584).

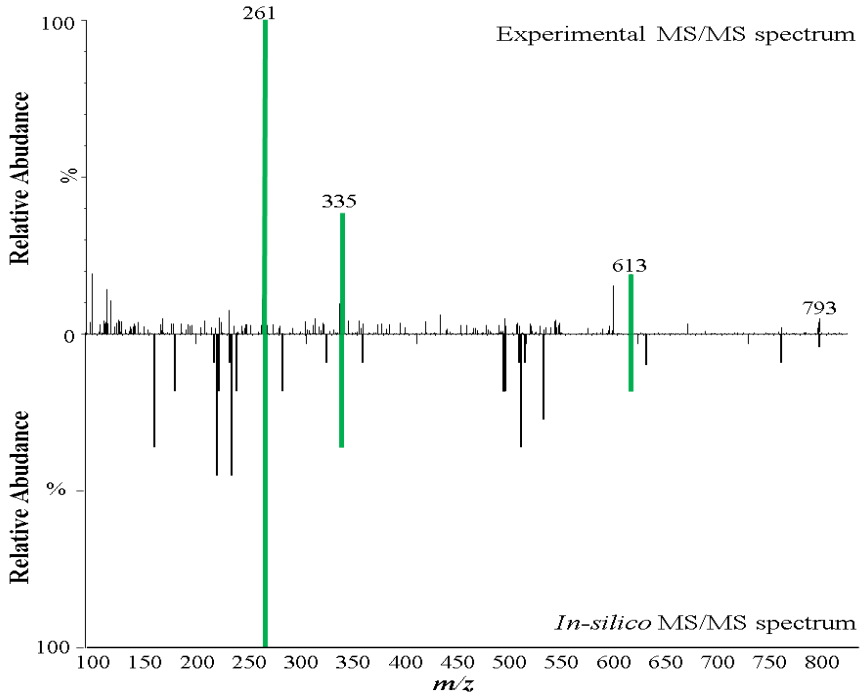


Figure 7. Identification of MCR-ALS resolved component number 5 for aerial part samples analyzed in positive mode as MGDG (36:6). The *m/z* value of the parent ion was 793. The experimental MS/MS spectra for this ion obtained at 20 eV CE (up signals) could be associated with the *in-silico* mass spectra of MGDG (36:6) obtained from Metlin database (down signals, METLIN ID 75584). The major product ions signals are colored in green.

Since the main goal of this work was not to perform a detailed biological interpretation of the observed lipid changes, only the identification of some of the lipids was attempted to confirm the reliability of the proposed methodology. Results of this identification are shown in Table 1, with the *m/z* values of parent ions, the name of the identified lipids, their elemental composition and METLIN ID, the product ions that allowed their identification and the CE of the MS/MS spectrum used for the identification.

As an example of the possibilities of the presented methodology, four of the five lipids shown in Figure 4 were finally identified. The lipid eluted at 100 minutes was identified Figure 7 as MGDG (36:6). The two lipids coleuting in the first-dimension column, with parent ions at 756 (red signal) and 758 (blue signal) *m/z* values were identified as PC(34:3) and PC(P-34:3), respectively. The MS/MS spectra of the first one showed two product ions at 184 and 237 *m/z* that could be correlated to the theoretical spectra. In the second case, the two product ions used for identifying the lipids were at 184 and 223 *m/z*. Finally, only one of the two totally coeluted lipids could be completely identified. The MS/MS spectrum of the one with the parent ion at 760 *m/z* (green signal) gave two major product ions at 184 and 263 *m/z*, which could be associated with the theoretical product ions of PC(P-34:2). The other one (purple signals in Figure 4), could be tentatively identified as PC(34:2) in agreement with the mass of the parent ion and its retention time. The mass of the parent ion was 1516 Da could be assigned to the $[2M+H]^+$ adduct of PC(34:2). This compound probably coelutes with

PC (P-34:2) (green signals in Figure 4A). Unfortunately, the obtained MS/MS spectra for this parent ion could not confirm this identification.

Table 1. Summary of results obtained for lipid identification.

	Parent Ion	Lipid name	Elemental composition	METLIN ID	Product Ions	CE (eV)
AERIAL PARTS ESI (+)	587	DAG (34:5)	C37H62O5	58746	209/231	40
	756	PC(34:3)	C42H78NO8P	39422	184/237	20
	758	PC(P-34:3)	C42H78NO7P	62042	184/223	20
	760	PC(P-34:2)	C42H80NO7P	59605	184/263	20
	786	PC(P-36:3)	C44H82NO7P	59607	184/263	20
	793	MGDG(36:6)	C45H74O10	75584	261/335/613	20
	796	TAG(48:6)	C51H86O6	98532	261/35	20
	815	PI(O-32:0)	C41H81O12P	81067	181/225	20
	873	TAG(52:4)	C55H98O6	4838	239/593	10
	875	PI(36:5)	C45H77O13P	80427	575/839	20
AERIAL PARTS ESI (-)	745	PA(36:2)	C39H73O8P	81923	375/470	20
	748	PG(34:1)	C40H77O10P	61862	265/493	20
	820	MGDG(36:6)	C45H74O10	75584	495/755	10
	837	PI(P-34:4)	C43H77O12P	80995	95/229	40
	841	PC(P-34:3)	C42H78NO7P	62042	182/259	20
ROOTS ESI (+)	464	Linolenyl laurate	C30H54O2	97081	183/247/447	20
	573	DAG(P-32:1)	C35H66O4	4697	113/265	20
	613	DAG(34:3)	C37H66O5	4319	237/263	20
	778	PC(36:6)	C44H76NO8P	39691	184/579	40
	781	PC(P-36:4)	C44H80NO7P	59640	184/567	40
	783	PC(36:4)	C44H80NO8P	59649	184/319	20
ROOTS ESI (-)	743	PG(34:3)	C40H73O10P	61852	237/259	20
	804	PC(34:1)	C42H82NO8P	39326	238/758	10
	815	PC(P-36:2)	C44H84NO7P	59543	263/768	10
	821	PC(36:3)	C44H82NO8P	39511	217/231	40
	962	Glycerol 2-(9Z,12Z-octadecadienoate) 1-hexadecanoate 3-O-[alpha-D-galactopyranosyl-(1->6)-beta-D-galactopyranoside]	C49H88O15	95895	635/897	20

Conclusions

A chemometrics-based data analysis strategy is proposed to gather all relevant information from untargeted lipidomic LC×LC-MS datasets. Despite the high complexity of untargeted LC×LC-MS datasets, the ROIMCR compression and resolution strategy allowed the determination of a large number of lipids and of the changes in their concentration in one single analysis from the LC×LC-MS data from rice samples exposed to As.

The main advantage of the proposed methodology is that it achieves a satisfactory resolution of complex lipidomics samples. In comparison with LC-MS based systems and with LC×LC-MS analysis followed by manual inspection of the detected signals, a higher number of lipids could be resolved. However, the difficulty in the identification of lipids is a relevant drawback. On the one hand, there is still a lack of theoretical lipid MS/MS spectra available in public databases. On the other hand, some of the obtained MS/MS spectra did not show daughter ions with intense enough *m/z* signals. This may be related to a poor detection sensitivity resulting from the dilution caused by two successive chromatographic steps or to the impossibility of optimizing CE. Potential solutions to overcome the detection-sensitivity limitations should be considered. For instance, the use of an active modulation as indicated in the recent work by Gargano [36]. Moreover, after the untargeted detection of the potential lipid biomarkers, the confirmation of these candidates may be performed using a targeted approach with better sensitivity and, consequently, provide a higher number of product ions after, for instance, the optimization of CE. Also, the combination with high resolution mass spectrometry (HRMS) would improve the detection-sensitivity and reduce the number of candidates to identify allowing a prior tentative identification by exact mass.

Considering the rice lipidomic study performed in this work, results showed that As exposure had significant effects on rice lipidome (specially at the high dose). However, the irrigation of rice plants with water that contains As at a concentration accepted by the European legislation did no show significant effects on rice lipidome.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement Number 320737.

References

(1) Lam, S. M.; Shui, G. *Journal of Genetics and Genomics* **2013**, *40*, 375-390.
(2) Sethi, S.; Brietzke, E. *Prostaglandins & Other Lipid Mediators* **2017**, *128-129*, 8-16.
(3) Checa, A.; Bedia, C.; Jaumot, J. *Analytica Chimica Acta* **2015**, *885*, 1-16.
(4) Fenaille, F.; Barbier Saint-Hilaire, P.; Rousseau, K.; Junot, C. *Journal of Chromatography A* **2017**, *1526*, 1-12.
(5) Li, M.; Zhou, Z.; Nie, H.; Bai, Y.; Liu, H. *Analytical and Bioanalytical Chemistry* **2011**, *399*, 243-249.
(6) Wolf, C.; Quinn, P. J. *Progress in lipid research* **2008**, *47*, 15-36.
(7) Navas-Iglesias, N.; Carrasco-Pancorbo, A.; Cuadros-Rodríguez, L. *TrAC Trends in Analytical Chemistry* **2009**, *28*, 393-403.
(8) Cajka, T.; Fiehn, O. *TrAC Trends in Analytical Chemistry* **2014**, *61*, 192-206.
(9) Vinayavekhin, N.; Saghatelian, A. *Current protocols in molecular biology* **2010**, *Chapter 30*, Unit 30.31.31-24.
(10) Fahy, E.; Subramaniam, S.; Murphy, R. C.; Nishijima, M.; Raetz, C. R. H.; Shimizu, T.; Spener, F.; van Meer, G.; Wakelam, M. J. O.; Dennis, E. A. *Journal of Lipid Research* **2009**, *50*, S9-S14.
(11) Baglai, A.; Gargano, A. F. G.; Jordens, J.; Mengerink, Y.; Honing, M.; van der Wal, S.;

Schoenmakers, P. J. *Journal of Chromatography A* **2017**, *1530*, 90-103.
(12) Stoll, D. R.; Carr, P. W. *Analytical Chemistry* **2017**, *89*, 519-531.
(13) Dugo, P.; Cacciola, F.; Kumm, T.; Dugo, G.; Mondello, L. *Journal of Chromatography A* **2008**, *1184*, 353-368.
(14) Carr, P. W.; Davis, J. M.; Rutan, S. C.; Stoll, D. R. *Advances in chromatography* **2012**, *50*, 139-235.
(15) Porter, S. E.; Stoll, D. R.; Rutan, S. C.; Carr, P. W.; Cohen, J. D. *Anal Chem* **2006**, *78*, 5559-5569.
(16) Navarro-Reig, M.; Jaumot, J.; Baglai, A.; Vivó-Truyols, G.; Schoenmakers, P. J.; Tauler, R. *Analytical Chemistry* **2017**, *89*, 7675-7683.
(17) Sinanian, M. M.; Cook, D. W.; Rutan, S. C.; Wijesinghe, D. S. *Anal. Chem.* **2016**, *88*, 11092-11099.
(18) Tistaert, C.; Bailey, H. P.; Allen, R. C.; Vander Heyden, Y.; Rutan, S. C. *J. Chemom.* **2012**, *26*, 474-486.
(19) Cook, D. W.; Rutan, S. C.; Stoll, D. R.; Carr, P. W. *Anal. Chim. Acta.* **2015**, *859*, 87-95.
(20) Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. *Journal of Lipid Research* **2008**, *49*, 1137-1146.
(21) Navarro-Reig, M.; Jaumot, J.; Piña, B.; Moyano, E.; Galceran, M. T.; Tauler, R. *Metallomics* **2017**, *9*, 660-675.
(22) 2006/118/EC, D., 2006.
(23) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC - Trends in Analytical Chemistry* **2016**, *82*, 425-442.
(24) De Juan, A.; Jaumot, J.; Tauler, R. *Anal. Chim. Acta* **2014**, *6*, 4964-4976.
(25) Jaumot, J.; de Juan, A.; Tauler, R. *Chemometrics Intell. Lab. Syst.* **2015**, *140*, 1-12.
(26) Wold, S.; Esbensen, K.; Geladi, P. *Chemometrics Intell. Lab. Syst.* **1987**, *2*, 37-52.
(27) Jansen, J. J.; Hoefsloot, H. C. J.; Van Der Greef, J.; Timmerman, M. E.; Westerhuis, J. A.; Smilde, A. K. *J. Chemometr.* **2005**, *19*, 469-481.
(28) Barker, M.; Rayens, W. *J. Chemometr.* **2003**, *17*, 166-173.
(29) Ruckebusch, C.; Blanchet, L. *Anal. Chim. Acta* **2013**, *765*, 28-36.
(30) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R. J. A. N.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *21*, 3043-3048.
(31) Vis, D. J.; Westerhuis, J. A.; Smilde, A. K.; van der Greef, J. *BMC Bioinformatics* **2007**, *8*.
(32) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
(33) Wold, S.; Johansson, A.; Cocchi, M. In *3D QSAR in Drug Design*, Kubiny, H., Ed.; ESCOM Science Publishers: Leiden, 1993, pp 583-618.
(34) Chong, I. G.; Jun, C. H. *Chemometrics Intell. Lab. Syst.* **2005**, *78*, 103-112.
(35) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nature Biotechnology* **2012**, *30*, 826.
(36) Gargano, A. F. G.; Duffin, M.; Navarro, P.; Schoenmakers, P. J. *Analytical Chemistry* **2016**, *88*, 1785-1793.

Informació Suplementària a la Publicació 7

An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis..

M. Navarro-Reig, J. Jaumot, R. Tauler

Enviat

Conditions of the environmental test chamber MLE-352H.

