UNIVERSITAT DE BARCELONA

Avaluació dels efectes de factors ambientals en organismes model mitjançant metodologies òmiques i quimiomètriques

Elena Ortiz Villanueva

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- CompartIgual 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - CompartirIgual 4.0. España de Creative* Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-ShareAlike 4.0. Spain License***.

UNIVERSITAT DE BARCELONA

Avaluació dels efectes de factors ambientals en organismes model mitjançant metodologies òmiques i quimiomètriques

Elena Ortiz Villanueva

AVALUACIÓ DELS EFECTES DE FACTORS AMBIENTALS EN ORGANISMES MODEL MITJANÇANT METODOLOGIES ÒMIQUES I QUIMIOMÈTRIQUES

Elena Ortiz Villanueva

FACULTAT DE QUÍMICA

DEPARTAMENT D'ENGINYERIA QUÍMICA I QUÍMICA ANALÍTICA

Programa de Doctorat:

Química Analítica i Medi Ambient

Avaluació dels efectes de factors ambientals en organismes model mitjançant metodologies òmiques i quimiomètriques

---

Memòria presentada per

Elena Ortiz Villanueva

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 Diagnosi Ambiental i Estudis de l'Aigua (IDAEA)

Consejo Superior de Investigaciones Científicas (CSIC)

Tutora:

Dra. Victoria Sanz-Nebot

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 Diagnosi Ambiental i Estudis de l'Aigua adscrit al Consejo Superior de Investigaciones Científicas, i el **Dr. Joaquim Jaumot Soler**, científic titular del mateix Departament,

FAN CONSTAR,

Que la present memòria titulada: "**Avaluació dels efectes de factors ambientals en organismes model mitjançant metodologies òmiques i quimiomètriques**", ha estat realitzada sota la nostra direcció per la Sra. **Elena Ortiz Villanueva** i que tots els resultats presentats són fruit de les experiències realitzades per la citada doctoranda en el Departament de Química Ambiental de l'Institut de Diagnosi Ambiental i Estudis de l'Aigua adscrit al Consejo Superior de Investigaciones Científicas.

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

Barcelona, Juliol de 2018

Dr. Romà Tauler Ferré

Dr. Joaquim Jaumot Soler

Als meus,

AGRAÏMENTS

Arribat el final d'aquesta bonica etapa, toca agrair a totes aquelles persones que al llarg d'aquests quatre anys i mig m'han ajudat a superar aquest repte i que d'alguna manera han aportat el seu granet de sorra. Moltes gràcies a tots de debò!

Primer de tot, voldria agrair als meus directors, el Dr. Romà Tauler i el Dr. Joaquim Jaumot, per donar-me l'oportunitat de realitzar aquesta Tesi. Romà, gràcies per la teva dedicació, per compartir els teus coneixements i pel teus consells i suport. A tu Joaquim, mil gràcies per estar dia rere dia, per guiar-me i aconsellar-me, pel teu suport i perquè sense tu aquesta Tesi no seria el mateix. També voldria agrair a la meva tutora de Tesi, la Dra. Victoria Sanz Nebot, per la seva ajuda durant aquest temps.

Aquesta Tesi tampoc hagués estat possible sense l'ajuda del Dr. Fernando Benavente. Gracias por transmitirme tu entusiasmo por la química analítica y por estar siempre dispuesto a ofrecerme tu ayuda y conocimientos. Al Dr. Benjamín Piña, per l'ajut prestat en la part més biològica d'aquesta Tesi. També a la Marta, a l'Eva i a la Laia, per introduir-me en el món dels embrions del peix zebra.

A tots els companys del projecte CHEMAGEB, amb qui he compartit tantes hores d'aquests últims anys. Als Postdocs (Carma, Cristian i Stefan), als nous doctorands (Marc i Miriam) i a tots aquells que han format part en algun moment d'aquest grup (Alejandro, Amrita, Igor, Marta, Mireia, Xin i Yahya), gràcies per oferir-me la vostra ajuda quan l'he necessitada, per respondre sempre a totes les meves preguntes i per tots els moments viscuts junts; cafès, sortides, congressos, etc. En especial, vull agrair als sis "nens i nenes" que van començar amb mi aquesta aventura: Elba, Eva, Francesc, Meri, Núria i Victor. Mai hauria pensat que la Tesi em donaria unes amistats tan grans. Vosaltres heu fet que aquesta hagi estat una de les millors etapes de la meva vida. Durant aquest temps hem compartit una infinitat de coses; riures constants, confidències, viatges, cerveses, berenars i, el més important, el dia a dia. No puc deixar de fer una menció especial a la meva "*twin*". Meri, saps molt bé que sense tu no hagués estat el mateix, gràcies pel teu suport i per ser-hi sempre.

I would also like to thank Dr. Thomas Hankemeier, Dra. Amy Harms and Dr. Alberto Pasamontes for giving me the opportunity of working in their laboratory and for their kindness during my stay in Leiden. En especial vull agrair a l'Alberto, el meu supervisor, per les nostres converses quimiomètriques i per fer-me sentir com a casa. No em podria deixar a l'altre gran puntal de l'estada, la Marian, gracias por ser mi compañera de batallas en el laboratorio y por acoger-me junto a Ramón sin pensarlo. A los tres, gracias por vuestra hospitalidad y por compartir conmigo esta gran experiencia.

Agrair també a totes aquelles persones que formen part de la meva vida des de fa temps. A les de sempre, Elisabet, Laura i Marta, gràcies per la vostra amistat incondicional i fer-me costat des de petita. A les nenes de la uni, Pili, Núria, Meri, Laura, Llucia i Helena, gràcies per continuar igual d'unides i viure amb mi els mals de cap d'aquesta Tesi. A les "barrienteras", Elisabet, Sílvia, Jennifer y Vic, gràcies pel vostre suport i confiança en les meves capacitats. I finalment, als nens del màster, Laura, Gerard, Meri, Anna, Stanis, Albert i Sergio, gràcies per la injecció d'energia i alegria que em doneu.

Per acabar, vull agrair i dedicar aquesta Tesi a la meva família. Als meus pares, mai us podré agrair tot el que heu fet i feu per mi. Gràcies per donar-me l'oportunitat d'arribar fins aquí i fer-me creure en mi mateixa. Gràcies també per donar-me la millor família de la qual no deixo mai de sentir-me orgullosa. Als meus germans, l'Albert i el Carles, perquè a part de fer de segons pares sempre m'heu fet costat en tot allò en el que crec. A l'Esteve, gràcies per ser el positivisme en persona i mostrar-me que s'ha de viure la vida d'una altra manera. Tampoc em vull oblidar de l'Anna i de l'Esther, gràcies per fer de germanes tan en els bons com en els mals moments i, sobretot, per donar-me els tresors més grans, la Mariona, el Nil i l'Ignasi, que sempre em fan treure un somriure en els dies grisos. Gràcies a tots vosaltres per fer-me sentir la persona més estimada i afortunada del món. A ti, Fran, que decirte que no sepas ya, sabes muy bien que esta Tesis la siento tanto tuya como mía. Gracias por enseñarme a quererme y a confiar en mí, por entregarme todo sin esperar nada a cambio, pero sobre todo, por llegar a mi vida, por convertir mis sueños en nuestros y por querer caminar junto a mí. GRACIAS.

ÍNDEX

RESUM

L'aplicació de tecnologies transcriptòmiques i metabolòmiques en estudis no dirigits té com a principal objectiu la caracterització global (funcional i estructural) dels transcrits de mRNA i dels metabòlits, respectivament, que conformen els sistemes biològics. En el camp mediambiental, aquestes dues aproximacions òmiques permeten avaluar i comparar els nivells d'aquestes molècules en els organismes vius en resposta a diferents estímuls o variacions en les condicions ambientals. D'aquesta manera, ambdues ciències proporcionen coneixement de les interaccions dels sistemes biològics amb el seu entorn a nivell molecular. En els estudis òmics no dirigits són imprescindibles les tècniques analítiques d'alt rendiment com la seqüenciació de RNA (RNA-Seq) i les tècniques de separació acoblades a l'espectrometria de masses, com per exemple la cromatografia de líquids i l'electroforesi capil·lar acoblades a l'espectrometria de masses (LC-MS i CE-MS, respectivament). Ara bé, els grans conjunts de dades generats en aquests estudis són complexos i fan necessari el desenvolupament i l'aplicació de mètodes estadístics i quimiomètrics multivariants d'anàlisi de dades, els quals permetin extreure informació biològica rellevant i facilitar-ne la seva interpretació. Aquesta Tesi s'ha centrat especialment en el desenvolupament de mètodes analítics i quimiomètrics que puguin ser útils en aquests tipus estudis i en la seva aplicació a diversos casos d'interès ambiental i toxicològic on es pren el peix zebra com a organisme model.

D'una banda, s'ha treballat en el desenvolupament i optimització de mètodes analítics de LC-MS i CE-MS i de tractament de dades multivariants per a estudis de metabolòmica no dirigida. S'ha avaluat la influència de diferents factors experimentals en la separació de metabòlits mitjançant la cromatografia de líquids d'interacció hidrofílica (HILIC) (per exemple, la fase estacionària, el pH, la força iònica i el modificador orgànic). També, s'han optimitzat les condicions experimentals per a l'anàlisi metabolòmica no dirigida mitjançant la tècnica de CE-MS. D'altra banda, s'han presentat diferents estratègies de tractament de dades metabolòmiques no dirigides basades en la resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS), les quals permeten la detecció i la identificació dels metabòlits de les dades de LC-MS i CE-MS. Finalment, s'ha aplicat la compressió

de les dades de cerca de regions d'interès (ROI) i la resolució per MCR-ALS (mètode ROIMCR) per a l'estudi simultani o fusió de conjunts de dades que provenen de diferents plataformes de MS. La idoneïtat de totes aquestes metodologies analítiques i quimiomètriques s'ha demostrat en estudis comparatius dels perfils metabòlics de mostres de llevat (*Saccharomyces cerevisiae*) en diferents condicions de creixement estressants.

Un segon aspecte d'aquesta Tesi ha estat l'aplicació de les metodologies prèviament esmentades en la investigació dels possibles efectes de diferents compostos disruptors endocrins, com ara el bisfenol A (BPA), el sulfonat de perfluorooctà (PFOS) i el tributilestany (TBT), en embrions de peix zebra (*Danio rerio*). En concret, s'ha detectat que els tres contaminants produeixen importants alteracions en el metabolisme dels embrions, produint efectes tòxics, estrès oxidatiu i alteracions en la proliferació cel·lular, a més d'efectes específics en vies de senyalització. Addicionalment, s'han postulat noves hipòtesis sobre els efectes toxicològics i morfològics adversos d'aquests compostos químics. En el cas del BPA s'ha realitzat, a més, un estudi transcriptòmic no dirigit de seqüenciació de RNA (RNA-Seq), que ha permès obtenir informació addicional a l'extreta a nivell metabolòmic, la qual cosa permet una comprensió més global i conjunta del mecanisme d'acció del BPA en el metabolisme dels embrions.

ABREVIATURES I ACRÒNIMS

ALS	Mínims quadrats alternats, *alternating least squares*
ANOVA	Anàlisi de la variància, *analysis of variance*
APCI	Ionització química a pressió atmosfèrica, *atmospheric pressure chemical ionization*
API	Ionització a pressió atmosfèrica, *atmospheric pressure ionization*
ASCA	ANOVA amb anàlisi simultània de components, *ANOVA-simultaneous component analysis*
AsLS	Mínims quadrats asimètrics, asymmetric least-squares
BFR	Retardant de flama bromat, *brominated flame retardants*
BPA	Bisfenol A, *bisphenol A*
BGE	Electròlit de separació, *background electrolyte*
CAWG	*Chemical analysis working group*
cDNA	Àcid desoxiribonucleic complementari, *complementary deoxyribonucleic acid*
CE	Electroforesi capil·lar, *capillary electrophoresis*
CE-MS	Electroforesi capil·lar acoblada a l'espectrometria de masses, *capillary electrophoresis coupled to mass spectrometry*
CI	Ionització química, *chemical ionization*
CID	Dissociació induïda per col·lisió, *collision-induced dissociation*
CMTF	Matriu acoblada i factorització tensora, *coupled matrix and tensor factorization*
CNAG	Centre nacional d'anàlisi genòmica
COW	*Correlation optimized warping*
CRF	Funció resposta cromatogràfica, *chromatographic response function*
DAD	Detector de díodes en línia, *diode-array detector*
DIMS	Espectrometria de masses d'injecció directa, *direct-injection mass spectrometry*
DISCO-SCA	Components distintius i comuns amb l'anàlisi simultània de components, *Distinctive and common components with simultaneous component analysis*
DNA	Àcid desoxiribonucleic, *deoxyribonucleic acid*
DoE	Disseny experimental, *design of experiments*

Dpf	Dies posteriors a la fertilització, *days post fertilization*
EC	Contaminant emergent, *emerging contaminants*
EDC	Compost disruptor endocrí, *endocrine disruptors compounds*
EI	Ionització per impacte electrònic, *electron ionization*
EIC	Cromatograma d'ions extrets, *extracted-ion chromatogram*
EOF	Flux electroosmòtic, *electroosmotic flow*
ESI	Ionització per electrosprai, *electrospray ionization*
FDR	Taxa de falsos descobriments, *false discovery rate*
FT-IR	Infraroig amb transformada de Fourier, *Fourier-transform infrared spectroscopy*
FT-ICR	Ressonància ciclotrònica d'ions amb transformada de Fourier, *Fourier transform ion cyclotron resonance*
FWHM	Amplada de pic a mitja alçada, *full width at half maximum*
GC	Cromatografia de gasos, *gas chromatography*
GC-MS	Cromatografia de gasos acoblada a l'espectrometria de masses, *gas chromatography coupled to mass spectrometry*
GMD	Base de dades del metaboloma de Golm, *Golm metabolome data base*
GST	Glutatió-S-Transferasa, g*lutathione S-transferase*
GSVD	Descomposició generalitzada de valors singulars, *generalized singular value decomposition*
HCD	*Higher-energy collisional dissociation*
HILIC	Cromatografia de líquids d'interacció hidrofílica, *hydrophilic interaction liquid chromatogry*
HILIC-MS	Cromatografia de líquids d'interacció hidrofílica acoblada a l'espectrometria de masses, *hydrophilic interaction liquid chromatogry coupled to mass spectrometry*
HMDB	*Human Metabolome Database*
Hpf	Hores posterior a la fertilització, *hours post fertilization*
HPLC	Cromatografia de líquids d'alta eficàcia, *high performance liquid chromatography*
HRMS	Espectrometria de masses d'alta resolució, *high resolution mass spectrometry*
H_0	Hipòtesi nul·la, *null hypothesis*
Id	Diàmetre intern, *internal diameter*
IPC	Cromatografia de bescanvi iònic, *ion pair chromatography*

IPLC	Cromatografia de fase invertida amb formadors de parells iònics, *ion-pair liquid chromatography*
IS	Estàndard intern, *internal standard*
IT	Trampa d'ions, *ion trap*
JIVE	Variació individual i conjunta explicada, *joint and individual variation explained*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
LC	Cromatografia de líquids, *liquid chromatography*
LC-DAD	Cromatografia de líquids acoblada a detector de díodes en línia, *liquid chromatography coupled to diode array detection*
LC-MS	Cromatografia de líquids acoblada a l'espectrometria de masses, *liquid chromatography coupled to mass spectrometry*
LC×LC-MS	Cromatografia de líquids bidimensional acoblada a l'espectrometria de masses, *comprehensive two-dimensional liquid chromatography coupled to mass spectrometry*
LV	Variable latent, *latent variable*
LOEC	Concentració més baixa en la que s'observen efectes, *lowest observed effect concentration*
Lof	Manca d'ajust, *lack of fit*
m/z	Relació de massa i càrrega, *mass-to-charge ratio*
MALDI	Ionització-desorció amb làser assistida per una matriu, *matrix-assisted laser desorption-ionization*
MANOVA	Anàlisi multivariant de la variància, *multivariat- ANOVA*
MCR-ALS	Resolució multivariant de corbes per mínims quadrats alternats, *multivariate curve resolution by alternating least squares*
mRNA	Àcid ribonucleic missatger, *messenger ribonucleic acid*
MS	Espectrometria de masses, *mass spectrometry*
MS/MS	Espectrometria de masses en tàndem, *tandem mass spectrometry*
MS^n	Fragmentació en etapes successives, *multistage mass spectrometry*
MSI	Imatges d'espectrometria de masses, *mass spectrometry imaging*
MRM	Monitorització de múltiples reaccions, *multiple reaction monitoring,*
NGS	Seqüenciació massiva de nova generació, *next generation sequencing*
NIST	*National institute of science and technology*
NPLC	Cromatografia de líquids de fase normal, *normal phase liquid chromatography*

oa-TOF	Temps de vol d'acceleració ortogonal, *orthogonal-acceleration time-of-flight*
OMI	Organització marítima internacional, *International maritime organization*
OMS	Organització mundial de la salut, *world health organization*
O2PLS	Projeccions ortogonals bidireccionals a estructures latents, *two-way orthogonal projections to latent structures*
OnPLS	Projeccions ortogonals multidimensionals a estructures latents, *multiblock orthogonal projections to latent structures*
PCA	Anàlisi per components principals, *principal component analysis*
PFC	Compost perfluorat, *perfluorinated compound*
PFOS	Sulfonat de perfluorooctà, *perfluorooctane sulfonate*
PLS	Mínims quadrats parcials, *partial least squares*
PLS-DA	Anàlisi discriminant per mínims quadrats parcials, *partial least squares-discriminant analysis*
ppm	Parts per milió, *parts per million*
q	Càrrega elèctrica, *electric charge*
Q	Quadrupol senzill, *quadrupole*
QC	Mostra control, *quality control*
qPCR	Reacció en cadena de la polimerasa quantitativa, *quantitative polymerase chain reaction*
QqQ	Triple quadrupol, *triple quadrupole*
Q-TOF	Quadrupol-temps de vol, *quadrupole-time-of-flight*
q/r	Càrrega/radi, *charge/radius*
r	Radi de l'ió, *ion radius*
RNA	Àcid ribonucleic, *ribonucleic acid*
RNA-Seq	Seqüenciació de RNA, *RNA sequencing*
rMANOVA	Anàlisi de la variància multivariant regularitzada, *regularized MANOVA*
RMN	Espectroscòpia de ressonància magnètica nuclear, *nuclear magnetic resonance*
ROI	Cerca de les regions d'interès, *regions of interest*
RPLC	Cromatografia de líquids de fase invertida, *reversed-phase líquid chromatography*
rRNA	Àcid ribonucleic ribosòmic, *ribosomal ribonucleic acid*
SCA	Anàlisi simultània de components, *simultaneous component analysis*

SiOH	Grups silanols, *silanol groups*
SNR_{Thr}	Relació senyal/soroll, *signal-to-noise ratio*
SVD	Descomposició en valors singulars, *singular value decomposition*
TBT	Tributilestany, *tributyltin*
TIC	Cromatograma total d'ions, *total ion chromatogram*
TIE	Electroferograma total d'ions, *total ion electropherogram*
TOF	Temps de vol, *time-of-flight*
tRNA	Àcid ribonucleic de transferència, *transfer ribonucleic acid*
UE	Unió Europea, *European Union*
UHPLC	Cromatografia de líquids d'ultraalta eficàcia, *ultra performance liquid chromatography*
UV-Vis	Ultravioleta visible, *ultraviolet-visible*
VIP	Variable important en la projecció, *variable importance in projection*
WLS	Mínims quadrats pesants, *weighted least square*
YMDB	*Yeast Metabolome Database*
ZFIN	*Zebrafish Information Network*

SÍMBOLS

μ_e	Mobilitat electroforètica, *electrophoretic mobility*
η	Viscositat, *viscosity*

NOTACIÓ

La notació matemàtica utilitzada en aquesta Tesi correspon a 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. La transposició d'una matrius s'indica amb una "T" com a superíndex (per exemple, $\mathbf{X}^T$).

CAPÍTOL 1.

Objectius i estructura de la Tesi

1.1. OBJECTIUS I CONTEXT

Avui en dia el canvi climàtic i la pol·lució causen un gran impacte sobre la salut humana i el medi ambient. En el camp mediambiental, les ciències òmiques tenen com a finalitat trobar una resposta dels efectes d'aquests canvis ambientals sobre els organismes biològics. En aquest context, l'objectiu principal d'aquesta Tesi és:

- El desenvolupament i l'aplicació de mètodes analítics d'espectrometria de masses (MS) i quimiomètrics que permetin l'avaluació dels efectes de factors ambientals estressants en organismes model mitjançant una aproximació metabolòmica no dirigida.

Per assolir aquest objectiu principal, aquesta Tesi presenta els següents objectius en funció del seu camp d'estudi:

Objectius analítics

- Avaluació i comparació del comportament de diferents fases estacionàries de cromatografia de líquids d'interacció hidrofílica (HILIC) pel seu ús en estudis de metabolòmica no dirigida.
- Estudi de la influència dels factors experimentals (modificador orgànic, força iònica i pH de la fase mòbil) en la separació cromatogràfica HILIC.
- Optimització de les condicions experimentals d'electroforesi capil·lar acoblada a l'espectrometria de masses (CE-MS) per a estudis de metabolòmica no dirigida.

Objectius quimiomètrics

- Desenvolupament i avaluació de diferents estratègies de tractament de dades metabolòmiques no dirigides (en particular, les obtingudes mitjançant CE-MS) basades en el mètode de la resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS).
- Desenvolupament i aplicació de mètodes de fusió de dades per a l'anàlisi simultània de conjunts de dades de diferents plataformes analítiques (HILIC-MS i CE-MS) i nivells òmics

(metabolòmica i transcriptòmica) per facilitar la compressió del comportament dels sistemes biològics estudiats en condicions estressants.

Objectius biològics

- Avaluació dels efectes de diferents compostos disruptors endocrins (bisfenol A, sulfonat de perfluorooctà i tributilestany) en embrions de peix zebra (*Danio rerio*) a partir d'estudis metabolòmics i trancriptòmics no dirigits.

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 ha estat el desenvolupament de nous mètodes analítics i quimiomètrics per avaluar els efectes del canvi climàtic i la pol·lució en diferents organismes model representatius dels ecosistemes. Aquests estudis s'han dut a terme considerant diferents nivells òmics, com el genoma, el transcriptoma o el metaboloma. Dins d'aquest projecte, aquesta Tesi s'ha centrat en l'ús de la metabolòmica i la transcriptòmica no dirigides per a l'estudi d'organismes model, com el llevat (*Saccharomyces cerevisiae*) i els embrions de peix zebra (*Danio rerio*), exposats a factors ambientals estressants.

1.2. ESTRUCTURA DE LA MEMÒRIA DE LA TESI DOCTORAL

La memòria d'aquesta Tesi s'estructura en cinc capítols, que es presenten a continuació.

En el primer capítol s'inclouen els objectius d'aquesta Tesi, la seva estructura i la relació dels treballs científics inclosos en la present memòria.

En el segon capítol es mostra una visió general de les ciències òmiques amb un especial interès en els estudis de transcriptòmica i metabolòmica ambiental, en les tècniques analítiques i en les estratègies d'anàlisi de dades. En concret, es presenten algunes de les metodologies analítiques no dirigides més emprades en el camp de la transcriptòmica i la metabolòmica, i es descriuen amb més detall aquelles utilitzades en aquesta Tesi. A més, s'introdueixen el organismes model emprats per a l'avaluació del risc ambiental associat a determinats factors estressants. Finalment, s'expliquen en detall els procediments de tractaments de les dades emprats en aquesta Tesi.

En el tercer capítol es presenten els resultats obtinguts en el desenvolupament i optimització de les metodologies analítiques i quimiomètriques proposades per dur a terme estudis de metabolòmica no dirigida de HILIC-MS i CE-MS. En HILIC-MS, s'han comparat diverses fases estacionàries HILIC i factors experimentals pel seu ús en estudis de metabolòmica no dirigida. En CE-MS, es presenta l'optimització de les condicions experimentals per a l'anàlisi no dirigida de metabòlits catiònics i aniònics en mostres biològiques fent servir capil·lars de sílice fosa. A més, es proposa una estratègia de tractament global de dades basada en MCR-ALS per a l'anàlisi de dades metabolòmiques no dirigides de MS. Finalment, s'introdueixen dues estratègies de fusió de dades òmiques obtingudes mitjançant diferents plataformes de MS amb la finalitat de millorar la interpretació biològica dels canvis observats en els organismes biològics sota condicions ambientals estressants.

En el quart capítol es mostra l'avaluació dels efectes de diferents disruptors endocrins (bisfenol A, el sulfonat de perfluorooctà i el tributilestany) en embrions de peix zebra. En primer lloc, s'estudien els efectes d'aquests tres compostos disruptors endocrins (EDCs) mitjançant un enfocament metabolòmic no dirigit. Seguidament, s'avaluen amb més profunditat els efectes del bisfenol A mitjançant la

integració dels resultats de dos nivells òmics diferents, el metaboloma i el transcriptoma, per tal d'obtenir una comprensió més completa de la disrupció causada per aquest compost químic en els embrions de peix zebra.

Finalment, en l'últim capítol es recullen les conclusions generals més importants d'aquesta Tesi.

1.3. RELACIÓ DELS TREBALLS CIENTÍFICS PRESENTATS EN LA MEMÒRIA

La recerca realitzada en aquesta Tesi ha donat lloc a les següents publicacions:

1. **Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites.**

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

   Revista: *Analytical Methods* 9 (2017) 774-785.

2. **Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling.**

   Autors: E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler.

   Revista: *Electrophoresis* 36 (2015) 2324-2335.

3. **Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data.**

   Autors: E. Ortiz-Villanueva, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler. J. Jaumot.

   Revista: *Analytica Chimica Acta* 978 (2017) 10-23.

4. **Assessment of endocrine disruptors effects on zebrafish (*Danio rerio*) embryos by untargeted LC-HRMS metabolomic analysis.**

   Autors: E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler.

   Revista: *Science of the Total Environment* 635 (2018) 156-166.

5. **Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach.**

   Autors: E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler.

   Revista: *Environmental Pollution* 231 (2017) 22-36.

CAPÍTOL 2. *Introducció*

2.1. METABOLÒMICA I TRANSCRIPTÒMICA D'ORGANISMES MODEL

En aquesta Tesi es proposa l'ús de la metabolòmica per a l'avaluació dels efectes produïts per diferents factors ambientals estressants en organismes biològics model, tals com el llevat (*Saccharomyces cerevisiae*) i els embrions de peix zebra (*Danio rerio*). Primer, s'han investigat els efectes de les condicions de creixement en el metaboloma del llevat. En segon lloc, s'han estudiat els efectes adversos causats en embrions de peix zebra per l'exposició a compostos disruptors endocrins. En aquest darrer cas es presenta la combinació de la metabolòmica i la transcriptòmica per a un estudi més profund dels efectes ocasionats en els embrions de peix zebra.

2.1.1. Origen de les ciències post-genòmiques

L'aparició i la posterior expansió de la terminologia "òmica" a finals del segle XX ha comportat una revolució en molts camps de la ciència i, especialment, en la biologia dels organismes vius a nivell molecular. El neologisme "òmica" prové del sufix "oma" que en llatí significa "totalitat o conjunt de". És a dir, les ciències òmiques fan referència a tots els camps d'estudi de la biologia que proporcionen una visió holística de la fisiologia cel·lular i del funcionament dels sistemes biològics, tals com la genòmica, la transcriptòmica, la proteòmica i la metabolòmica (**Figura 2.1**).

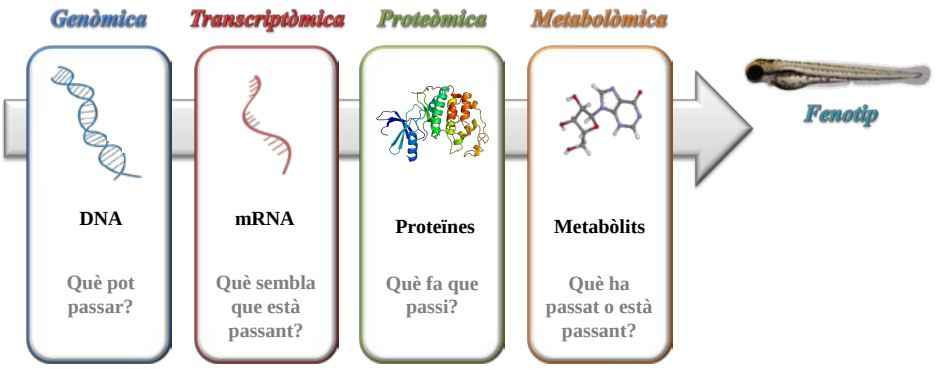


Figura 2.1. Descripció general de les principals ciències de la cascada òmica fins al fenotip.

En concret, la recerca sobre els diferents nivells de la "cascada òmica" és un pilar fonamental per a l'estudi de la biologia dels sistemes per la seva relació amb el dogma fonamental de la biologia

molecular, el qual reflecteix la transmissió de la informació dels gens fins als metabòlits. L'anàlisi integrat de la resposta dels organismes a una pertorbació des del nivell del genoma, del transcriptoma, del proteoma i del metaboloma fins al fenotip dóna un millor comprensió dels mecanismes bioquímics i biològics implicats en els organismes vius. A la **Figura 2.1** es mostra la base i la relació de les disciplines tradicionals de la cascada òmica.

La genòmica és la ciència que implica l'estudi sistemàtic del genoma dels organismes, com l'estructura i les funcions dels gens o la regulació i la transformació de la informació molecular de les seqüències d'àcid desoxiribonucleic (DNA) [1, 2]. La genòmica ha evolucionat fins a l'actual era post-genòmica. En l'etapa pre-genòmica s'investigaven els gens de forma independent: la seva localització cromosòmica i la seva funció i relació amb patologies o estímuls específics. En canvi, l'etapa post-genòmica es centra en el genoma complet i en els seus canvis a diferents condicions [3, 4]. Així, la genòmica actual esta constituïda per dues branques principals: la "genòmica estructural" orientada a la caracterització de les seqüències que conformen el DNA i que permet la generació de mapes genètics dels organismes, i la "genòmica funcional" que es basa en l'obtenció d'informació sistemàtica de les funcions dels gens. Tot i la gran quantitat d'informació que ofereix la genòmica sobre els sistemes biològics i de les seves respostes enfront a condicions específiques, és necessària una caracterització funcional i quantificació més precisa de la transcripció de l'àcid ribonucleic (RNA) (transcriptòmica) [5, 6], de la síntesi de proteïnes (proteòmica) [7, 8] i dels seus productes metabòlics (metabolòmica) [9, 10]. Tant és així que sovint les disciplines de la transcriptòmica, la proteòmica i la metabolòmica es poden agrupar dins de la "genòmica funcional" atès que estudien els productes de l'expressió dels gens. A la **Figura 2.2**. es mostra una representació detallada de la relació entre les diferents ciències òmiques tradicionals i l'expressió gènica. Així doncs, la transcriptòmica, la proteòmica i la metabolòmica conjuntament amb la genòmica són les ciències òmiques essencials per adquirir nous coneixements sobre la funcionalitat cel·lular dels organismes vius i els marcadors biològics de diagnòstic i pronòstic de malalties o de reacció a estímuls externs.