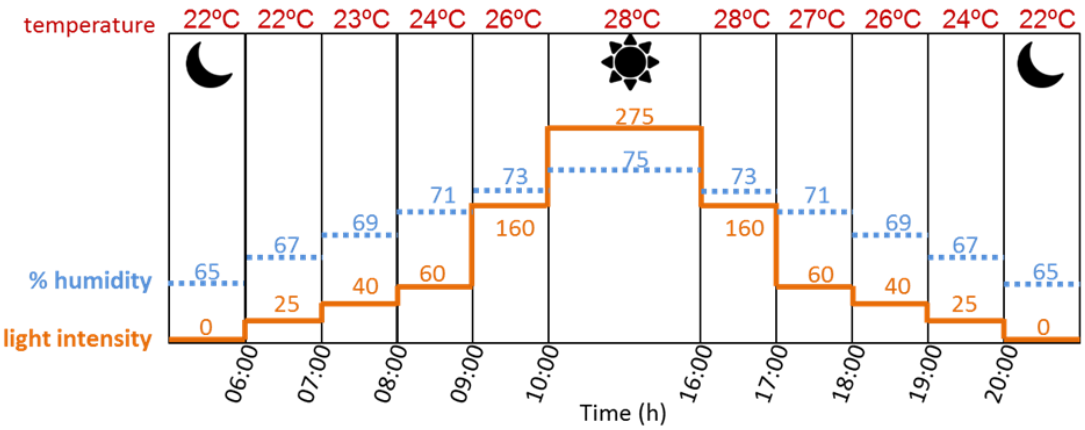


Figure S1. Experimental temperature, relative humidity and light long-day rice cultivation conditions at the growth chamber.

Chemometric tools: ROIMCR

Regions of interest (ROI) strategy

ROI strategy allows for the selection of mass traces, which means *m/z* values whose intensity signals are higher than a fixed signal-to-noise ratio threshold (SNR_{Thr}) and, also, with a number of occurrences that allows the proper definition of a chromatographic peak.

Figure S2 shows the general scheme of ROI strategy. First (step 1), for every retention in a chromatogram the ROI *m/z* values are searched. This search generates a couple of vectors containing the found *m/z* values and their related intensities for every retention time. Logically, as different compounds are detected at different retention times, these found *m/z* values are different among retention times. Therefore, the obtained vectors at different retention times have distinct lengths (depending on the number of ROIs found at each retention time). Finally, these vectors are reorganized into a matrix grouping the common ROIs among all the retention times. The final *m/z* values of each ROI are calculated as the mean of the *m/z* values obtained for that specific ROI (step 2). In the case that a particular ROI is not present at a retention time, the intensity for this ROI at this retention time is set to a low random intensity value at the noise level. This approach allows preserving all the information detected in both common and uncommon ROIs in the finally obtained matrix for each sample.

For the lipidomics study, the whole set of samples should be simultaneously considered. Consequently, a superaugmented data matrix has to be build up from the previous 15 individual compressed matrices obtained by the ROI search described above. Every compressed matrix was then normalized to correct the instrumental intensity drifts among injections. Normalization was done by dividing each matrix by the mean chromatographic area of the internal standards and surrogates added to the lipidome extract of that sample. After normalization, individual compressed matrices were arranged in a column-wise augmented data matrix.

Since compressed individual data matrices have a different number of ROI *m/z* values (each sample has a different number of columns/ROI *m/z* values, left side of step 3), a search of ROIs among different samples is performed. Again, common and uncommon ROI *m/z* values are evaluated considering both of them. This strategy allows reducing the number of ROI *m/z* values by grouping those with an *m/z* difference below the mass error tolerance. When an individual compressed matrix has not a significant intensity at a particular ROI *m/z* value, low random intensity values at the noise level were assigned. In this work, the strategy for column-wise augmentation consists of the independent augmentation of control and treated samples. More details about ROI strategy can be found at the works of Gorrochategui *et. al*. [1, 2].

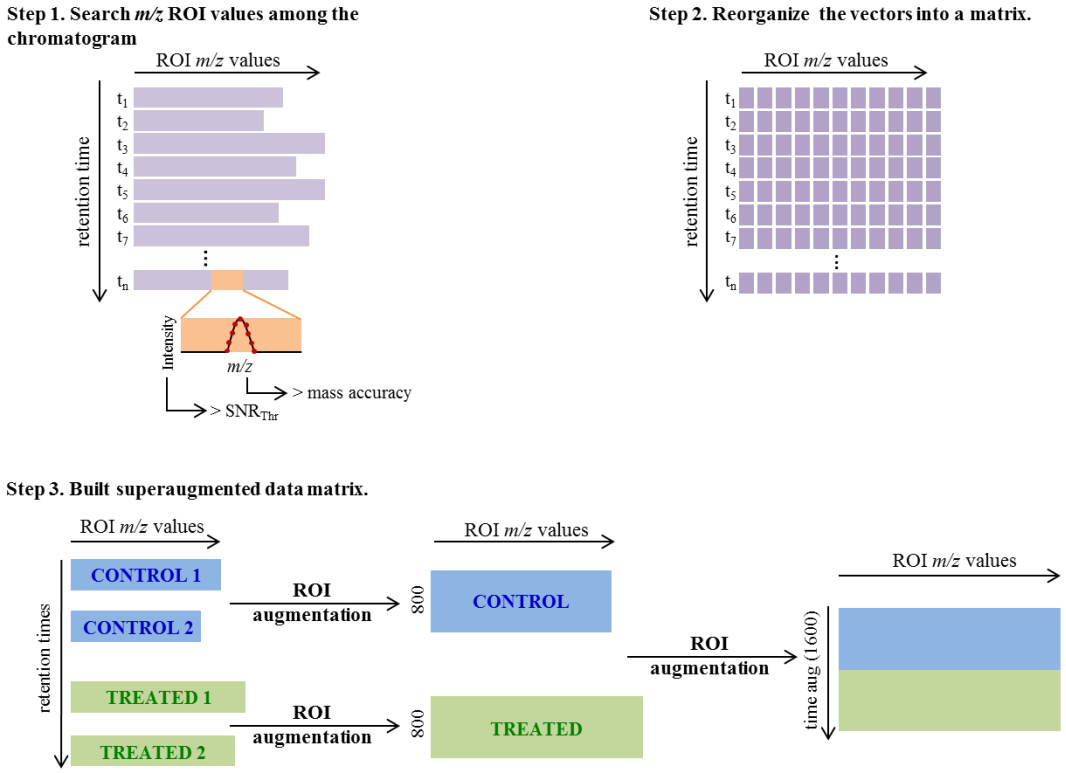


Figure S2. General Scheme of ROI strategy. Step 1: ROI *m/z* values are searched for every retention in a chromatogram. Step 2: Reorganization of the vectors into a matrix grouping the common ROIs among all the retention times. Step 3: Building of the superaugmented data matrix.

MCR-ALS for feature detection

MCR-ALS decomposes experimental data sets arranged in a data matrix according to the following bilinear model:

$$\mathbf{D}_K = \mathbf{C}_K\mathbf{S}^{\mathbf{T}} + \mathbf{E}_K \qquad \text{Equation (1)}$$

I the case of this work, $\mathbf{D_K}$ (size $I \times J$) is the experimental LC-MS matrix corresponding to one of the second dimension column modulation taken from the first dimension column. Rows of this matrix are the MS spectra at all second dimension retention times, and columns are the second dimension chromatograms at all *m/z* ROI values. $\mathbf{C_K}$ (size $I \times N$) is the matrix containing the resolved second dimension elution profiles for this modulation and $\mathbf{S^T}$ (size $N \times J$) is the matrix containing their corresponding mass spectra. N represents the number of resolved components using the MCR-ALS method. Finally, $\mathbf{E_K}$ (size $I \times J$) is the matrix of the residuals not explained by the MCR model.

This data analysis strategy can be easily extended to the simultaneous analysis of several chromatographic runs (several second dimension column modulations and several samples). The same bilinear model used in Equation 1 can be extended as follows:

$$\mathbf{D_{saug}} = \begin{bmatrix} \mathbf{D_{1,1}} \\ \mathbf{D_{1,2}} \\ \vdots \\ \mathbf{D_{l,k}} \\ \vdots \\ \mathbf{D_{15,72}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{1,1}} \\ \mathbf{C_{1,2}} \\ \vdots \\ \mathbf{C_{l,k}} \\ \vdots \\ \mathbf{C_{15,72}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_{1,1}} \\ \mathbf{E_{1,2}} \\ \vdots \\ \mathbf{E_{l,k}} \\ \vdots \\ \mathbf{E_{15,72}} \end{bmatrix} = \mathbf{C_{saug}S^T} + \mathbf{E_{saug}} \qquad \text{Equation (2)}$$

Where $\mathbf{D_{saug}}$ is the column-wise superaugmented data matrix build using the previously described ROI augmentation strategy. Since *L* is the number of analyzed samples and *K* is the number of the second dimension modulations taken from the first column, the number of rows for $\mathbf{D_{saug}}$ is equal to $I \times L \times K$. Decomposition of matrix $\mathbf{D_{saug}}$ generates $\mathbf{C_{saug}}$ (size $ILK \times N$) which has the second dimension resolved elution profiles at each modulation (*K*) and at each sample (*L*) for the *N* components. In addition, $\mathbf{S^T}$ (size $N \times J$) contains the mass spectra for the resolved *N* components. According to Equation 2, resolved mass spectra ($\mathbf{S^T}$) were forced to be the same for the common components (lipids) in the different modulations from all the analyzed rice samples. However, elution profiles of the same component resolved in the column-wise augmented concentration matrix $\mathbf{C_{saug}}$ were allowed to be different for each one of the chromatographic runs. Fist dimension elution profiles for every component in each sample can be obtained by refolding appropriately every column in $\mathbf{C_{saug}}$ to give a matrix of dimensions ($I \times K$) for each sample. Every column of the refolded matrix gave the first dimension elution profile size ($1 \times K$). Therefore, a matrix of the first dimension elution profiles (size $N \times K$)) was obtained for each sample.

The percentage of explained variance (R^2, equation 3) and the lack of fit (LOF, equation 4) were used as figures of merit for the MCR-ALS resolution.

$$R^2 = \left(\frac{\sum_{i,j} d_{i,j}^2 - \sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}\right) \times 100 \qquad \text{Equació (3)}$$

$$\text{lof}\,\% = 100 \times \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \qquad \text{Equació (4)}$$

Where di,j is the element of the original matrix at row *i* and column *j*, $e_{i,j}$ is the related residual, which is obtained from the difference between $d_{i,j}$ and its analogous reproduced with the MCR-ALS model ($\hat{d}_{i,j}$).

An initial guess of the number of components of Dsaug matrix was obtained using singular value decomposition (SVD) algorithm [3]. This number was only an initial approximation, as the final number of components was decided by taking into account data fitting results and the reliability of the resolved profiles. Initial estimates of pure spectra (ST) for the iterative optimization should be provided, which were computed using a purest spectra detection method based on the SIMPLISMA approach [4, 5]. Finally, ALS optimization was carried out applying non-negativity (for chromatographic elution and spectra profiles of every component) and spectral normalization (equal height). This application of constraints provided chemical meaning to the pure mathematical solution.

Chemometric analysis.

1. MCR-ALS resolution

In the case of matric $\mathbf{D_{AN}}$, the obtained MCR-ALS model had R^2 equal to 96.5% and LOF equal to 18.6%. $\mathbf{D_{RP}}$ was resolved with R^2 of 96.8% and LOF equal to 18.4%. Finally, in the case of matrix $\mathbf{D_{RN}}$, R^2 was 95.8% and LOF was equal to 19.2%.

2. Statistical assessment of As effects on rice lipidome

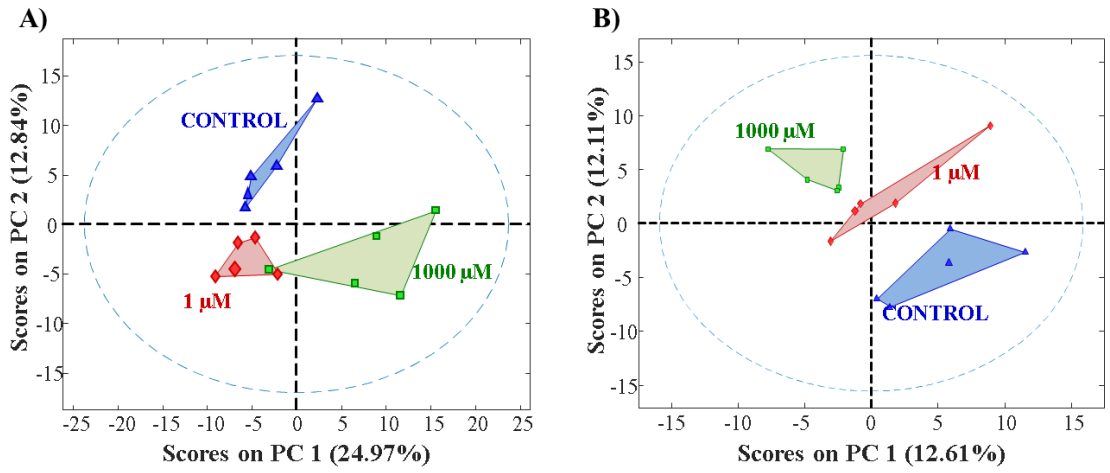


Figure S3. PCA scores plot obtained for A) aerial part samples and B) root samples analyzed in negative mode.

3. Feature Selection

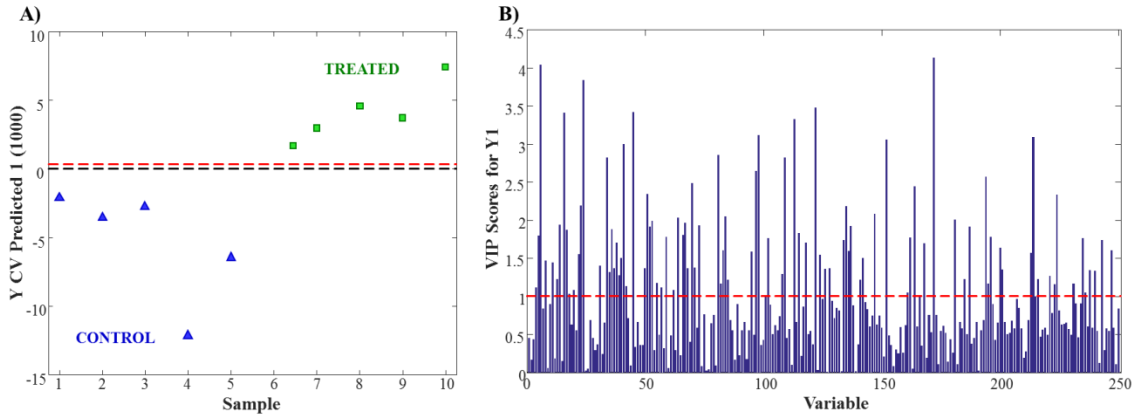


Figure S4. PLS-DA results for root samples analyzed in negative mode. A) Scores plot showing the prediction of control and treated samples. B) VIP scores plot.

References.

1. Gorrochategui, E.;Jaumot, J.;Lacorte, S.;Tauler, R., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC, Trends Anal. Chem.*, 2016, **82**, 425-442.
2. Gorrochategui, E.;Jaumot, J.;Tauler, R., A protocol for LC-MS metabolomic data processing using chemometric tools, 2015.
3. Golub, G. H.;Loan, C. F. V., *Matrix computations*, Johns Hopkins University Press, Baltimore, third edn., 1996.
4. Windig, W.;Guilment, J., Interactive self-modeling mixture analysis, *Anal. Chem.*, 1991, **63**, 1425-1432.
5. Windig, W.;Stephenson, D. A., Self-modeling mixture analysis of second-derivative near-infrared spectral data using the simplisma approach, *Anal. Chem.*, 1992, **64**, 2735-2742.

5.5. Discussió conjunta dels resultats

A continuació es descriuen els resultats obtinguts en aquest capítol en referència a l'estudi de l'estructura de les dades de LC×LC-MS i la seva modelització, així com les estratègies de compressió utilitzades. Finalment, es presenten els resultats obtinguts en les aplicacions que s'han fet en aquesta Tesi emprant LC×LC-MS.

5.5.1. Estructura i modelització de les dades LC×LC-MS

En les publicacions 5 i 6 es descriu l'estructura de les dades de LC×LC-MS i s'expliquen les dues possibles formes d'ordenar aquestes dades quan es treballa amb una sola mostra (publicació 5) o amb més d'una mostra (publicació 6).

En la Figura 5.2A es representen aquests dos possibles arranjaments de les dades de LC×LC-MS quan s'analitza una única mostra, en funció de com s'ordenen les matrius de LC-MS de cada modulació (fraccions provinents de la primera columna recollides pel modulador). Per cada modulació s'obté una separació cromatogràfica individual a la segona columna, i per tant, s'obté una matriu de dades de LC-MS per cadascuna de les modulacions (matrius $\mathbf{B_m}$ en la Figura 5.2A). Les files d'aquesta matriu de dades d'una sola modulació contenen els espectres de masses a cada temps de retenció de la segona dimensió cromatogràfica, mentre que les columnes contenen els cromatogrames de la segona dimensió per cada valor de *m/z*. Quan les matrius $\mathbf{B_m}$ es col·loquen de forma paral·lela, una darrera de l'altre, mantenint en comú els valors de *m/z* i els temps de retenció en la segona columna (^2t_R), les dades de LC×LC-MS s'ordenen en una estructura en forma de cub (cub $\underline{\mathbf{B}}$ en la Figura 5.2A). En l'eix *x* (files de cada matriu de dades) d'aquest cub es troben els diferents espectres de masses, mentre que sobre els eixos *y* (columnes de cada matriu de dades) i *z* (diferents matrius de dades) es troben les dades de retenció de les dues dimensions cromatogràfiques respectivament (primera i segona columna). En canvi, un altre possible arranjament de les dades s'obté quan les matrius $\mathbf{B_m}$ es col·loquen una sota de l'altra, mantenint els seus valors de *m/z* en comú entre elles. En aquest cas, les dades de LC×LC-MS per a una mostra tenen una estructura en forma de matriu augmentada en la direcció de les columnes (matriu $\mathbf{B_{aug}}$ en la Figura 5.2A). Les files d'aquesta matriu augmentada contenen els espectres de masses per cada combinació de les dues dimensions cromatogràfiques i les seves columnes contenen els cromatogrames bidimensionals desplegats un darrere l'altre, agrupats successivament segons les diferents modulacions a la segona dimensió per cada canal de *m/z*.

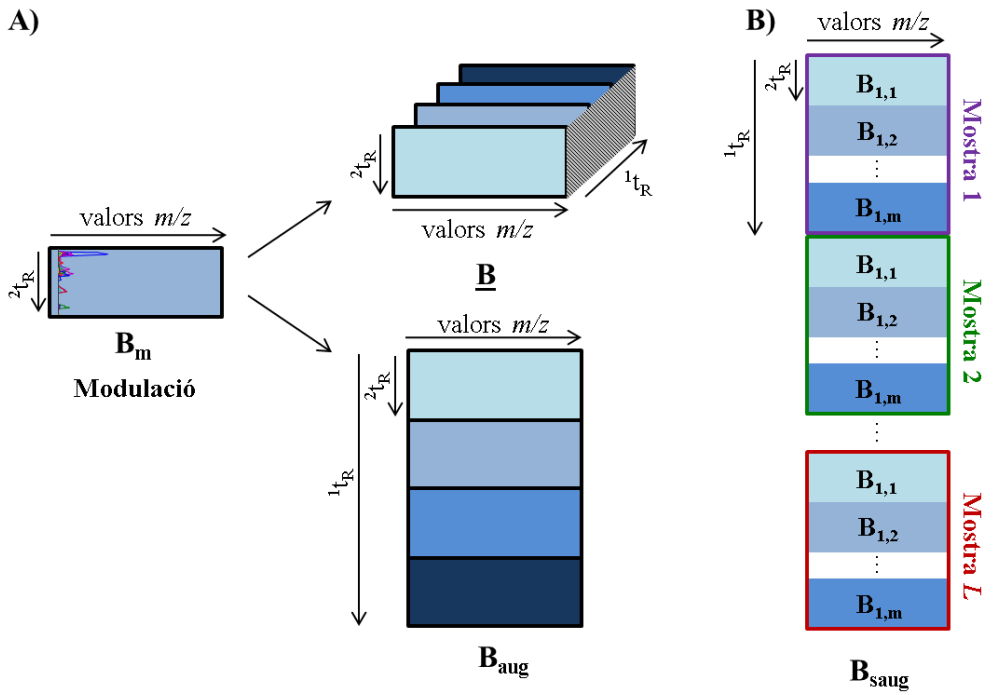


Figura 5.2. Descripció gràfica dels arranjaments de dades LC×LC-MS. Les matrius $\mathbf{B_m}$ són les matrius LC-MS que contenen la separació en la segona columna cromatogràfica per cada modulació provinent de la primera columna cromatogràfica. A) Una única mostra ordenada en forma de cub (**B**) i en forma de matriu augmentada en la direcció de les columnes ($\mathbf{B_{aug}}$). B) Matriu superaugmentada de dades LC×LC-MS per un conjunt de *L* mostres ($\mathbf{B_{saug}}$).

En la publicació 6 es mostra que quan s'analitza simultàniament un conjunt de diverses mostres, el conjunt de dades sencer es pot representar també per un conjunt de cubs de dades o per una matriu superaugmentada en la direcció de les columnes (matriu $\mathbf{B_{saug}}$ en la Figura 5.2B). Aquesta matriu es construeix col·locant cadascuna de les matrius augmentades obtingudes per cadascuna de les diferents mostres una sota de l'altra, mantenint sempre els espectres MS en les files d'aquesta matriu superaugmentada i amb el seus valors de *m/z* en comú (a les columnes de la matriu).

Un aspecte a tenir en compte quan es treballa amb dades multidireccionals és la seva possible modelització a partir de mètodes bilineals o amb un grau superior de multilinealitat. Per aquest motiu, en la publicació 5 es va avaluar quina era la modelització més adequada per les dades LC×LC-MS. En el cromatograma bidimensional estudiat en aquesta publicació, es detectaven set regions cromatogràfiques importants (veure Figura 2 de la publicació 5). A mode d'exemple, una d'aquestes regions LC×LC-MS es va utilitzar per a avaluar les diferents opcions existents per la seva modelització. La regió seleccionada, que es mostra en la Figura 5.3, era molt interesant ja que contenia dos parells d'isòmers posicionals: SLO/SOL (solapats en la segona columna) i PLO/POL (solapats en les dues dimensions). Les abreviatures dels àcids grassos utilitzades són: S àcid esteàric, O àcid oleic i P àcid palmític.

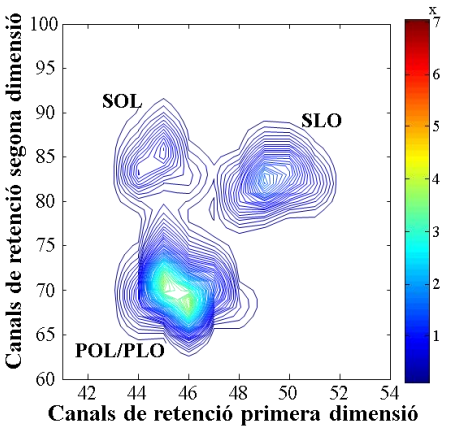


Figura 5.3. Regió cromatogràfica de l'anàlisi de triacilglicerols per LC×LC-MS seleccionada per l'estudi de la trilinealitat de les dades. La barra de colors indica la intensitat dels pics cromatogràfics.

El primer pas per avaluar el tipus de model més adequat per a les dades de LC×LC-MS va ser realitzar la descomposició en valors singulars (SVD) de la matriu de dades corresponent a la regió seleccionada arranjada segons les tres possibles direccions o modes de les dades generades:

- Matriu augmentada en la direcció de les columnes (els valors de *m/z* en el mode comú, les matrius $\mathbf{B_m}$ de cada modulació una a sota de l'altra).
- Matriu augmentada en la direcció de les files (els temps de retenció de la segona dimensió cromatogràfica en el mode comú, les matrius $\mathbf{B_m}$ de cada modulació una al costat de l'altra).
- Matriu augmentada en forma de tub (cada modulació en un vector fila).

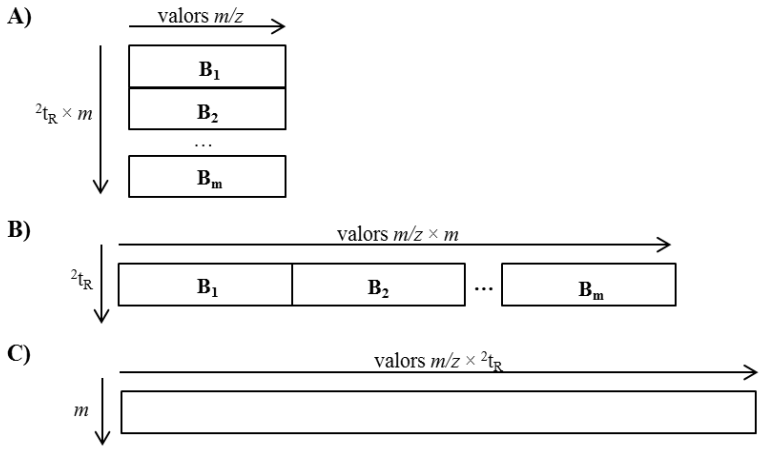


Figura 5.4. Representació gràfica dels tres tipus d'arranjament per a una matriu de dades de LC×LC-MS. Les matrius $\mathbf{B_m}$ corresponen a les matrius de LC-MS de cada modulació i *m* és el nombre de modulacions. A) Augmentada en la direcció de les columnes, valors de *m/z* en el mode comú. B) Augmentada en la direcció de les files, el temps de retenció de la segona dimensió cromatogràfica (^2t_R) és el mode comú. C) Augmentada en forma de tub, cada modulació en un vector fila.

En la Figura 5.4 es mostra una representació d'aquests tres tipus d'arranjament de les dades segons les seves tres direccions o modes. Els resultats obtinguts van mostrar que el nombre de components

linealment independents en absència de soroll o error experimental ('pseudo-rang' químic de la matriu o 'rang' matemàtic de la matriu sense considerar el soroll experimental) per explicar la mateixa variància era sempre més petit en el cas de la matriu augmentada en la direcció de les columnes (es necessitaven 5 valors singulars o components linealment independents) que per les altres dues (es necessitaven 6 valors singulars o components linealment independents). El fet de que les tres matrius no tinguessin el mateix "pseudo-rang" químic indica que les dades de LC×LC-MS presenten una estructura que no es pot modelitzar correctament amb models de multilinealitat superior al de la bilinealitat, emprat generalment per matrius de dades o estructures de dades ordenades en dues direccions o modes (*two-way* or *two-mode* data). Per tant es corrobora que quan aquestes mateixes dades s'ordenen en forma de cub (*three-way* or *three-mode* data), els models trilineals de descomposició de dades no seran adequats i els perfils resolts dels diferents components en les diferents direccions o modes (espectres MS i perfils d'elució en les dues columnes) no seran correctes.

A continuació, per tal d'estudiar millor aquestes desviacions del model trilineal en la regió seleccionada es van comparar els resultats obtinguts mitjançant els models bilineals i trilineals explicats en la secció 2.3.3. (Resolució dels pics cromatogràfics i detecció de variables característiques) de la introducció d'aquesta Tesi: MCR-ALS bilineal [18], MCR-ALS trilineal [19] i MCR-ALS trilineal amb llibertat de desplaçament dels perfils d'elució en la direcció temporal [19], PARAFAC [20] i PARAFAC2 [21]. En la Taula 5.1 es resumeixen els resultats obtinguts per aquests cinc mètodes.