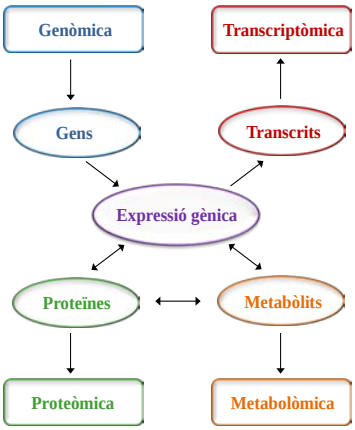


Figura 2.2. Relacions entre el genoma i les diferents tecnologies per avaluar els canvis en l'expressió gènica a nivell de RNA (transcriptòmica), en els nivells de proteïnes (proteòmica) i en les petites molècules (metabolòmica).

La transcriptòmica (la ciència relacionada amb el que pot estar passant davant una determinada situació) ha permès completar la informació gènica que no es podia explicar a través de la genòmica i ha donat lloc als primers estudis multi-òmica [11]. La comparació de les seqüències de DNA i de l'expressió dels transcrits de RNA facilita la identificació dels elements estructurals del genoma i del transcriptoma [11]. En canvi, la proteòmica (la ciència que il·lustra els rols funcionals executats per les proteïnes) ofereix informació complementària a la genòmica i a la transcriptòmica crucial per a la comprensió a nivell molecular de processos biològics més complexos. L'estudi de les proteïnes i de la seva estructura i funció és vital per al coneixement de la biologia dels organismes vius ja que són les molècules responsables del funcionament cel·lular i de les diferents rutes metabòliques [12]. Finalment, la metabolòmica (la ciència que explica el que ha passat o està passant en l'organisme en un moment determinat) permet estudiar els productes finals (metabòlits) de qualsevol funció molecular i ruta metabòlica (manifesta els resultats finals de la informació continguda en els gens). Aquesta disciplina ajuda a comprendre els canvis a nivell de proteoma i reflecteix el comportament de les cèl·lules al nivell més proper al fenotip. Així, permet desxifrar el que ocorre a nivell molecular entre el genoma i el fenotip, el qual descriu les característiques físiques i bioquímiques de l'organisme, com la morfologia, les propietats fisiològiques i el seu desenvolupament i comportament [13]. Aquests diferents trets observables a unes condicions específiques estan determinats per la informació genètica total d'un organisme (genotip) i per la influència dels diferents factors ambientals, així com de la seva interacció [14]. En resum, la "cascada òmica" representa el flux

d'informació biològica que va des del genotip al fenotip, en resposta als possibles estímuls, com són les malalties o les pertorbacions ambientals.

Avui en dia, s'han descrit al voltant de 200 ciències òmiques, moltes de les quals són subdisciplines de les òmiques tradicionals (genòmica, trasncriptòmica, proteòmica i metabolòmica) que estan o poden estar també relacionades amb la biologia dels sistemes [15]. En el marc de òmica ambiental, les subdisciplines de les ciències òmiques tradicionals permeten avaluar les alteracions que pateixen els organismes en resposta al seu entorn físic. Aquesta informació condueix a ampliar el coneixement sobre la fisiologia dels organismes i de les rutes bioquímiques potencialment actives en condicions ambientals determinades [16, 17]. Aquesta Tesi es centrarà en les subdisciplines òmiques ambientals de la transcriptòmica i de la metabolòmica per investigar els efectes de diversos factors i contaminants ambientals.

2.1.2. La transcriptòmica

El projecte del genoma humà (*The Human Genome Project*) va fixar com el seu principal objectiu determinar la seqüència de DNA del genoma humà [18]. La seqüenciació d'aquest genoma va representar el final d'una etapa en l'estudi de la biologia i l'inici de la nova era post-genòmica. Els avenços en la bioinformàtica i el desenvolupament de tècniques de seqüenciació del DNA van permetre que al 2001 es pogués disposar d'un primer esborrany dels 3200 milions de parells de bases de DNA (3200 Mb) que formen el genoma humà [19, 20]. En comparació, per exemple, el genoma del llevat (*Saccharomyces cerevisiae*) presenta una mida petita (~12 Mb) i poc DNA no informatiu [21]. El llevat és el primer microorganisme (unicel·lular) eucariota que va ser seqüenciat. Més endavant, es va seqüenciar el genoma d'altres organismes pluricel·lulars (vertebrats, invertebrats, plantes, etc.), la qual cosa va impulsar el desenvolupament d'altres ciències òmiques que van permetre obtenir un millor coneixement del funcionament d'aquests organismes a partir de la transcriptòmica [22].

La transcriptòmica consisteix en l'anàlisi de tot el conjunt de molècules de RNA expressades (transcriptoma) codificades pel genoma. El terme transcriptoma engloba les molècules de RNA

missatger (mRNA), de RNA ribosòmic (rRNA), de RNA de transferència (tRNA) i possibles molècules de RNA no codificant produïdes per una cèl·lula o organisme en unes condicions concretes. És a dir, la transcriptòmica estudia l'expressió gènica a nivell de RNA i ofereix informació estructural i funcional dels gens. Així, la transcriptòmica permet conèixer les molècules involucrades en els processos biològics mitjançant l'estudi del conjunt total de transcrits de mRNA derivats dels gens que s'expressen en una cèl·lula en un moment concret i en unes condicions fisiològiques determinades [23].

Amb el temps la transcriptòmica ha esdevingut un camp fonamental en el món de la recerca científica per diverses raons [24]. D'una banda, l'anàlisi del transcriptoma reflecteix la dinàmica de l'expressió gènica. La majoria de les cèl·lules comparteixen el mateix conjunt de gens però els patrons de transcripció d'aquests gens poden ser temporals i espacials, cosa que genera una gran diversitat de tipus de cèl·lules amb diferents funcions segons la situació específica. D'altra banda, l'estudi del transcriptoma contribueix a explicar la incoherència entre el nombre de gens codificats i el nombre de proteïnes traduïdes. A més, també és el punt de partida per entendre la regulació de la traducció de mRNA que contenen la informació genètica per a la síntesi de proteïnes [22].

En el context ambiental, la comparació del transcriptoma sota diferents estímuls o condicions ambientals permet la identificació dels gens que s'expressen de forma diferent com a resposta als factors ambientals estressants considerats. D'aquesta forma, la comprensió dels canvis en el transcriptoma és crucial per esdeveniments com la proliferació cel·lular o el desenvolupament dels organismes en condicions adverses.

2.1.3. La metabolòmica

A diferència de la genòmica i de la proteòmica, que són camps d'elevada complexitat degut als controls homeostàtics o processos de regulació, la metabolòmica proporciona informació directament relacionada amb la fisiològia o l'estat patològic (fenotip) de l'organisme [17, 25, 26]. La genòmica o la proteòmica en particular, no proporcionen evidències clares del que està passant en l'organisme en un moment determinat. En canvi, la metabolòmica permet un seguiment qualitatiu i quantitatiu de

"l'empremta química", és a dir, de les respostes metabòliques dels organismes vius enfront factors externs o modificacions genètiques [27, 28]. En conseqüència, la metabolòmica ha esdevingut una de les disciplines primordials per al diagnòstic molecular i per a la detecció de biomarcadors.

La metabolòmica té com a principal objectiu l'estudi de les molècules de baix pes molecular (<1500 Da) (metabòlits) d'una cèl·lula, teixit o organisme en unes condicions determinades [15, 29]. Aquestes condicions abasten diferents pertorbacions, modificacions genètiques, estats patològics i respostes a estímuls externs. Així, els metabòlits es poden considerar com els productes o els intermediaris de processos biològics que donen informació sobre l'activitat bioquímica i que conformen el que és coneix com a metaboloma. El metaboloma és el reflex tant dels mecanismes reguladors ambientals com dels biològics (com l'epigenètica, transcripció, modificacions posttraduccionals), i proporciona una perspectiva molt valuosa del fenotip.

La metabolòmica ha demostrat el seu potencial en àrees molt diverses, com ara en els estudis d'estrès ambiental [9], de toxicologia [30], de nutrició [31], de manipulació genètica [32] i en el diagnòstic de malalties [33], entre d'altres. La metabolòmica ambiental és una subdisciplina recent que es basa en la caracterització de les interaccions dels organismes vius amb el seu entorn [9, 10]. Aquesta ciència presenta molts avantatges en la investigació de la interacció organisme-medi ambient i en l'avaluació de les funcions i l'estat dels sistemes biològics a nivell molecular. Actualment, la metabolòmica és una disciplina que està creixent de forma molt ràpida en el context mediambiental, començant per la investigació de les respostes dels organismes a agents abiòtics, fins als estudis més exhaustius de les respostes dels organismes als cicles estacionals, a l'alimentació, als fàrmacs, a la contaminació o a la simple adaptació a un altre entorn ambiental o biota [9].

Tot i que el metaboloma de la majoria dels organismes conté un nombre relativament reduït de metabòlits endògens (~5000) respecte al nombre total de gens, transcrits o proteïnes d'una cèl·lula o organisme, aquests metabòlits presenten una àmplia gamma de propietats fisicoquímiques (com polaritat, acidesa, solubilitat, etc.) i una gran diversitat de concentracions, les quals dificulten la seva

anàlisi simultània. Això, implica un gran repte en el desenvolupament i aplicació de mètodes analítics robustos i reproduïbles per a la caracterització i quantificació d'aquests compostos [34].

2.1.4. Anàlisi dirigida i no dirigida

En el camp de les ciències òmiques existeixen dues estratègies diferents d'anàlisi: l'anàlisi dirigida (*target*) i l'anàlisi no dirigida o global (*untarget* o *non-target*) [26, 35]. Els objectius d'aquestes dues estratègies són diferents i ambdues presenten avantatges i inconvenients.

L'anàlisi dirigida és freqüentment impulsada per una qüestió o hipòtesi bioquímica concreta que dóna lloc a investigar unes rutes metabòliques específiques. Els mètodes dirigits es basen en l'obtenció d'informació de l'abundància o concentració d'un grup definit i sovint reduït de metabòlits o transcrits de RNA que pertanyen a una o més rutes metabòliques relacionades d'interès.

Aquest enfocament dirigit permet el desenvolupament de mètodes analítics específics per a la detecció i quantificació d'aquests compostos predeterminats mitjançant estàndards comercials [36]. Aquesta estratègia ha estat àmpliament utilitzada, com per exemple en el diagnòstic de malalties [37, 38], en el descobriment de fàrmacs [39, 40] o l'avaluació de l'impacte dels contaminants ambientals [41, 42]. Durant moltes dècades, els mètodes analítics s'han basat únicament en un enfocament dirigit, ja que ofereixen bona sensibilitat i resultats més fiables a l'hora de quantificar compostos diana. Tradicionalment, els mètodes dirigits són preferibles als no dirigits per a l'anàlisi de compostos que es troben a molt baixes concentracions en els sistemes o matrius biològiques. Però aquests mètodes presenten l'inconvenient de no permetre el descobriment de nous biomarcadors [43, 44]. Cal tenir en compte que tots aquells compostos que no hagin estat seleccionats *a priori* no seran investigats. Aquest fet és de crucial importància quan els compostos *target* investigats no proporcionen suficient informació per explicar el que està passant en un determinat organisme sotmès a un estímul o estrès extern. És per això que sovint s'ha de recórrer a l'estratègia no dirigida.

Actualment, la tècnica més emprada per a l'anàlisi dirigida del transcriptoma és la reacció en cadena de la polimerasa quantitativa (qPCR). En canvi, per a l'anàlisi del metaboloma són l'espectroscòpia de ressonància magnètica nuclear (RMN) i l'espectrometria de masses (MS). A la **Figura 2.3a** es

mostra un exemple de les diferents fases o etapes de treball dels estudis de metabolòmica dirigida basats en LC-MS. En aquest cas, s'analitzen primer els estàndards dels metabòlits diana, que s'utilitzen per al desenvolupament i optimització del mètode analític i l'obtenció de les corbes de calibratge corresponents. A continuació, s'analitzen les mostres d'interès i es quantifiquen els metabòlits escollits prèviament. En aquests estudis, generalment s'empren analitzadors de triple quadrupol (QqQ) i es fa l'adquisició de les dades en mode de monitorització de múltiples reaccions (*multiple reaction monitoring*, MRM) per tal d'aconseguir la màxima sensibilitat i selectivitat.

L'anàlisi no dirigida consisteix en l'estudi simultani i global del major nombre possible de transcrits de RNA o de metabòlits presents en una determinada mostra biològica. Aquest enfocament té com a principal objectiu determinar els transcrits o metabòlits del sistema biològic que presenten canvis en l'expressió gènica o en la concentració [45, 46]. Per tant, és una estratègia ambiciosa que intenta fer l'anàlisi simultània d'un gran nombre de molècules amb propietats i concentracions diferents, per així, dilucidar possibles nous biomarcadors i obtenir una visió més completa dels efectes dels agents o estressants externs sobre l'organisme, com ara per exemple els contaminants ambientals.

A diferència de l'estratègia dirigida, l'anàlisi no dirigida genera grans conjunts de dades d'elevada complexitat. Per exemple, en els estudis no dirigits de RNA-Seq (transcriptòmica) i en les tècniques d'espectrometria de masses d'alta resolució (metabolòmica), les dades adquirides arriben a ocupar diversos gigabytes de memòria per cada mostra analitzada. En aquests casos, la inspecció manual de les dades no és factible i requereix del desenvolupament de mètodes de tractament de les dades adquirides que permetin el seu processament i comprensió. Tot i els primers obstacles que es van presentar en el processament i la interpretació de les dades no dirigides procedents d'aquestes anàlisis, els últims avenços han donat lloc al desenvolupament de noves metodologies que han permès avançar en l'extracció de la informació valuosa dels sistemes biològics estudiats. A la **Figura 2.3b** es resumeixen les fases de treball dels estudis de metabolòmica no dirigida de LC-MS i CE-MS d'aquesta tesi.

Si bé l'anàlisi no dirigida és menys sensible que l'anàlisi dirigida, la cobertura de la diversitat de molècules és molt més amplia i els requisits de preparació de mostra (extracció) i de desenvolupament de la metodologia analítica són menors. En aquesta tesi s'ha emprat la perspectiva no dirigida per a l'anàlisi de metabòlits i transcrits de RNA biomarcadors dels sistemes biològics estudiats, el llevat i el embrions de peix zebra.

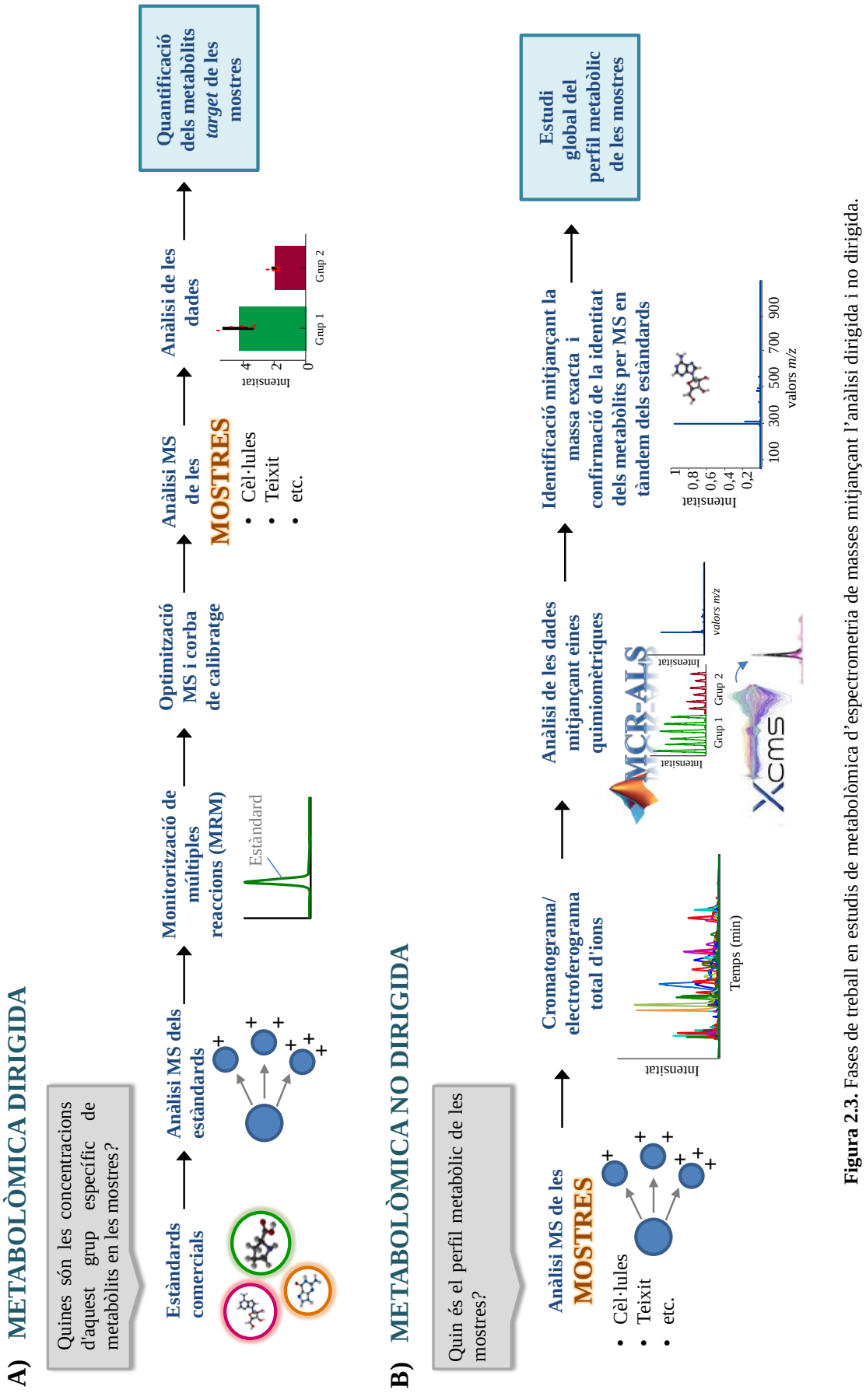


Figura 2.3. Fases de treball en estudis de metabolòmica d'espectrometria de masses mitjançant l'anàlisi dirigida i no dirigida.

2.1.5. Fases de treball en la transcriptòmica i metabolòmica no dirigida

Les diferents fases de treball en l'anàlisi no dirigida o *untarget* del transcriptoma i del metaboloma tenen com a objectiu poder comparar diferents grups biològics que permetin identificar els transcrits (gens) i/o metabòlits que modifiquen significativament la seva concentració en unes condicions especifiques. Per obtenir resultats fiables s'ha de tenir en consideració diversos factors experimentals, els quals es detallen a continuació. A la **Figura 2.4** es representa de manera esquemàtica les diferents fases de treball que s'han seguit en aquesta Tesi doctoral per a l'anàlisi no dirigida del transcriptoma i del metaboloma.

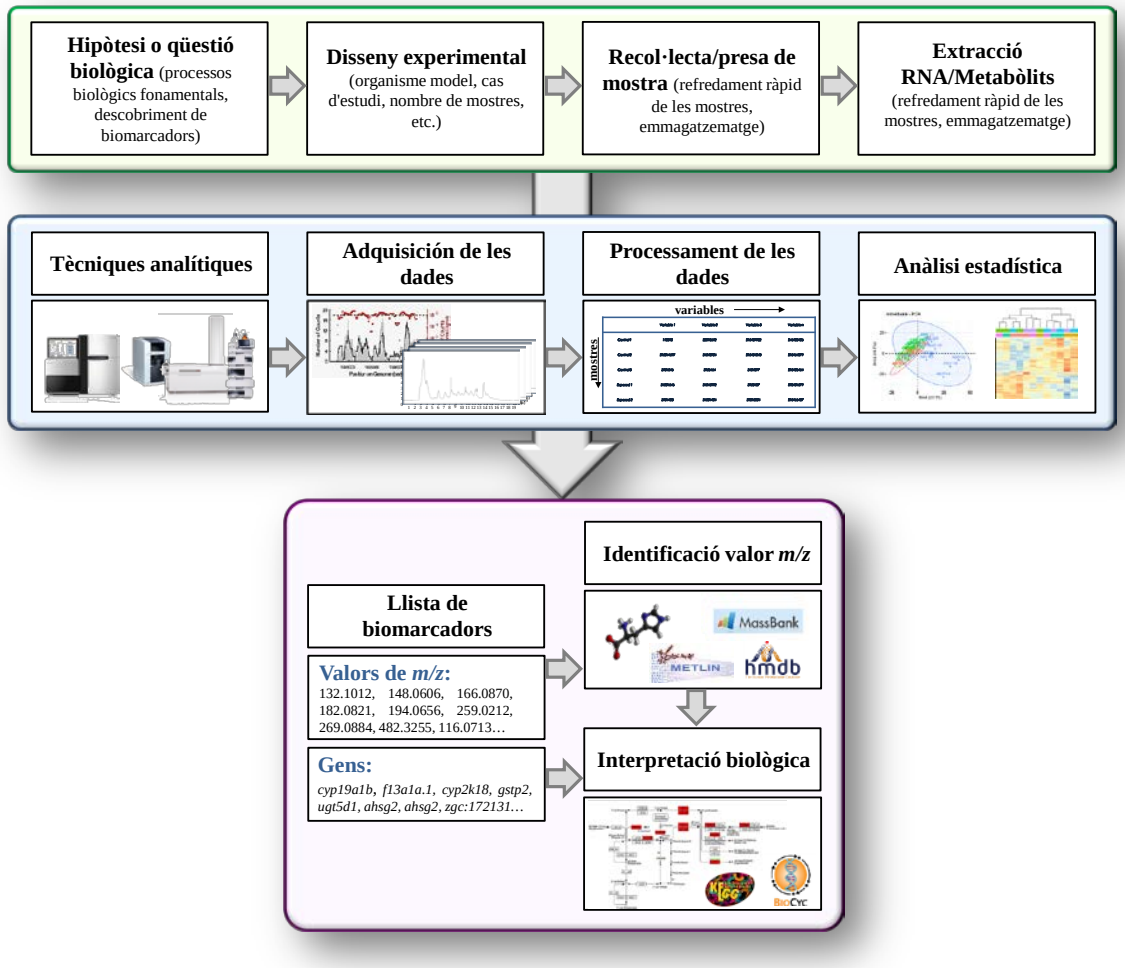


Figura 2.4. Fases de treball dels estudis transcriptòmics i metabolòmics no dirigits.

Primer de tot cal realitzar un disseny experimental minucións basat en una hipòtesi o qüestió a resoldre. El disseny experimental inclou aspectes com:

1. El tipus d'experiment (per exemple, organisme model emprat, cas d'estudi).
2. Els factors experimentals (per exemple, temps d'exposició, dosi).
3. Les comparacions que es volen investigar (per exemple, mostres control i tractades o mostres a diferents dosis o exposicions).

Aquest etapa de disseny experimental porta acompanyada moltes vegades estudis preliminars com per exemple, estudis per conèixer els efectes toxicològics de determinats compostos químics.

Una vegada s'ha concebut l'experiment, les principals etapes que permeten determinar els canvis en els processos biològics dels sistemes estudiats en les investigacions de transcriptòmica i metabolòmica són:

1. Recollida i preparació de les mostres.
2. Implementació de la metodologia analítica.
3. Processament de les dades generades i anàlisi estadística.
4. Identificació dels biomarcadors.
5. Interpretació biològica dels resultats.

Un altre aspecte important en el disseny experimental d'estudis òmics és el de minimitzar les fonts de variació no relacionades amb els processos biològics d'interès. Per tal de reduir les fonts de variabilitat no desitjades, cal tenir en compte algunes consideracions o mesures de control del procés analític. En primer lloc, cal determinar un nombre mínim de replicats biològics per a poder obtenir informació de la variabilitat global de les dades amb la potència estadística desitjada. Sovint es requereix d'un nombre mínim de tres replicats biològics però és aconsellable tenir almenys cinc replicats per generar resultats consistents. En segon lloc, cal realitzar l'anàlisi de les mostres en un ordre aleatori, fixant també el nombre de replicats tècnics (d'anàlisi instrumental) de les mostres per a controlar l'error experimental. A més, en els estudis de metabolòmica no dirigida, també s'aconsella dur a terme l'anàlisi de blancs, l'anàlisi de controls positius i negatius i l'anàlisi de mostres control (*quality control samples*, QC) al llarg de l'experiment que permetin detectar variacions instrumentals i experimentals durant el procés analític [47].

Els controls positius i negatius s'afegeixen a les mostres per així poder normalitzar després les concentracions dels metabòlits detectats. Aquests controls poden ser de diferents tipus. Els patrons subrogats (*surrogates*) són afegits a les mostres abans de l'extracció dels anàlits (metabòlits) de la mostra i s'utilitzen per avaluar i corregir l'efecte de la matriu de la mostra biològica i la precisió en la preparació experimental de les mostres a analitzar. Els patrons o estàndards interns (IS) s'afegeixen a les mostres després de l'extracció i s'utilitzen per compensar les possibles variacions en la resposta instrumental al llarg del temps. Generalment, tant els subrogats com els estàndards interns són compostos químicament similars als metabòlits d'interès, de forma que es comporten de manera similar durant la preparació de mostra i la seva anàlisi instrumental. No han d'interferir amb els anàlits i, a més, no es troben ni en les mostres ambientals ni biològiques analitzades (és el cas per exemple, dels estàndards marcats isotòpicament).

En canvi, les mostres de control de qualitat, QC, són una mescla representativa de tots els metabòlits que es troben a les mostres investigades. S'utilitzen habitualment per controlar l'estabilitat de la plataforma analítica i garantir una bona qualitat de les dades en estudis metabolòmics [48-50]. Idealment, les QC es preparen a partir d'alíquotes de cadascuna de les mostres analitzades [51] però quan les quantitats de les mostres són limitades o es tracta d'estudis a gran escala (de centenar o milers de mostres) on les anàlisis han de durar diversos mesos, cal utilitzar mostres QC alternatives. En aquestes situacions, sovint s'utilitza una mostra control externa comercial (mostra biològica comercial) [48]. En cas que no es disposi de cap de les dues opcions esmentades, la solució més comuna és emprar una mescla de metabòlits sintètics, que contingui compostos de cadascuna de les famílies que s'espera trobar en les mostres i estigui preparada sota les mateixes condicions experimentals que les mostres estudiades.

2.1.6. Organismes model i estrès ambiental

L'òmica ambiental aprofita les eines de la genòmica, la transcriptòmica, la proteòmica i la metabolòmica per a identificar noves rutes metabòliques o indicadors precoços de trastorns o alteracions biològiques (biomarcadors) amb la finalitat d'entendre els mecanismes d'acció d'agents

externs i predir els riscos ecològics associats a la seva exposició. En concret, en aquesta tesi s'han aplicat la transcriptòmica i metabolòmica ambiental per a la cerca de nous biomarcadors relacionats amb agents externs o factors ambientals estressants.

La transcriptòmica i la metabolòmica ajuden a resoldre qüestions crucials, com els mecanismes de prevenció dels possibles riscos associats a contaminants químics, la identificació de nous biomarcadors ambientals o la optimització de recursos naturals per protegir la salut humana, la fauna i el medi ambient de forma sostenible [52, 53]. Per intentar respondre a aquestes preguntes, s'han establert una sèrie d'organismes model *in vivo* que faciliten dur a terme els estudis experimentals en el camp de la investigació microbiana, animal i vegetal [54].

Els organismes model són un component essencial en la recerca biològica, biomèdica i mediambiental per a estudiar processos biològics. El coneixement d'aquests organismes model permet entendre els seus sistemes i processos biològics particulars, amb l'expectativa de proporcionar informació del funcionament d'aquestes espècies i extrapolar aquests coneixements a altres organismes, com l'ésser humà [55]. Quan l'experimentació amb humans no pot seguir els requeriments ètics [56], els models *in vivo* són una alternativa per a l'estudi, per exemple, de les causes de malalties o els efectes d'estressants ambientals.

Generalment, l'ús d'organismes model presenta diversos avantatges experimentals en el laboratori. Les principals característiques perquè una espècie sigui considerada un organisme model són [56]:

1. Fàcil creixement i manutenció.
2. Temps de reproducció ràpid.
3. Creixement ben establert i desenvolupament conegut.
4. Semblança a altres organismes d'interès.
5. Informació disponible dels diferents nivells òmics.

No obstant això, la selecció de l'organisme model adequat en cada cas no és una tasca senzilla i cal tenir en compte diverses consideracions. Primer, els processos biològics d'interès que es volen

comparar han d'estar clarament identificats. Finalment, aquests processos han d'estar ben caracteritzats en els diversos organismes model [57, 58].

En l'actualitat, la majoria d'investigacions en el món de l'òmica i la biologia cel·lular utilitzen només un grup petit d'organismes model representatius dels diferents regnes, que incloen el *Saccharomyces cerevisiae*, l'*Arabidopsis thaliana*, la *Drosophila melanogaster*, el *Caenorhabditis elegans*, el *Danio rerio*, el *Mus musculus*, entre d'altres [59]. En concret, la *Metabolomics Society's Model Organism Metabolomes* aconsella els organismes model que es mostren a la **Taula 2.1** [54].

Taula 2.1. Llista d'organismes model recomanats en el camp de la metabolòmica.

Regne	Nom llatí	Nom comú
Bacteri	*Escherichia coli*	-
Fong	*Saccharomyces cerevisiae*	Llevat
Animals (invertebrats)	*Caenorhabditis elegans*	Cuc nematode
	Daphnia magna	Puça d'aigua
	Drosophila melanogaster	Mosca comuna de la fruita
Animals (vertebrats)	*Danio rerio*	Peix zebra
	Mus musculus	Ratolí
Plantes	*Arabidopsis thaliana*	Arabidopsis
	Medicago truncatula	Melgó truncat
	Oryza sativa	Arròs
	Solanum lycopersicum	Tomàquet

En aquesta Tesi, s'han escollit *Saccharomyces cerevisiae* i *Danio rerio* per a l'estudi dels efectes de factors o contaminants ambientals a nivell de metaboloma i transcriptoma. A continuació es descriuen els dos organismes model mencionats.