Taula 5.1. Comparació dels resultats obtinguts per MCR-ALS, MCR-ALS trilineal, MCR-ALS trilineal amb llibertat en la direcció temporal, PARAFAC i PARAFAC2.

Mètode	**Nombre de components**	***LOF* (%)**	**Consistència del nucli (%)**
MCR-ALS bilineal	5	3	-
MCR-ALS trilineal amb llibertat en la direcció temporal	5	17	-
MCR-ALS trilineal	5	21	-
PARAFAC	4	14	80
PARAFAC2	5	3	99

La desviació de la trilinealitat es fa visible sobretot en els valors de *lack of fit* (*LOF*, veure secció 2.3.2 del capítol 2) ja que els ajustos de les dades quan es va fer la modelització amb els models estrictament trilineals (MCR-ALS trilineal, MCR-ALS trilineal amb llibertat de desplaçament dels perfils d'elució en la direcció temporal i PARAFAC) van ser considerablement pitjors que en el cas de l'aplicació de MCR-ALS bilineal i PARAFAC2. Tal com s'explica en la introducció de la Tesi (capítol

2), el mètode MCR-ALS assumeix per defecte un model bilineal de les dades analitzades, mentre que el mètode PARAFAC2 està basat en un mètode trilineal que tolera certes derives en els temps de retenció i/o petites diferències en la forma dels pics cromatogràfics entre les diverses modulacions. S'observa, a més, que el valor de la consistència del nucli [22] (veure secció 2.3.3 del capítol 2) obtingut en el model de PARAFAC va ser del 80%, si les dades complissin totalment el model trilineal aquest valor hauria de ser més proper al 100%. D'altra banda, el fet que el MCR-ALS trilineal amb llibertat de desplaçament dels perfils d'elució en la direcció temporal tingués una manca d'ajust superior que el MCR-ALS bilineal va indicar que possiblement es produïen també petits canvis en la forma dels pics cromatogràfics, especialment en el cas de condicions de coelució de pics. En resum, els resultats obtinguts van mostrar que les desviacions del model trilineal eren degudes tant a derives en els temps de retenció, com a canvis en la forma dels pics cromatogràfics d'un mateix compost entre les diferents modulacions cromatogràfiques.

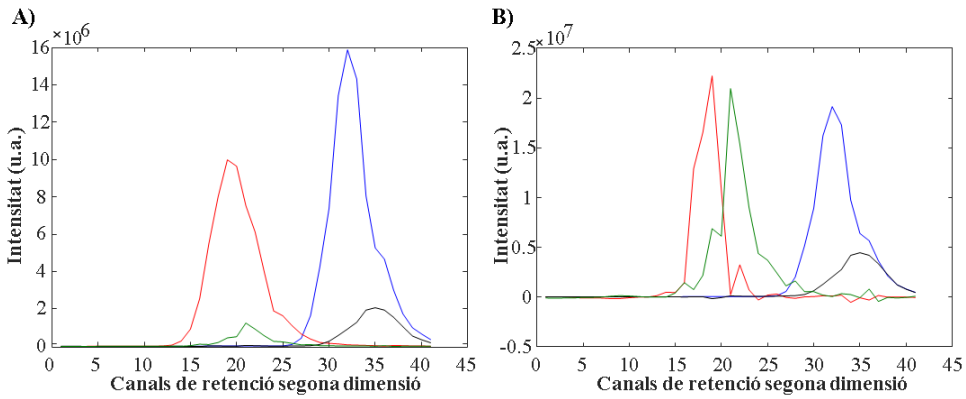


Figura 5.5. Perfils d'elució a la segona dimensió cromatogràfica obtinguts mitjançant A) MCR-ALS i B) PARAFAC2. Els colors indiquen els quatre triacilglicerols resolts: El blau és el SLO, el negre el SOL, el vermell el POL i el verd el PLO.

En comparar el mètode de MCR-ALS amb el de mètode PARAFAC2, cal destacar que un aspecte important a tenir en compte del PARAFAC2 és que no permet l'aplicació de restriccions (per exemple, la de no-negativitat) en els perfils d'elució de la segona dimensió cromatogràfica. Per aquest motiu, els perfils obtinguts amb aquest mètode van ser menys versemblants que els obtinguts mitjançant MCR-ALS. En la Figura 5.5 es mostren els perfils d'elució en la segona dimensió cromatogràfica obtinguts pels dos mètodes. En aquesta figura es pot observar que els perfils obtinguts pel mètode de PARAFAC2 presenten més deformacions no associades a l'elució cromatogràfica i tenen valors negatius, la qual cosa no té sentit químic. Els resultats obtinguts per MCR-ALS bilineal sota restriccions en canvi no presenten aquests

defectes i descriuen la separació més adequadament. Cal destacar en aquest sentit la forta coelució present en alguns dels components resolts que es troben totalment incrustats (*embedded*) dins dels altres.

Els resultats obtinguts en aquesta Tesi són coherents amb els que es van assolir en publicacions anteriors del grup de recerca on s'ha realitzat aquesta Tesi en l'estudi de mètodes de resolució tridireccionals [23, 24]. També coincideixen amb els que han exposat, més recentment, Bortolato i Olivieri en el treball en el qual comparen MCR-ALS i PARAFAC2 mostrant les limitacions d'aquest darrer mètode [25]. En la introducció de la Tesi (capítol 2) s'ha comentat que el mètode PARAFAC2 requereix que els productes creuats de les diferents matrius dels perfils d'elució resolts, $\mathbf{G_k}$, s'han de mantenir constants al llarg de totes les modulacions per així mantenir-se les condicions de model trilineal i de solució única [21, 26]. Bortolato i Olivieri van demostrar que aquest requeriment implica que el mètode PARAFAC2 és capaç de modelar els canvis en els perfils cromatogràfics entre modulacions només si es compleixen les següents condicions cromatogràfiques: 1) quan les derives en el temps de retenció es produeixen sense coelució forta i en absència d'interferències, i 2) quan no hi ha diferències significatives en la forma dels pics cromatogràfics entre diferents modulacions. Si aquestes dues condicions no es compleixen, el resultats obtinguts per PARAFAC2 no són del tot fiables [25].

Tenint en compte els resultats anteriors, es pot concloure que el millor model per a resoldre les dades experimentals estudiades de LC×LC-MS és el MCR-ALS bilineal. En treballs anteriors que han utilitzat el mètode PARAFAC, ha estat necessari realitzar un pas previ d'alineament dels pics cromatogràfics entre les diferents modulacions i així corregir-ne les possibles derives amb el temps [27]. Aquest pas en canvi, no és necessari si s'utilitza MCR-ALS, la qual cosa és un avantatge significatiu que estalvia imprecisions i temps. Altres treballs de la bibliografia que estudien dades de LC×LC-DAD també arriben a la conclusió de que el MCR-ALS bilineal és el mètode més adient per a la resolució d'aquest tipus de dades cromatogràfiques multidireccionals [2, 16, 17]. A més, a la bibliografia també es troben treballs semblants realitzats amb dades de cromatografia de gasos bidimensional acoblada a espectrometria de masses (GC×GC-MS) [28-31]. Els resultats d'aquests treballs mostren que les dades de GC×GC tenen un comportament més trilineal que les de LC×LC, ja que els canvis entre els pics cromatogràfics d'un mateix compost en les diferents modulacions són menors. Malgrat això, aquests treballs també conclouen que el mètode de MCR-ALS bilineal és una bona possibilitat per a la resolució de dades cromatogràfiques multidimensionals.

5.5.2. Estratègies de compressió de les dades LC×LC-MS

Un cop avaluada l'estructura de les dades de LC×LC-MS i mostrats els avantatges de la seva anàlisi mitjançant el mètode MCR-ALS bilineal, es va proposar una estratègia de compressió de dades que permetés superar la limitació que suposa la gran mida dels conjunts de dades de LC×LC-MS. Sobretot quan s'analitzen un nombre gran de mostres simultàniament, com és el cas dels estudis metabolòmics.

En les publicacions 6 i 7 es proposa l'ús de l'estratègia basada en la cerca de ROIs per a comprimir les dades en la direcció dels espectres de masses. En tots dos casos, aquesta aproximació va permetre considerar només un nombre relativament baix de senyals de *m/z*, aproximadament 700 per mostra. Cal destacar que el principal avantatge d'utilitzar la cerca de ROIs en la compressió de la direcció dels espectres de masses és que es manté la resolució espectral de les dades originals. És a dir, quan s'utilitza un espectròmetre d'alta resolució (com és el cas de la publicació 6) els valors de *m/z* mesurats conserven la seva exactitud original. A més, l'aplicació de l'augmentació de ROIs va permetre construir fàcilment les matrius superaugmentades en la direcció de les columnes, en forçar un mateix eix de *m/z* per a totes les mostres.

En el cas de la publicació 7, les matrius superaugmentades obtingudes per 15 mostres tenien una mida final de 63060 files (corresponents als temps de retenció, 4204 per mostra) i aproximadament 1400 columnes (corresponents als valors de *m/z* de les ROIs considerades en totes les mostres). Això vol dir que es va necessitar pel seu emmagatzematge un total de 0,7 Gb (63063 × 1400 elements). Per tant, aquestes matrius de dades un cop comprimides pel procediment de cerca de ROIs ja tenien una mida apta per la seva resolució mitjançat el mètode de MCR-ALS en un ordinador de laboratori amb una capacitat de memòria estàndard. En canvi, en el cas de la publicació 6, el mètode d'anàlisi utilitzat va ser més llarg (250 minuts), la qual cosa va implicar que la matriu augmentada de LC×LC-MS d'una sola mostra obtinguda després de la compressió per cerca de ROIs tingués una mida de 32144 files i aproximadament 700 columnes. A més, en la publicació 6 es van analitzar simultàniament un total 80 mostres (la matriu superaugmentada tindria 2571520 files). L'emmagatzematge necessari en aquest cas seria d'aproximadament 14 Gb (2571520 × 700 elements). En conseqüència, en aquest cas no era possible la seva anàlisi mitjançant un ordinador convencional i va ser necessària una segona etapa de compressió de dades abans de construir la matriu superaugmentada amb tot el conjunt de mostres analitzades. Aquesta segona compressió es va realitzar en la direcció temporal a partir de la transformació d'ondetes. En la Taula 5.2 es mostren els Gb totals dels conjunts de mostres de les publicacions 6 i 7 originals, després de

la compressió espectral per cerca de ROIs i, finalment, després de la compressió temporal per transformada d'ondetes

Taula 5.2. Mida dels conjunts de dades de les publicacions 6 i 7 expressada en nombre total de Gb per les dades originals, les dades comprimides en la direcció espectral i les dades comprimides en la direcció temporal.

	Dades originals (Gb)	**Dades comprimides per cerca de ROIs (Gb)**	**Dades comprimides per transformada d'ondetes (Gb)**
Publicació 6	90,0	14,4	1,8
Publicació 7	3,4	0,7	-

En la Figura 5.6 es mostra el resultat d'aplicar aquesta compressió temporal en el cromatograma bidimensional desplegat d'una mostra d'arròs. En aquesta figura s'aprecia que la transformació d'ondetes (*wavelets*) no va ocasionar cap canvi rellevant en les dades, ja que tots els pics del cromatograma es mantenen. El resultat de combinar la compressió de la direcció espectral mitjançant la cerca de ROIs amb la compressió temporal mitjançant la transformació d'ondetes va ser que la mida de les dades es va poder comprimir 50 vegades (fins a 1,8 Gb per tot el conjunt de 80 mostres) sense perdre informació important ni en la direcció cromatogràfica ni en l'espectral. Tot i així, per tal de reduir encara més la mida de les dades i així accelerar els càlculs del mètode MCR-ALS, es va també dividir el cromatograma en tres finestres de temps.

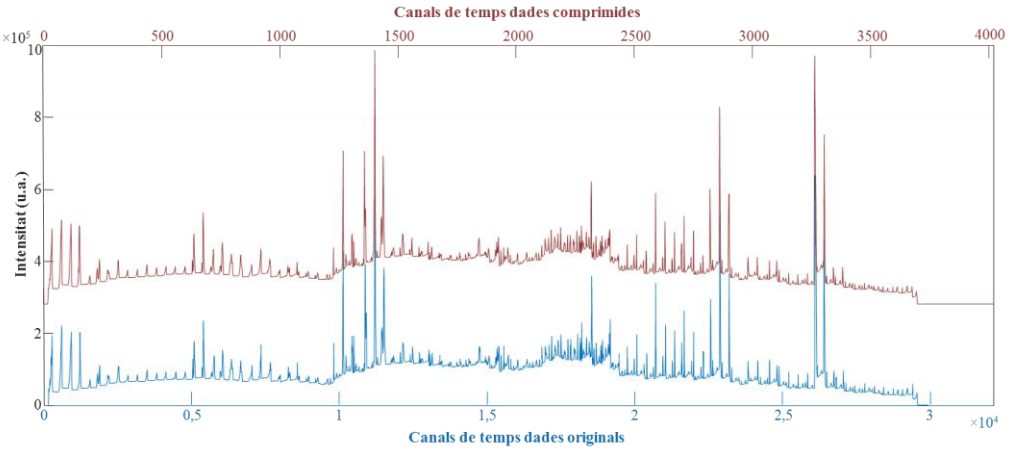


Figura 5.6. Exemple de compressió de dades per transformació d'ondetes en un cromatograma bidimensional obtingut en la publicació 6. La línia blava representa les dades originals (32144 canals de temps). La línia vermella representa les dades comprimides (4018 canals de temps). No s'observen diferències rellevants en els dos cromatogrames.

Tenint en compte tots els resultats mostrats fins el moment, l'estratègia de tractament de dades que s'ha proposat en aquesta tesi per a tractar els conjunts de dades (multimostra) de LC×LC-MS consisteix en la combinació de:

1) La compressió de dades en la direcció espectral mitjançant la cerca de ROIs i en la direcció temporal amb el procediment de transformació d'ondetes (*wavelets*).
2) La resolució d'aquestes dades comprimides mitjançant MCR-ALS.

Aquesta combinació s'anomena ROIMCR [32] i la publicació 6 és la primera en la que s'ha utilitzat aquest procediment per dades de LC×LC-HRMS.

5.5.3. Aplicació de les metodologies LC×LC-MS acoblades amb l'anàlisi quimiomètrica de les dades

En aquesta secció es presenten les aplicacions realitzades durant aquesta Tesi en les que s'ha emprat l'anàlisi de diferents tipus de mostres per LC×LC-MS combinada amb l'estratègia de tractament de dades descrita en la secció anterior.

Caracterització de triacilglicerols en oli de blat de moro

En la publicació 5 es presenten els resultats obtinguts en l'anàlisi de triacilglicerols (TAGs) d'una mostra d'oli de blat de moro per LC×LC-MS. La caracterització dels TAGs en olis vegetals és complicada ja que existeixen diferents isòmers estructurals. Aquests isòmers tenen propietats químiques pràcticament idèntiques i, per tant, són molt difícils de separar cromatogràficament.

Per tal de demostrar l'eficàcia del mètode quimiomètric MCR-ALS per a resoldre les dades de LC×LC-MS d'aquest cas difícil, en la publicació 5 es van analitzar tres regions del cromatograma bidimensional obtingut (com, per exemple, la regió mostrada en la Figura 5.3) on diversos isòmers estructurals es trobaven coeluïts (veure regions 1, 2 i 3 de la Figura 2 de la publicació 5). Quan aquests isòmers coelueixen, no es poden distingir per eines tradicionals, ja que el seus espectres de masses presenten els mateixos valors de massa pels seus ions moleculars ($[M+H]^+$) i alguns dels seus fragments diacilglicerols ($[DG+H]^+$). Mitjançant el mètode MCR-ALS es van poder resoldre els perfils d'elució i espectres de masses purs de cada isòmer en les tres regions analitzades (veure Figures 4, 5 i 6 de la publicació 5). Gràcies a la resolució simultània dels seus espectres de masses purs es van poden identificar tots els isòmers. En tots els casos, el percentatge de variància explicada (R^2) pel model MCR-ALS va ser superior al 99,5% i el percentatge de falta d'ajust (LOF) inferior al 6,8%. Per exemple, en la

Figura 6.7 es mostren els perfils d'elució en les dues dimensions cromatogràfiques i els espectres de masses resolts pels isòmers posicionals PLO i POL. Com que les abundàncies relatives de les masses corresponents als fragments dels diacilglicerols $[DG+H]^+$ eren diferents (575,5 per LP i 577,5 pel PO), el mètode MCR-ALS va aconseguir la resolució correcta dels dos isòmers i la seva posterior identificació.

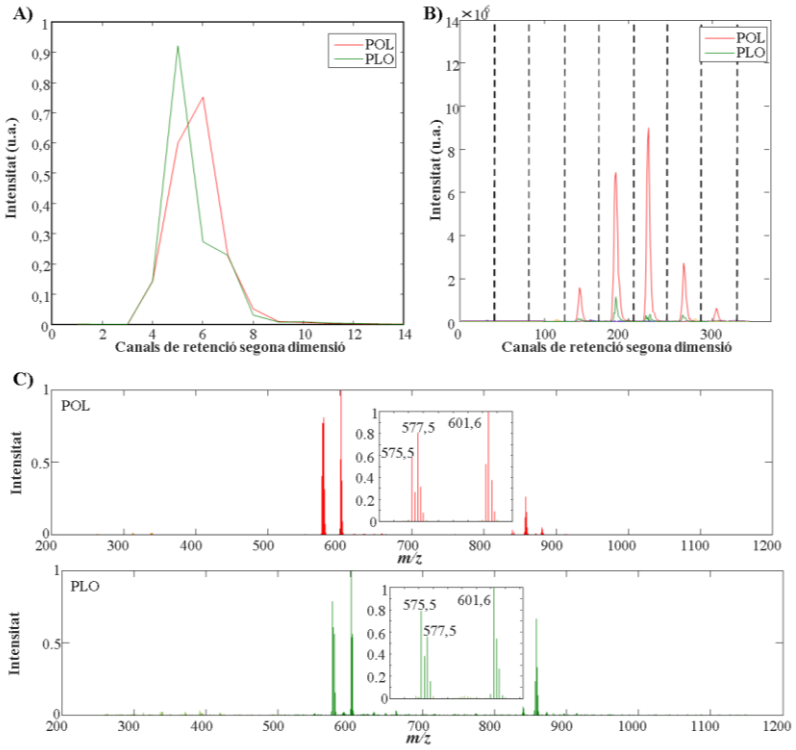


Figura 6.7. Resolució per MCR-ALS dels perfils d'elució i dels espectres de masses dels isòmers posicionals POL i PLO. A) Perfils d'elució a la primera dimensió cromatogràfica. B) Perfil d'elució a la segona dimensió cromatogràfica. C) Espectres de masses resolts pels dos isòmers. Les vistes augmentades mostren la regió de l'espectre de masses corresponen als fragments diacilglicerols de cada TAGs.

Metabolòmica i lipidòmica no dirigides en l'estudi dels efectes de diferents factors ambientals sobre el cultiu de l'arròs

En les publicacions 6 i 7 es presenta, respectivament, una aproximació de metabolòmica i lipidòmica no dirigides en l'estudi dels efectes de diferents factors ambientals sobre els cultius d'arròs. En aquests casos es va analitzar simultàniament un conjunt gran de mostres, de manera que va ser necessària l'aplicació de l'etapa prèvia de compressió de dades.

D'una banda, en la publicació 6 es va avaluar la possible influència de l'hora del dia (cicle circadiari) de la collita i de la quantitat d'aigua disponible en el creixement de l'arròs mitjançant un estudi de metabolòmica no dirigida. Per a poder avaluar l'efecte del temps (hora del dia) de collita, les mostres d'aquest estudi es van recollir a deu punts temporals diferents al llarg d'un cicle diari de 24 h. L'estudi de

l'efecte de la quantitat d'aigua disponible es va analitzar regant només la meitat de les mostres abans del primer temps de recollida. D'altra banda, en la publicació 7 es van estudiar els efectes de la presència de concentracions significatives d'arsènic (As) en l'aigua de regadiu sobre els lípids de l'arròs mitjançant una aproximació lipidòmica no dirigida. Amb aquest objectiu, les mostres d'arròs es van regar amb solucions d'As 1 i 1000 µM. El nivell baix de tractament es va fixar a 1 µM perquè aquest és el límit de concentració d'As en aigua acceptat per la legislació de la Unió Europea (*Groundwater Directive* 2006/118/EC) [33]. D'altra banda, el nivell alt de tractament es va establir a 1000 µM perquè es va trobar que aquest era el nivell màxim de concentració d'As que provocava canvis apreciables a nivell de fenotip (canvi de color de parts aèries i arrels i disminució del creixement foliar) sense causar la mort de les plantes. A continuació es descriuen i es comparen els resultats obtinguts en aquests estudis.

Desenvolupament dels procediments analítics LC×LC en l'anàlisi de les mostres naturals complexes

El desenvolupament del mètode de LC×LC utilitzat en la publicació 6 per a l'anàlisi dels metabòlits de l'arròs es va realitzar durant una estada de recerca a la Universitat d'Amsterdam, en el grup de recerca de Química Analítica. Aquest mètode consistia en la utilització d'una columna HILIC (TSK gel amida-80, 250 mm × 2.0 mm i.d.; 5 µm) en la primera dimensió cromatogràfica i una columna de fase inversa (RP) en la segona (KINETEX C18, 50 mm × 2.1 mm i.d.; 5 µm). Aquests dos tipus de modes cromatogràfics van oferir una separació complementària dels metabòlits, és a dir que les dues dimensions cromatogràfiques de l'anàlisi van ser gairebé ortogonals. Això va fer que el poder de separació del mètode desenvolupat fos elevat, ja que la capacitat cromatogràfica total de pic s'aproximava al producte de les capacitats individuals de pic de les dues dimensions cromatogràfiques. L'únic inconvenient que presenta aquesta combinació de columnes és que pot donar lloc al trencament de pic (*breakthrough*). Durant els primers minuts d'anàlisi, la mostra eluïa de la columna HILIC (primera dimensió) amb un percentatge elevat de solvent orgànic a la fase mòbil. Conseqüentment, en entrar a la columna RP (segona dimensió) els anàlits més polars patien trencament de pic (una part de l'anàlit eluïa amb el volum mort i la resta al temps d'elució corresponent). Per poder solucionar això, es va instal·lar al sistema una bomba addicional que afegia aigua a la mostra abans de ser injectada a la segona columna. D'aquesta manera disminuïa el percentatge de solvent orgànic en la fase mòbil evitant doncs aquest efecte. En la Figura 5.8 es mostra el cromatograma bidimensional obtingut en l'anàlisi dels metabòlits d'una mostra de part aèria d'arròs. En aquesta figura es pot observar que els pics es trobaven ben distribuïts al llarg de tota l'àrea del cromatograma i que, per tant, la separació obtinguda es pot considerar satisfactòria.

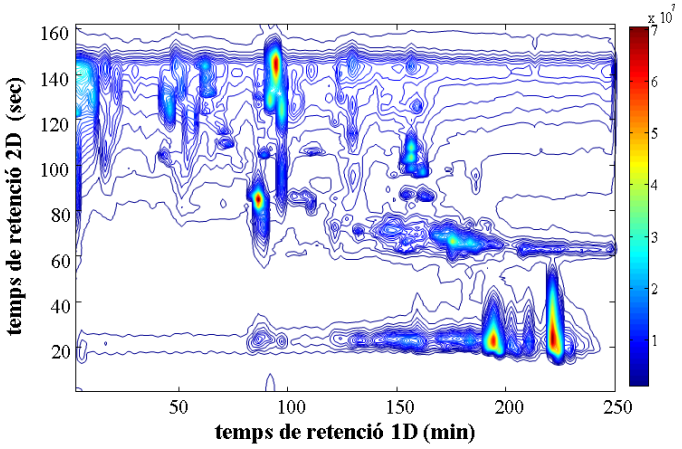


Figura 5.8. Cromatograma bidimensional obtingut en l'anàlisi dels metabòlits d'una mostra d'arròs. La barra de colors indica la intensitat dels pics cromatogràfics.

Com que en la publicació 6 es va demostrar l'efectivitat de combinar una columna HILIC amb una columna RP, en la publicació 7 es va tornar a fer servir aquesta combinació de columnes. En el mode cromatogràfic de RP els lípids es separen segons les seves principals famílies, per la qual cosa, la columna de RP (ZORBAX Eclipse XDB-C18, 150 mm × 2,1 mm i.d.; 5µm) es va escollir per la primera dimensió cromatogràfica. En la segona dimensió la columna emprada va ser una KINETEX HILIC (30 mm × 3 mm i.d.; 2,6 µm). Un avantatge d'utilitzar la columna RP en la primera dimensió és que no es produeix el trencament de pic i, per tant, no és necessari utilitzar una bomba addicional en el sistema. A més, s'han trobat alguns treballs en la bibliografia que també obtenen resultats satisfactoris utilitzant aquesta combinació de columnes en l'anàlisi de lípids [34, 35]. Cal destacar que aquest mètode es va desenvolupar i posar a punt a l'IDAEA-CSIC, on s'ha instal·lat un equip de LC×LC gràcies als coneixements obtinguts prèviament a la Universitat d'Amsterdam.

Resolució per MCR-ALS de les dades de LC×LC

En tots els casos l'aplicació de MCR-ALS va permetre resoldre els espectres de masses (informació qualitativa) i els perfils d'elució en les dues dimensions cromatogràfiques (informació quantitativa) dels compostos presents en les mostres analitzades. Tots els models de MCR-ALS obtinguts van mostrar bons ajustos (percentatges de variància (R^2) explicada superiors a 97% i percentatges de falta d'ajust (LOF) inferiors al 16%). En el cas de la publicació 6, en total es van poder resoldre simultàniament en una sola anàlisi MCR-ALS fins a 200 components, 154 dels quals es van associar a metabòlits i 46 a contribucions del senyal del solvent i de soroll instrumental. En la publicació 7 es van resoldre fins a 200 lípids per cadascuna de les quatre matrius de dades analitzades (parts aèries i arrels analitzades en mode positiu i en mode negatiu). Aquest nombre de compostos (metabòlits i lípids) resolts va ser superior a l'aconseguit en altres treballs de la bibliografia, en els quals s'utilitzava un mètode de separació per LC×LC similar, però

en els quals s'aplicava una estratègia de tractament de dades més convencional, basada en una inspecció manual dels pics després de l'ús dels softwares MZmine [34] o LipidBlast [35]. A més, aquest nombre de metabòlits resolts també era superior a l'aconseguit en altres estudis previs de la bibliografia en els que s'analitzaven els metabòlits mitjançant 1D-LC, però on se seguia una estratègia de tractament de dades similar a la que es proposava en aquest capítol [36-44].

En la Figura 5.9 es mostra un exemple de la resolució de MCR-ALS aconseguida en la publicació 6 per tres metabòlits fortament coeluïts en les dues dimensions cromatogràfiques. Els perfils d'elució obtinguts (Figures 9A i 9B) van permetre observar els canvis de concentració dels components resolts en els diferents tipus de mostres, i d'aquesta manera es van poden avaluar els efectes dels factors ambientals estudiats. Els espectres de masses resolts (Figura 5.9C), es van utilitzar en la identificació temptativa d'aquests metabòlits.