Llevat (*Saccharomyces cerevisiae*)

El llevat (*S. cerevisiae*) és un microorganisme eucariota unicel·lular del regne dels fongs (*Fungi*) conegut pel procés de fermentació, propietat que s'ha explotat des de l'antiguitat en la producció del pa i de begudes com la cervesa i el vi. *S. cerevisiae* pertany al fílum *Ascomycota*, a la classe *Saccharomycetes*, a l'ordre *Saccharomycetales* i a la família *Saccharomycetaceae*. Tot i que s'han

descrit aproximadament unes 1.000 espècies dins del regne dels *Fungi*, l'espècie *S. cerevisiae* ha guanyat protagonisme en els laboratoris convertint-se en un poderós model biològic dels organismes eucariotes pels seus múltiples avantatges. *S. cerevisiae* posseeix pràcticament totes les característiques que defineixen un organisme model i d'altres que el fan gairebé únic, com un genoma compacte amb una genètica senzilla. Aquestes propietats fan que sigui un organisme especialment adequat per a la genòmica, la transcriptòmica, la proteòmica i la metabolòmica. A més, des de que el llevat va ser introduït com a organisme experimental ha esdevingut un excel·lent organisme model en estudis fonamentals en bioquímica, biologia molecular i cel·lular i recerca aplicada, fins i tot més que *E. coli* (model procariota) [60, 61]. *S. cerevisiae* és habitualment l'organisme d'elecció per desenvolupar i validar noves tecnologies i per investigar processos biològics fonamentals [62]. La principal demostració de la utilitat d'aquest organisme en el món científic és que va ser el primer organisme eucariota del qual es va seqüenciar completament el seu genoma, conformat aproximadament per 12 milions de bases (Mb) que codifiquen aproximadament 6.600 gens.

Condicions de creixement de *S. cerevisiae*

El cicle vital i l'estratègia reproductiva de *S. cerevisiae* es veuen condicionats per diversos factors, com la disponibilitat de nutrients o les condicions ambientals, alguns dels quals s'han investigat en aquesta Tesi. Els factors més importants són la presència d'oxigen, el pH, la temperatura i la composició del medi de creixement. Aquests paràmetres produeixen canvis en els perfils metabòlics, transcripcionals, traduccionals i posttraduccionals de les cèl·lules, així com en el seu desenvolupament [63].

S. cerevisiae es caracteritza per ser un organisme anaeròbic, capaç de desenvolupar un metabolisme oxidatiu en presència d'oxigen i metabolisme fermentatiu en la seva absència. Així, l'oxigen és essencial per a un bon creixement del llevat i és la força impulsora de molts aspectes del seu metabolisme, inclosa la fermentació. L'oxigen és absorbit ràpidament pel llevat i s'utilitza per a sintetitzar àcids grassos insaturats i esterols que formen la membrana cel·lular. Aquestes molècules són importants tant pel creixement com per a la fermentació i serveixen com a mitjà

d'emmagatzematge d'oxigen dins de la cèl·lula. L'esgotament d'oxigen porta a la interrupció de la respiració i dóna lloc a la fermentació.

Les condicions de pH i de temperatura són també paràmetres rellevants en el creixement del llevat i el seu metabolisme. Les condicions òptimes de creixement per *S. cerevisiae* són pH neutre o lleugerament àcid i una temperatura al voltant de 30-35°C. En aquesta Tesi s'han estudiat els canvis en els perfils metabòlics del llevat a 30°C (temperatura òptima) i a 37°C (temperatura estressant). A aquestes temperatures elevades, les condicions no són òptimes per a la fermentació dels llevats ja que es produeixen un gran nombre d'èsters els quals afecten a la seva viabilitat i estabilitat [64, 65].

D'altra banda, la disponibilitat de nutrients, com sucres, aminoàcids i compostos de nitrogen, és clau en el desenvolupament i creixement de les cèl·lules de llevat. *S. cerevisiae* també pot presentar un canvi de resposta similar a la manca d'oxigen degut al nivell de nutrients del medi. Les cèl·lules de llevat creixen en un medi pobre en carboni i sals que proporcionen nitrogen, fòsfor i traces de metalls. Aquest creixement és molt més ràpid en presència d'un medi ric que contingui extracte de llevat i bactopeptona. La font de carboni s'obté majoritàriament a partir de sucres d'hexosa, com ara la glucosa o la fructosa. En medis rics en font de carboni (glucosa o altres sucres fermentables), la fermentació i la reproducció asexual per gemmació representen avantatges evolutius. En canvi, quan la glucosa o els sucres fermentables s'esgoten, es produeix el pas d'un metabolisme fermentatiu a un de respiratori. En la respiració, els sucres i l'etanol produït són metabolitzats a diòxid de carboni amb consum d'oxigen, un procés molt més eficient que la fermentació però més lent. Per exemple en el cas estudiat en aquesta Tesi, quan una població de cèl·lules en creixement exponencial en medi de glucosa (medi ric) es transferida a un medi on la font de carboni és pobra (pocs àtoms de carbonis) o no és fermentable, com l'acetat (compost amb dos àtoms de carboni, compost C2), el llevat es desplaça des d'un metabolisme fermentatiu fins a un aeròbic [64, 66].

Peix zebra (*Danio rerio*)

El peix zebra (*D. rerio*) és un peix tropical d'aigua dolça de la família dels ciprínids (*Cyprinidae*) proposat a la dècada dels anys 70 com a organisme model vertebrat en investigacions biològiques. En

gran part, les principals virtuts de *D. rerio* es van reconèixer gràcies a l'esforç i dedicació de George Streinsinger i els seus companys de la Universitat d'Oregon [67]. No obstant això, no va ser fins a les dècades del 1980 i del 1990 quan el peix zebra va aconseguir definitivament un lloc de privilegi en els laboratoris. *D. rerio* forma part del regne *Animalia*, del fílum *Chordata*, de la classe *Actinopterygi*, de l'ordre *Cypriniformes*, de la família dels *Cyprinidae* i del gènere *Danio*. Aquest ciprínid actinopterigi (amb aletes radiades) rep també comunament el nom de peix zebra degut al seu aspecte ratllat. A la **Figura 2.5** es pot veure que aquest peix adult presenta als laterals entre 5 i 9 bandes uniformes de color blavós superposades al color de fons que, en els mascles, és daurat i, en les femelles, és platejat. La zona ventral és de color blanquinós i rosat.

Figura 2.5. Fotografia del (a) dorsal i (b) lateral d'un embrió de 5 dies i d'un (c) peix adult.

El peix zebra és natiu del sud-est asiàtic i es troba principalment al voltant de les conques dels rius Ganges i Brahmaputra al nord-est de l'Índia, al Pakistan, a Bangladesh, al Nepal i a Myanmar [68]. *D. rerio* es troba àmpliament distribuït en aigües poc profundes. Els seus hàbitats naturals més freqüentment són els canals, rierols, sèquies i basses de conreu d'arròs [69]. Els peixos zebra són omnívors però s'alimenten principalment de zooplàncton i d'insectes. En els aquaris de laboratori, necessiten aigua a pH neutre (al voltant de 7), lleugerament dura i a una temperatura entre 22 i 30 ºC.

Morfològicament, els *D. rerio* són peixos petits i allargats. Als 3 o 4 mesos d'edat, adquireixen maduresa sexual i les femelles adultes poden reproduir-se per aparellament generant fins a 200 o 300 ous per setmana. La fecundació de l'ou és iniciada per un període de llum després d'un període de foscor. En les etapes embrionàries i larvàries, el peix zebra té aproximadament una longitud de 1 a 4 mm.

En aquesta Tesi s'ha treballat amb embrions de peix zebra (veure **Figura 2.5a** i **b**) per la seva similitud genètica als éssers humans. El desenvolupament embrionari de *D. Rerio* segueix unes fases ben caracteritzades i un estricte patró temporal definit per les primeres hores de desenvolupament en unitats de hores posteriors a la fertilització (hpf) i, posteriorment, en unitats de dpf (**Figura 2.6**) [70]. Primer, hi ha un estadi unicel·lular (0-0,7 hpf) (1), on una vegada els ous han estat fertilitzats inicien una primera fase zigòtica. Una segona fase de clivellament (0,7-2,2 hpf) (2) que comença després de la primera divisió, on les cèl·lules o blastòmers es divideixen cada 15 minuts aproximadament fins a un total de sis divisions. Després es produeix la formació de la blàstula (2,25-5,25 hpf) (3), nom que rep el blastodisc des de la vuitena divisió zigòtica fins a la fase de gàstrula a la catorzena divisió cel·lular del zigot. En la gastrulació (5,25-10 hpf) (4) l'embrió adquireix una orientació axial amb tres capes germinals (endoderma, ectoderma i mesoderma). La capa més interna o endoderma formarà l'aparell respiratori i digestiu excepte la boca i la faringe. La capa exterior o ectoderma formarà el cervell i el sistema nerviós. Entre l'endoderma i l'ectoderma es troba la mesoderma que formarà la notocorda, el sistema circulatori i l'aparell excretor entre d'altres. A continuació, en la fase de segmentació (10-24 hpf) (5), s'inicia una organogènesi primària on es formen els somites i la cua es comença a desenvolupar. Després en l'estadi de faríngula (24-48 hpf) (6), s'inicia la formació de les aletes pectorals, apareix la pigmentació en general i s'acaba produint l'eclosió (7). En l'eclosió (48-72 hpf) els embrions s'alliberen del còrion. Des d'aquest moment fins a l'etapa larvària primerenca (5 dpf), en que l'organisme pot alimentar-se per si mateix aprofitant els nutrients del sac vitel·lí [71, 72], es coneix com eleuteoroembrió (8). Finalment, hi ha la fase larvària (5-30 dpf) (9) que dura fins a l'entrada de la fase juvenil i és quan es produeixen els canvis cap a la fase adulta (10).

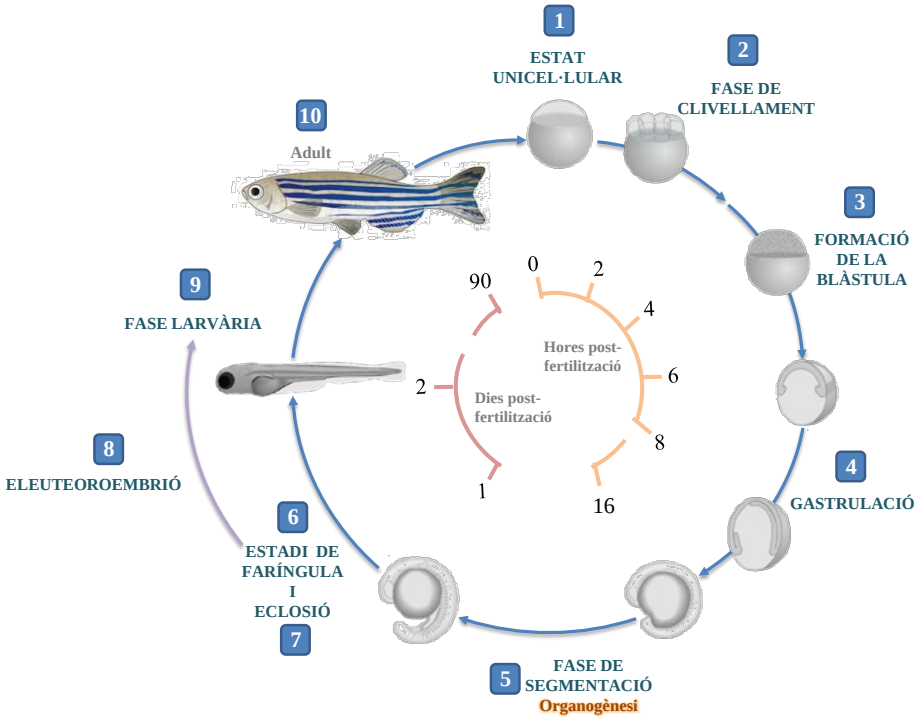


Figura 2.6. Desenvolupament embrionari de *D. Rerio*, adaptada de [73].

Tot i que des dels anys setanta el *D. rerio* ha estat àmpliament utilitzat en investigacions de la biologia del desenvolupament i en l'embriologia, en els darrers anys ha esdevingut un organisme comú en investigacions del camp de la genètica, de les ciències ambientals, de toxicologia i, sobretot, en el descobriment de nous fàrmacs *in vivo* [74, 75]. Així, s'ha investigat la rellevància i predictibilitat de la resposta del peix zebra i dels humans exposats a fàrmacs, i s'ha demostrat que el peix zebra presenta una alta predictibilitat (al voltant del 75%) de la resposta dels humans enfront el consum de medicaments [72, 76].

D. Rerio posseeix totes les propietats i avantatges dels organismes model i d'altres que el fan únic en el món de la recerca, com la gran quantitat de cries, la transparència dels embrions i l'accés a la manipulació experimental amb finalitats científiques fins les 120 hpf d'acord amb l'estricta legislació europea sobre la protecció dels animals [77]. L'estructura morfològica i els òrgans interns es poden visualitzar fàcilment mitjançant microscòpia atesa la transparència dels embrions i del còrion que els embolcalla. A més, les característiques fisiològiques de *D. Rerio* són homòlogues als éssers humans. De fet, el peix zebra representa una alternativa viable a altres models mamífers utilitzats, per exemple,

en proves de toxicitat ja que la seva manutenció és més fàcil i menys costosa que en els models de rosegadors. A més, la inducció enzimàtica i l'expressió gènica de *D. rerio* poden monitoritzar-se fàcilment i molts biomarcadors intracel·lulars, com el glutatió, la formació de proteïnes i adductes de DNA es poden estudiar en tot l'animal. Els sistema cardiovascular, el sistema nerviós i les vies metabòliques són molt semblants als dels mamífers a nivell anatòmic, fisiològic i molecular [72, 78].

Els avenços tecnològics en la biotecnologia molecular i genètica han desembocat en un canvi en el camp de la toxicologia des de les caracteritzacions de les respostes moleculars a les pertorbacions químiques. D'aquesta manera, s'ha arribat a una comprensió més profunda del que passa entre el genotip i el fenotip. Les propietats extraordinàries del peix zebra han permès una millor comprensió dels canvis en els processos bioquímics, moleculars i funcionals.

Recursos com ara l'anotació i seqüenciació del genoma del peix zebra, faciliten el seu ús extensiu com a model vertebrat en diversos nivells òmics, com la transcriptòmica, la proteòmica i la metabolòmica [79]. A dia d'avui s'ha arribat a seqüenciar el 85-90% del genoma de *D. rerio*, el qual s'estima que conté 1.700 milions de parells de bases (Mb) que corresponen a uns 14.000 gens. El genoma de *D. rerio* és molt similar al genoma humà (aproximadament un 87% de similitud). La informació genòmica d'aquesta espècie es reuneix en diferents plataformes, com la base de dades *The Zebrafish Information Network* (ZFIN) la qual permet la caracterització d'espècies mutants [80].

Els principal avantatges esmentats anteriorment juntament amb l'extraordinària permeabilitat de *D. rerio* fan que aquest organisme sigui també un model adequat per finalitats toxicològiques relacionades amb el risc mediambiental derivat de l'ús de productes químics [81, 82]. Fins als 3 dpf la principal ruta d'exposició química de *D. Rerio* és la capa dèrmica. A partir d'aquí, també es pot dur a terme mitjançant la ingesta dels contaminants.

El peix zebra no s'ha utilitzat únicament com a organisme model vertebrat de toxicitat [83, 84] sinó que serveix també com a espècie d'assaig ecotoxicològic per determinar els efectes de substàncies químiques sobre la supervivència dels peixos, el creixement i la reproducció [85, 86]. Actualment, les proves de toxicitat amb embrions de peix zebra són considerades una gran alternativa per reduir o

reemplaçar estudis de toxicitat *in vivo* amb peixos [87]. Altres aplicacions del peix zebra en toxicologia inclouen la investigació de mecanismes tòxics de fàrmacs i productes químics [84]. El peix zebra s'utilitza per a l'anàlisi dels efectes subletals de contaminants emergents (ECs) que poden ser greus per a la supervivència de les espècies vives o la salut humana. Els compostos disruptors endocrins (*endocrine disruptors compounds*, EDCs), per exemple, generen especial preocupació. En concret, en aquesta Tesi s'han emprat embrions de peix zebra per avaluar els efectes de diferents EDCs.

Compostos disruptors endocrins (EDCs)

La presència de contaminants en el medi ambient és un tema de gran rellevància social. Entre els compostos freqüentment trobats en el medi ambient, els EDCs han suscitat una elevat interès ja que està demostrat que poden interferir amb el sistema endocrí dels éssers vius [88]. L'ús d'organismes model en combinació amb assajos toxicològics confereix una opció molt valuosa en les línies de recerca ambientals destinades a l'estudi dels EDCs. Els models aquàtics, en especial els crustacis com la *Daphnia magna* [89, 90] o els vertebrats com el peix zebra (*D. rerio*) [91, 92], són extensament emprats en aquest tipus d'investigacions. Per exemple, el coneixement de les propietats moleculars del sistema endocrí i les vies de senyalització hormonals del peix zebra fan que aquesta espècie sigui l'organisme model per excel·lència per a la investigació dels mecanismes d'acció dels EDCs [84, 93, 94].

Avui en dia, la preocupació pels efectes perjudicials dels EDCs en la salut humana, la fauna i el medi ambient ha adquirit especial rellevància. Els EDCs són compostos exògens que causen danys directes als animals o inicien processos anormals del sistema endocrí degut a desequilibris hormonals que afecten els mecanismes homeostàtics, el creixement, el desenvolupament i el sistema reproductor dels animals [94, 95]. A més, les concentracions d'EDCs presents en el medi ambient poden contribuir en desordres metabòlics com l'obesitat, la diabetis i la disfunció metabòlica [96, 97]. Entre els EDCs, hi ha una amplia gamma de compostos que inclouen, organoclorats, dioxines, organotins, compostos perfluorats (PFCs), retardants de flama bromats (BFRs), alquilfenols, bisfenol A (BPA) i ftalats [98].

La majoria d'aquest compostos estan totalment prohibits o el seu ús està regulat sota condicions estrictes. La prohibició o restricció d'aquests compostos depèn dels efectes adversos o toxicològics i del seu potencial sobre la salut humana i la fauna. Cal destacar que la prohibició d'aquests compostos químics limita immediatament l'exposició, però es requereixen molts anys perquè desapareguin completament del medi ambient. A la **Taula 2.2** es descriu la legislació aplicada als tres EDCs (BPA, sulfonat de perfluorooctà (PFOS) i tributilestany (TBT)) investigats en aquesta Tesi.

Taula 2.2. Legislació aplicada als diferents EDCs investigats en aquesta Tesi.

EDC	Tipus/origen	Legislació
Bisfenol A	Plastificants	2009: Canadà es converteix en el primer país en prohibir el BPA en ampolles de beure per nadons i la Organització Mundial de la Salut (OMS) comença a avaluar els riscos del BPA. 2014: La EU estableix un límit de migració del BPA en joguines infantils, el qual es torna a revisar el 2017.
PFCs (per exemple, sulfonat de perfluorooctà, PFOS)	Plastificants	2009: Restringits pel Conveni d'Estocolm i la Unió Europea (UE).
Organotins (per exemple, tributilestany, TBT)	Contaminants ambientals en aliments	2003: Prohibits com *antifouling* en vaixells per l'Organització Marítima Internacional (OMI). 2008: Prohibició completa per OMI.

Des del punt de vista de la recerca, en els darrers anys s'han dut a terme molts esforços per establir noves tecnologies per a la detecció d'EDCs en teixits humans, animals o matrius ambientals, especialment en dosis baixes i durant el període de desenvolupament dels organismes [94, 99]. Actualment, nous enfocaments que combinen les diferents òmiques, la biologia de sistemes i la modelització computacional ajuden a comprendre la complexitat dels efectes i les conseqüències de l'exposició a aquests compostos en les diferents etapes de la vida dels éssers vius [79, 100, 101]. En concret, en aquesta Tesi s'han avaluat els efectes del BPA, del PFOS i del TBT en l'organisme model vertebrat *D. rerio* mitjançant una anàlisi transcriptòmica i metabolòmica.

2.2. TÈCNIQUES ANALÍTIQUES

En els camps de la transcriptòmica i la metabolòmica diverses tècniques han estat extensament emprades per a la caracterització de sistemes bioquímics, la identificació de biomarcadors i l'estudi de processos fisiològics. Durant l'última dècada, ha augmentat la demanda de tècniques analítiques d'alt rendiment per obtenir mètodes més sensibles i selectius per a l'anàlisi de mostres complexes, i poder així abastar una millor comprensió dels processos moleculars.

2.2.1. Anàlisi transcriptòmica

La informació genètica codificada en el DNA i emmagatzemada en els gens s'expressa a través dels mecanismes de transcripció i traducció, a partir dels quals es produeixen molècules de mRNA i proteïnes, respectivament. El mRNA té un paper clau en la regulació de l'expressió gènica i en importants processos de la biologia cel·lular. En conseqüència, la transcriptòmica té com a objectiu quantificar els nivells d'expressió dels gens mitjançant tècniques que permeten analitzar el conjunt de mRNA resultant de la transcripció del DNA del genoma.

L'estudi i l'anàlisi del transcriptoma és crucial per entendre la funció dels gens. De forma general, es pot establir que si un gen s'expressa en una condició o cèl·lula determinada és perquè efectua alguna funció. L'estudi global del transcriptoma també permet establir patrons de regulació gènica coordinada, que contribueixen a dilucidar la funció i l'agrupament de diversos gens sota un estímul o condició específica i a identificar elements promotors comuns en diversos gens.

Actualment, els tres mètodes més comuns d'anàlisi del transcriptoma són el mètode dirigit o *target* de la reacció en cadena de la polimerasa quantitativa (qPCR) [102] i els mètodes no dirigits o *untarget* dels xips de DNA [103, 104] i de la seqüenciació de RNA (RNA-Seq) [105, 106]. La tècnica de qPCR és una metodologia dirigida que utilitza seqüencies curtes de DNA anomenades encebadors per amplificar els transcrits coneguts d'una mostra. La quantificació ràpida de centenars de transcrits s'obté de forma econòmica però es requereix d'un coneixement previ de les seqüències dels transcrits d'interès. L'aparició dels xips de DNA i de la RNA-Seq han fet que la transcriptòmica ofereixi una visió més global i extensa dels transcrits de RNA. Aquestes metodologies no dirigides són actualment

les tècniques dominants en el camp de la transcriptòmica, ja que permeten avançar en el coneixement de la dinàmica de l'expressió dels gens i de la seva influència en el desenvolupament de molts processos biològics [107]. A més, ambdues tècniques milloren la comprensió de la interacció entre els factors genètics i els factors mediambientals, la qual cosa proporciona una millor informació a nivell global d'expressió gènica [108]. Els xips de DNA permeten l'anàlisi simultània de l'expressió de milers de gens a nivell de mRNA de forma econòmica mitjançant seqüències específiques de DNA, conegudes com a sondes d'hibridació (*probes*). Cadascuna de les sondes d'hibridació es localitza de forma inequívoca en un punt del xip i representa un gen del qual es vol mesurar la seva expressió [23, 24, 109, 110]. Tanmateix, aquesta tècnica inclou moltes fonts de variabilitat i requereix de l'ús d'eines estadístiques en el disseny experimental i en l'anàlisi de les dades [111]. A més, al ser una tècnica basada en el concepte d'hibridació presenta una moderada cobertura del transcriptoma i necessita de coneixement previ de les seqüències de transcrits existents en l'organisme per a la seva implementació. En canvi, la tècnica de RNA-Seq al no necessitar transcrits de referència ha revolucionat el camp de la transcriptòmica donant lloc a una perspectiva totalment holística.

La metodologia de RNA-Seq ha estat la tècnica emprada en aquesta Tesi per a l'estudi del transcriptoma d'embrions de peix zebra exposats a bisfenol A.

RNA-Seq

Actualment, gràcies als avenços en les tècniques de seqüenciació de DNA a través de les tecnologies de seqüenciació massiva de nova generació, NGS (*Next Generation Sequencing*), s'han revolucionat els camps de la genòmica i la transcriptòmica. Aquestes tecnologies de nova generació permeten generar informació d'alta resolució i, també, obrir nous horitzons en la comprensió detallada i global dels processos d'expressió gènica [112]. La caracterització completa de l'expressió gènica en una cèl·lula o teixit i el seu anàlisi global és possible a través de la implementació de la seqüenciació de cDNA o, més recentment, de la seqüenciació directa de RNA mitjançant la tècnica de RNA-Seq [105, 106]. Aquesta tècnica ha canviat la manera d'entendre i estudiar el transcriptoma ja que és una tècnica capaç de donar una cobertura completa dels transcrits. Així, permet generar informació no només de

la seqüència de RNA, sinó també de l'estructura d'exons i possibles esdeveniments d'empalmament (*splicing*) alternatiu [113, 114].

En els últims anys, la tècnica RNA-Seq ha tingut un desenvolupament accelerat i ha esdevingut el principal mètode d'elecció per a l'estudi i la caracterització del transcriptoma desbancant poc a poc els àmpliament emprats fins ara xips de DNA. La RNA-Seq és una eina de seqüenciació basada en NGS que captura i quantifica els transcrits presents en un extracte de RNA [106]. Així, aquesta tècnica contribueix en una estimació més precisa de l'expressió gènica i de les diferents isoformes dels transcrits. A més, RNA-Seq no depèn de l'anotació prèvia del genoma per la selecció de sondes com en els xips, per tant, evita les possibles fonts de variabilitat que es produeixen durant el procés d'hibridació [115].

L'anàlisi del transcriptoma per RNA-Seq emprat en aquesta Tesi consta de tres etapes [116] (**Figura 2.7**):

1. Construcció d'una llibreria.
2. Seqüenciació mitjançant una plataforma de NGS específica.
3. Anàlisi de les dades generades.

Primer, es captura el RNA total o mRNA, el qual es fragmenta i es converteix en una llibreria de cDNA. Una de les etapes fonamentals d'aquesta tècnica és l'obtenció de RNA de bona qualitat de la mostra que representi tots els transcrits que es produeixen en les condicions estudiades. Per a l'extracció del RNA sovint es fan servir kits d'extracció de mRNA que aïllen l'RNA de l'organisme abans d'analitzar els transcrits. La fragmentació del RNA o del cDNA es realitza per nebulització, per digestió amb enzims de restricció o a través de l'ús de cations divalents sota condicions elevades de pressió [105]. Generalment, el fraccionament es realitza posteriorment a la síntesi de cDNA. Aquesta síntesi es realitza mitjançant procediments estàndards ben establerts que utilitzen l'enzim transcriptasa inversa.

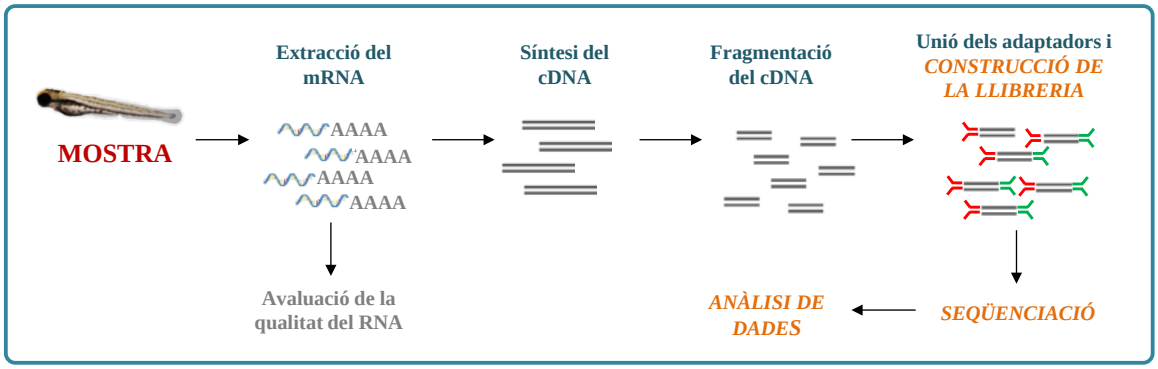


Figura 2.7. Esquema general del procés de seqüenciació de RNA-Seq.

Una vegada obtingut el cDNA, s'uneixen adaptadors (per activar la seqüenciació) de tal manera que cada fragment generat contindrà un adaptador unit en els seus extrems 3' i 5'. Les seqüències d'aquests adaptadors es coneixen i seran necessàries per a la seqüenciació de cada fragment, i en alguns casos es poden també utilitzar per diferenciar altres grups de fragments procedents de mostres de cDNA diferents. Les seqüències conegudes es poden unir directament a la mostra de RNA, prèviament a la síntesi de cDNA [117, 118] o es pot afegir directament a la cadena de cDNA [118, 119].

Les dades generades per RNA-Seq tenen més precisió que els nivells d'expressió gènica que es poden obtenir per qPCR i els resultats són molt més reproduïbles [105, 120]. D'altra banda, les seqüències de nucleòtids generades són aproximadament de 100 parells de bases (pb) de longitud però poden variar entre 30 pb i més de 10.000 pb segons el mètode de seqüenciació emprat. El rang dinàmic de RNA-Seq és de 5 ordres de magnitud, un avantatge clau respecte els transcriptomes obtinguts mitjançant xips de DNA de 3 a 4 ordres de magnitud. A més, la quantitat de mostra de RNA que és necessita és molt menor (quantitat de ng) en comparació amb la necessària en els xips (quantitat de µg) [23].

Hi ha diverses plataformes comercials de seqüenciació per RNA-Seq, tals com Roche 454 (Roche), Solexa (Illumina), SOLiD (Thermo Fisher Scientific), Ion Torrent (Thermo Fisher Scientific) i PacBio (PacBio) [6, 23]. Les tecnologies de NGS més comunament emprades són Roche 454 (Roche) i Solexa (Illumina) [121]. En aquesta Tesi s'ha utilitzat la plataforma d'Illumina del Centre Nacional

d'Anàlisi Genòmica (CNAG) per portar a terme les anàlisis de RNA-Seq d'embrions de peix zebra exposat a bisfenol A.

2.2.2. Anàlisi metabolòmica

En les últimes dècades s'han emprat moltes tècniques analítiques per aplicacions metabolòmiques, com la ressonància magnètica nuclear (RMN) [122], l'espectroscòpia d'infraroig amb transformada de Fourier (FT-IR) [123] [124] i l'espectrometria de masses (MS) acoblada a tècniques de separació o mitjançant la injecció directa (*direct-injection mass spectrometry*, DIMS) [125]. L'espectrometria de masses (MS) i la ressonància magnètica nuclear (RMN) són les tècniques més freqüents i juguen un paper molt rellevant en l'estudi del metaboloma [126-128]. L'espectroscòpia de RMN és una tècnica quantitativa recomanada per l'alta reproductibilitat dels resultats i la seva capacitat d'elucidar estructures químiques, però presenta una baixa sensibilitat respecte a l'espectrometria de masses [127].

L'espectrometria masses ha esdevingut una eina clau en la metabolòmica, degut també a la precisió, selectivitat i potencial d'identificació de compostos desconeguts. A més, el nombre de metabòlits detectats en un cas concret es pot millorar aprofitant l'ampli ventall de metodologies de MS existents gràcies a les diferents fonts d'ionització i analitzadors comercialment disponibles [129, 130].

Ara bé, l'ús de tècniques de separació prèviament a l'espectrometria de masses és sovint essencial degut a la complexitat de les mostres biològiques. Entre les metodologies basades en tècniques de separació acoblades a l'espectrometria de MS més emprades per dur a terme estudis de metabolòmica, cal destacar les tècniques de GC-MS, de LC-MS i de CE-MS, cadascuna de les quals presenta avantatges i inconvenients.

Tècniques de separació en el camp de la metabolòmica

L'èxit de la metabolòmica no dirigida ha estat en gran part possible degut a la combinació de tècniques de separació acoblades a l'espectrometria de masses, les quals faciliten l'anàlisi de les múltiples possibles famílies de metabòlits. La cromatografia de gasos (GC) acoblada a MS ha estat

molt emprada en l'anàlisi metabolòmica a causa de la seva alta eficiència en la separació de compostos i de l'àmplia disponibilitat de bases de dades per la seva posterior identificació, com les bases de dades del National Institute of Science and Technology (NIST) [131], la Golm Metabolome Database (GMD) [132] i la Fiehn/Binbase library [133]. Però, GC-MS és únicament idònia per a l'anàlisi de metabòlits petits i volàtils llevat que utilitzi procediments de derivatització laboriosos per poder estudiar compostos poc volàtils. En canvi, l'electroforesi capil·lar (CE) i la cromatografia de líquids (LC) permeten l'anàlisi directa de metabòlits no volàtils. CE és especialment adequada per a l'anàlisi de compostos polars i carregats, ja que es basa en la separació de compostos segons la relació càrrega/radi. LC permet la separació d'una àmplia gamma de metabòlits, independentment de la seva naturalesa hidrofílica o hidrofòbica [49, 134]. Ambdues tècniques han guanyat molt d'interès a causa de la seva alta eficàcia, sensibilitat, automatització i possibilitat d'acoblament en línia a l'espectrometria de MS. Aquestes qualitats fan que CE-MS i LC-MS siguin tècniques d'alta resolució que garanteixen una anàlisi reproduïble i fiable del metaboloma.

Electroforesi capil·lar (CE)

La CE és una tècnica d'alta resolució molt versàtil basada en les diferents velocitats de migració dels anàlits (metabòlits) segons la relació càrrega/radi (q/r) de l'ió solvatat quan s'aplica un camp elèctric dins d'un capil·lar de sílice fosa de diàmetre intern petit (id, 25-75 µm) ple d'un electròlit de separació (BGE) [135]. La mobilitat electroforètica (μ_e) és constant per a cada ió i es pot expressar com:

$$\mu_e = \frac{q}{6\pi\eta r} \qquad \text{Equació 2.1}$$

on q és la càrrega de l'ió, η és la viscositat del BGE i r és el radi de l'ió solvatat. Per tant, si tenim una mostra amb diferents anàlits, en aplicar el camp elèctric, les diferents μ_e fan que els anions migrin cap al pol positiu (ànode) i els cations cap al pol negatiu (càtode). A més, els ions més petits i més carregats migraran a major velocitat que els de mida més gran i càrrega menor.

Un fenomen important en la separació per CE és el flux electroosmòtic (EOF) [136, 137]. Quan s'empren capil·lars de sílice fosa la superfície de la paret interna es troba carregada negativament

degut als grups silanols (SiOH) que es troben dissociats a pH superiors a 2. Per tant, quan s'utilitza un BGE aquós de pH específic, aquests grups silanols de la paret interna (de certa càrrega negativa) atreuen els contraions catiònics del BGE i es forma una doble capa de cations que origina una diferència de potencial coneguda com a potencial zeta. En el moment que s'aplica el voltatge a través del capil·lar de separació, els cations de la doble capa es mouen cap al càtode, generant un EOF catòdic. La formació del EOF provoca que gairebé totes les espècies, ja siguin positives, negatives o neutres, siguin arrossegades cap a l'extrem de sortida del capil·lar de sílice fosa (càtode), on està el detector (polaritat normal) (**Figura 2.8**). Per tant, la mobilitat efectiva dels anàlits ve donada per la seva μ_e i l'EOF, que depenen principalment de les característiques del BGE.

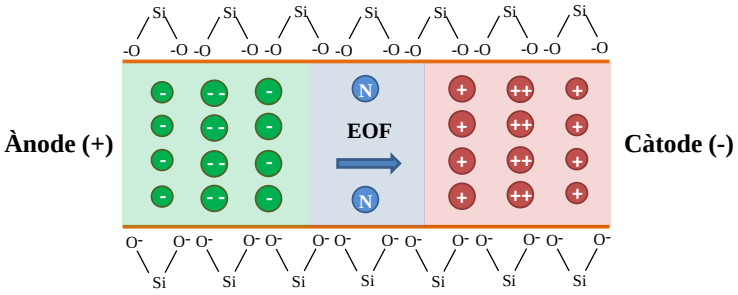


Figura 2.8. Representació esquemàtica de la migració electroforètica dels cations, anions i molècules neutres en el capil·lar de sílice fosa en condicions de polaritat normal.

Tot i que l'ús de la CE en el camp de la metabolòmica no és molt extens, el seu potencial i les diferents aplicacions d'aquesta tècnica han estat àmpliament reconegudes [138-140]. Una de les aplicacions més exitoses de la CE és la seva participació en el projecte del genoma humà per a la millora del coneixement de les seqüències de DNA [20, 141]. En l'actual era post-genòmica, la CE és una tècnica important per a la separació i caracterització dels metabòlits relacionats amb el metabolisme primari, com aminoàcids, àcids carboxílics, sucres i àcids nucleics [142-145]. Per exemple, Mischak ha manifestat reiteradament que CE-MS es pot aplicar de forma exitosa i rutinària per al descobriment de nous biomarcadors d'interès biomèdic [146, 147]. Encara que és una tècnica només utilitzada per un nombre reduït de grups de recerca presenta un ampli ventall d'aplicacions en el marc de la metabolòmica. Avui en dia, CE-MS s'empra per a l'anàlisi metabolòmica dirigida i,

principalment, no dirigida de mostres clíniques [148, 149], microbianes [150, 151], vegetals [152, 153] o alimentàries [154, 155].

Els principals avantatges de la CE són la seva elevada eficàcia i la rapidesa de les anàlisis sense necessitat d'extensos pretractaments de mostra. Altres avantatges de la CE són els petits volums de mostra, la instrumentació relativament senzilla i el baix consum de reactius i solvents. Per contra, els inconvenients d'aquesta tècnica són la baixa sensibilitat en unitats de concentració deguda principalment a les reduïdes dimensions del capil·lar de separació que limiten el volum de mostra (en nL) i els possibles problemes de reproductibilitat que ocasionen una alta variabilitat en els temps de migració i les àrees dels pics. No obstant això, la sensibilitat es pot millorar fàcilment acoblant la CE amb l'espectrometria de MS, la qual permet obtenir també informació estructural dels compostos separats [156]. Així doncs, l'ús de CE-MS amb ionització per electrosprai (ESI) és especialment útil per a l'anàlisi de compostos polars i carregats. A més, la baixa reproductibilitat en els temps de migració o les àrees dels pics es poden corregir adequadament mitjançant pretractaments de les dades o mètodes d'alineament de pics.

Cromatografia de líquids (LC)

La cromatografia de líquids ha esdevingut la tècnica de referència per a la separació i caracterització d'un gran nombre de molècules per nombroses aplicacions en la recerca. Especialment, ha crescut en popularitat en el camp de la metabolòmica a causa de l'alta resolució i la capacitat de separació de metabòlits de gran diversitat estructural [47, 157].

La LC és una tècnica molt versàtil ja que al llarg dels anys s'han desenvolupat diversos modes de separació segons la fase estacionària i el tipus d'interacció amb l'anàlit. Exemples són la cromatografia en fase invertida, en fase normal, d'interacció hidrofílica, d'exclusió per mida, de bescanvi iònic i d'afinitat.

A més, recentment s'han portat a terme grans avenços en la separació de LC per tal de millorar la capacitat de retenció i l'eficàcia de les columnes i reduir el temps d'anàlisi. D'una banda, una de les tendències en la cromatografia de líquids ha estat la reducció de la mida de partícula del rebliment per

tal d'augmentar l'eficàcia de la separació cromatogràfica. Ara bé, la disminució de la mida de partícula dóna lloc a un considerable increment de la pressió i a factors extra-columna, com la longitud i el diàmetre de les connexions del sistema i els volums morts en l'injector i el detector, que contribueixen fortament a l'eixamplament de banda cromatogràfica quan es treballa amb aquesta mida de partícula amb els sistemes tradicionals de cromatografia d'alta eficàcia (*high performance liquid chromatography*, HPLC). En efecte, els instruments han tingut que evolucionar paral·lelament a la disminució del diàmetre de les partícules del rebliments de les columnes cromatogràfiques, desenvolupant-se sistemes de cromatografia de líquids d'ultraalta eficàcia (*ultra performance liquid chromatography*, UHPLC) que suporten pressions de treball superiors als 1.000 bars i minimitzen els efectes extra-columna que afecten a l'eficàcia de la separació. La tècnica d'UHPLC es caracteritza per oferir una millor eficàcia, velocitat i sensibilitat en comparació a l'HPLC gràcies a que els sistemes tenen volums morts molt petits i que disposen de detectors molt ràpids per tal de poder aconseguir suficients punts d'adquisició per a reconstruir el pic cromatogràfic adequadament [158].

D'altra banda, en els darrer anys també s'han dut a terme avenços relacionats amb el desenvolupament de noves fases estacionàries per tal d'ampliar les aplicacions d'aquesta tècnica. Les columnes de fase invertida, com la C8 i C18, durant molts anys han estat les fases estacionàries més populars atesa la seva versatilitat en l'anàlisi d'un ampli ventall de compostos. De fet, la LC de fase invertida (*reversed-phase líquid chromatography*, RPLC) ha donat peu al desenvolupament d'una gran varietat de columnes de sílice d'alta resolució, robustes i resistents per a una anàlisi eficient i ràpida [159]. Però en el camp de la metabolòmica, les columnes de fase invertida presenten una limitació degut a la baixa retenció dels compostos molt polars i hidròfils que s'estudien habitualment. En conseqüència, s'ha proposat l'ús d'altres modes de separació, com la cromatografia de bescanvi iònic (*ion-pair chromatography*, IPC) [160], la cromatografia de fase invertida amb formadors de parells iònics (*ion-pair liquid chromatography*, IPLC) [161] o la cromatografia de líquids d'interacció hidrofílica (*hydrophilic interaction liquid chromatography*, HILIC) [162] per superar aquest important inconvenient de les columnes de fase invertida.

Cromatografia de líquids d'interacció hidrofílica (HILIC)

La cromatografia HILIC va ser proposada per Alpert el 1990 [163] i ha captat l'atenció de molts científics durant els últims anys. Actualment, el potencial de la separació en mode HILIC és àmpliament reconegut en el món de la metabolòmica per a separar compostos polars [164-166].

La cromatografia HILIC es pot considerar una variant de la cromatografia de líquids de fase normal (*normal phase liquid chromatography*, NPLC) considerant que ambdues metodologies utilitzen fases estacionàries polars i que la retenció augmenta amb la polaritat dels anàlits [167, 168]. El mecanisme de retenció HILIC proposat per Alpert per columnes de sílice pura es basa en la partició dels anàlits entre la fase mòbil (amb almenys un 2-3 % d'aigua en un contingut elevat de modificador orgànic) i una capa superficial d'aigua adsorbida sobre la fase estacionària polar (veure **Figura 2.9**). Ara bé, el mecanisme de separació en HILIC és més complex que en NPLC, ja que hi poden intervenir fenòmens d'adsorció sobre la superfície, interaccions electrostàtiques per bescanvi iònic i dipol-dipol i interaccions per pont d'hidrogen amb el suport de sílice.

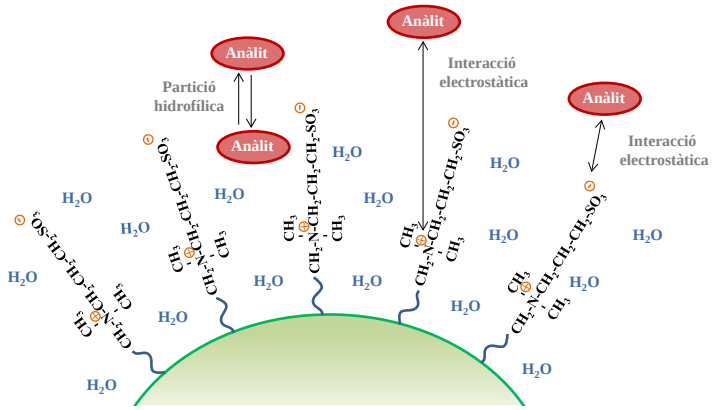


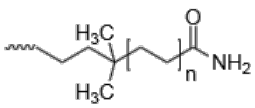
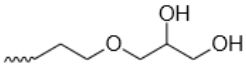
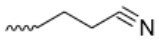
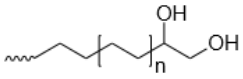
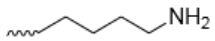
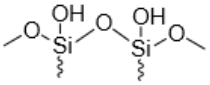
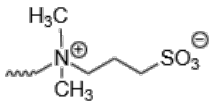
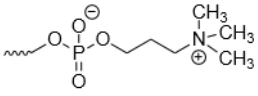
Figura 2.9. Representació esquemàtica del mecanisme de retenció de les fases estacionàries HILIC (exemple de la fase estacionària zwitteriònica ZIC-HILIC).

La cromatografia d'interacció hidròfilica utilitza fases mòbils comuns en RPLC amb un percentatge de modificador orgànic generalment superior al 40%. En aquestes condicions, la cromatografia HILIC ofereix una bona retenció dels compostos extremadament polars sense la necessitat d'addicionar formadors de parells iònics a la fase mòbil. A causa de la composició de les fases mòbils HILIC amb un elevat contingut de modificador orgànic i, per tant, de menor viscositat, aquest mode proporciona una millor eficàcia i una separació més ràpida dels compostos polars que quan es treballa amb RPLC emprant fases mòbils riques en aigua [169, 170]. A més, la cromatografia HILIC és molt fàcil

d'acoblar a diverses tècniques de detecció, com detectors ultravioleta-visible (UV-Vis) i d'espectrometria de masses (MS) [171]. En especial, la idoneïtat de les columnes HILIC per acoblar-se a l'espectrometria de MS dóna una excel·lent sensibilitat en ESI-MS, degut a l'elevat contingut de modificador orgànic de la fase mòbil. Tots aquests avantatges han causat l'èxit d'aquestes fases estacionàries, guanyant gran popularitat en la recerca i, especialment, en els estudis metabolomics [172].

La separació HILIC destaca per la seva versatilitat a conseqüència de l'àmplia gama de fases estacionàries que s'han desenvolupat en els últims anys, amb diferents materials de suport i gran diversitat química dels grups funcionals de la superfície. Les estructures i els grups funcionals més habituals en les columnes HILIC s'indiquen a la **Taula 2.3** destacant les principals aplicacions de cada tipus de fase estacionària.

Taula 2.3. Fases estacionàries HILIC més habituals.

Tipus	Fase estacionària	Grup funcional	Aplicacions
Neutres	Amida		Metabòlits polars, pèptids, glicoproteïnes, oligosacàrids i carbohidrats.
	Diol		Metabòlits polar, proteïnes, vitamines, compostos fenòlics, oligonucleòtids.
	Ciano		Compostos polars, fàrmacs.
	Mode mixt (diol)		Nucleòsids, compostos polars i no polars.
Carregades	Amino		Metabòlits polars, nucleòsids, carboxílic àcids, aminoàcids.
	Sílice pura		Metabòlits, fàrmacs, toxines, contaminants.
Zwitteriòniques	Sulfobetaïna		Metabòlits polars, sucres, pèptids, fàrmacs, compostos fenòlics (especial afinitat amb cations).
	Fosforilcolina		Metabòlits polars, pèptids, fàrmacs, compostos fenòlics (especial afinitat amb anions).

Les fases estacionàries HILIC es classifiquen segons la càrrega dels grups funcionals de la fase estacionària (neutres, carregades o zwitteriòniques). Les fases estacionàries neutres contenen grups funcionals polars que no estan carregats a valors de pH entre 3 i 8. La majoria de fases estacionàries HILIC pertanyen a aquesta categoria i reuneixen una gran varietat de grups funcionals, com amida, aspartamida, diol, ciano, ciclodextrina i sacàrids. A més, s'han sintetitzat fases estacionàries neutres que poden treballar en mode mixt (*mixed-mode chromatography*). Aquestes fases estacionàries en mode mixt permeten treballar amb més d'un mode d'interacció ja que contenen fases estacionàries alquíliques amb grups polars i/o iònics, de manera que es poden donar interaccions hidrofòbiques, hidrofíliques i de bescanvi iònic. D'aquesta manera, es poden analitzar simultàniament tant compostos polars i iònics com compostos no polars. En canvi, les fases estacionaries carregades presenten grups funcionals polars amb carrega positiva o negativa segons el pH de la fase mòbil. Les fases amino i de sílice són les més conegudes d'aquesta família. Finalment, les fases estacionàries zwitteriòniques són aquelles que contenen quantitats iguals de grups de càrrega oposada units a la superfície de sílice. Generalment, els grups funcionals zwitteriònics presenten una funcionalitat fortament àcida i fortament bàsica que no depèn del pH [168, 170, 173].

Hi ha molts factors que poden influenciar en l'eficàcia i la resolució de la cromatografia HILIC, tals com les dimensions de la columna, el diàmetre de partícula de fase estacionària i la seva porositat. Tanmateix, la interacció dels anàlits amb els grups funcionals de la fase estacionària depèn en gran mesura de les condicions cromatogràfiques (per exemple, tipus de fase mòbil o la presència d'additius en la fase mòbil). A conseqüència de les possibles interaccions entre els anàlits i la fase estacionària, els principals paràmetres que afecten en el mecanisme de retenció de HILIC són el modificador orgànic, la força iònica i el pH de la fase mòbil que donen lloc a selectivitats molt diferents [174]. Una visió global de la relació entre els paràmetres cromatogràfics i les interaccions HILIC pot permetre una millor optimització de la separació cromatogràfica.

Tècniques de detecció

Detecció ultravioleta-visible (UV-Vis)

La detecció UV-Vis és àmpliament emprada en LC i destaca per la seva universalitat, ja que són molts els anàlits que tenen grups cromòfors que absorbeixen en l'interval de longitud d'ona entre 190 i 600 nm. L'ús del detector de díodes en línia (DAD) enlloc de la detecció d'una o un nombre reduït de longituds d'ona presenta molts avantatges, tals com la visualització de l'espectre UV-Vis durant tot el temps d'anàlisi o la determinació de l'espectre i màxim d'absorbància per cada metabòlit permetent la identificació d'aquests en estudis complexos, com per exemple en dissenys experimentals. A més, el detector DAD és també el mètode més emprat en CE degut al seu baix cost, rapidesa d'anàlisi i la gran quantitat d'informació espectral que genera. Tot i així, els principals inconvenient d'aquesta tècnica són la baixa sensibilitat i selectivitat. Aquestes limitacions són de gran importància en els estudis metabolòmic on es tracten mostres biològiques d'elevada complexitat i, a més, la concentració dels metabòlits és molt baixa.

En la metabolòmica, aquesta tècnica permet trobar diferències en els perfils metabòlics [175, 176] però no permet la identificació ni caracterització individual dels metabòlits. És només capaç de detectar aquells compostos que disposin de grups cromòfors i que es trobin a una concentració elevada.

Detecció per espectrometria de masses (MS)

Les excel·lents prestacions de l'espectrometria MS tant a nivell de sensibilitat, selectivitat i quantitat d'informació estructural que ofereix fan que avui en dia sigui l'eina de referència en l'òmica, especialment en la proteòmica i la metabolòmica [12, 177, 178]. L'espectrometria de MS d'injecció directa (DIMS) és útil per a la caracterització de mostres biològiques en el marc de l'òmica [179]. Malgrat això, el potencial de la detecció per MS augmenta considerablement quan s'acobla a tècniques de separació d'alta resolució com la CE o la LC.

Els espectròmetres de masses són instruments molt complexos que generalment consten d'una font d'ionització, un analitzador i un detector [180]. La selectivitat i la sensibilitat dels mètodes CE-MS i

LC-MS depenen bàsicament de l'adequada selecció de la font d'ionització, de l'analitzador i el mode d'adquisició que permeti aprofitar al màxim les capacitats del sistema. A més, alguns analitzadors permeten experiments de MS en tàndem (MS/MS) o fragmentació en etapes successives (MS^n) de forma que es pot dur a terme l'elucidació i la identificació inequívoca dels compostos analitzats mitjançant estudis de fragmentació [181]. En aquesta Tesi, s'han portat a terme diferents estudis de metabolòmica no dirigida de CE-MS i LC-MS en els quals s'ha emprat la font d'ionització d'electrosprai (ESI) i diferents analitzadors de MS.

Tècniques d'ionització. Electrosprai (ESI)

En els estudis de metabolòmica, la font d'ionització d'electrosprai (ESI) és la font d'ionització més emprada a pressió atmosfèrica (*atmospheric pressure ionization*, API) [47]. L'ESI és la font d'ionització per excel·lència, ja que és compatible per a l'anàlisi d'un gran ventall de compostos de polaritat mitjana, alta i baix i d'alt pes molecular. A més, ESI és una font d'ionització "tova", és a dir, produeix una baixa fragmentació a la font i, generalment, s'obté un senyal (ió) o grups de senyals (clúster isotòpic o d'ions múltiplement carregats) per a cada molècula de metabòlit (normalment la molècula protonada o desprotonada). La ionització dels metabòlits en ESI es porta a terme mitjançant l'aplicació d'un camp elèctric, diferència de potencial (2-5 kV), entre l'extrem de l'elèctrode o l'agulla per on surt la mostra líquida i el contra elèctrode situat a l'entrada de l'espectròmetre de masses. La ionització es dóna en fase líquida on els metabòlits es protonen (ESI positiu) o desprotonen (ESI negatiu) via equilibris àcid/base. A la sortida de l'agulla es genera un esprai de gotes carregades que acaba formant el que es coneix com a con de Taylor. La formació de l'esprai s'assisteix amb un gas nebulitzador (N_2). En les gotes carregades d'ions positius o negatius (segons la polaritat emprada) els ions passen a fase gas mitjançant un procés d'evaporació iònica. A mesura que el solvent s'evapora la repulsió de les càrregues dins de les gotes augmenta i s'ocasiona la formació de gotes carregades més petites. Aquestes gotes es dispersen i continuen fent-se més petites fins que la repulsió de les càrregues de les gotes és tan gran que exploten i es formen ions en fase gas que es transfereixen al sistema de buit de l'espectròmetre de masses [182-184] (**Figura 2.10**).

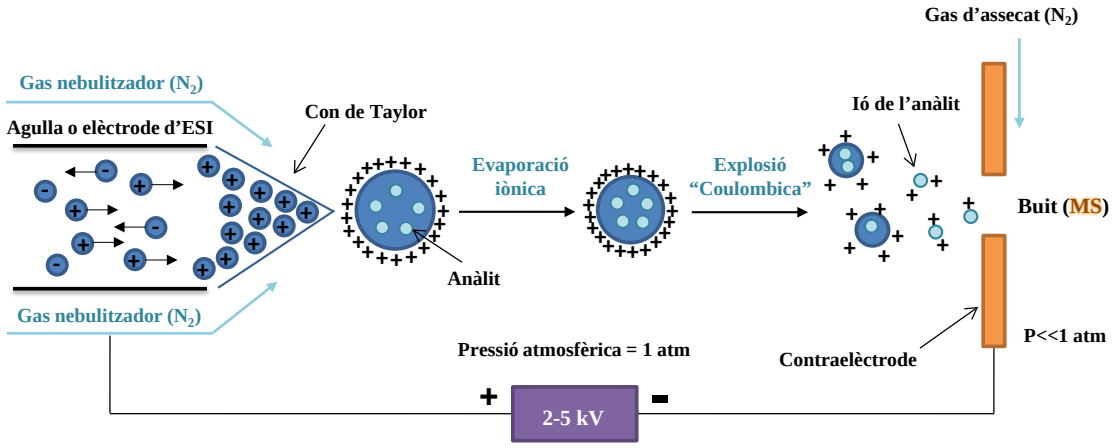


Figura 2.10. Formació de l'esprai en ESI positiu.

L'eficàcia de la ionització en ESI depèn de les propietats dels metabòlits, del pH i de la composició de la solució on es troben dissolts. Per exemple, la presència de substàncies que coelueixen (LC) o comigren (CE) amb els metabòlits o els mateixos additius dels solvents que s'utilitzen en la separació (sodi, amoníac, formiat o acetat, etc.) poden disminuir la intensitat dels ions $[M+H]^+$ o $[M-H]^-$ emprats normalment com a ions diagnòstic.

Analitzadors de masses

Els analitzadors dels espectròmetres de masses tenen la finalitat de separar els ions en funció de la seva relació de massa i càrrega (*m/z*) i marquen en gran mesura la seva sensibilitat, selectivitat i els possibles modes de treball. Actualment, els principals analitzadors de masses són el quadrupol senzill (*quadrupole*, Q), la trampa d'ions (*ion trap*, IT), l'analitzador de temps de vol (*time-of-flight*, TOF), l'Orbitrap i la ressonància ciclotrònica d'ions amb transformada de Fourier (FT-ICR). Els analitzadors esmentats, es diferencien majoritàriament per les seves prestacions de sensibilitat, resolució, exactitud, interval de masses d'escombrat, interval dinàmic, velocitat d'escombratge i la possibilitat de realitzar MS en tàndem. Els analitzadors de masses com la IT i la FT-ICR permeten realitzar en el mateix dispositiu la fragmentació d'ions mitjançant la prèvia selecció de l'ió precursor i experiments en tàndem amb el temps. En canvi, el Q, el TOF i l'Orbitrap recorren a la fragmentació a la font d'ionització o a sistemes híbrids on s'acoblen dos analitzadors en sèrie per portar a terme fragmentacions de tàndem en l'espai (per exemple, el triple quadrupol (*triple quadrupole*, QqQ) i el

quadrupol-temps de vol (*quadrupole-time-of-flight*, Q-TOF)). En aquesta Tesi s'han emprat els analitzadors de masses TOF i Orbitrap per dur a terme els corresponents estudis de metabolòmica.

Analitzador de temps de vol (TOF)

L'analitzador de temps de vol (TOF) és un dels analitzadors més antics i més emprats avui en dia. Aquest analitzador es basa en la diferent velocitat que adquireixen els ions per recórrer un tub de 1-2 m que es troba entre la font d'ionització i el detector al aplicar un camp elèctric (veure **Figura 2.11a**). El temps que triguen els ions en travessar el tub depèn de la seva relació *m/z*. Considerant que la càrrega i l'energia cinètica dels ions formats en la font són constants, el temps de vol permet determinar de forma molt precisa la massa de cada ió [185, 186]. Per tant, els ions de major càrrega i menor massa arribaran abans al detector que els ions de menor càrrega i major massa (**Figura 2.11a**).

El desenvolupament de la tecnologia d'acceleració ortogonal (oa-TOF) ha millorat considerablement el poder de resolució i l'exactitud de massa d'aquest instrument [187]. Aquesta distribució ortogonal consisteix en fer passar els ions a través d'un reflector que inverteix la direcció del seu vol (**Figura 2.11b**) per tal de poder compensar les diferències d'energia cinètica dels ions que provenen de la font d'ionització. Els ions de més velocitat (menor *m/z*) entraran més en el reflector i tardaran més en arribar al detector. D'aquesta manera, si un paquet d'ions d'una determinada relació *m/z* conté ions amb diferents energies cinètiques, el reflector disminuirà la dispersió dels temps de vol dels ions [188]. Així, aquest tipus d'analitzador permet reduir la velocitat i la distribució en l'espai dels ions aconseguint resolucions de fins als 60.000 FWHM (*full width half màximum*, amplada de pic a mitja alçada) a *m/z* 1222 i una bona exactitud en la mesura de la massa [189]. Aquesta configuració ortogonal proporciona una millora en la resolució de masses i, també, permet l'ús de fonts d'ionització contínues, com la ionització ESI. A més, el TOF és un analitzador relativament simple i econòmic que ofereix un rang de masses en teoria il·limitat.

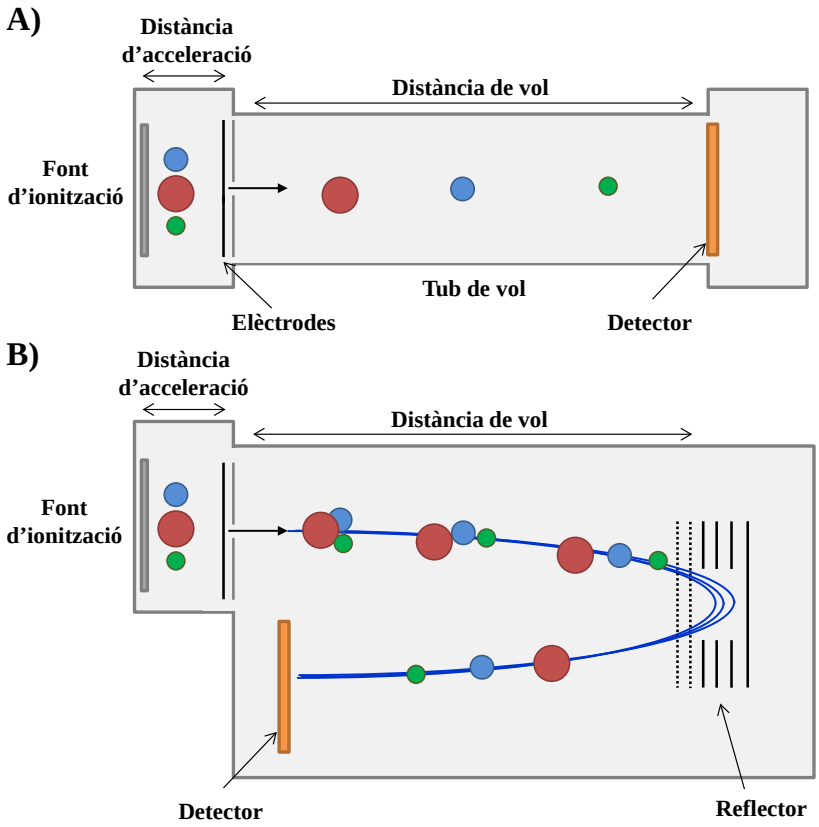


Figura 2.11. Representació esquemàtica de la separació dels ions en l'analitzador TOF; (a) lineal i (b) ortogonal.