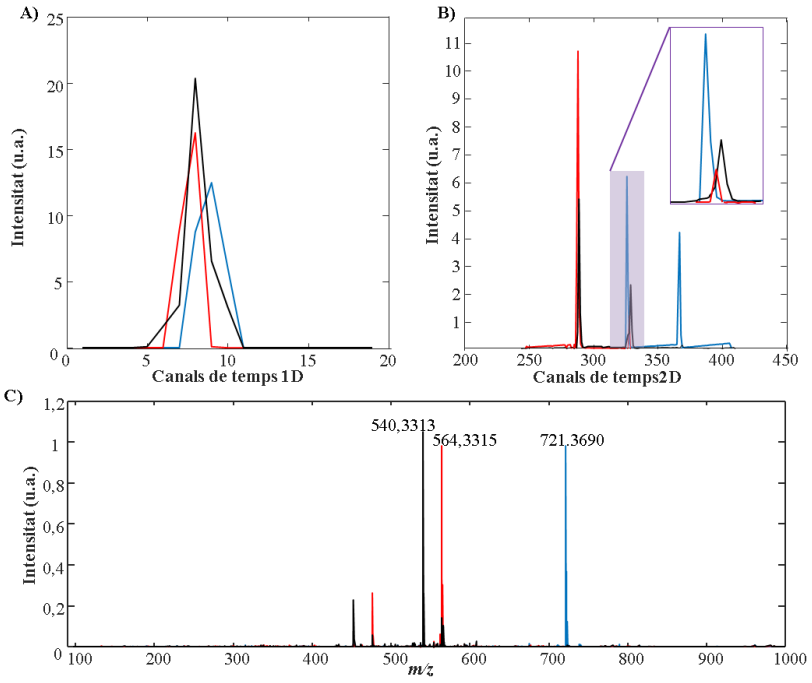


Figura 5.9. Exemple de la resolució de MCR-ALS en la publicació 6 de tres metabòlits coeluïts en les dues dimensions cromatogràfiques. A) Perfils d'elució en la primera columna. B) Perfils d'elució en la segona columna, l'insert mostra una vista augmentada d'una modulació en la qual coleueixen els tres metabòlits. C) Espectres de masses resolts.

Estudi quimiomètric dels factors ambientals avaluats

Les àrees cromatogràfiques dels compostos (metabòlits o lípids) trobats en les diferents mostres analitzades per LC×LC-MS es van obtenir directament a partir dels components resolts per MCR-ALS.

L'aplicació de diversos mètodes quimiomètrics a aquestes àrees cromatogràfiques va permetre l'avaluació dels efectes dels diversos factors ambientals estudiats.

En la publicació 6, l'aplicació del procediment d'ASCA a les àrees cromatogràfiques dels metabòlits resolts va permetre determinar que els dos factors (quantitat d'aigua i hora del dia de la collita) estudiats tenien un efecte significatiu en el metabolisme de l'arròs. El nivell de significació *p* obtingut per la quantitat d'aigua va ser igual a 0,0001 i l'obtingut per l'hora del dia de la collita va ser 0,0032, considerant en ambdós casos 10000 replicats en el test de permutacions.

Els efectes d'aquests dos factors es van estudiar també de manera independent. D'una banda l'efecte de la quantitat d'aigua es va avaluar mitjançant el mètode PLS-DA. El model de PLS-DA obtingut va permetre diferenciar clarament les dues classes de mostres analitzades (controls i mostres sotmeses a una quantitat baixa d'aigua, veure Figura 5 de la publicació 6), amb un 90% de mostres ben classificades. A més, el gràfic de les variables més importants en projecció (VIPs) va mostrar quins dels metabòlits resolts tenien una major influència en aquesta diferenciació dels dos tipus de mostres. Algunes de les famílies identificades com a rellevants, com ara els glicòsids, els flavonoides i algunes hormones, ja s'havien trobat que eren influents en estudis previs sobre els efectes de la sequera en el metabolisme de les plantes [45-47]. En canvi, d'altres com els nucleòsids, no s'havien detectat encara com a metabòlits importants en aquest tipus d'estudis d'escassetat d'aigua. D'altra banda, l'efecte del període de collita es va estudiar per separat en les plantes regades en condicions normals i en les plantes sotmeses a manca d'aigua, aplicant en cada cas el mètode MCR-ALS a les dues matrius d'àrees respectives per separat. En els dos casos es van obtenir models de MCR-ALS amb quatre components, que explicaven un 85,8% de la variància en el cas de les mostres que havien crescut en condicions normals i un 90,1% en el cas de les mostres sotmeses a sequera. Els perfils obtinguts pel mètode de MCR-ALS van revelar que els principals canvis en el perfil metabòlic de l'arròs al llarg del dia estaven relacionats amb els canvis en la intensitat lumínica. Els resultats obtinguts coincideixen amb les tendències observades en estudis anteriors de transcriptòmica de l'arròs [48, 49]. A més, també s'han observat resultats similars en treballs sobre el cicle circadiari d'altres plantes, com l'*A. thaliana* o la planta del té (*Camelia sinensis* L.) [47, 50, 51].

En la publicació 7, l'avaluació per PCA i ASCA de les àrees cromatogràfiques dels lípids resolts en les diferents mostres va permetre determinar que la contaminació per As en l'aigua de regadiu té un efecte significatiu en la concentració dels lípids de l'arròs (nivells de significació de *p* entre 0,003 i 0,001 (considerant 1000 permutacions) en les quatre matrius analitzades (parts aèries i arrels de l'arròs

analitzades en mode de ionització positiu i en mode negatiu). Els models de PCA obtinguts van permetre distingir les mostres controls de les mostres regades amb As 1000 µM en les quatre matrius analitzades. En tots els casos, aquesta distinció es va observar en els dos primers PCs. En canvi, les mostres regades amb As 1 µM no es van poder distingir amb claredat de les mostres control. En la Figura 5.10 es mostren els diferents tipus de mostres analitzades sobre el gràfic d'*scores* obtingut per les parts aèries (Figura 5.10A) i les arrels (Figura 5.10B) analitzades en mode positiu.

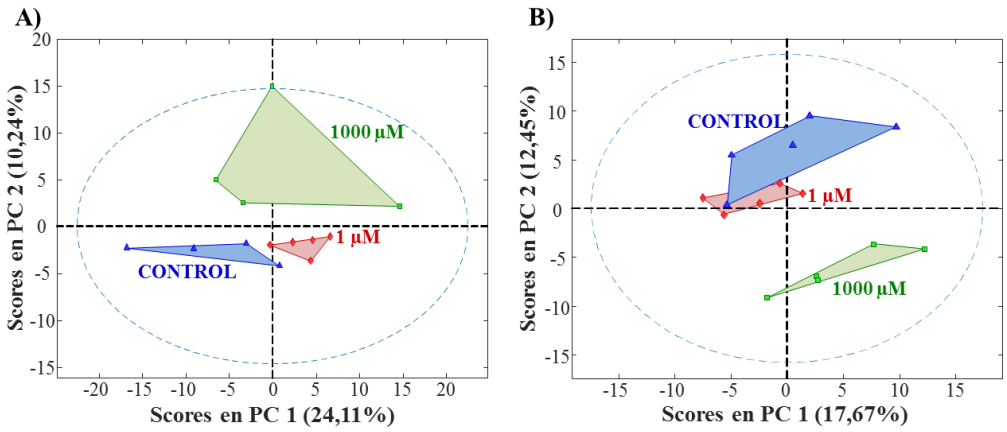


Figura 5.10. Gràfic d'*scores* obtingut per les mostres analitzades en mode positiu. A) Parts aèries. B) Arrels.

Finalment, es va utilitzar també el mètode PLS-DA per seleccionar els lípids més influïts pel tractament amb As. En les quatre matrius analitzades es va aconseguir distingir completament les mostres controls de les mostres regades amb As 1000 µM. A partir del gràfic de VIPs dels models de PLS-DA es van poder seleccionar entre 54 i 77 lípids, la concentració dels quals canviava significativament entre aquests dos tipus de mostra pels quatre casos analitzats (parts aèries i arrels analitzades en mode de ionització positiu i negatiu). Alguns dels lípids seleccionats com a importants van poder ser identificats tal com s'explica en la secció següent.

Identificació dels metabòlits i lípids resolts

En la publicació 6 es va emprar un espectròmetre de masses d'alta resolució que permetia identificar els metabòlits resolts comparant els valors de massa exacta obtinguts experimentalment amb valors de massa teòrics disponibles en bases de dades públiques com MassBank[52], METLIN[53] o HMDB[54]. Seguint aquest procediment, es van poder identificar 139 metabòlits amb un error de massa inferior a 10 ppm, tal com es recomana quan s'utilitza un analitzador de temps de vol (TOF). Els metabòlits identificats van ser de 15 famílies diferents: aminoàcids, sucres, nucleòsids, nucleòtids, àcids carboxílics, hormones, cofactors, compostos aromàtics, àcids grassos, lípids, glicòsids, flavonoides, alcaloides,

terpens i altres metabòlits secundaris. Cal comentar que en un estudi en el qual l'objectiu principal sigui la interpretació biològica dels efectes estudiats caldria completar aquesta identificació temptativa realitzada amb procediments complementaris. Per a poder fer això, seguint les indicacions de la directiva europea 2002/657/CE [33], caldria enregistrar per exemple els espectres de MS/MS dels metabòlits d'interès i obtenir-ne com a mínim dos ions fills. Alternativament, també es podria comparar els metabòlits identificats amb els seus patrons de referència si es troben disponibles comercialment.

La identificació dels lípids en la publicació 7 es va realitzar a partir del seu espectre de MS/MS, ja que l'analitzador emprat ho permetia (triple quadrupol). L'obtenció d'aquests espectres de fragmentació es va realitzar en dues sèries d'anàlisi. En la primera es van enregistrar les mostres en mode d'escombratge d'ions (*full scan*) en mode de ionització positiu i negatiu. Seguidament es van tractar les dades mitjançant el protocol ROIMCR i el posterior estudi quimiomètric mostrat en l'apartat anterior, a partir dels quals es va arribar a una llista de lípids afectats pel tractament d'As. En la segona sèrie d'anàlisi es van enregistrar els espectres de MS/MS dels lípids d'aquesta llista. Al tractar-se d'un estudi no dirigit, l'optimització de l'energia de col·lisió (CE) dels lípids d'interès no es va realitzar, de manera que els espectres de MS/MS es van enregistrar en quatre valors de CE diferents (10, 20, 30 i 40 eV). La identificació es va dur a terme comparant el seu espectre de MS/MS experimental amb els espectres MS/MS disponibles en la base de dades METLIN [53]. La Figura 6 de la publicació 7 mostra un exemple de com es va realitzar aquesta identificació i a la Taula 1 de la mateixa publicació es mostra la llista final dels lípids identificats per aquest procediment. En total es van identificar 11 fosfatidilcolines (PC), tres fosfatidilinositols (PI), tres diacilglicerols (DAG), tres triacilglicerols (TAG), dos monogalactosilglicerols (MGDG), dos fosfatidilglicerols (PG), un àcid fosfatídic (PA), un àcid gras i un glicerol. En un treball futur caldria millorar el procés d'identificació dels lípids, ja que alguns dels espectres de MS/MS obtinguts en el present treball no presentaven un senyal prou intens dels ions fills. Aquest problema podria ser degut a la dilució que pateix la mostra en les separacions LC×LC. Una possible solució podria ser l'ús d'una modulació activa, com proposen A. Gargano i els seus col·laboradors, que consisteix en utilitzar columnes de trampa (*trap columns*) enlloc de bucles (*loops*) en el modulador per a concentrar les fraccions l'elució provinents de la primera columna [55]. També, un cop finalitzada la detecció no dirigida dels possibles biomarcadors, la confirmació d'aquests es podria resoldre mitjançant una aproximació dirigida (*target*), la qual cosa permetria optimitzar la CE de cada lípid i millorar així la

sensibilitat del mètode. Igual que en el cas anterior, la utilització de patrons de referència també ajudaria a completar una identificació dels lípids més acurada.

Interpretació biològica

L'últim pas en els estudis de metabolòmica i lipidòmica no dirigides és la interpretació biològica dels resultats obtinguts. En el cas de les dues publicacions incloses en aquest capítol aquesta interpretació no era l'objectiu principal, ja que els dos treballs tenien una finalitat sobretot metodològica, de posada a punt del procediment no dirigit d'anàlisi de mostres metabolòmiques per LC×LC-MS i de l'avaluació quimiomètrica de la gran quantitat de dades generades. Malgrat això, en la publicació 6 es van avaluar els canvis de concentració observats en els metabòlits de l'arròs segons quina era l'hora de la seva collita i de la quantitat d'aigua disponible dels cultius.

En l'avaluació dels efectes que tenia la quantitat d'aigua de regadiu sobre els cultius d'arròs, les famílies de metabòlits amb un valor de VIP més elevat van ser els glicòsids i els flavonoides. Aquestes dues famílies s'acumulen a les fulles de les plantes quan aquestes creixen en condicions adverses, com la falta d'aigua. L'emmagatzematge (increment de concentració) d'aquests metabòlits secundaris està relacionat amb funcions de protecció de les plantes. Per exemple, en el cas dels flavonoides l'enzim que inicia la seva via biosintètica (fenilalanina amoni-liasa, PAL) augmenta la seva activitat quan els organismes vegetals es troben en situacions d'estrès hídric [45, 46, 56]. També es van trobar algunes hormones amb un valor de VIP elevat, com per exemple l'àcid abscísic (ABA). L'ABA té un paper important en la resposta de les plantes a l'estrès hídric, ja que regula el tancament dels estomes i això ajuda a reduir la pèrdua d'aigua per transpiració. A més, l'ABA també indueix la síntesi dels osmòlits, els quals baixen el potencial hídric de les cèl·lules i així fan que captin més aigua circumdant i/o que retinguin la que ja tenen [57-59].

En investigar els canvis en la concentració dels metabòlits al llarg del dia es va trobar, per exemple, que els glicòsids no s'alteraven i, en canvi, els sucres presentaven una contribució alta en les hores de més llum, però disminuïa la seva concentració a la posta de sol. Aquestes tendències també s'han observat en un estudi previ similar basat en la transcriptòmica de l'arròs [48].