Orbitrap

Aquest analitzador és especialment útil per a l'anàlisi metabolòmica no dirigida, ja que facilita molt l'elucidació estructural i la identificació dels metabòlits desconeguts. L'Orbitrap és un analitzador de masses d'alta resolució que consta d'un elèctrode extern en forma de barril i un elèctrode central en forma de fus (**Figura 2.12**). Els ions són injectats tangencialment per paquets dins d'un camp elèctric generat entre els dos elèctrodes. Els ions hi queden atrapats degut a que l'atracció electrostàtica cap a l'elèctrode interior es contraresta amb la força centrifuga. Així, els ions circulen en orbites al voltant de l'elèctrode central i a la vegada es desplacen al llarg de l'eix d'aquest mateix elèctrode de tal manera que els ions amb una mateixa relació *m/z* es mouen en anells que oscil·len al voltant del fus central. La freqüència d'aquestes oscil·lacions harmòniques permet determinar el valor *m/z* dels ions, ja que és independent de la velocitat dels ions i inversament proporcional a l'arrel quadrada de la relació *m/z*. Els ions són transferits per una petita trampa d'ions en forma de C (*C-trap*) que estabilitza

mitjançant un gas, compacta i finalment injecta els ions dins l'analitzador. A més, aquest analitzador disposa d'una cel·la de dissociació (*higher-energy collisional dissociation*, HCD) que permet dur a terme la fragmentació de tots els ions generats a la font d'ionització (*all-ion fragmentation*).

L'Orbitrap és probablement l'analitzador més jove de l'espectrometria de masses, però ha tingut un alt impacte en el camp de la proteòmica i la metabolòmica des dels seus començaments per les seves prestacions. Aquest analitzador ofereix resolucions de 100.000-240.000 FWHM (*m/z* 200), permet treballar en un rang *m/z* de 50-6.000 i proporciona una molt bona exactitud en la mesura de la massa (<5 parts per milió (ppm)) [189, 190].

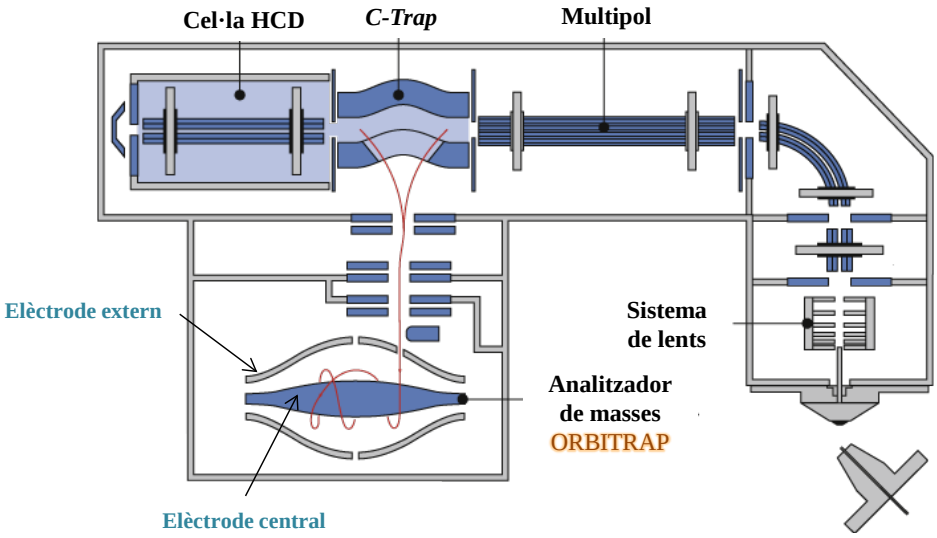


Figura 2.12. Esquema de l'Exactive Plus Orbitrap (Thermo Fisher).

2.3. ANÀLISI DE LES DADES METABOLÒMIQUES

En aquest apartat es revisen les tècniques quimiomètriques emprades en aquesta Tesi pel tractament de les dades òmiques. En concret, s'ofereix una visió general de les eines de preprocessament, de visualització i d'anàlisi aplicades a les dades metabolòmiques de LC-DAD, LC-MS i CE-MS. A continuació, s'introdueix també la integració o fusió de la informació de diferents plataformes i nivells òmics. No es descriuran en canvi, els tractaments necessaris per a l'anàlisi de les dades transcriptòmiques de RNA-Seq. Aquestes dades han estat obtingudes i processades en el Centre

Nacional d'Anàlisi Genòmica (CNAG) en col·laboració amb el grup d'investigació del Departament de Toxicologia Ambiental de l'Institut de Diagnosi Ambiental i Estudis de l'Aigua (IDAEA) dirigit pel Dr. Benjamín Piña.

Els avenços aconseguits en els últims anys en la bioanalítica i en la bioinformàtica han donat lloc a una visió integral i global dels processos biològics que estudien les ciències òmiques. El desenvolupament constant de noves metodologies analítiques, eines d'anàlisi i de tractament de dades ha obert noves perspectives en les ciències ambientals i biològiques que han facilitat un canvi en el paradigma de la recerca, des d'un concepte minimalista en el que només s'estudia la química d'un únic compost o procés, fins a un concepte més holístic que permet la caracterització molt més completa del conjunt de compostos químics i processos dels sistemes biològics. Aquest nou concepte globalitzador es veu reflectit en l'aparició dels denominats mètodes analítics no dirigits (descrits en l'apartat 2.1.4), els quals es basen en la detecció simultània del major nombre possible de compostos del sistema analitzat.

Donada la gran complexitat i grandària dels conjunts de dades òmiques no dirigides, existeix una necessitat urgent d'implementar mètodes quimiomètrics d'anàlisi de dades amb l'objectiu de millorar i automatitzar el seu anàlisi.

En el context de la metabolòmica no dirigida, les tècniques d'espectrometria de masses proporcionen conjunts de dades multi- o megavariants extremadament complexos. En conseqüència, la quimiometria aplicada a aquests estudis ha esdevingut una eina essencial per a extreure informació més rigorosa i objectiva dels processos biològics que ocorren en el organismes vius en unes condicions determinades.

A continuació es presenta una breu descripció de la naturalesa i estructura de les dades estudiades i dels diferents preprocessaments i mètodes d'anàlisi de les dades emprats en aquesta Tesi.

2.3.1. Naturalesa de les dades d'espectrometria de masses

Els avenços tecnològics en l'espectrometria de masses han permès l'obtenció massiva d'informació química d'alta resolució en molts camps de la ciència (biologia, salut ambiental, química, etc.). Actualment aquestes tècniques fan possible i viable obtenir un coneixement molt precís i global del funcionament dels sistemes biològics estudiats.

En el context de la metabolòmica no dirigida, es realitza generalment un escombratge no selectiu (*full scan*) de tots els ions de la mostra. Les dades d'espectrometria de masses es poden adquirir en mode *centroide* (només mostra el valor de *m/z* del màxim de la distribució dels diferents ions presents en un espectre) o en mode *profile* (mostra un perfil complet de la distribució dels diferents ions presents en un espectre) (veure **Figura 2.13**). Durant molts anys el mode *centroide* ha estat el mode d'adquisició per excel·lència i una opció sovint obligatòria en la química analítica degut a la limitada capacitat d'emmagatzematge i processament dels antics equips informàtics. Els arxius *centroide* són significativament menors ja que emmagatzemen menys informació per cada senyal. Tot i així, la identificació es veu millorada considerablement en mode *profile* ja que permet determinar la distribució isotòpica dels ions de forma més acurada. Aquesta informació és especialment útil quan es vol diferenciar possibles candidats amb múltiples fórmules moleculars similars [191].

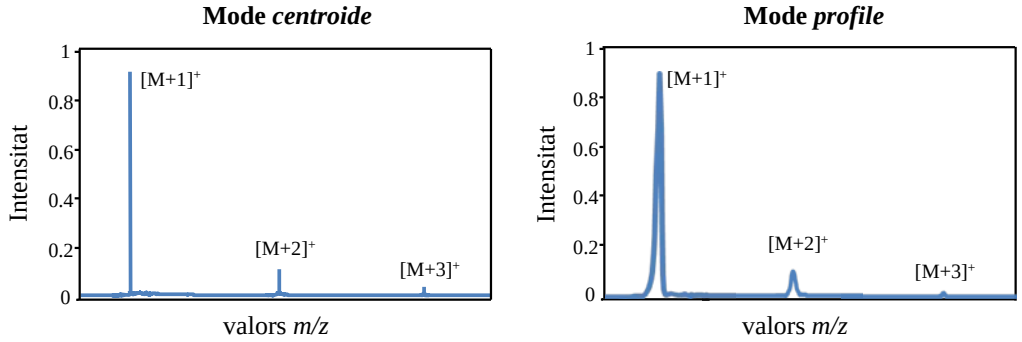


Figura 2.13. Exemple d'un espectre de masses en mode *centroide* i en mode *profile*. L'espectre en mode *profile* conté la informació màxima i correspon a les dades en brut (originals), però ocupa un espai d'emmagatzematge considerable. L'espectre en *centroide* és una variant compacta i processada de l 'espectre en *profile*, que captura la major informació espectral.

L'adquisició en mode *profile* ha esdevingut avui en dia la millor opció i la més emprada en el camp de l'òmica gràcies als beneficis que ofereix en la identificació dels compostos i a l'ampliada capacitat d'emmagatzematge dels ordinadors.

Ara bé, els espectres de masses en *full scan* procedents d'estudis de metabolòmica no dirigida són complexos i voluminosos atès el gran nombre de valors *m/z* detectats en mode *profile* (fins a centenars de vegades més grans que en mode *centroide*), sobretot, quan s'utilitza espectrometria de masses d'alta resolució (*high resolution mass spectrometry*, HRMS). Així doncs, l'anàlisi de les dades es converteix en un gran repte degut a la dificultat que comporta l'extracció de la informació útil d'aquests volums massius de dades experimentals.

2.3.2. Estructura de les dades metabolòmiques

Les dades generades en els estudis de metabolòmica mitjançant LC-DAD, LC-MS i CE-MS obtingudes en aquesta Tesi es poden classificar segons la seva complexitat en dades ordenades en una direcció (*one-way data*), en dues direccions (*two-way data*) o en tres direccions (*three-way data*). Així, les dades que es troben ordenades en una direcció corresponen a dades vectorials, mentre que les dades que es troben ordenades en dues direccions s'organitzen en taules o matrius de dades de valors numèrics. Finalment, les dades ordenades en tres direccions, donen lloc al que s'anomena cub de dades.

En concret, en aquesta Tesi s'han analitzat cromatogrames totals d'ions (*total ion chromatograms,* TICs) o electroferogrames totals d'ions (*total ion electropherograms*, TIEs) que es representen en forma de vectors (*one-way data*, **Figura 2.14a**). Els TICs o TIEs s'obtenen com a resultat de la suma de les intensitats de tot el rang de masses per a cada temps de retenció o migració. Quan es treballa amb més d'un vector a la vegada s'obté una matriu de dades (*two-way data*) on a les files es troben les mostres. També, s'han considerat cromatogrames o electroforogrames amb la detecció a múltiples longituds d'ona o *m/z* que estan organitzats en taules o matrius de valors numèrics (*two-way data,* **Figura 2.14b**). Aquestes taules de dades s'organitzen en matrius on en la dimensió de les *x* es troben els temps de retenció o migració i en la dimensió *y* el rang de longituds d'ona (DAD) o *m/z*

analitzades (MS). Però, quan es treballa amb conjunts de cromatogrames o electroforogrames organitzats en matrius s'obté un cub de dades on a la dimensió *z* es troben les diferents mostres considerades (*three-way data*, **Figura 2.14c**).

Generalment, en els estudis òmics no s'analitza tan sols una única mostra sinó diverses mostres que proporcionen diferents vectors o matrius de dades. L'estudi de tot el conjunt de dades experimentals és el què proporcionarà informació consistent dels canvis en els perfils metabòlics entre els diferents factors o condicions investigades. Per tant, la recopilació de tots els vectors o matrius d'un determinat experiment dóna lloc a estructures d'ordre superior on a les columnes hi ha les variables mesurades (valors *m/z*) i en les files les diverses mostres analitzades. Aquestes taules o matrius de dades es representaran per una lletra majúscula en negreta, per exemple **D**.

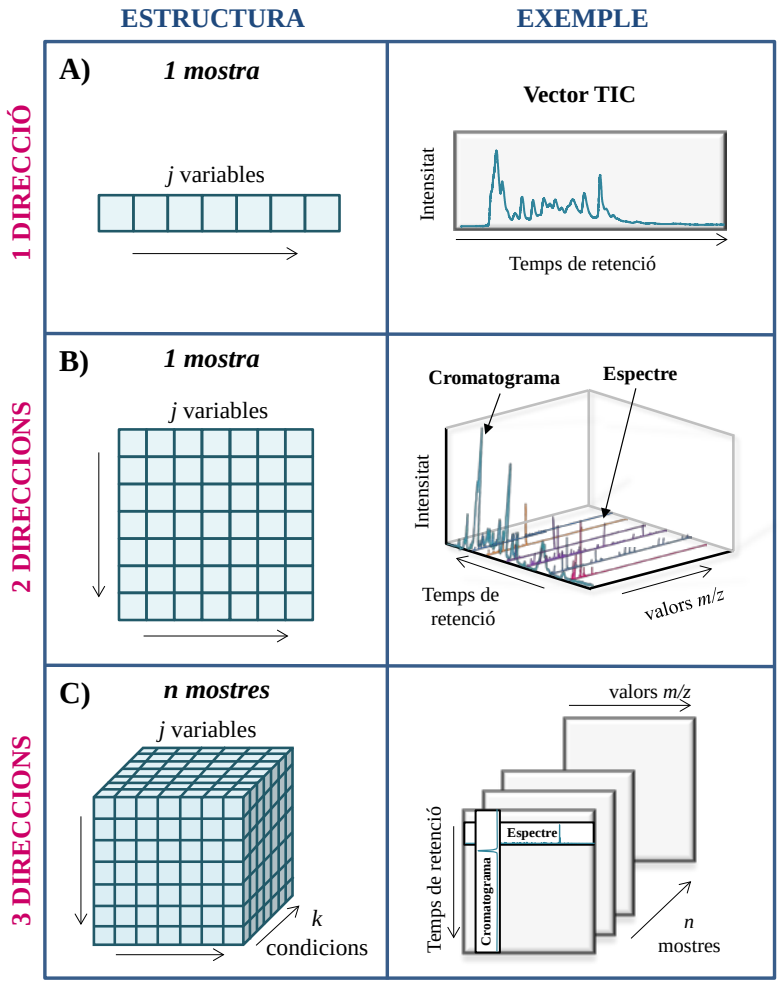


Figura 2.14. Tipus d'estructures de les dades metabolòmiques estudiades: (a) d'una direcció, (b) dues direccions i (c) tres direccions.

L'estratègia d'augmentació o formació de matrius augmentades, $\mathbf{D}_{aug}$, consisteix en concatenar els diferents vectors o matrius de dades en la direcció de les files (una al costat de l'altra, *row-wise*), de les columnes (una sota de l'altra, *column-wise*) o en les dues direccions *(row- and column-wise*). En aquests casos, només cal que els vectors o matrius de dades experimentals obtingudes tinguin el mateix nombre de files, columnes o ambdues, respectivament.

En aquesta Tesi, s'ha emprat l'augmentació en la direcció de les columnes (veure **Figura 2.15**), ja que permet l'anàlisi simultània de diversos vectors o matrius de dades corresponents a sistemes químics amb variables comunes estudiats amb una única tècnica analítica. Les mostres TIC/TIE (vectors), s'han augmentat en la direcció de les columnes, aprofitant que els temps de retenció o migració són comuns per totes les mostres, donant lloc a una matriu (**Figura 2.15a**). En canvi, els cromatogrames/electroferogrames, que es representen en forma de matrius, s'han augmentat també en la direcció de les columnes, ja que són iguals per tots els cromatogrames LC-MS i electroferogrames CE-MS (rang *m/z* comú) (**Figura 2.15b**) i per tots els cromatogrames LC-DAD (longituds d'ona comunes), generant matrius augmentades. A la **Figura 2.15** es mostra com els vectors TIC i els cromatogrames de LC-MS són augmentats en la direcció de les columnes aprofitant que tenen la direcció de les columnes en comú (temps de retenció i valors de *m/z*, respectivament).

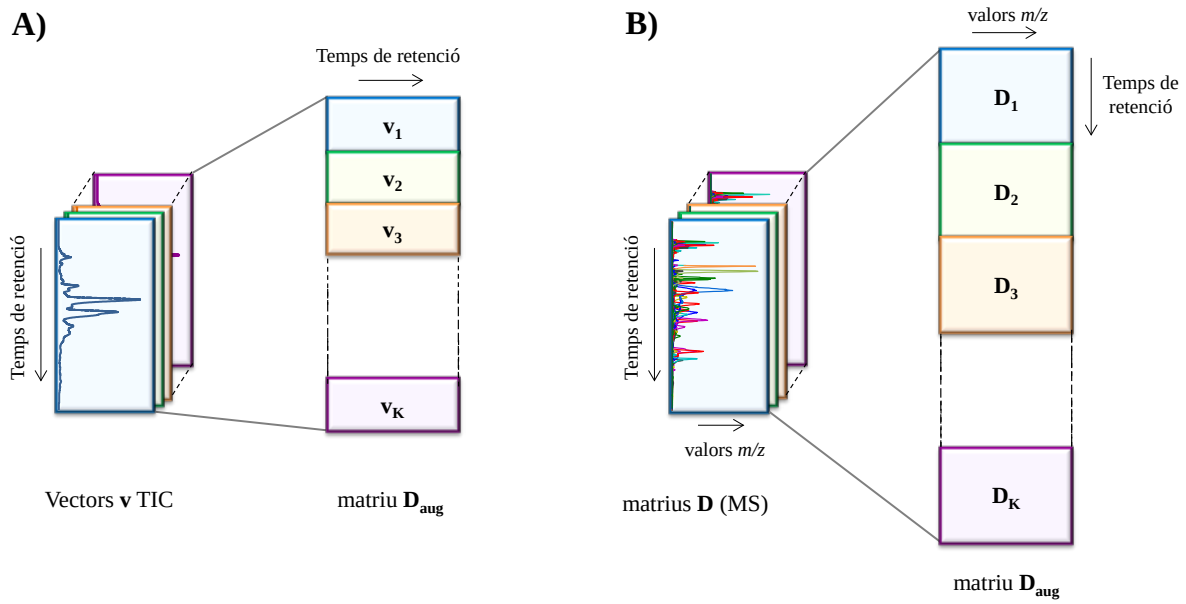


Figura 2.15. Exemples d'augmentació de (a) vectors TIC i (b) matrius de LC-MS en la direcció de les columnes.

En els darrers anys, també ha crescut l'interès per la fusió de dades procedents de diferents tècniques analítiques que també requeriran d'aquest tipus d'arranjaments en la direcció de les columnes o de les files per a la seva anàlisi simultània.

2.3.3. Tractament preliminar de les dades experimentals

Abans del preprocessament i l'anàlisi quimiomètrica de les dades, els arxius de DAD o MS necessiten un tractament previ, que acostuma a ser una de les parts més laborioses, per tal de millorar la qualitat dels senyals experimentals. Aquesta preparació de les dades constarà d'etapes de transformació de les dades, de reducció de mida o estructuració, depenent del format en què s'hagin obtingut inicialment, de la seva mida, etc.

Conversió i importació de les dades

Una vegada adquirides les dades cromatogràfiques o electroforètiques, el primer pas necessari per la seva anàlisi implica la conversió d'aquestes del format original (dades *raw*), que són difícils d'emprar fora dels programes de la casa comercial, en formats que es poden llegir a la majoria de plataformes de càlcul numèric (per exemple, MATLAB o R). Entre els formats de dades estàndard existents, els més populars són els formats basats en arxius XML (és a dir, *mzXML*, *mzData* i *mzML*), *netCDF* i en fitxers de text clàssic com el format *txt*. La majoria dels programes dels fabricants dels instruments de LC-MS o CE-MS tenen eines específiques per a la conversió de les dades en formats oberts. Per exemple, el programa d'adquisició i anàlisi de dades de Thermo Fisher (Xcalibur) presenta l'opció "*File converter*" la qual permet la conversió directa de les dades crues en arxius *netCDF* o fitxers *txt*. En el cas de les dades obtingudes mitjançant instruments Agilent (arxius .d) amb detecció DAD, el programa ChemStation també disposa d'aquesta opció de conversió de les dades directament. Per contra, els fitxers de MS adquirits amb el programa MassHunter d'Agilent no tenen aquesta utilitat. Així, es necessita l'ús d'eines externes com ProteoWizard [192, 193], el qual conté una eina de conversió anomenada *msconvert* que permet la transformació a formats oberts, com *mzXML*, *mzML* i *txt*.

Per a l'anàlisi de dades en entorns de MATLAB o R, és possible la importació de les dades utilitzant diferents estratègies. Quan es treballa amb MATLAB, per importar les dades de CE-MS i LC-MS s'utilitzen les rutines i funcions disponibles a la Bioinformatics Toolbox que permeten convertir les dades de format estàndard (com el *netCDF o mzXML*) en variables de MATLAB. En canvi, en l'entorn R, les dades de MS normalment s'importen mitjançant el paquet *mzR* disponible en el projecte Bioconductor. Aquest paquet *mzR* proporciona una interfície unificada per a la majoria dels formats de dades oberts descrits anteriorment com *mzXML*, *mzML*, *mzData* i *netCDF*. En aquesta Tesi s'ha treballat en l'entorn de programació MATLAB pel tractament de dades posterior.

Compressió de les dades

La compressió de les dades és un pas necessari (sovint obligatori) per reduir la mida de les dades i, en conseqüència, els temps d'anàlisi. Aquest pas permet evitar problemes associats amb la limitada capacitat de memòria i processament dels ordinadors intentant evitar la pèrdua d'informació durant la seva anàlisi.

Les dades de LC-DAD no són molt voluminoses (generalment de l'ordre de pocs megabytes per mostra). A més, sovint s'eliminen aquelles longitud d'ona on els compostos d'interès no absorbeixen per tal de disminuir el soroll i agilitzar el temps d'anàlisi. En canvi, en el cas de les dades d'espectrometria de masses, la compressió de les dades és de gran importància, ja que facilitarà la seva anàlisi posterior. S'han descrit diferents metodologies que permeten comprimir les dades d'espectrometria de masses i arranjar-les en la seva forma bidimensional. Els procediments d'interpolació, d'agrupament en caixetes (*binning*) i de cerca de les regions d'interès (*regions of interest*, ROI) són els més emprats. En els diferents treballs de metabolòmica presentats en aquesta Tesi s'han emprat aquestes tres estratègies per a la compressió dels conjunts de dades experimentals. Les estratègies esmentades per a la compressió de dades d'espectrometria de masses es descriuen a continuació amb més detall. Les dues primeres estratègies d'interpolació i *binning* s'utilitzen habitualment combinades amb una altra estratègia de compressió per finestres de temps que ajuda a reduir encara més la mida de les dades.

Interpolació i *binning* en combinació amb la divisió de finestres de temps

La interpolació i, en especial, el *binning* són procediments que poden emprar-se per a la compressió de les dades *full scan* de LC-MS i CE-MS. Aquests procediments de compressió transformen les dades experimentals en matrius de dades on els espectres de MS als diferents temps de retenció o migració es troben en les seves files (dimensió *x*) i els cromatogrames o electroferogrames als diferents valors *m/z* en les seves columnes (dimensió *y*). La compressió de les dades per interpolació o *binning* s'aplica sobre els espectres de masses enregistrats a cada temps de retenció o migració (fila) de les dades experimentals. D'aquesta manera, els espectres originals d'alta resolució en els diferents temps de retenció o migració (mesurats a valors diferents de *m/z* no equidistants) es converteixen en espectres de baixa resolució, on els valors *m/z* estan separats segons el grau d'interpolació o mida de *bin* prèviament definit. En concret, en el cas del *binning* es sumen les lectures dins d'un interval espectral definit per la mida de la 'caixeta' o '*bin*', i així formar una única mesura (veure **Figura 2.16a**). Per tant, la compressió de dades amb aquestes estratègies es realitza en la dimensió espectral (eix *m/z*). Un inconvenient rellevant d'aquests procediments és la dificultat en la selecció adequada del nivell d'interpolació o de la mida de *bin*. Per exemple, si el nivell d'interpolació o mida de *bin* seleccionats són massa petits es pot perdre la forma del pic cromatogràfic o electroforètic i, per tant, no serà detectat. En canvi, si pel contrari els valors són massa grans poden existir múltiples coelucions o comigracions entre pics fent que els pics petits puguin desaparèixer per l'augment del nivell de soroll. Un altre inconvenient molt important d'aquests mètodes, és la pèrdua de resolució espectral que es produeix en la seva aplicació en comparació a la resolució espectral ortogonal de l'espectròmetre de masses [194].

En la majoria d'estudis de metabolòmica es requereix de l'anàlisi simultània de les diferents mostres d'un mateix experiment. Quan la mida de *bin* o el nivell d'interpolació escollit és petit, aquests procediments de compressió poden no ser suficients per al processament conjunt de totes les mostres degut al seu gran volum. Per exemple, una mostra interpolada amb una resolució de 0,01 Da/e pot ocupar aproximadament un gigabyte i farà que sigui intractable tot el conjunt de mostres amb ordinadors estàndard. En aquest context, seria necessària la divisió dels cromatogrames o

electroferogrames en diferents finestres o regions de temps (*time windowing*, **Figura 2.16b**), les quals poden ser aleshores analitzades de forma separada [195, 196]. Tot i així, si les dades són encara molt voluminoses es poden subdividir els espectres de masses en diferents finestres o regions de l'espectre (*spectral windowing*, **Figura 2.16c**), la qual cosa no obstant, pot representar un augment del temps d'anàlisi. Aquest pas addicional a les estratègies d'interpolació i *binning* permet analitzar simultàniament totes les mostres experimentals ja que redueix significativament la mida de la matriu augmentada.

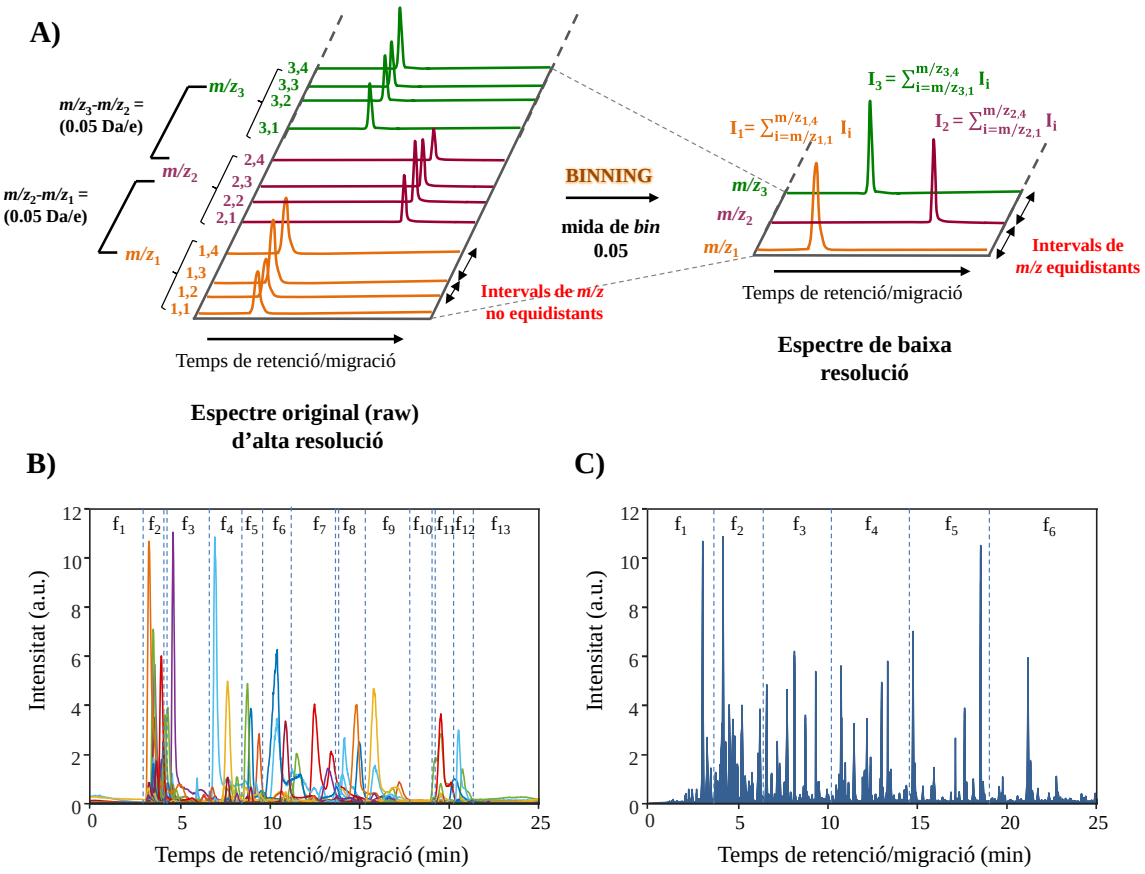


Figura 2.16. Representació gràfica de la compressió de les dades; (a) *binning*, (b) *time windowing* i (c) *spectral windowing*.

Regions d'interès (ROI)

La compressió de les dades basada en la cerca de les regions d'interès (*regions of interest*, ROI) és una aproximació alternativa als procediments d'interpolació i *binning*. Aquest mètode va ser inicialment descrit per Stolt i col·laboradors l'any 2006 [197]. Posteriorment, es va introduir a

l'algorisme *centWave* de la plataforma XCMS [198], i més recentment ha estat adaptat a l'entorn MATLAB [199]. L'estratègia ROI es basa en el concepte de compressió de les dades de MS a partir de la cerca de les regions d'interès (ROI) dels cromatogrames o electroferogrames. És a dir, permet la selecció d'aquells valors *m/z* que tenen una intensitat de senyal més alta que un valor llindar de relació senyal/soroll (*signal-to-noise ratio*, SNR_{Thr}) preseleccionat i que estan definits per un nombre de punts que descriguin la correcta resolució d'un pic cromatogràfic o electroforètic. Així, un valor ROI ha de contenir un nombre mínim de punts consecutius que estiguin en un interval de valors de *m/z* amb una amplitud definida per la precisió en la determinació del valor de massa de l'espectròmetre amb el que es treballa. Els valors *m/z* dels ROIs es busquen a cada temps de retenció o migració d'un cromatograma, i s'obtenen vectors de diferent mida per cada espectre a cada temps de retenció o migració (veure **Figura 2.17**). Finalment, aquells ROIs comuns a diferents temps de retenció o migració es combinen. Quan no apareixen en algun temps de retenció o migració es consideren que són pràcticament zero (o un valor pròxim al soroll instrumental), per així obtenir el mateix nombre final de valors ROI per tots els espectres MS del cromatograma o electroferograma considerat. Per a cada ROI, els valors de *m/z* es calculen a partir de la mitjana de tots els valors *m/z* de la sèrie de punts de dades agrupades en un mateix ROI dins d'un mateix interval d'error de masses (error o desviació *m/z*). Finalment, mitjançant aquesta compressió s'obté una matriu de dades reduïda per a cada cromatograma o electroferograma corresponent a una mostra analitzada amb els espectres de masses als diferents temps de retenció o migració a les files (dimensió *x*) i els cromatogrames al diferents valors *m/z* dels ROIs escollits a les columnes (dimensió *y*) [199, 200]. El nivell de compressió aconseguit amb l'estratègia de ROI és molt elevat i és suficient per poder analitzar a la vegada diversos cromatogrames o electroferogrames d'un mateix experiment sense pèrdua de resolució espectral (exactitud de massa) respecte a la mesura experimental (instrumental). El nombre de ROIs obtinguts per diferents mostres pot variar entre desenes i centenars i, per tant, en l'anàlisi simultània d'aquestes mostres s'hauran de considerar tots aquells ROIs (valors *m/z*) que han estat obtinguts en totes elles, tant els que són ROIs comuns entre mostres com els que no ho són. Actualment, s'ha incrementat l'ús de l'estratègia de ROI en els paquets de tractament de dades, substituint els mètodes clàssics d'interpolació o *binning*.

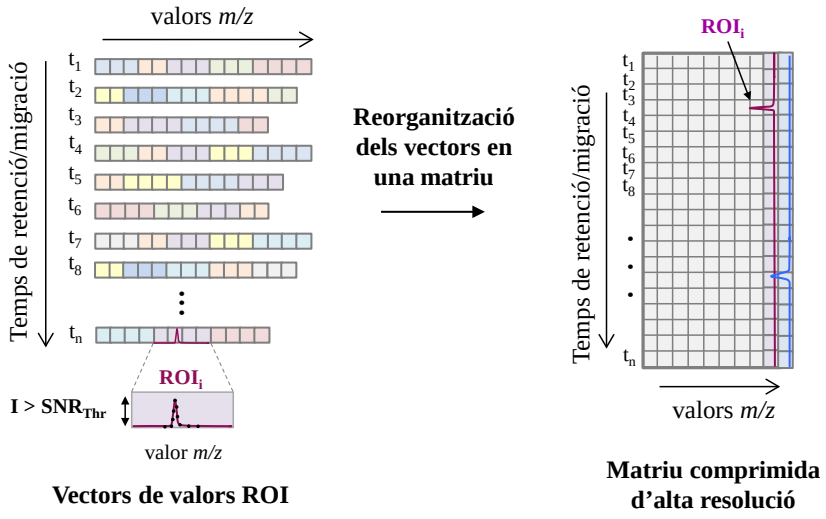


Figura 2.17. Representació gràfica de la compressió ROI d'una matriu de dades.

2.3.4. Preprocessament de les dades experimentals

Per tal d'extreure el màxim d'informació útil del conjunt de dades metabolòmiques i fer possible la interpretació dels efectes dels factors o condicions investigades, sovint és necessària l'aplicació de tècniques de preprocessament de les dades per millorar la seva qualitat. No existeix un únic mètode de pretractament que es pugui utilitzar per totes les dades generades en metabolòmica. Aquests mètodes de preprocessament permeten eliminar o minimitzar les fonts de variació de les dades que poden emmascarar la informació rellevant o d'interès biològic. Cal tenir en compte que els estudis òmics generen conjunts de dades complexos que contenen informació de tres tipus: 1) variació química o biològica causada pel factor estudiat; 2) variació natural entre organismes no causada per l'efecte estudiat; i 3) errors experimentals, instrumentals i variació estocàstica (variació aleatòria). La separació entre aquests tipus de variació és important per poder avaluar els canvis produïts pels factors estudiats. La variació aleatòria que es produeix en totes les determinacions experimentals afecta la qualitat dels models obtinguts. A més, els errors sistemàtics en el procés experimental també poden influir en els models. Per tant, l'ús de mètodes de preprocessament robustos és molt important per a la correcta interpretació dels resultats obtinguts.

La selecció de l'eina o eines adequades de preprocessament dependrà majoritàriament de diferents factors, com la naturalesa de les dades, la informació que es pretén obtenir a partir d'elles i dels

mètodes d'anàlisi que s'empraran. Al llarg d'aquesta Tesi els principals mètodes de pretractament de dades que s'han aplicat són la correcció de la línia base, l'alineament dels senyals i la normalització de les dades.

Correcció de la línia base

La correcció de la línia base i el soroll de fons és fonamental per a l'anàlisi de les dades experimentals. Aquestes variacions en les dades són ocasionades principalment per la influència de petits canvis en la resposta del detector o en la composició dels solvents de separació que s'utilitzen en les condicions d'anàlisi instrumental. El principal problema d'aquestes variacions durant l'anàlisi és que introdueixen fonts d'error en les dades. Els mètodes de correcció de línia base permeten reduir o eliminar la influència dels factors esmentats en la variació de les dades. És a dir, tenen com a objectiu separar del senyal mesurat la contribució que hi pot haver d'aquestes interferències procedents de l'anàlisi instrumental i dels solvents emprats.

El mètode de mínims quadrats ponderats (*weighted least squares*, WLS) [201] és una eina molt eficaç per la correcció automàtica de la línia base. Es pot emprar per a la correcció de dades cromatogràfiques i electroforètiques, tot i que des de els seu origen s'ha emprat freqüentment en aplicacions espectroscòpiques. L'algorisme WLS es basa en una estimació automàtica de la línia base tenint en compte els punts que considera que provenen de la contribució únicament de la línia base real. Consisteix en un procés iteratiu que ajusta la línia de base de cada espectre i determina quines variables estan clarament per sobre o per sota del senyal teòric de la línia base. El resultat d'aquesta correcció és l'eliminació automàtica del soroll de fons, però sempre evitant crear valors negatius. Generalment, la línia base s'ajusta per un polinomi d'ordre baix. Aquest algorisme s'ha emprat per a la correcció de la línia base de les dades TIC i TIE abans de la seva exploració mitjançant mètodes quimiomètrics multivariants.

D'altra banda, també s'ha emprat el mètode de mínims quadrats asimètrics (*asymmetric least-squares*, AsLS) [202, 203] per a corregir la línia base dels cromatogrames de LC-DAD. L'algorisme AsLS és una eina comunament emprada per a la correcció del soroll de fons i el suavitzat de la línia base

d'espectres i cromatogrames. Aquest algorisme es basa en un ajustament iteratiu de la línia base dels cromatogrames. En la primera iteració, tots els punts del cromatograma tenen el mateix pes en l'ajust, per la qual cosa una part del cromatograma queda per sobre de la línia de base ajustada (residuals positius) i una altra per sota (residuals negatius). En les següents iteracions es penalitzen els punts del cromatograma per sobre de la línia de base ajustada (amb residuals positius) fins que es compleix el criteri de convergència de l'ajust. El cromatograma corregit s'obté restant del cromatograma original la línia base ajustada.

Alineament de pics

Un dels principals problemes en la cromatografia de líquids i, especialment, en l'electroforesi capil·lar són les fluctuacions (*shift*) en els temps de retenció o migració dels pics. Els mètodes d'alineació milloren o redueixen aquestes variacions de les dades de CE i LC (en els seus electroferogrames TIE i cromatogrames TIC, respectivament) abans de procedir a la seva anàlisi. Aquest aspecte és de gran importància en les dades de CE on la desviació en el temps de migració sovint és superior al 1-2%. Aquesta desincronització de les dades pot afectar a la fiabilitat i a la robustesa dels models. A més, tant en CE com en LC aquest desajust dels temps de migració o retenció pot dificultar l'assignació de la identitat del mateix compost en diferents electroferogrames o cromatogrames, és a dir, si el temps no és el mateix no es podrà confirmar que aquell pic correspon al mateix anàlit o metabòlit.

En aquesta Tesi, s'ha utilitzat el mètode *correlation optimized warping* (COW) [204, 205] per minimitzar al màxim el desplaçament dels pics en els cromatogrames TIC i en els electroferogrames TIE. El procediment d'alineació COW, s'aplica per dur a terme alineaments locals a vectors de dades (*one-way data*), tot i que s'ha adaptat per a l'alineament de matrius de dades completes de LC-MS i CE-MS (*two-way data*) [206]. El mètode COW permet alinear un vector corresponent a una mostra amb un altre vector de referència que és representatiu de tot el conjunt de mostres que es vol alinear. El procés d'alineament es porta a terme mitjançant la divisió del vector mostra en segments o finestres de mida definida que poden augmentar o disminuir de longitud per tal de trobar la correlació òptima d'aquests respecte els segments del vector de referència. Amb aquesta metodologia cal definir, per tant, el vector de referència, la longitud dels segments i el grau d'augment o disminució de la longitud

(*slack*) dels segments de la mostra. La selecció adient d'aquests paràmetres es realitza automàticament aplicant l'algorisme descrit per Skov i col·laboradors [207], el qual interpola linealment cada segment de la mostra per tal de crear segments d'igual longitud als del vector de referència. Així, s'aconsegueix el màxim de correlació entre segments i s'arriba a la millor correcció de cada segment. En el cas de les dades cromatogràfiques i electroforètiques, l'alineació s'efectua en la direcció del temps de retenció o migració, respectivament. Cal destacar que aquesta necessitat d'alineament no existeix en el cas d'analitzar les dades de CE i LC en mode *full scan* (quan s'obté en l'anàlisi de cada mostra una taula de dues dimensions o matriu de dades) pel mètode MCR-ALS, tal com s'explicarà més endavant.

Normalització i escalat de les dades

La normalització i escalat de les dades tenen com a objectiu principal la reducció de la variació sistemàtica entre mostres i les diferències d'escala entre les variables mesurades. En les dades no dirigides, es poden utilitzar dues estratègies diferents per eliminar aquesta variació no desitjada en les mesures: normalització de les mostres (en la direcció de les files) i escalat i transformació de les variables (en la direcció de les columnes).

La normalització de les mostres és necessària per ajustar les diferències entre mostres. Aquesta normalització és molt senzilla i pot ser química o matemàtica. En aquesta Tesi, s'ha utilitzat la normalització química a través de subrogats, patrons interns i mostres control (QCs) (veure secció 2.1.5.). Els subrogats i els patrons interns 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 [208].

En canvi, l'escalat i la transformació de les dades permet la comparació entre les variables a les diferents mostres. Aquesta etapa és crucial en el preprocessament de les dades metabolòmiques on el rang de concentracions del metabòlits és molt ampli. Els mètodes de normalització de les variables (metabòlits) emprats en aquesta tesi han estat el centrat (*mean-center*), l'autoescalat i la transformació MinMax de les dades.

El centrat de les dades consisteix en restar a cada valor original de les variables el valor de la seva mitjana. D'aquesta manera, en la nova matriu de dades (matriu centrada) cada variable té una mitjana igual a 0 i realitza una translació de l'origen de coordenades de les dades des de 0 fins al valor de la mitjana de les dades [209]. Així, aquest pretractament permet visualitzar les variacions respecte al valor mitjà i descarta aquella informació de les variables que no varia i que es pot considerar constant. Aquest preprocessament s'ha aplicat als cromatogrames TIC i als electroferogrames TIE.

L'autoescalat de les dades es basa en restar a cada valor original de les variables el valor de la seva mitjana i dividir-lo per la seva desviació estàndard. En aquest cas, les dades també experimenten una translació del seu origen a causa del centrat amb la mitjana de les dades i, a més, un escalat degut a la seva divisió per la desviació estàndard. D'aquesta manera, totes les variables tenen la mitjana igual a 0 i la desviació estàndard a 1 [209]. Aquest pretractament és adient per tal de facilitar la interpretació dels resultats quan les variables tenen escales amb ordres de magnitud diferents. S'utilitza sobretot en l'aplicació de mètodes d'anàlisi de dades que se centren en l'avaluació de la variància observada per les variables mesurades entre diferents mostres. Per exemple, s'ha aplicat sobre les matrius d'àrees de metabòlits prèviament a la seva anàlisi per components principals, PCA (veure secció 2.3.6.)

Per últim, la transformació MinMax augmenta el pes de les variables que presenten un valor més baix respecte aquelles de valor més alt, de forma similar a l'autoescalat. És un procediment que sovint es recomana per conjunts esbiaixats de dades on els valors d'un gran nombre de variables són baixos i un nombre menor de variables tenen valors alts, com és el cas de les dades metabolòmiques de MS. La transformació MinMax reescala cada columna d'una matriu de dades mitjançant la subtracció del valor mínim de cada variable de la columna i la seva divisió pel seu rang (diferencia entre el valor màxim i el valor mínim de la variable). Per evitar contribucions del soroll de fons, s'aplica únicament a aquelles variables que estiguin per sobre d'un valor llindar (per exemple, valor de soroll de fons) [210]. En les dades metabolòmiques, la normalització MinMax és una alternativa que permet una millor resolució de les variables (metabòlits) de baixa intensitat que poden estar fàcilment emmascarades per altres variables amb un senyal més elevat (de major concentració). Per exemple, aquest pretractament s'ha aplicat a les dades de CE-MS després de la seva compressió per interpolació

i prèviament a la seva anàlisi mitjançant la resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS).

2.3.5. Resolució dels pics cromatogràfics i electroforètics amb detecció multivariant

La resolució de dades cromatogràfiques i electroforètiques té com a objectiu separar els components purs d'una mescla associats a un espectre (DAD o MS) i a un perfil d'elució (LC) o migració (CE) concret, per aconseguir diferenciar-los entre ells i d'altres contribucions lligades al soroll. La resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS) [211] és un procediment àmpliament emprat en l'anàlisi dels components d'una mescla a partir de mesures multivariants agrupades en una taula de dues dimensions o matrius de dades, com és el cas de les mesures *full-scan* obtingudes mitjançant LC-MS i CE-MS. El mètode de MCR-ALS no requereix d'un alineament previ de les dades multivariants en els dos modes o direccions de mesurament.

Resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS)

El mètode de resolució multivariant de corbes per mínims quadrats alternats (*multivariate curve resolution alternating least squares*, MCR-ALS [211]) és una eina quimiomètrica àmpliament utilitzada amb la finalitat de resoldre les diferents contribucions dels diferents components (espècies químiques) existents en mescles complexes. Alguns exemples de sistemes analitzats mitjançant MCR-ALS són l'anàlisi de dades espectroscòpiques [212], d'imatges espectrals [213, 214], de dades òmiques obtingudes per RMN [215] i de tècniques de separació acoblades a l'espectrometria de masses cromatogràfiques [195, 196], entre d'altres. En aquesta Tesi, el mètode MCR-ALS s'ha aplicat a l'anàlisi de les dades metabolòmiques obtingudes per LC-DAD, LC-MS i CE-MS. Aquest mètode ha estat àmpliament descrit en treballs previs i només els aspectes més importants relacionats amb aquesta tesi es descriuen a continuació.

MCR-ALS es basa en la descomposició matemàtica de la informació continguda en una matriu de dades **D**, seguint el següent model bilineal:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equació 2.2}$$

En el cas de les dades LC-DAD, LC-MS o CE-MS, la matriu de dades **D** *(I×J)* conté els espectres experimentals als *i* temps de retenció o migració (*i*=1,...,*I*) en les files i els cromatogrames o electroferogrames a les *j* longituds d'ona (DAD) o valors de *m/z* (MS) (*j*=1,...,*J*) en les columnes. MCR-ALS descompon aquesta matriu **D** en el producte de dues matrius, **C** i $\mathbf{S}^{\mathbf{T}}$. La matriu **C**(*I*×*N*) descriu els perfils de concentració (d'elució o migració) dels N (*n*=1,…,*N*) components purs de la matriu **D** i la matriu $\mathbf{S}^{\mathbf{T}}$ (*N*×*J*) conté els espectres UV o MS corresponents pels N components (veure **Figura 2.18**). Per exemple, en els cas de l'espectrometria de masses, cada component MCR-ALS està associat als diferents valors *m/z* que descriuen un mateix perfil d'elució o migració. Finalment, la matriu **E** conté la part de la variància de la matriu **D** no explicada pel model.

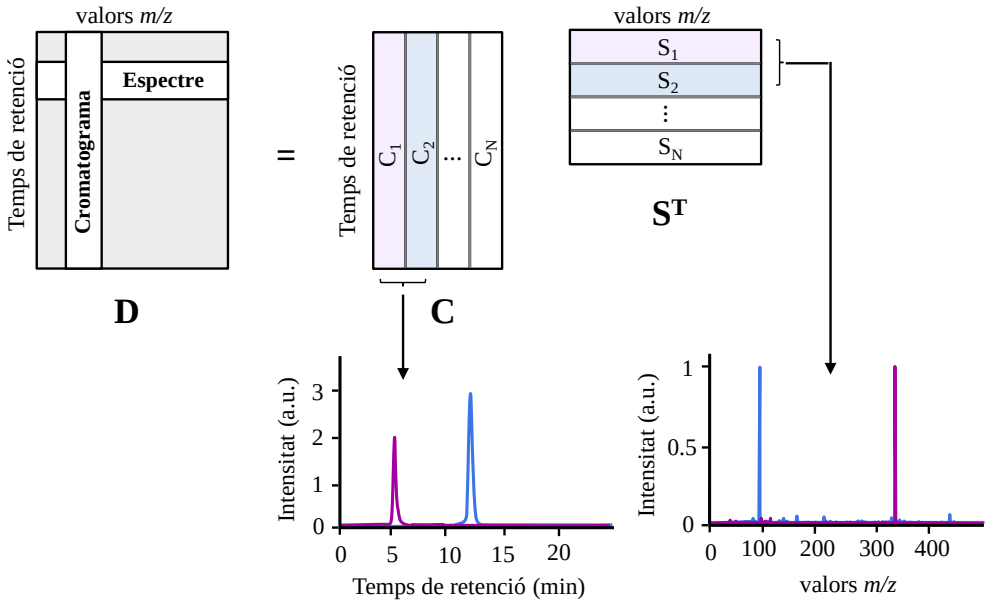


Figura 2.18. Representació del model bilineal de MCR-ALS per a una matriu de dades metabolòmiques de LC-MS.

El procediment de descomposició de la matriu **D** assumeix el model bilineal de l'equació 2.2 i es caracteritza per les següents etapes principals (**Figura 2.19**):

1. Determinació del nombre de components presents en una mostra analitzada per LC o CE) (nombre de metabòlits d'una matriu de dades experimentals **D**).
2. Estimació inicial dels perfils de concentració o dels espectres.

3. Càlcul de les matrius **C** i $\mathbf{S^T}$ mitjançant l'aplicació d'un procés d'optimització iteratiu de mínims quadrats alternats (*alternating least squares*, ALS).

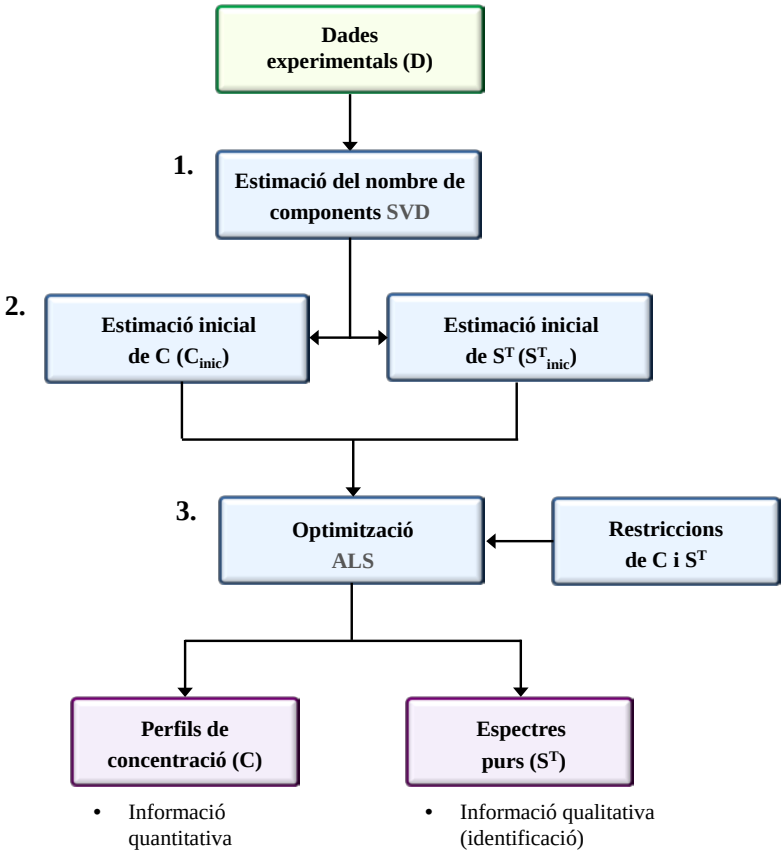


Figura 2.19. Diagrama resum de les etapes del procés de resolució de MCR-ALS.

La resolució MCR-ALS comença amb una selecció del nombre de components que es consideraran en els passos posteriors. Habitualment, el nombre de components necessaris per explicar la variació química (no el soroll experimental) de la matriu **D** es determina mitjançant la utilització de mètodes com l'anàlisi per components principals (*principal component analysis*, PCA) [216] o la descomposició en valors singulars (*singular value decomposition*, SVD) [217]. La forma més senzilla i ràpida és l'obtenció dels valors propis (*eigenvalues*) o valors singulars (arrel quadrada dels valors propis) trobats per SVD. Els valors propis indiquen la quantitat de variació retinguda per cada component. Tant els valors propis com la variància explicada per cada component va disminuint en afegir més components fins a un nivell petit en que només s'explica soroll experimental. A la pràctica, la selecció del nombre de components N de la matriu **D** a partir d'aquests mètodes pot no ser tan

evident degut a la complexitat de la mostra i al soroll inherent a la mesura, com en el cas de les dades de MS emprades en els estudis metabolòmics. Per tant, en aquests casos es recomana que es porti a terme el procés de resolució variant el nombre de components. D'aquesta manera, es comparen els resultats dels diferents models i s'adopta com a solució final aquell nombre de components que millor descriu les dades seguint el principi de parsimònia, és a dir, el menor nombre de components que ajusti les dades de forma acceptable i proporcioni perfils de concentració i espectres amb sentit químic.

Un cop determinat el nombre de components que es relaciona amb les espècies o anàlits presents en el sistema, es realitza l'estimació inicial de les matrius dels perfil de concentració ($\mathbf{C}_{inic}$) o dels espectres purs ($\mathbf{S}^{T}_{inic}$) les quals s'empraran en l'optimització iterativa posterior. En aquesta Tesi, s'ha emprat un mètode de detecció de variables pures basat en l'algorisme SIMPLISMA (*simple-to-use iterative self-modeling mixture analysis*) proposat per Windig [218, 219] per dur a terme aquesta etapa d'estimació de $\mathbf{C}_{inic}$ o $\mathbf{S}^{T}_{inic}$.

Després de determinar el nombre de components i d'obtenir les seves estimacions inicials (dels perfils d'elució, $\mathbf{C}$, o espectres, $\mathbf{S}^{T}$), el següent pas és la seva optimització iterativa a partir d'un algorisme iteratiu de mínims quadrats alternats (ALS). Les estimacions dels perfils de concentració o espectrals s'optimitzen a més aplicant restriccions (*constraints*) que aportin significat químic a les solucions purament matemàtiques. La utilització adequada d'aquestes restriccions és indispensable ja que permeten donar sentit químic a les solucions obtingudes i disminuir l'efecte de les ambigüitats inherents a la resolució de l'equació 2.2 [211, 220, 221].

En el context de les dades metabolòmiques, les restriccions que s'han aplicat en l'optimització ALS dels perfils en aquesta Tesi han estat la *No-negativitat* (els perfils d'elució i espectres només poden tenir valors positius), la *Unimodalitat* (els perfils d'elució només tenen un pic) i la *Normalització dels espectres* (els espectres purs resolts tenen la mateixa àrea o alçada) [222]. En concret, les dades LC-DAD s'han analitzat sota restriccions de no-negativitat tant pels perfils l'elució com espectrals, d'unimodalitat pels perfils d'elució i de normalització d'igual àrea dels espectres UV resolts. En canvi,

en les dades d'espectrometria de masses (LC-MS i CE-MS), s'han aplicat les restriccions de no-negativitat pels perfils de concentració i espectres, i de normalització d'igual alçada dels espectres MS resolts. En el cas de les dades MS, les possibles ambigüitats associades a la resolució dels perfils de concentració i dels espectres de MS de cada metabòlit mitjançant el procediment MCR-ALS queden fortament reduïdes degut a l'elevada selectivitat de les mesures d'espectrometria de masses i a la presència d'un nombre molt elevat d'elements pràcticament nuls (*sparsity*) que sovint acaben sent substituïts per zeros [223, 224].

Finalment, el procés iteratiu es dóna per finalitzat quan es compleix el criteri de convergència establert a l'inici de l'optimització ALS. El criteri de convergència es basa en la diferència relativa de les desviacions estàndard dels residuals entre els valors experimentals i els ajustats en l'optimització ALS, la qual s'ha de trobar per sota d'un valor llindar o preestablert (normalment fixat al 0,1%) [225].

La qualitat de l'optimització ALS es pot avaluar a partir de dos paràmetres, el percentatge de variància explicada (R^2) i el percentatge de manca d'ajust de les dades del model (*lack of fit*, Lof) [220].

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

$$\text{Lof}\,(\%) = 100\,\sqrt{\frac{\sum_{i,j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i,j} d_{ij}^2}} \qquad \text{Equació 2.4}$$

on d_{ij} correspon a un valor de la matriu de dades original **D** i e_{ij} correspon als residuals obtinguts a partir dels valors de la matriu original (d_{ij}) i dels valors de la matriu ajustada ($\hat{d}_{ij}$) per ALS. Aquest valor de manca d'ajust hauria de tenir idealment un valor similar al soroll de fons o error experimental present a les dades originals. Per trobar aquest valor adequat es realitza la optimització ALS amb un nombre creixent de components. Si la manca d'ajust millora significativament vol dir que el model anterior no tenia la suficient informació per descriure el sistema. En canvi, si la manca d'ajust empitjora o no varia de forma significativa, indica que l'addició de components al model no és necessària. És important no considerar un nombre de components més gran que el real ja que es corre

el perill de sobreajustar les dades (*overfitting*), és a dir, introduir soroll experimental en els perfils de concentració i espectrals resolts.

En el context de les dades metabolòmiques, després de l'optimització ALS s'obtenen:

1. Els perfils de concentració purs dels diferents components (metabòlits) del sistema biològic estudiat. A partir de les àrees de pic (o de la intensitat dels màxims) d'aquests perfils es pot conèixer l'abundància o concentració relativa de cadascun dels components.
2. Els espectres purs dels components del sistema biològic estudiat. A partir d'aquests perfils es poden identificar els metabòlits (en particular, quan s'utilitza l'espectrometria de masses).

MCR-ALS aplicat a l'anàlisi simultània de múltiples mostres (matrius augmentades)

En els estudis òmics es generen diversos conjunts de matrius de dades experimentals corresponents a replicats biològics o a mostres analitzades sota diferents condicions. Un dels majors avantatges del mètode MCR-ALS és la seva versatilitat a l'hora de poder analitzar simultàniament aquests conjunts de dades. Això és possible a partir de l'estratègia d'augmentació de les matrius de dades individuals.

En el cas de dades LC-DAD, l'extensió del model bilineal de MRC-ALS aplicat a l'anàlisi de matrius augmentades ($\mathbf{D}_{\mathbf{aug\_DAD}}$) en la direcció de les columnes es pot expressar en la següent forma matricial:

$$\mathbf{D}_{\mathbf{aug\_DAD}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} = \mathbf{C}_{\mathbf{aug}}\mathbf{S}^T + \mathbf{E}_{\mathbf{aug}} \qquad \text{Equació 2.5}$$

on la matriu augmentada $\mathbf{D}_{\mathbf{aug\_DAD}}$, formada per la concatenació de diferents matrius experimentals ($\mathbf{D}_k$, k=1,...K) es descompon en el producte de la matriu augmentada dels perfils de concentració $\mathbf{C}_{\mathbf{aug}}$ amb la matriu d'espectres d'absorció $\mathbf{S}^T$. La matriu residual $\mathbf{E}_{\mathbf{aug}}$ conté la variància no explicada pel model MCR. Així doncs, el model bilineal assumeix que els components identificats a les matrius individuals $\mathbf{D}_k$ arranjades en la matriu augmentada $\mathbf{D}_{\mathbf{aug\_DAD}}$, comparteixen els mateixos espectres d'absorció $\mathbf{S}^T$ mentre que els perfils de concentració de $\mathbf{C}_k$ a les mostres poden ser diferents en cadascuna de les k condicions.

Anàlogament, la resolució de les matrius augmentades de MS ($\mathbf{D}_{aug\_MS}$) (**Figura 2.20**) proporciona una única matriu d'espectres, $\mathbf{S}^T$, equivalent als espectres de masses, que descriu la composició (valors *m/z*) dels perfils metabòlics, la qual és comuna a totes les matrius $\mathbf{D}_k$ considerades. En canvi, la matriu augmentada resolta $\mathbf{C}_{aug}$ descriu els canvis en els perfils de migració o elució de cadascun dels components resolts en cadascuna de les diferents mostres analitzades (cadascuna de les matrius $\mathbf{D}_k$). Cal per tant remarcar que en el model descrit per l'equació 2.5, els perfils de migració o elució (pics) d'un mateix component en les diferents matrius $\mathbf{C}_k$ poden variar en posició (*shift*) i en forma. Així, es permet la resolució dels pics sense l'ús previ de procediment d'alineació dels pics, ja que d'acord amb l'equació 2.5, l'alineació només és necessària en la dimensió espectral que és comuna a totes les mostres. Aquest aspecte és realment útil en la resolució de dades de LC-MS i, en especial, de CE-MS, on el temps de migració i la forma del pics electroforètics d'un mateix metabòlit pot diferir considerablement entre injeccions però els espectres es mantenen iguals. En l'anàlisi de matrius augmentades s'apliquen també diferents tipus de restriccions per tal d'obtenir perfils de components que tinguin sentit químic i que, per tant, puguin ser fàcilment interpretables. La implementació i el tipus de restriccions en el model segueixen la mateixa estratègia descrita per a l'anàlisi MCR-ALS de les matrius individuals.

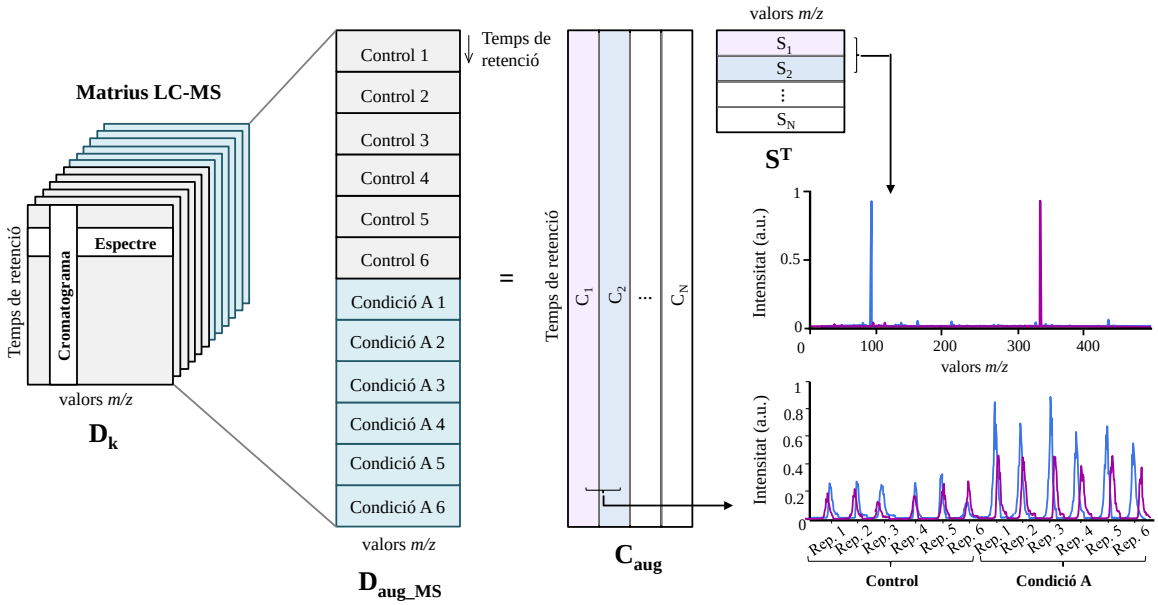


Figura 2.20. Representació gràfica de l'extensió del model MCR-ALS a l'anàlisi d'una matriu augmentada formada per 12 mostres analitzades mitjançant LC-MS (6 controls i 6 tractades).