Tenint en compte tots els resultats exposats en aquest capítol es pot concloure que la metodologia proposada ha estat adequada i que es pot recomanar per a estudis de metabolòmica i lipidòmica no dirigides. En la Figura 5.11 es presenta un esquema dels principals passos a seguir quan s'utilitza una

anàlisi de LC×LC-MS combinada amb l'estratègia quimiomètrica de tractament de dades proposada en aquest capítol.

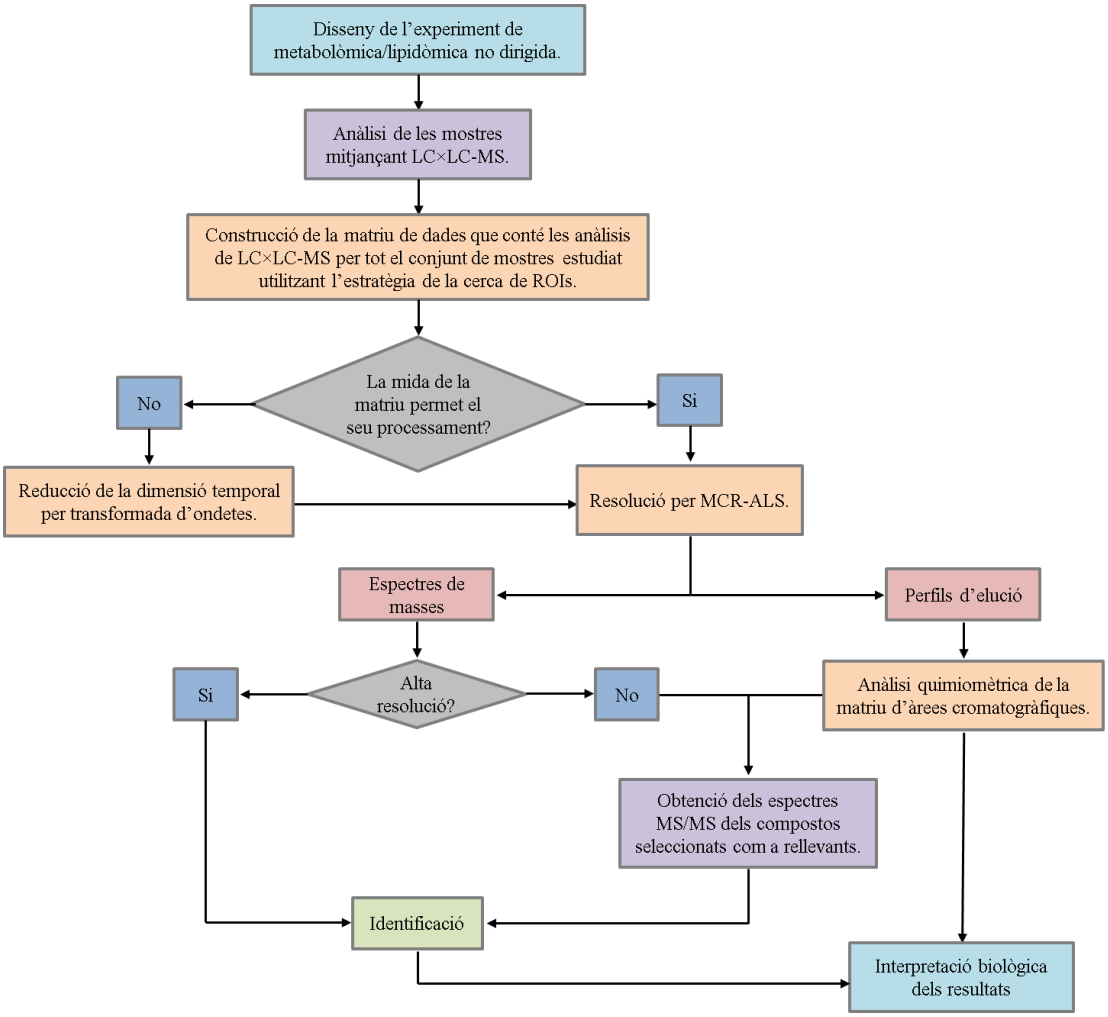


Figura 5.11. Resum dels passos a seguir en un estudi de metabolòmica o lipidòmica no dirigida en el que es combina l'anàlisi de les mostres per LC×LC-MS amb l'estratègia de tractament quimiomètric de dades basada en el procediment ROIMCR.

5.6. Referències

1. Stoll, D. R.;Carr, P. W., Two-Dimensional Liquid Chromatography: A State of the Art Tutorial, *Analytical Chemistry*, 2017, **89**, 519-531.
2. Cook, D. W.;Burnham, M. L.;Harmes, D. C.;Stoll, D. R.;Rutan, S. C., Comparison of multivariate curve resolution strategies in quantitative LCxLC: Application to the quantification of furanocoumarins in apiaceous vegetables, *Analytica Chimica Acta*, 2017, **961**, 49-58.
3. Ouyang, X.;Leonards, P.;Legler, J.;van der Oost, R.;de Boer, J.;Lamoree, M., Comprehensive two-dimensional liquid chromatography coupled to high resolution time of flight mass spectrometry for chemical characterization of sewage treatment plant effluents, *Journal of chromatography. A*, 2015, **1380**, 139-145.
4. Pirok, B. W. J.;Gargano, A. F. G.;Schoenmakers, P. J., Optimizing separations in online comprehensive two-dimensional liquid chromatography, *Journal of Separation Sciences*, 2018, **41**, 68-98.
5. Kilz, P.;Radke, W., Application of two-dimensional chromatography to the characterization of macromolecules and biomacromolecules, *Analytical and Bioanalytical Chemistry*, 2015, **407**, 193-215.

6. Stoll, D.;Danforth, J.;Zhang, K.;Beck, A., Characterization of therapeutic antibodies and related products by two-dimensional liquid chromatography coupled with UV absorbance and mass spectrometric detection, *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, 2016, **1032**, 51-60.
7. Li, Z.;Chen, K.;Guo, M.-z.;Tang, D.-q., Two-dimensional liquid chromatography and its application in traditional Chinese medicine analysis and metabonomic investigation, *Journal of Separation Science*, 2016, **39**, 21-37.
8. Lee, C.;Zang, J.;Cuff, J.;McGachy, N.;Natishan, T. K.;Welch, C. J.;Helmy, R.;Bernardoni, F., Application of Heart-Cutting 2D-LC for the Determination of Peak Purity for a Chiral Pharmaceutical Compound by HPLC, *Chromatographia*, 2013, **76**, 5-11.
9. Kivilompolo, M.;Hyötyläinen, T., Comprehensive two-dimensional liquid chromatography in analysis of Lamiaceae herbs: Characterisation and quantification of antioxidant phenolic acids, *Journal of Chromatography A*, 2007, **1145**, 155-164.
10. Cacciola, F.;Farnetti, S.;Dugo, P.;Marriott, P. J.;Mondello, L., Comprehensive two-dimensional liquid chromatography for polyphenol analysis in foodstuffs, *Journal of Separation Science*, 2017, **40**, 7-24.
11. Donato, P.;Rigano, F.;Cacciola, F.;Schure, M.;Farnetti, S.;Russo, M.;Dugo, P.;Mondello, L., Comprehensive two-dimensional liquid chromatography-tandem mass spectrometry for the simultaneous determination of wine polyphenols and target contaminants, *Journal of chromatography. A*, 2016, **1458**, 54-62.
12. Patti, G. J.;Yanes, O.;Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy, *Nature Reviews Molecular Cell Biology*, 2012, **13**, 263-269.
13. Checa, A.;Bedia, C.;Jaumot, J., Lipidomic data analysis: Tutorial, practical guidelines and applications, *Analytica Chimica Acta*, 2015, **885**, 1-16.
14. Porter, S. E.;Stoll, D. R.;Rutan, S. C.;Carr, P. W.;Cohen, J. D., Analysis of four-way two-dimensional liquid chromatography-diode array data: application to metabolomics, *Analytical Chemistry*, 2006, **78**, 5559-5569.
15. Monteiro, M. S.;Carvalho, M.;Bastos, M. L.;Guedes de Pinho, P., Metabolomics analysis for biomarker discovery: advances and challenges, *Current medicinal chemistry*, 2013, **20**, 257-271.
16. Cook, D. W.;Rutan, S. C.;Stoll, D. R.;Carr, P. W., Two dimensional assisted liquid chromatography - a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, *Analytica Chimica Acta*, 2015, **859**, 87-95.
17. Tistaert, C.;Bailey, H. P.;Allen, R. C.;Heyden, Y. V.;Rutan, S. C., Resolution of spectrally rank-deficient multivariate curve resolution: alternating least squares components in comprehensive two-dimensional liquid chromatographic analysis, *Journal of Chemometrics*, 2012, **26**, 474-486.
18. Tauler, R., Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, 1995, **30**, 133-146.
19. Tauler, R.;Marqués, I.;Casassas, E., Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, *Journal of Chemometrics*, 1998, **12**, 55-75.
20. Bro, R., PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, 1997, **38**, 149-171.
21. Kiers, H. A. L.;ten Berge, J. M. F.;Bro, R., PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, *Journal of Chemometrics*, 1999, **13**, 275-294.
22. Bro, R.;Kiers, H. A. L., A new efficient method for determining the number of components in PARAFAC models, *Journal of Chemometrics*, 2003, **17**, 274-286.
23. De Juan, A.;Tauler, R., Comparison of three-way resolution methods for non-trilinear chemical data sets, *Journal of Chemometrics*, 2001, **15**, 749-772.
24. Zachariassen, C. B.;Larsen, J.;van den Berg, F.;Bro, R.;de Juan, A.;Tauler, R., Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system, *Chemometrics and Intelligent Laboratory Systems*, 2006, **83**, 13-25.
25. Bortolato, S. A.;Olivieri, A. C., Chemometric processing of second-order liquid chromatographic data with UV–vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, *Analytica Chimica Acta*, 2014, **842**, 11-19.
26. Bro, R.;Andersson, C. A.;Kiers, H. A. L., PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, *Journal of Chemometrics*, 1999, **13**, 295-309.
27. Allen, R. C.;Rutan, S. C., Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets, *Analytica Chimica Acta*, 2012, **723**, 7-17.

28. Parastar, H.;Radović, J. R.;Jalali-Heravi, M.;Diez, S.;Bayona, J. M.;Tauler, R., Resolution and Quantification of Complex Mixtures of Polycyclic Aromatic Hydrocarbons in Heavy Fuel Oil Sample by Means of GC × GC-TOFMS Combined to Multivariate Curve Resolution, *Analytical Chemistry*, 2011, **83**, 9289-9297.
29. Parastar, H.;Jalali-Heravi, M.;Tauler, R., Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, *Chemometrics and Intelligent Laboratory Systems*, 2012, **117**, 80-91.
30. Parastar, H.;Radović, J. R.;Bayona, J. M.;Tauler, R., Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares, *Analytical and Bioanalytical Chemistry*, 2013, **405**, 6235-6249.
31. Parastar, H.;Tauler, R., Multivariate Curve Resolution of Hyphenated and Multidimensional Chromatographic Measurements: A New Insight to Address Current Chromatographic Challenges, *Analytical Chemistry*, 2014, **86**, 286-297.
32. Gorrochategui, E.;Jaumot, J.;Lacorte, S.;Tauler, R., Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC Trends in Analytical Chemistry*, 2016, **82**, 425-442.
33. Implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, 2002, **Directive 2002/657/CE**.
34. Baglai, A.;Gargano, A. F. G.;Jordens, J.;Mengerink, Y.;Honing, M.;van der Wal, S.;Schoenmakers, P. J., Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: Two-dimensional liquid chromatography–mass spectrometry vs. liquid chromatography–trapped-ion-mobility–mass spectrometry, *Journal of Chromatography A*, 2017, **1530**, 90-103.
35. Berkecz, R.;Tömösi, F.;Körmöczi, T.;Szegedi, V.;Horváth, J.;Janáky, T., Comprehensive phospholipid and sphingomyelin profiling of different brain regions in mouse model of anxiety disorder using online two-dimensional (HILIC/RP)-LC/MS method, *Journal of Pharmaceutical and Biomedical Analysis*, 2018, **149**, 308-317.
36. Marques, A. S.;Bedia, C.;Lima, K. M. G.;Tauler, R., Assessment of the effects of As(III) treatment on cyanobacteria lipidomic profiles by LC-MS and MCR-ALS, *Analytical and Bioanalytical Chemistry*, 2016, **408**, 5829-5841.
37. Bedia, C.;Dalmau, N.;Jaumot, J.;Tauler, R., Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors, *Environmental Research*, 2015, **140**, 18-31.
38. Dalmau, N.;Jaumot, J.;Tauler, R.;Bedia, C., Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells, *Molecular BioSystems*, 2015, **11**, 3397-3406.
39. Gorrochategui, E.;Casas, J.;Porte, C.;Lacorte, S.;Tauler, R., Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells, *Analytica Chimica Acta*, 2015, **854**, 20-33.
40. Gorrochategui, E.;Li, J.;Fullwood, N. J.;Ying, G. G.;Tian, M.;Cui, L.;Shen, H.;Lacorte, S.;Tauler, R.;Martin, F. L., Diet-sourced carbon-based nanoparticles induce lipid alterations in tissues of zebrafish (Danio rerio) with genomic hypermethylation changes in brain, *Mutagenesis*, 2017, **32**, 91-103.
41. Ribeiro, H. C.;Klassen, A.;Pedrini, M.;Carvalho, M. S.;Rizzo, L. B.;Noto, M. N.;Zeni-Graiff, M.;Sethi, S.;Fonseca, F. A. H.;Tasic, L.;Hayashi, M. A. F.;Cordeiro, Q.;Brietzke, E.;Sussulini, A., A preliminary study of bipolar disorder type I by mass spectrometry-based serum lipidomics, *Psychiatry Research*, 2017, **258**, 268-273.
42. Vinayavekhin, N.;Sueajai, J.;Chaihad, N.;Panrak, R.;Chokchaisiri, R.;Sangvanich, P.;Suksamrarn, A.;Piyachaturawat, P., Serum lipidomics analysis of ovariectomized rats under Curcuma comosa treatment, *Journal of Ethnopharmacology*, 2016, **192**, 273-282.
43. Farrés, M.;Piña, B.;Tauler, R., LC-MS based metabolomics and chemometrics study of the toxic effects of copper on Saccharomyces cerevisiae, *Metallomics*, 2016, **8**, 790-798.
44. Ortiz-Villanueva, E.;Navarro-Martín, L.;Jaumot, J.;Benavente, F.;Sanz-Nebot, V.;Piña, B.;Tauler, R., Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach, *Environmental Pollution*, 2017, **231**, 22-36.