2.3.6. Anàlisi dels resultats obtinguts per MCR-ALS i selecció de biomarcadors

Com en molts altres casos de dades analítiques, una propietat característica de les dades òmiques és que el nombre de mostres és significativament inferior al nombre de variables, el que es coneix com ($n<p$). Així, l'anàlisi de les dades metabolòmiques requereix de mètodes de tractament de dades que apliquen eines matemàtiques, estadístiques univariants i multivariants per tal de maximitzar la informació que es pot extreure a partir d'elles [226].

Anàlisi estadística univariant

Després del preprocessament de les dades, el conjunt d'àrees de les variables (o metabòlits) de les mostres metabolòmiques obtingudes bé mitjançant la resolució MCR-ALS o a través de la seva integració directa amb els programes instrumentals de la casa comercial es poden avaluar estadísticament mitjançant diverses eines que tracten independentment cadascuna de les variables per a la selecció de potencials biomarcadors.

Existeixen múltiples mètodes d'anàlisi univariants per a l'anàlisi de dades metabolòmiques. En aquesta Tesi s'han emprat el test *t* [227], el test *U* de Mann-Whitney [228] i l'anàlisi de la variància (ANOVA) [229], els quals permeten saber si la mitjana d'una variable (que habitualment representa la concentració d'un metabòlit) mostra una diferència significativa entre grups de mostres o poblacions. La selecció entre aquests mètodes ve definida pel nombre de grups a considerar i la distribució dels valors de la intensitat/senyal de cada metabòlit [230]. Per exemple, quan s'avaluen les diferències entre dos o més grups de mostres s'utilitzen tests paramètrics (que assumeixen que les dades segueixen una distribució normal) com el test *t* o l'ANOVA. Per tant, abans d'emprar qualsevol d'aquests tests per comparar les mitjanes de les poblacions, s'ha de verificar la distribució normal de les dades mitjançant, per exemple, els tests de Shapiro-Wilk o Kolmogorov-Smirnov [231].

D'una banda, la família dels *t*-tests s'utilitza per comparar les mitjanes entre dues poblacions o grups de mostres [232]. Segons les hipòtesis que compleixen les diferents poblacions pel que fa a la seva distribució normal i la igualtat de la variància, es poden aplicar diferents tipus de test *t*, com el *Student's t*-test (quan les variàncies de les dues poblacions es poden considerar iguals) o el test de

Welch (quan les variàncies de les dues poblacions no es poden considerar iguals). Per contra, en aquells casos que no es pot assumir la normalitat de les dades són preferibles els tests no paramètrics, com el test *U* de Mann-Whitney.

D'altra banda, l'anàlisi de la variància (ANOVA) és pot considerar com una generalització del test *t* per més de dos grups de mostres [232]. Es poden distingir diferents enfocaments d'ANOVA, com l'ANOVA d'una entrada (una única variable i un únic factor o tractament, *one-way ANOVA*) i l'ANOVA de dues entrades (una única variable i dos factors, *two-way ANOVA*). En totes aquestes variants, l'efecte dels factors s'avalua a partir d'una resposta única o univariant, però en l'últim cas es pot extreure més informació degut a que no només s'estima la importància dels factors sinó també la possible interacció entre ells. Quan no es pot assumir la normalitat de les dades, el test no paramètric Kruskal-Wallis és una bona alternativa a l'ANOVA per comprovar si la mitjana d'una variable mostra una diferència significativa entre grups de mostres o poblacions [232].

En el context de la metabolòmica no dirigida, el principal avantatge del mètodes univariants és la seva facilitat d'ús i interpretació dels resultats. Malgrat això, sovint aquests mètodes s'apliquen en paral·lel per als centenars o milers de metabòlits (variables) d'un estudi (depenent de les condicions experimentals emprades) [230, 233]. A mesura que augmenta la quantitat d'hipòtesis a estudiar en paral·lel, augmenta la probabilitat de rebutjar equivocadament una hipòtesi nul·la a l'atzar i, per tant, cometre errors de tipus I (falsos positius). Aquesta possibilitat d'acumular falsos positius es coneix com el problema de les comparacions múltiples (*multiple hypothesis testing*). Per poder controlar aquest problema associat a les dades amb un nombre elevat de variables, cal tenir en consideració la utilització de mètodes de comprovació addicionals als tests estadístics univariants. Així, es distingeixen diferents aproximacions per a la correcció dels valors *p* de significació obtinguts mitjançant aquestes anàlisis estadístiques univariants. Aquestes correccions es caracteritzen per evitar una sobreestimació (falsos positius) i subestimació (falsos negatius) de variables significatives. En cada estudi cal escollir entre diferents mètodes de correcció considerant el nombre total de variables i si es vol seguir un enfocament més o menys conservador. Per exemple, la correcció de Bonferroni s'utilitza per reduir l'error de tipus I quan múltiples comparacions es porten a terme en un únic conjunt

de dades [234]. El valor *p* corregit es calcula dividint el valor *p* pel nombre de comparacions simultànies que s'estan realitzant. El principal inconvenient de la correcció de Bonferroni és que és un enfocament massa conservador ja que redueix el nombre de falsos positius a costa de reduir també la quantitat de veritables positius. En conseqüència, s'han desenvolupat mètodes de correcció menys conservadors que es basen en la necessitat de minimitzar la taxa de falsos descobriments (*false discovery rate*, FDR) quan es porta a terme l'avaluació de múltiples hipòtesis. Per exemple, en estudis metabolòmics no dirigits, on s'analitzen simultàniament un gran nombre de metabòlits, l'ús de mètodes de correcció FDR menys estrictes, com per exemple el mètode Benjamini i Hochberg [235], és de gran utilitat per a la correcció dels valors *p* i evitar els falsos positius.

A la **Figura 2.21** es mostra el diagrama de flux de les diferents estratègies estadístiques univariants i multivariants que es poden emprar en els estudis de metabolòmica. A més, quan es comparen més de dos grups de mostres els test paramètrics poden ser mètodes univariants d'ANOVA (es compara una variable) i diferents mètodes multivariants (es comparen múltiples variables simultàniament). Hi ha diversos mètode multivariants per a l'anàlisi de la variància de les dades, com l'anàlisi ANOVA multivariant de la variància (*multivariant*-ANOVA, MANOVA), l'ANOVA amb anàlisi simultània de components (*ANOVA-simultaenous component analysis*, ASCA*)* i l'anàlisi de la variància multivariant regularitzada (*regularized MANOVA*, rMANOVA). Aquest mètodes multivariants es descriuran amb detall a la secció següent.

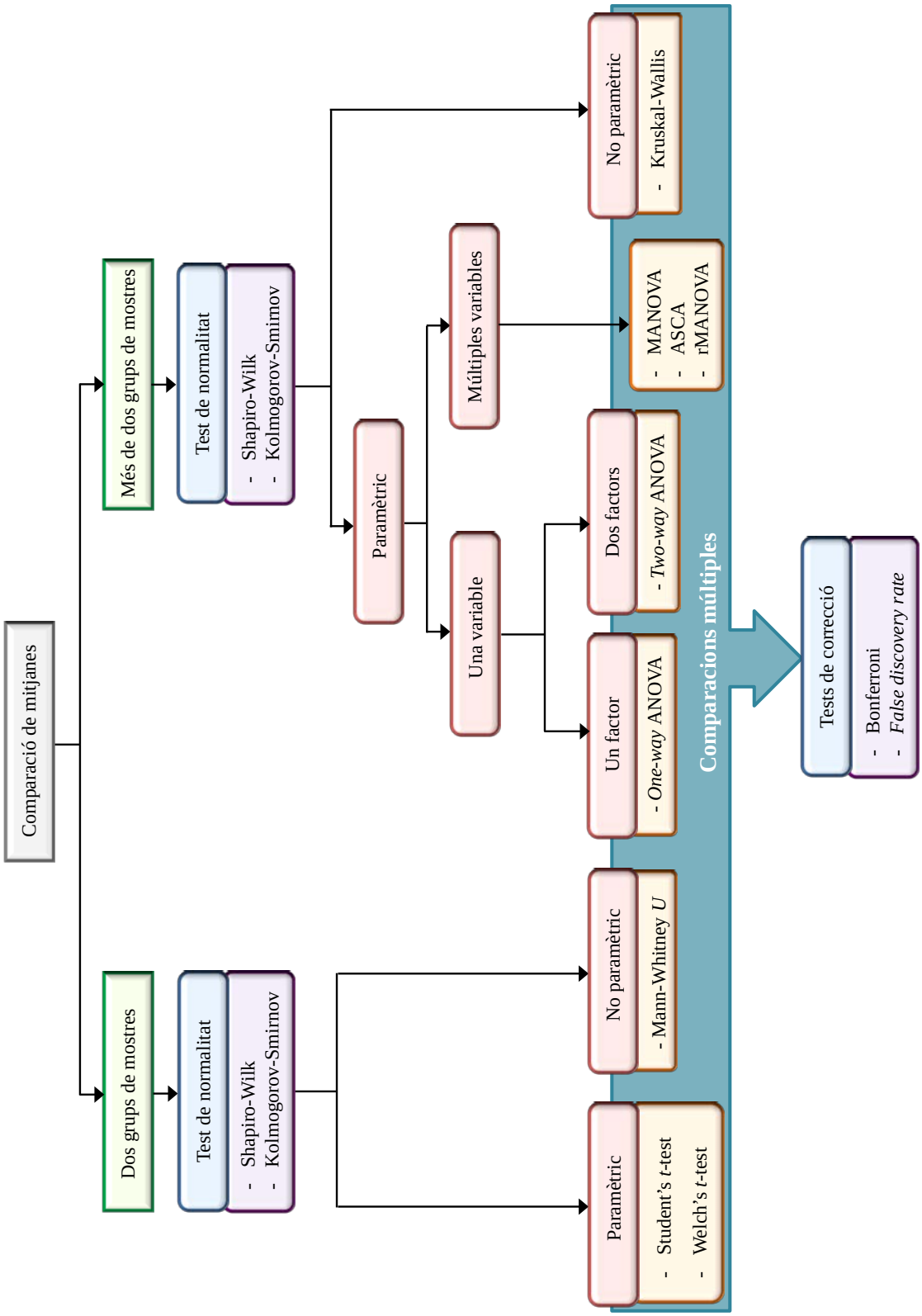


Figura 2.21. Diagrama de flux de l'anàlisi estadística per dur a terme la comparació de mitjanes entre dos o múltiples grups.

Anàlisi estadística multivariant

L'anàlisi estadística multivariant determina el comportament i/o la contribució de les múltiples variables mesurades en els canvis de la composició química de les mostres analitzades [236]. En el cas de la metabolòmica, l'anàlisi multivariant de les dades permet la diferenciació entre tipus de mostres i la detecció d'aquelles variables estudiades que causen la major part de la variació experimental.

Els mètodes d'anàlisi multivariant s'han desenvolupant amb l'objectiu d'obtenir informació a nivell de reconeixement de patrons, classificació i representació gràfica que expliqui la variació general de les dades i ajudi a la detecció de les variables més rellevants. A més, també s'han desenvolupat mètodes multivariants d'anàlisi de la variància de les dades per tal d'avaluar a nivell estadístic els efectes d'un o diversos factors (tractaments o exposicions) a través dels canvis observats en les respostes de les variables mesurades.

Mètodes d'agrupacions i classificació de mostres

Una de les principals finalitats quan s'analitzen dades metabolòmiques és l'agrupació i la classificació de les mostres estudiades segons els diferents factors investigats. D'aquesta manera, s'intenta classificar les mostres en diferents grups per tal de disminuir la complexitat de les dades i aconseguir extreure similituds i diferencies entre elles de la forma més intuïtiva possible. En conseqüència, el principal objectiu dels mètodes d'agrupació de mostres és definir grups de mostres de forma que es minimitzi la variació dins de cada grup al mateix temps que es maximitza la variació entre grups o classes diferents [237]. Es distingeixen dos tipus de classificació depenent si es fa servir coneixement previ (anàlisi d'agrupacions supervisada) o no (anàlisi d'agrupacions no supervisada). Els mètodes d'anàlisi d'agrupacions no supervisats donen una visió general de les dades per a la detecció de grups d'objectes o *clusters* sense cap informació prèvia. Per tant, la creació del nombre i tipus de grups es farà en base a les similituds entre mostres o variables de les dades. En canvi, els mètodes d'anàlisi d'agrupacions supervisats es basen en classificar les mostres a partir d'informació coneguda prèvia. Per exemple, es pot indicar el nombre de grups de mostres que s'espera que siguin presents al conjunt de dades analitzat.

A continuació, es presenten els diferents mètodes d'agrupacions emprats en aquesta Tesi. Primer, es fa una descripció detallada dels mètodes no supervisats de l'anàlisi d'agrupacions jeràrquica (HCA) i l'anàlisi de components principals (PCA). Finalment, s'explica el fonament del mètode supervisat de l'anàlisi discriminant per mínims quadrats parcials (PLS-DA).

Anàlisi d'agrupacions jeràrquica i mapes de calor

L'anàlisi d'agrupacions jeràrquica (*hierarchical cluster analysis*, HCA) és un mètode d'anàlisi multivariant no supervisat que dur a terme una exploració de les dades experimentals i permet obtenir una visió general de la relació entre les mostres o les variables.

L'agrupament jeràrquic transforma les distàncies multidimensionals entre objectes d'una matriu de dades en un conjunt de particions ordenades jeràrquicament [231, 238]. Aquesta jerarquització es pot representar gràficament mitjançant un dendrograma en forma d'arbre. La jerarquització es basa en una determinada mesura de proximitat (distància) entre els grups d'objectes en que cadascun d'aquests grups queda inclòs en un grup més gran fins arribar a un únic grup (**Figura 2.22**). Els grups s'obtenen de l'agrupació de parells d'objectes d'acord a una matriu de correlació que avalua cada parell d'objectes possibles. El parell que produeixi una major intercorrelació formarà un grup nou i així successivament. D'aquesta manera, es mostra les relacions entre grups i el criteri d'agrupació associat a cadascun. Hi ha diferents formes de mesurar les distàncies i d'enllaçar els diferents grups [231]. Per mesurar les distàncies es fan servir, per exemple, la distància euclídea, la distància de Manhattan, la distància de Mahalanobis o el coeficient de correlació de Pearson. Per enllaçar els diferents grups hi ha mètodes com l'enllaç simple, l'enllaç complet, l'enllaç intermedi, l'enllaç al veí més proper o el mètode de Ward.

El mètode d'agrupacions jeràrquic és un dels mètodes més emprats per a l'anàlisi de les dades òmiques [239-242]. Un dels principals avantatges que presenta és que no requereix de cap tipus d'informació prèvia sobre les dades. Malgrat això, presenta dificultats per a treballar amb conjunts de dades molt grans degut als recursos computacionals necessaris per poder calcular les correlacions entre milers d'objectes.

En aquesta Tesi, l'anàlisi jeràrquic de les dades òmiques s'ha dut a terme mitjançant la funció *clustergram* de la Bioinformatics Toolbox dins l'entorn de programació i visualització de MATLAB (The Mathworks, MA, USA). Aquesta funció s'ha emprat fent servir la distància euclídea i l'enllaç intermedi per a generar l'arbre jeràrquic, el qual es visualitza mitjançant dendogrames i un mapa de calor (veure **Figura 2.22**). Un mapa de calor és una representació bidimensional de les dades experimentals en què els valors de la matriu són representats per colors [243]. Així doncs, es mostra la correlació entre els valors de la matriu amb escales de colors, proporcionant una visió general de la informació de les dades de forma molt senzilla i intuïtiva.

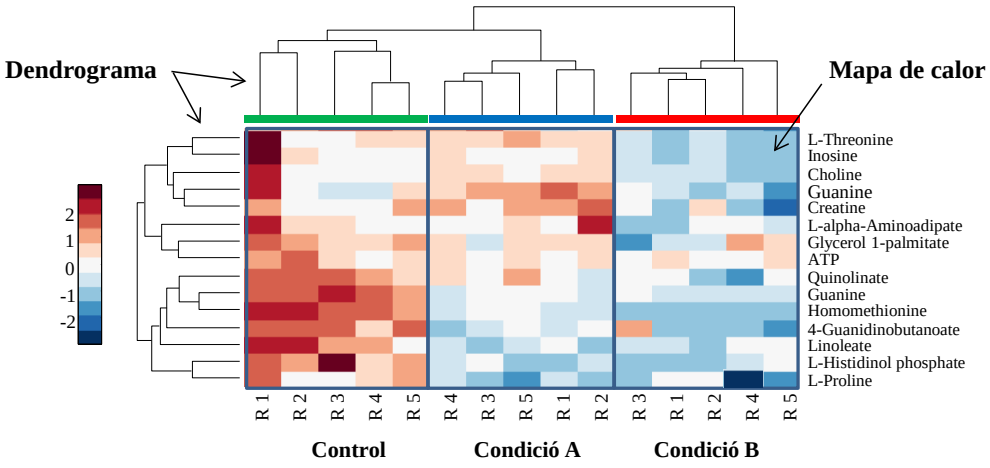


Figura 2.22. Dendrograma i mapa de calor aplicats a les dades de metabolòmica. Cada cel·la del mapa de calor indica l'abundància de cada variable (metabòlit) de menys (blau) a més (vermell).

Anàlisi de components principals (PCA)

L'anàlisi de components principals (*principal component analysis*, PCA) [216, 244] és probablement el mètode exploratori més àmpliament emprat dins de la família de mètodes d'agrupaments no supervisats i es troba disponible gairebé en tots el paquets d'anàlisi estadística. L'anàlisi per PCA és extensament conegut en diferents àmbits de la química i ,en particular, en la química ambiental [245, 246] i l'òmica [240, 247, 248].

El principal objectiu de PCA és reduir la dimensionalitat del conjunt de dades en què hi ha un gran nombre de variables interrelacionades, sempre i quan es conservi el màxim possible d'informació

present a les dades originals. Així, el mètode de PCA aconsegueix condensar aquesta informació sobre el sistema utilitzant un nombre menor de variables per explicar la variància observada en les dades i eliminar variables que aporten informació redundant. La simplificació de la informació del sistema és conseqüència de la combinació lineal de les variables originals en un nou conjunt de variables ortogonals no correlacionades entre sí, anomenades components principals o factors que proporcionen una millor i més fàcil interpretació de les dades (**Figura 2.23**).

Matemàticament, PCA es basa en la descomposició de la informació continguda en la matriu de dades **D**, seguint un model bilineal similar a l'emprat en el cas de MCR-ALS. Així, el model bilineal de l'anàlisi de components principals descompon la matriu de dades experimentals **D** en el producte de dues matrius, **T** i $\mathbf{P^T}$:

$$\mathbf{D} = \mathbf{TP^T} + \mathbf{E} \qquad \text{Equació 2.6}$$

on les matrius **T** i $\mathbf{P^T}$ defineixen, respectivament, la matriu d'*scores* (projecció de les mostres) i la matriu de *loadings* (projecció de les variables) i **E** és la matriu de residuals no explicats pel model i que es pot relacionar amb el soroll experimental. Per tant, els components principals defineixen un nou espai de coordenades on es poden visualitzar tant les mostres com les variables. La visualització de la matriu d'*scores* s'anomena gràfic d'*scores* (*scores plot*) i dóna, per tant, una visió general de la distribució i agrupació de les mostres analitzades segons la seva similitud (clústers) (veure **Figura 2.23**). A més, d'establir relacions entre les mostres, també permet detectar mostres amb valors atípics (*outliers*). En canvi, el gràfic de *loadings* (*loadings plot*) descriu la contribució de cadascuna de les variables originals de la matriu **D** en cada component descrit pel model de PCA. Les variables amb valors grans de *loadings* sobre el mateix component, covarien. Si tenen el mateix signe, covarien en la mateixa direcció o positivament mentre que si tenen signes oposats, ho fan en direccions oposades o negativament (covarien inversament). Els components dels gràfics d'*scores* i *loadings* són els mateixos, per tant serà fàcil correlacionar les característiques generals de les mostres (agrupacions) amb les seves fonts de variació experimental.

Quan es tracta de dades metabolòmiques, la representació simultània dels gràfics d'*scores* i *loadings* pot indicar quines són les agrupacions de les mostres i obtenir informació sobre els perfils metabòlics. Així, PCA permet determinar el nombre de grups de mostres i quins metabòlits contribueixen més en la seva diferenciació [249].

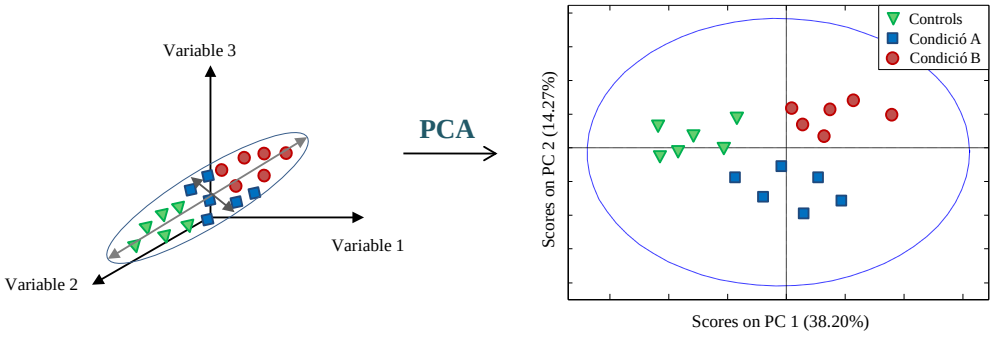


Figura 2.23. Projecció de les dades originals (mostres) en els nous eixos ortogonals (components principals).

La descomposició matemàtica de l'equació 2.6 és realitza sota restriccions de màxima variància explicada i ortogonalitat. Aquest fet implica que els components principals s'ordenin sempre en funció de la variància explicada en ordre descendent. Per tant, el primer component es trobarà en la direcció que expliqui el màxim percentatge de variància, i el mateix pel segon component que al ser ortogonal al primer explica la variància restant no descrita pel primer, i així successivament. D'aquesta manera, només els primers components principals, els quals descriuen la major part de la variància, seran seleccionats per explicar la variació experimental observada en la matriu **D**. És per això, que en els estudis de metabolòmica de MS els components posteriors poden ser omesos sense una pèrdua significativa d'informació rellevant ja que només explicaran una petita part de la variància, la qual estarà relacionada amb el soroll experimental.

En aquesta Tesi, l'anàlisi quimiomètrica de les dades per PCA s'ha realitzat utilitzant la PLS Toolbox (Eigenvector Research Ltd., Manson, WA, USA).

Anàlisi discriminant per mínims quadrats parcials (PLS-DA)

El mètode de mínims quadrats parcials (*partial least squares*, PLS) [250] és un mètode multivariant de regressió que avalua la relació entre una matriu de variables predictores **X** i un vector resposta de variables dependents **y**, per un mateix conjunt de mostres seguint la següent equació:

$$\mathbf{y} = \mathbf{bX} \qquad \text{Equació 2.7}$$

on **b** és el vector que conté els coeficients de regressió calculats durant el calibratge (construcció) del model.

El mètode PLS s'utilitza principalment en el camp del calibratge multivariant on el vector resposta **y** és quantitatiu. Aquest mètode s'ha modificat posteriorment per finalitats de classificació donant lloc a l'anomenat anàlisi discriminant per mínims quadrats parcials (*partial least squares-discriminant analysis*, PLS-DA), on el vector resposta és qualitatiu [251].

L'anàlisi discriminant de PLS-DA és avui en dia un dels mètodes d'agrupament supervisats més emprats en el marc de la metabolòmica [252] i la transcriptòmica [253]. Com a mètode de regressió lineal multivariant, el mètode PLS-DA permet trobar models de correlació entre la matriu de dades experimental **X** (conjunt de variables predictores) i el vector **y** (variables a predir) (**Figura 2.24a**) que conté informació de les classes o categories de mostres de la matriu **X** [254]. El mètode PLS-DA redueix la dimensionalitat del conjunt de dades experimentals mitjançant la descomposició en un conjunt de components o factors anomenats variables latents (*latent variables*, LVs) que representen la màxima covariància entre **X** i **y**. Aquestes noves variables (LVs) s'obtenen de la combinació lineal dels valors originals i independents de la matriu **X** i es correlacionen de manera òptima amb les classes de la variable dependent **y**. Així doncs, la regressió construeix un conjunt de vectors de coeficients de pes (*weights*, **w**) que s'agrupen en una matriu **W**, la qual descriu la relació entre **X** i **y** durant el procés de regressió i s'utilitza per calcular el vector de regressió **b** a partir de les següents equacions [255]:

$$\mathbf{b} = \mathbf{X}^{+}\mathbf{y} \qquad \text{Equació 2.8}$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{T}\mathbf{W})^{-1}\mathbf{Q}^{T}$$ Equació 2.9

Les fórmules matricials que descriuen el model matemàtic de PLS es descriuen a continuació (equació 2.10 i 2.11):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E}$$ Equació 2.10

$$\mathbf{y} = \mathbf{U}\mathbf{Q}^{T} + \mathbf{F}$$ Equació 2.11

T i **U** són les matrius d'*scores* de **X** i **y**. **P** i **Q** són les matrius de *loadings*, i **E** i **F** són les matrius de residuals no explicats pel model de PLS. Els superíndexs T i + indiquen la transposició i la pseudoinversa de les matrius, respectivament.

En la construcció del model PLS-DA és molt important definir correctament el nombre de LVs necessàries. Per aquest motiu, és essencial un procés de validació durant la generació del model. Si la quantitat de mostres experimentals no és suficient per a obtenir conjunts de dades de calibratge i de validació, és necessari l'ús d'un mètode de validació creuada adequat per evitar errors en el model. La qualitat del model PLS-DA s'expressa a través de la seva sensibilitat (probabilitat de classificar correctament una mostra a la classe a la qual pertany) i especificitat (probabilitat d'assignar correctament una mostra que no pertany a la classe) que s'obtenen a partir de la matriu de confusió del model. La matriu de confusió per dues classes es pot estructurar de la següent manera (la classe condició A s'identifica com positiva, P, i la classe control com negativa, N):

Taula 2.4. Diagnòstic del model PLS-DA.

		Classe predita	
		Condició A (P)	**Control (N)**
Classe experimental	**Condició A**	VP	FN
	Control	FP	VN

Sensibilitat = VP / (VP + FN) Equació 2.12

Especificitat = VN / (VN + FP) Equació 2.13

on VP són els veritables positius, FN els falsos negatius, VN els veritables negatius i FP els falsos positius.

A partir d'un model PLS-DA, es pot obtenir informació diversa. D'una banda, el model de PLS-DA ofereix gràfics d'*scores* (visualització de les mostres) (**Figura 2.24b**) i gràfics de *loadings* (visualització de les variables) que proporcionen informació similar a la de l'anàlisi per PCA. D'altra banda, l'anàlisi multivariant de PLS-DA disposa de mètodes de selecció de variables que permeten la tria d'un conjunt més petit de variables de la matriu **X** que són les que estan realment correlacionades amb la variable resposta o efecte investigat (variable dependent **y**). Exemples de mètodes de selecció de variables són els quocients de selectivitat (*selectivity ratio*, SR) [256] o el mètode de detecció de variables importants en la projecció (*variable importance in projection*, VIP) [257].

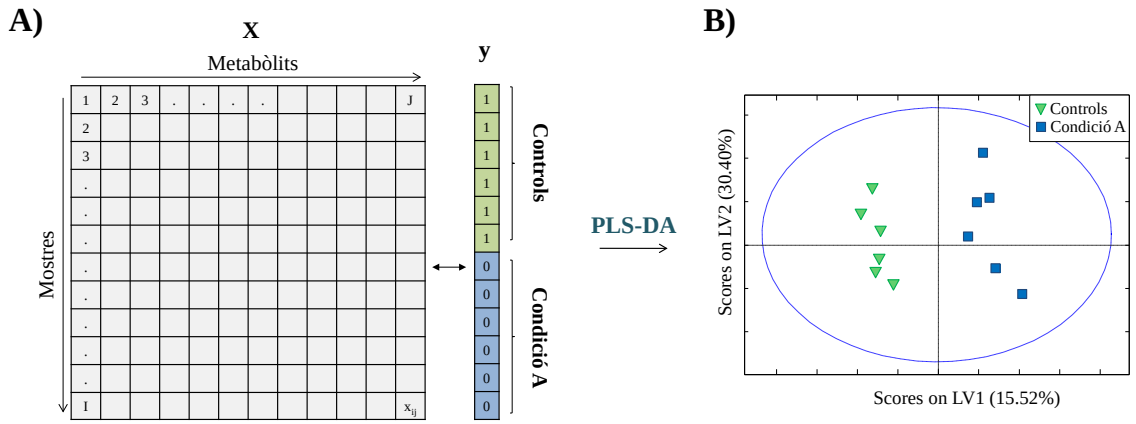


Figura 2.24. Representació (a) del model PLS-DA i (b) del gràfic d'*scores* obtingut a partir de l'anàlisi per PLS-DA per la discriminació de dues classes.

En aquesta Tesi, el procediment emprat per a la selecció de les variables rellevants relacionades amb la variació de **y** s'ha basat en l'aplicació del mètode de selecció de variables importants en la projecció (VIP).

Variables importants en la projecció (VIP)

El mètode de selecció de les variables importants en la projecció (VIP) va ser proposat per Wold el 1993 [257]. Aquest mètode dóna una mesura de la influència o importància (valor VIP) de cadascuna de les variables de la matriu **X** en la construcció del model PLS per predir **y**. D'aquesta manera, es pot seleccionar quines són les variables que contribueixen en l'explicació de la variància de la resposta **y**. Com més gran sigui el valor VIP d'una variable, més important és en la construcció del model.