45. Phelix, C. F.;Feltus, F. A., Plant stress biomarkers from biosimulations: the Transcriptome-To-Metabolome (TTM) technology - effects of drought stress on rice, *Plant biology*, 2015, **17**, 63-73.
46. Mohanty, B.;Kitazumi, A.;Cheung, C. Y. M.;Lakshmanan, M.;de los Reyes, B. G.;Jang, I.-C.;Lee, D.-Y., Identification of candidate network hubs involved in metabolic adjustments of rice under drought stress by integrating transcriptome data and genome-scale metabolic network, *Plant Science*, 2016, **242**, 224-239.
47. Chow, B. Y.;Kay, S. A., Global approaches for telling time: Omics and the Arabidopsis circadian clock, *Seminars in Cell & Developmental Biology*, 2013, **24**, 383-392.
48. Izawa, T.;Mihara, M.;Suzuki, Y.;Gupta, M.;Itoh, H.;Nagano, A. J.;Motoyama, R.;Sawada, Y.;Yano, M.;Hirai, M. Y.;Makino, A.;Nagamura, Y., Os-GIGANTEA confers robust diurnal rhythms on the global transcriptome of rice in the field, *The Plant cell*, 2011, **23**, 1741-1755.
49. Matsuzaki, J.;Kawahar, Y.;Izawa, T., Punctual transcriptional regulation by the rice circadian clock under fluctuating field conditions, *The Plant cell*, 2015, **27**, 633-648.
50. Wu, Q.;Chen, Z.;Sun, W.;Deng, T.;Chen, M., De novo sequencing of the leaf transcriptome reveals complex light-responsive regulatory networks in Camellia sinensis cv. Baijiguan, *Frontiers in Plant Science*, 2016, **7**, 332.
51. Augustijn, D.;Roy, U.;Van Schadewijk, R.;De Groot, H. J. M.;Alia, A., Metabolic profiling of intact Arabidopsis thaliana leaves during circadian cycle using 1H high resolution magic angle spinning NMR, *PLoS ONE*, 2016, **11**, e0163258.
52. Horai, H.;Arita, M.;Kanaya, S.;Nihei, Y.;Ikeda, T.;Suwa, K.;Ojima, Y.;Tanaka, K.;Tanaka, S.;Aoshima, K.;Oda, Y.;Kakazu, Y.;Kusano, M.;Tohge, T.;Matsuda, F.;Sawada, Y.;Hirai, M. Y.;Nakanishi, H.;Ikeda, K.;Akimoto, N.;Maoka, T.;Takahashi, H.;Ara, T.;Sakurai, N.;Suzuki, H.;Shibata, D.;Neumann, S.;Iida, T.;Tanaka, K.;Funatsu, K.;Matsuura, F.;Soga, T.;Taguchi, R.;Saito, K.;Nishioka, T., MassBank: A public repository for sharing mass spectral data for life sciences, *Journal of Mass Spectrometry*, 2010, **45**, 703-714.
53. Tautenhahn, R.;Cho, K.;Uritboonthai, W.;Zhu, Z.;Patti, G. J.;Siuzdak, G., An accelerated workflow for untargeted metabolomics using the METLIN database, *Nature Biotechnology*, 2012, **30**, 826.
54. Wishart, D. S.;Knox, C.;Guo, A. C.;Eisner, R.;Young, N.;Gautam, B.;Hau, D. D.;Psychogios, N.;Dong, E.;Bouatra, S.;Mandal, R.;Sinelnikov, I.;Xia, J.;Jia, L.;Cruz, J. A.;Lim, E.;Sobsey, C. A.;Shrivastava, S.;Huang, P.;Liu, P.;Fang, L.;Peng, J.;Fradette, R.;Cheng, D.;Tzur, D.;Clements, M.;Lewis, A.;de souza, A.;Zuniga, A.;Dawe, M.;Xiong, Y.;Clive, D.;Greiner, R.;Nazyrova, A.;Shaykhutdinov, R.;Li, L.;Vogel, H. J.;Forsythei, I., HMDB: A knowledgebase for the human metabolome, *Nucleic Acids Research.*, 2009, **37**, D603-D610.
55. Gargano, A. F. G.;Duffin, M.;Navarro, P.;Schoenmakers, P. J., Reducing Dilution and Analysis Time in Online Comprehensive Two-Dimensional Liquid Chromatography by Active Modulation, *Analytical Chemistry*, 2016, **88**, 1785-1793.
56. del Pozo, L. G.;Salces, E. D., Actividad fenilalanina-amonio liasa (PAL) inducida por el PECTIMORF© en protoplastos de naranjo agrio (Citrus aurantium L.), *Revista del Jardin Botánico Nacional*, 2002, **23**, 303-305.
57. Cutler, S. R.;Rodriguez, P. L.;Finkelstein, R. R.;Abrams, S. R., Abscisic acid: emergence of a core signaling network, *Annual review of plant biology*, 2010, **61**, 651-679.
58. Verslues, P. E.;Bray, E. A., Role of abscisic acid (ABA) and Arabidopsis thaliana ABA-insensitive loci in low water potential-induced ABA and proline accumulation, *Journal of Experimental Botany*, 2006, **57**, 201-212.
59. Yancey, P. H.;Clark, M. E.;Hand, S. C.;Bowlus, R. D.;Somero, G. N., Living with water stress: evolution of osmolyte systems, *Science (New York, N.Y.)*, 1982, **217**, 1214-1222.

Capítol 6

Conclusions

Els mètodes de cromatografia de líquids unidimensional o bidimensional acoblats a espectrometria de masses desenvolupats en aquesta Tesi són adequats en l'anàlisi no dirigida dels metabòlits i lípids de l'arròs (*Oryza sativa* L.). En aquest context, els mètodes d'anàlisi multivariant de dades han permès extreure la informació analítica i bioquímica present en les dades obtingudes en els estudis de metabolòmica no dirigida.

A continuació es presenten les conclusions específiques referents a les estratègies cromatogràfiques i quimiomètriques estudiades i als efectes dels factors estressants avaluats en aquesta Tesi:

Estratègies cromatogràfiques i quimiomètriques:

1) Les fases estacionàries HILIC són molt adequades en els estudis de metabolòmica no dirigida. Quan s'analitzen metabòlits polars, les fases amida (especialment la TSK-gel) i zwitteriònica proporcionen millors resultats que la de mode mixt diol. No obstant, quan s'analitzen simultàniament metabòlits polars i apolars (o de polaritat intermèdia) és recomanable l'ús de la fase estacionària de mode mixt diol. Per poder confirmar i generalitzar aquests resultants caldria realitzar estudis addicionals als que s'han presentat en aquesta Tesi, on s'analitzessin mescles amb un nombre més gran de metabòlits de polaritats diferents i s'optimitzessin les condicions per cadascuna de les fases estacionàries emprades. Els factors experimentals que més influeixen en l'anàlisi cromatogràfica dels metabòlits han estat el tipus de fase estacionària HILIC i el pH de la fase mòbil. A més, cal destacar que la força iònica de la fase mòbil no influeix significativament en les separacions dels metabòlits estudiats.
2) Tant el procediment XCMS com el mètode MCR-ALS han resultat útils en el processament de les dades obtingudes per LC-MS en estudis de metabolòmica no dirigida. Les dues estratègies aporten resultats comparables. Per una banda, el XCMS és fàcil d'utilitzar per a un usuari no expert i el seu temps de computació és relativament curt. D'altra banda, el mètode basat en MCR-ALS és un procediment robust i flexible, ja que no necessita d'una optimització dels paràmetres associats a cada tipus d'instrument LC-MS, ni requereix la modelització i alineament previs dels pics cromatogràfics. A més, el mètode basat en MCR-ALS permet treballar amb dades més complexes on es produeixen canvis en els temps de retenció i en la forma dels pics cromatogràfics dels metabòlits resolts, com passa freqüentment en els mètodes cromatogràfics, sobretot quan hi ha fortes coelucions, i també en altres tècniques com l'electroforesi capil·lar

acoblada a espectrometria de masses (CE-MS) i l'espectroscòpia d'imatges d'espectrometria de masses (MSI).

3) La compressió de les dades de LC-MS en la direcció espectral mitjançant la cerca de ROIs és més efectiva que els procediments tradicionals com el *binning*. L'aplicació del procediment de compressió per cerca de ROIs disminueix de forma considerable la mida i emmagatzematge dels fitxers de dades LC-MS i permet el seu estudi mitjançant el mètode MCR-ALS. En canvi, quan s'aplica el procediment de *binning* les dades resultants encara generen fitxers massa grans i cal dividir-les en diferents finestres de temps abans de processar-les amb MCR-ALS. A més, la compressió mitjançant la cerca de ROIs manté la resolució espectral de les dades originals, mentre que en el procediment de *binning* les dades perden la seva resolució espectral, la qual cosa dificulta la posterior identificació dels metabòlits i lípids.
4) L'aplicació de models que relacionen de forma quantitativa l'estructura dels compostos amb la seva retenció cromatogràfica (models QSRR) en estudis de metabolòmica no dirigida pot resultar una eina útil en l'etapa d'identificació dels diferents metabòlits detectats. L'aplicació d'aquests mètodes permet descartar alguns dels possibles metabòlits proposats per les bases de dades on es recullen els seus valors de massa exacta i, per tant, reduir el nombre total de candidats. Els models QSRR presentats en aquesta Tesi poden encara millorar-se, per exemple, a partir de l'anàlisi d'un nombre més gran de metabòlits en les mescles d'entrenament i validació externa emprades en la seva elaboració.
5) Les dades de cromatografia de líquids unidimensional i bidimensional acoblada a espectrometria de masses presenten generalment derives en els temps de retenció i canvis en la forma dels pics cromatogràfics resolts (per exemple en LC×LC entre modulacions consecutives). El mètode MCR-ALS bilineal és especialment adequat en aquests casos. L'estratègia de tractament de dades LC×LC-MS basada en MCR-ALS és senzilla i els resultats obtinguts són fàcils d'interpretar de forma quantitativa i qualitativa a partir dels perfils d'elució i dels espectres de masses resolts dels metabòlits presents en les mostres analitzades.
6) La combinació de LC×LC-MS amb el procediment de ROIMCR (compressió ROI més resolució MCR-ALS) és una estratègia molt útil en els estudis de metabolòmica i lipidòmica no dirigides. Mitjançant aquesta estratègia combinada ha estat possible detectar, resoldre i identificar un nombre elevat de metabòlits i lípids (més de 150), en una sola anàlisi.

7) La metodologia de LC×LC-MS/MS proposada en aquesta Tesi ha estat capaç de resoldre satisfactòriament mostres lipidòmiques complexes. És encara necessària una millora dels procediments d'identificació dels lípids. La dilució causada per les dues separacions cromatogràfiques consecutives i el fet de no poder optimitzar l'energia de col·lisió per a cada ió en la seva detecció per espectrometria de masses provoca una falta de sensibilitat en el mètode i, per tant, dificulta la identificació dels compostos. Possibles solucions a aquesta limitació podrien ser l'ús d'una modulació activa i també la confirmació dels lípids candidats mitjançant estratègies dirigides.

Efectes dels factors estressants estudiats:

8) S'ha demostrat que la presència de metalls pesants (Cd i Cu) en l'aigua de regadiu provoca canvis significatius en el metaboloma de l'arròs. El tractament amb Cd va alterar significativament la concentració de 54 metabòlits, mentre que el de Cu en va alterar la de 23 metabòlits. Les principals rutes metabòliques afectades pels dos metalls van ser les mateixes, la qual cosa indica que l'arròs respon de manera similar als dos metalls. Bàsicament, aquesta resposta metabòlica es relaciona amb una disminució del creixement i de l'activitat fotosintètica i amb la inducció d'un mecanisme de desintoxicació per a disminuir el dany cel·lular.
9) L'augment de temperatura i l'estrès hídric de manca d'aigua tenen un efecte significatiu en el lipidoma de l'arròs. En tots dos casos, la principal resposta de l'arròs es relaciona amb el manteniment de la fluïdesa òptima de les membranes cel·lulars. En el cas de l'augment de temperatura, el grau d'insaturació dels àcids grassos, els diglicèrids (DAGs) i els triglicèrids (TAGs) va disminuir, provocant així un augment en la rigidesa de les membranes. En canvi, en el cas de l'estrès hídric el grau d'insaturació dels lípids detectats va augmentar, ocasionant així un augment en la fluïdesa de les membranes.
10) S'ha comprovat que l'hora de collita i la quantitat d'aigua de regadiu tenen també un efecte significatiu en el metaboloma de l'arròs. L'alteració de la quantitat d'aigua va afectar a les concentracions dels flavonoides, els glicòsids i les hormones. En el cas de l'hora de collita es va trobar que els perfils temporals dels metabòlits detectats evolucionaven segons els canvis de la intensitat lumínica al llarg del dia (cicle circadiari).