El valor o coeficient VIP (*VIP scores*) de cada variable (metabòlit) es calcula a partir de la suma ponderada dels quadrats dels coeficients de pes (**w**) obtinguts en la construcció del model PLS [258]. Per exemple, el valor de VIP de la variable (metabòlit) *j* es pot descriure segons l'expressió següent:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}} \qquad \text{Equació 2.14}$$

$$SSY_f = b_f^2 t_f' t_f \qquad \text{Equació 2.15}$$

$$SSY_{total} = \mathbf{b}^2\mathbf{T'T} \qquad \text{Equació 2.16}$$

on w_{jf} és el valor del pes de la variable *j* en la variable latent *f*, SSY_f la suma de quadrats de la variància explicada per la variable latent *f*, J el nombre de variables (metabòlits) en la matriu de dades **X**, SSY_{total} la suma de quadrats total de la variable dependent **y**, F el nombre total de variables latents considerades, T els *scores* de la matriu **X** i **b** el vector de coeficients del model PLS. VIP_j és una mesura de la contribució de la variable *j* en funció de la variància explicada per cada variable latent del model PLS, en la qual w_{jf}^2 representa la importància de la variable *j* en la variable latent *f*.

Generalment, es designa arbitràriament la unitat com a valor llindar a partir del qual les variables o metabòlits es consideren importants per a la construcció del model de PLS, atès que aquest valor coincideix amb la mitjana dels valors dels VIP al quadrat [258]. A la **Figura 2.25** es mostra un exemple de gràfic d'*scores* VIP on els valors superiors a la unitat serien escollits com a metabòlits rellevants per a la discriminació entre classes i, en conseqüència, avaluats com potencials biomarcadors metabòlics.

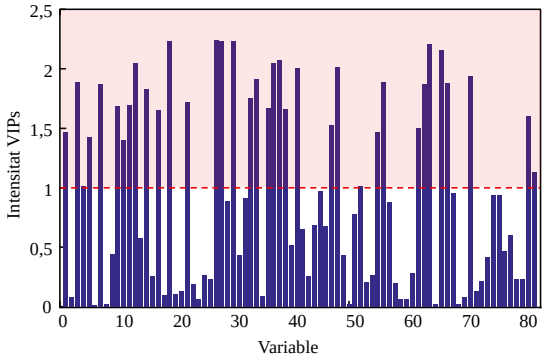


Figura 2.25. Exemple de gràfic d'*scores* VIP de l'anàlisi PLS-DA.

En aquesta Tesi, l'anàlisi de PLS-DA s'ha realitzat utilitzant la PLS Toolbox (Eigenvector Research Ltd., Manson, WA, USA).

Mètodes d'anàlisi de la variància

L'ANOVA és un mètode estadístic molt emprat en l'anàlisi de dades que segueixen un disseny experimental [259]. Quan s'avalua una única variable en funció dels factors experimentals, ANOVA és una eina estadística que es basa en la comparació dels valors d'una variable (metabòlit) dins d'un mateix grup de mostres respecte als valors que pren la mateixa variable entre diferents grups de mostres. Quan es mesuren múltiples variables en cadascun dels objectes (mostres), l'ANOVA multivariant no té en compte les possibles correlacions (covariància) entre les diferents variables mesurades i es proposa emprar altres estratègies alternatives.

L'extensió multivariant d'ANOVA coneguda com anàlisi multivariant de la variància (*multivariate-ANOVA*, MANOVA) [260, 261] és una primera possible opció quan es treballa amb aquests tipus de dades multivariants. En aquest mètode es pren en consideració tant el disseny experimental com la covariància entre de les variables mesurades. Amb aquest objectiu, s'avalua tant la dispersió dins d'un grup de mostres com la que hi ha entre els diferents grups de mostres. A l'equació 2.17, la matriu **B** mesura la diferència entre les mitjanes dels diferents grups d'un factor específic i la matriu **W** mesura la dispersió de les mostres dins d'un mateix grup respecte la seva mitjana. Així, MANOVA estableix una relació entre la dispersió inter- i intra-grup per tal de determinar si un efecte és significatiu mitjançant la següent equació:

$$\mathbf{J} = \mathbf{W}^{-1}\,\mathbf{B} \qquad \text{Equació 2.17}$$

No obstant, quan el nombre de mostres és molt més petit que el nombre de variables correlacionades, com és el cas de les dades òmiques, la inversa de la matriu **W** no es pot calcular (deficiència de rang) i s'han d'emprar altres alternatives [262]. És per això, que en els últims anys s'han desenvolupat altres models que permetin l'anàlisi de conjunts de dades amb un gran nombre de variables molt comuns en quimiometria i en els estudis d'òmica, com l'ANOVA amb anàlisi simultània de components

(*ANOVA-simultaenous component analysis*, ASCA) [263] o l'anàlisi de la variància multivariant regularitzada (*regularized MANOVA*, rMANOVA) [264].

ANOVA amb anàlisi simultània de components (ASCA)

L'ANOVA amb anàlisi simultània de components (*ANOVA simultaneous component analysis*, ASCA) [263] es va proposar l'any 2005 com una alternativa al mètode MANOVA en el camp de metabolòmica. ASCA és un mètode multivariant especialment útil per a l'anàlisi de conjunts de dades complexes provinents de dissenys experimentals, com les dades metabolòmiques. En particular, aquest mètode combina els avantatges de l'ANOVA amb l'anàlisi simultània de components (*simultaneous component analysis*, SCA), que és un mètode similar al PCA. En la metodologia ASCA, SCA s'aplica sobre les matrius dels efectes de cada factor per separat, segons el disseny experimental emprat, conjuntament també amb totes les possibles matrius d'interacció entre factors.

En aquesta Tesi el mètode ASCA s'ha aplicat a dissenys experimentals equilibrats (*balanced*). És a dir, quan per cada nivell dels factors investigats té exactament el mateix nombre de replicats. Per exemple, el model ANOVA considerant tres factors (A, B i C) és pot descriure matemàticament de la següent manera:

$$\mathbf{X} = \overline{\mathbf{X}} + \mathbf{X_A} + \mathbf{X_B} + \mathbf{X_C} + \mathbf{X_{AB}} + \mathbf{X_{AC}} + \mathbf{X_{BC}} + \mathbf{X_{ABC}} + \mathbf{E} \qquad \text{Equació 2.18}$$

on $\mathbf{X}$ és la matriu experimental i $\overline{\mathbf{X}}$ la matriu de mitjanes de la matriu experimental. Les matrius $\mathbf{X_A}$, $\mathbf{X_B}$, $\mathbf{X_C}$ corresponen a les variacions observades degudes als efectes individuals dels diferents factors considerats. Les matrius $\mathbf{X_{AB}}$, $\mathbf{X_{AC}}$, $\mathbf{X_{BC}}$, $\mathbf{X_{ABC}}$ són les matrius de les variacions observades degudes als efectes d'interacció de les diferents combinacions possibles dels factors considerats. Finalment, $\mathbf{E}$ és la matriu residual que representa la variació que no esta explicada pel model ANOVA. En les dades òmiques, aquesta matriu $\mathbf{E}$ conté la variació del soroll degut a la mesura experimental i a la variació individual no comuna entre els replicats.

En els dissenys experimentals equilibrats, s'utilitza la suma de quadrats de tipus I per avaluar els efectes dels diferents factors estudiats en l'anàlisi ANOVA. La suma de quadrats tipus I es basa en

l'assignació seqüencial de la part de la variància explicada a cadascun dels factors principals considerats; posteriorment a les interaccions bidireccionals i, així successivament, incrementant l'ordre de les interaccions. Degut a aquest procés de subtracció successiva (deflació), la suma de quadrats corresponents a cadascun dels factors, a les seves interaccions i a la variància residual és ortogonal. En el cas de dissenys experimentals no equilibrats (*unbalanced*), l'algoritme d'ASCA és més complex degut a que els factors del disseny es correlacionen entre si, és a dir, no són ortogonals.

Quan s'aplica SCA sobre cada matriu de factors i d'interacció, el model matemàtic d'ASCA (*balanced*) de l'equació 2.18 es pot descriure de la següent manera:

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{T_A P^T_A} + \mathbf{T_B P^T_B} + \mathbf{T_C P^T_C} + \mathbf{T_{AB} P^T_{AB}} + \mathbf{T_{AC} P^T_{AC}} + \mathbf{T_{BC} P^T_{BC}} + \mathbf{T_{ABC} P^T_{ABC}} + \mathbf{E} \qquad \text{Equació 2.19}$$

on cadascuna de les matrius **T** i **P** són les matrius d'*scores* i de *loadings* pels diferents efectes dels factors i interaccions del model de l'equació 2.18, respectivament. Aquesta combinació d'ANOVA i SCA permet obtenir informació sobre la distribució de les mostres per cada factor individual estudiat i sobre quina és la influència de cada factor sobre les diferents variables o metabòlits [265]. A més, el gràfic d'*scores* permet observar el comportament dels diferents tipus de mostres en relació a cada factor, representant el valor de la mitjana de les mostres corresponents a cada nivell del factor estudiat.

En el mètode ASCA la significació estadística dels efectes de cada factor i de les seves interaccions es pot investigar mitjançant un test de permutació [265, 266]. Aquest test s'aplica de manera que es permuten un nombre elevat de vegades totes les mostres (les files de les matrius de dades originals) segons tots els factors i es recalculen les sumes de quadrats. La determinació de la importància dels factors estudiats consisteix en avaluar la freqüència dels valors de la suma de quadrats de totes les matrius permutades per un factor determinat (**Figura 2.26**). Quan la suma de quadrats de les matrius permutades és menor a la suma de quadrats de la matriu original es rebutja la hipòtesi nul·la de no presència d'efecte del factor considerat (H_0) (**Figura 2.26a**) i quan és igual s'accepta que no hi ha hagut efecte del factor considerat H_0 (**Figura 2.26b**). Així doncs, la H_0 del test de permutació

reflecteix que el factor considerat no produeix cap efecte sobre les dades experimentals. La significació estadística dels factors es quantifica mitjançant el valor *p*. El valor *p* de la significació estadística es calcula dividint el nombre de casos en el qual la suma de quadrats de les matrius permutades és major a l'original entre el nombre total de permutacions realitzades. Aquest valor *p* representa el nivell de significació més petit pel que la mostra obtinguda obligaria a rebutjar la H_0. Si el valor *p* és menor que el nivell de significació prefixat es rebutja la H_0 (el factor té efecte), en canvi, si és major aquesta hipòtesi s'accepta (el factor no té cap efecte). Es necessari per tant que el nombre de permutacions fetes sigui elevat. En aquesta Tesi s'ha emprat generalment un nombre de permutacions igual a 10.000.

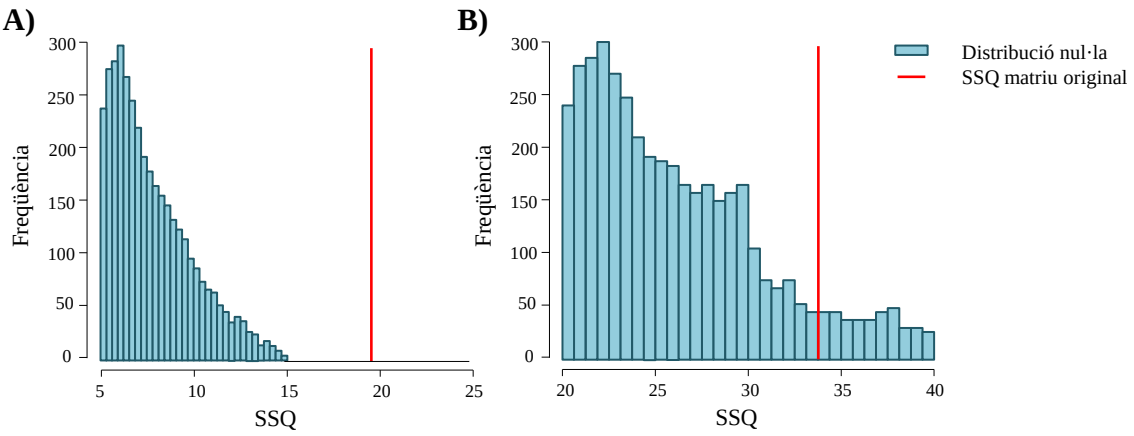


Figura 2.26. Exemple d'histograma d'un estudi metabolòmic amb un disseny experimental (DoE) equilibrat de dos factors: (a) on el factor A mostra un efecte significatiu, la suma de quadrats de la matriu original (línia vermella) no es troba superposada a la matriu permutada (barres blaves) i (b) on el factors B no mostra un efecte significatiu, la matriu original (línia vermella) està superposada a la matriu permutada (barres blaves).

Anàlisi de la variància multivariant regularitzada (rMANOVA)

L'anàlisi de la variància multivariant regularitzada (rMANOVA) [264] va ser proposada per Engel *et al*. el 2015 com a mètode alternatiu als models de MANOVA i ASCA, i té un especial interès la seva possible aplicació en estudis òmics. Aquest mètode és descrit pels propis autors com una mitjana ponderada dels models d'ASCA i de MANOVA. Contràriament a MANOVA, rMANOVA es pot emprar en els conjunts de dades on el nombre de variables o metabòlits (*p*) és considerablement

superior al de mostres (*n*) de la mateixa manera que l'ASCA. Ara bé, a diferència de l'ASCA, aquest mètode té en compte la correlació entre les variables en l'etapa de construcció del model [264, 267]. El model rMANOVA es basa en una equació molt similar a l'equació 2.17 del model MANOVA (**J = $\mathbf{W^{-1}}$ B**):

$$\mathbf{J^* = ([1-\delta]W + \delta T)^{-1}B} \qquad \text{Equació 2.20}$$

on la matriu **T** és la matriu objectiu (o estructura anterior a la matriu **W**). La regularització s'utilitza en molts mètodes multivariants per obtenir una bona estimació quan es tracta dades de grans dimensions. El paràmetre $\delta \in [0,1]$ s'utilitza per sospesar entre el biaix i la variància. El model ASCA ($\delta = 1$, T = **I**) considera que tot és biaix i no hi ha variància. Per contra, MANOVA ($\delta = 0$) té en compte que tot és variància i no existeix cap biaix. El valor òptim de δ (δ_{opt}) es calcula de forma ràpida segons el teorema de Ledoit-Wolf [268].

rMANOVA presenta un altre aspecte de gran importància en els estudis metabolòmics, permet detectar els metabòlits (*feature detection*) més importants per a la construcció del model. A més, l'avaluació de la significació dels factors es pot determinar mitjançant un test de permutació que estima la distribució nul·la de les dades de forma anàloga al que s'ha descrit en el mètode d'ASCA.

2.3.7. Fusió o integració de dades òmiques

Avui en dia, hi ha una tendència creixent a que els estudis òmics integrin simultàniament mesures de diferents nivells òmics, de diferents tècniques analítiques o de diferents matrius biològiques per a obtenir una visió més acurada del sistema estudiat. Malgrat que l'anàlisi dels diferents conjunts de dades (més o menys) heterogenis és un repte, aquests enfocaments multi-òmics han donat lloc a una nova era en la biologia dels sistemes amb una millor comprensió de la dinàmica dels sistemes cel·lulars i els principis funcionals dels processos biològics subjacents [269, 270].

En l'actualitat hi ha la necessitat de desenvolupar noves aproximacions de fusió de dades, les quals constitueixen un dels principals objectes d'investigació en els camps de l'estadística computacional i de la bioinformàtica.

La integració de les dades es pot aplicar de maneres diferents i a diferents nivells. El 2009, T. Ebbels i R. Cavill [271] van suggerir la possibilitat de dur a terme tres tipus diferents d'integració de dades: integració conceptual o de nivell alt, integració de nivell mig i integració de nivell baix (**Figura 2.27**) [272, 273]. Aquestes tres opcions es diferencien segons la quantitat de modelatge matemàtic, les hipòtesis requerides i la facilitat d'implementació.

La integració de dades més senzilla des d'un punt de vista quimiomètric és la integració conceptual o fusió de nivell alt. En aquest cas, cada bloc de dades òmiques s'analitza de forma independent, i després, les conclusions resultants es consideren conjuntament per tenir una visió més global de la informació proporcionada per les dades, és a dir, per aconseguir fer una interpretació biològica conjunta de les dades òmiques (**Figura 2.27a**). Encara que aquest enfocament pot aportar coneixement molt valuós, no pot evitar que es puguin perdre associacions importants que només es poden trobar quan tots els conjunts de dades s'analitzen simultàniament, com és en el cas de fer una integració de nivell mig i de nivell baix. La integració o fusió de nivell mig és una forma comuna d'integració de dades en el camp de la òmica on es busquen associacions estadístiques entre variables pre-seleccionades dels diferents blocs de dades per tal de generar una hipòtesi conjunta d'aquests (**Figura 2.27b**). Tot i que la reducció de la dimensió de les dades facilita la seva anàlisi estadística, la informació resultant d'aquesta estratègia sovint és incompleta i/o difícil d'interpretar. Finalment, la integració o fusió de nivell baix fa referència a la situació en la qual un únic model computacional o matemàtic és capaç de descriure la variància conjunta dels diferents blocs de dades i, a més, el model pot modelitzar i predir cada nivell d'organització biològica (**Figura 2.27c**). Diferents mètodes proposats des de la quimiometria permeten integrar els blocs de dades de diferents orígens. Qualsevol que sigui el tipus de fusió emprat es pot aplicar a la integració de diferents tipus de dades, com per exemple de diferents plataformes (*inter-platform integration*) (per exemple, CE i LC), de diferents tipus de matrius biològiques (*inter-sample integration*) (per exemple, múscul i cervell) i/o de diferents nivells moleculars (*inter-omic integration*) (per exemple, transcriptòmica i metabolòmica) [269, 271].

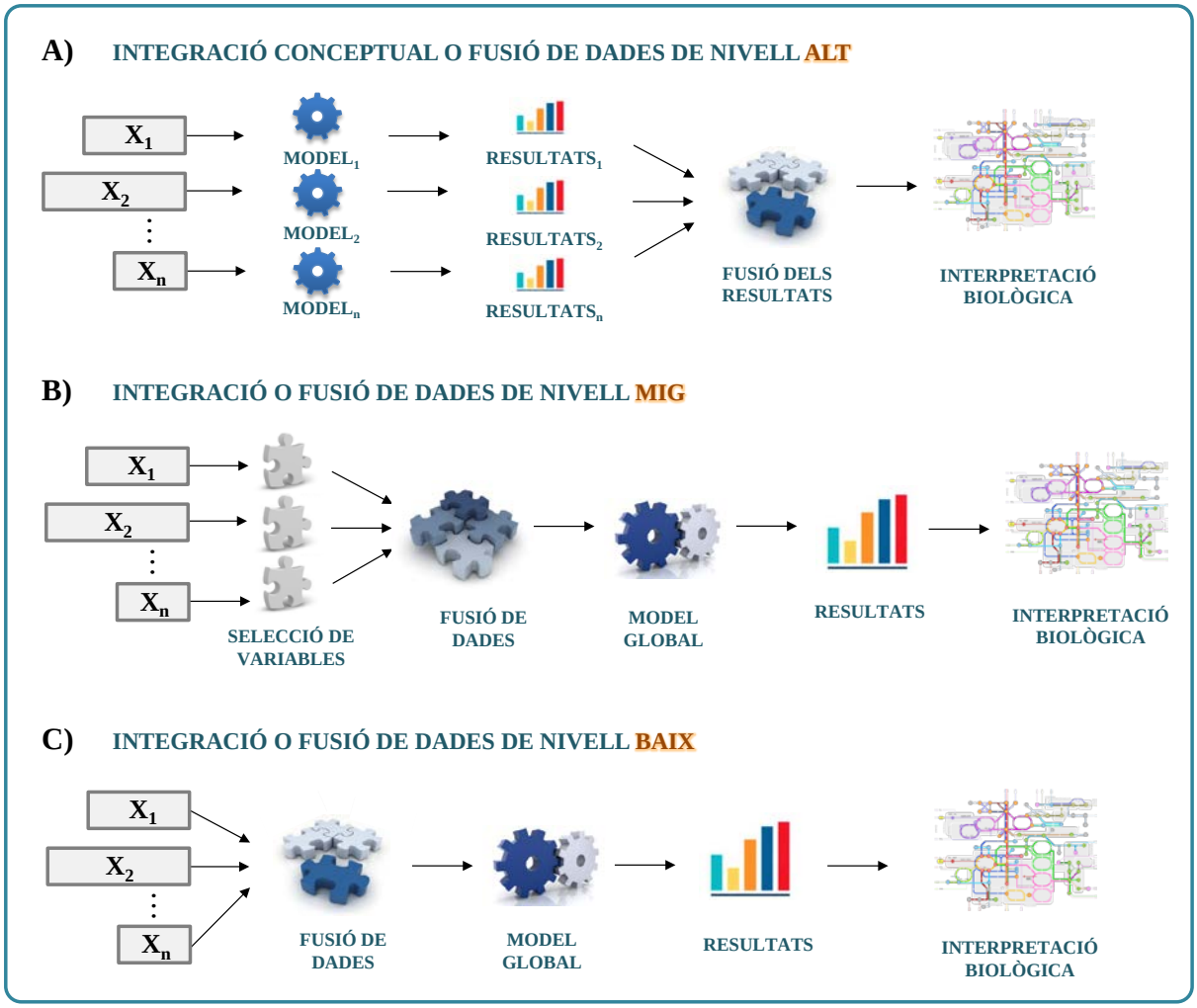


Figura 2.27. Principals estratègies de fusió de dades, adaptada de [273].

En el cas de la fusió de dades de nivell alt (**Figura 2.27a**), cada conjunt de dades s'analitza de forma independent utilitzant mètodes multivariants per a la classificació de mostres i la detecció de les variables rellevants [274, 275]. En canvi, la fusió de dades de nivell mig ha necessitat d'esforços addicionals per al desenvolupament de metodologies apropiades de fusió. Els mètodes més emprats en aquests estudis tracten d'identificar la variància comuna i específica de cadascun dels blocs analitzats després de la selecció de les variables més rellevants de cada bloc, de tal manera que es redueix la dimensió dels conjunts de dades. S'han presentat nombroses estratègies per a la integració de dades de diferents fonts a aquest nivell. Entre elles, destaca el mètode de descomposició generalitzada de valors singulars (*generalized singular value decomposition*, GSVD) [276], el mètode de projeccions ortogonals bidireccionals a estructures latents (*two-way orthogonal projections to latent structures*,

O2PLS) [277], el mètode de projeccions ortogonals multidimensionals a estructures latents (*multiblock orthogonal projections to latent structures*, OnPLS) [278], el mètode de components distintius i comuns amb l'anàlisi simultània de components (*distinctive and common components with simultaneous component analysis*, DISCO-SCA) [274], el mètode de variació individual i conjunta explicada (*joint and individual variation explained*, JIVE) [279] i el mètode de matriu acoblada i factorització tensora (*coupled matrix and tensor factorization*, CMTF) [280]. Algunes d'aquestes estratègies de fusió de nivell mig també es poden utilitzar per a la fusió de dades de nivell baix però cal prèviament preescalar les dades experimentals originals. Tanmateix, el mètode de MCR-ALS abans descrit permet també a través de l'augmentació de matrius de dades [195, 196] fer en moltes ocasions l'anàlisi conjunta de dades d'orígens diversos, és a dir, permet fer l'anàlisi de dades fusionades. En aquesta Tesi, es presenta un mètode de fusió de dades basat en MCR-ALS per a la integració de dades metabolòmiques de CE-MS i LC-MS.

2.4. IDENTIFICACIÓ DELS METABÒLITS I INTERPRETACIÓ BIOLÒGICA

L'assignació i la identificació d'aquells metabòlits, la concentració dels quals canvia significativament davant d'un estímul o condicions estressants, és una etapa essencial dels estudis metabolòmics, ja que permet esbrinar els seus possibles efectes sobre els organismes vius i generar una interpretació biològica coherent. Tanmateix, la identificació i la caracterització inequívoca de l'estructura dels metabòlits afectats pel procés o tractament estudiat és un procés complex que requereix de gran esforç. En els estudis de metabolòmica no dirigida, la dificultat s'incrementa dràsticament per l'elevat nombre de metabòlits estudiats simultàniament, a diferencia de l'aproximació dirigida on només s'estudia un nombre petit de metabòlits coneguts. En els enfocaments no dirigits, sovint desenes o centenars de metabòlits biològicament rellevants (estadísticament significatius) s'han d'identificar per assolir aquest nou coneixement. El nivell d'identificació dels metabòlits pot variar depenent de la tècnica analítica i de la robustesa del mètode aplicat, així com de les bases de dades i recursos informàtics utilitzats.

En els darrers anys, la identificació dels metabòlits d'interès ha donat peu a un procés de discussió molt actiu entre els membres del camp de la metabolòmica per tal d'acordar els diferents nivells d'identificació permesos. Per exemple, la Societat de Metabolòmica està desenvolupant un seguit de protocols i normes per arribar a un major consens en la forma de procedir en les investigacions dins d'aquest context [281-283]. A partir de la iniciativa d'estandarditzar els estudis metabolòmics (*Metabolomics Standards Initiative*), el grup de treball d'anàlisi química (*chemical analysis working group*, CAWG) ha establert un conjunt de normes que cal complir en la identificació dels compostos de les anàlisis metabolòmiques [284]. S'han distingit quatre nivells d'identificació vàlids a través d'aquesta iniciativa. D'entrada, la identificació de nivell 1 (o definitiva) requereix que almenys dues propietats ortogonals del metabòlit temptatiu es confirmin amb un patró pur analitzat sota les mateixes condicions experimentals. En els nivells 2 i 3 (o identificació temptativa), la identitat del metabòlit s'assoleix quan una o més propietats del metabòlit coincideixen amb les del patró pur també analitzat o extret de bases de dades o dades bibliogràfiques (per exemple, la coincidència de la massa exacta). Aquests dos nivells ofereixen una identificació específica del metabòlit en concret (nivell 2) o bé de la classe/família a la qual pertany (nivell 3). Per últim, el nivell 4 fa referència als metabòlits que no s'han pogut identificat (*unknow*s). Actualment, es recomana indicar el grau de rigor en la identificació dels metabòlits en els diferents estudis seguint aquest sistema de quatre nivells.

En la majoria de publicacions metabolòmiques d'aquesta Tesi, les assignacions dels metabòlits s'ajusten a la massa exacta de l'ió detectat en comparació a la massa exacta d'estructures químiques conegudes descrites en bases de dades (nivell 2 o 3). Cal destacar a més que en un dels estudis realitzats (article IV), s'aconsegueixen tots els requisits d'identificació, fins al nivell 1. En aquest cas, la identificació temptativa assignada per massa exacta del metabòlit es contrasta amb l'espectre de MS/MS del patró pur procedent de les bases de dades d'espectres de masses. Així doncs, les bases de dades d'espectrometria de masses juguen un paper clau en la identificació dels metabòlits. El desenvolupament de bases de dades d'espectrometria de masses cada vegada més àmplies i ben anotades és de gran importància per a la identificació de nous compostos o de possibles biomarcadors [285]. Tot i que són constants els esforços per augmentar i millorar el contingut d'aquestes bases de

dades, aquests repositoris encara estan lluny de contenir dades experimentals de tots els metabòlits. La millora de les bases de dades es veu limitada en gran part per la quantitat reduïda de patrons purs disponibles comercialment i, per tant, de l'accessibilitat als seus espectres de fragmentació de MS/MS o MS en tàndem. No obstant això, en l'actualitat s'estan desenvolupant noves eines computacionals per a poder predir els patrons de fragmentació de MS i, així, poder identificar millor aquells metabòlits dels quals no es disposen els seus espectres de MS en tàndem a les bases de dades [286-292].

Les bases de dades disponibles de manera gratuïta o comercialment en el camp de la metabolòmica proporcionen informació sobre les estructures químiques, propietats fisicoquímiques, espectres de MS, funcions biològiques i mapatge dels metabòlits. Fiehn i col·laboradors [293] han classificat les bases de dades actuals en dues categories:

- Bases de dades d'identificació de compostos, com PubChem [294], ChemSpider [295], METLIN metabolite database [296], MassBank [297], Human Metabolome Database (HMDB) [298] o National Institute of Science and Technology (NIST) data base.
- Bases de dades de rutes metabòliques, com Kyoto Encyclopedia of Genes and Genomes (KEGG) [299], Reactome [300], Wikipathways [301] o Biocyc [302].

Encara que PubChem, ChemSpider i Chemical Abstracts Service (CAS) disposen de repositoris de milers de compostos químics, no solen emprar-se en el marc de la metabolòmica atès a l'escassa informació biològica i d'espectrometria de masses de la gran majoria de compostos. Per contra, METLIN, MassBank, HMDB i NIST ofereixen aquesta informació d'espectrometria de masses i permeten l'assignació dels metabòlits gràcies a l'accés a espectres de patrons de referència.

2.4.1. Identificació dels metabòlits en les análisis de LC-MS i CE-MS

En la metabolòmica no dirigida normalment es compara l'abundància relativa dels metabòlits presents en les diferents mostres analitzades sense una identificació prèvia d'aquests. L'anàlisi de les mostres biològiques mitjançant tècniques de separació de CE i LC acoblades a l'espectrometria de masses d'alta resolució (*high resolution mass spectrometry*, HRMS), com el temps de vol (TOF) i l'Orbitrap,

permeten detectar entre centenars i milers de metabòlits simultàniament. Després de la resolució i l'anàlisi estadística dels metabòlits rellevants per a l'estudi, es procedeix a la seva identificació i caracterització a partir de la seva informació espectral. En aquesta Tesi, l'assignació dels metabòlits s'ha dut a terme a través de la seva massa exacta i dels seus possibles patrons de fragmentació de MS.

La massa exacta (és a dir, la relació de *m/z*) de l'ió s'utilitza per a la seva cerca en les bases de dades adequades per a la identificació dels compostos. Els resultats obtinguts a partir de la massa exacta només permeten fer una primera identificació temptativa del metabòlit la qual és sovint confirmada mitjançat el seu temps de retenció i/o el seu anàlisi per MS/MS. Quan no es disposa del patró pur corresponent al metabòlit estudiat sota les mateixes condicions experimentals, l'ús dels espectres de MS/MS de referència de bases de dades sol ser la manera més concloent d'identificar un metabòlit.

Les bases de dades metabolòmiques de LC-MS i CE-MS són generalment les mateixes degut a que es treballa sota les mateixes condicions instrumentals d'ionització d'electrosprai (ESI) i d'analitzadors de masses (per exemple, TOF i Orbitrap). En aquesta Tesi, s'han emprat les bases de dades HMDB [298], METLIN [303, 304], MassBank [297] i NIST [305]. En el cas del metaboloma de *S.cerevisiae*, també s'ha utilitzat la base de dades específica del llevat Yeast Metabolome Database (YMDB) [306]. A continuació es fa una breu descripció de les bases de dades esmentades.

Human Metabolome Database (HMDB)

La HMDB és una base de dades electrònica (http://www.hmdb.ca) i gratuïta que conté informació detallada dels metabòlits que es troben en el cos humà. Des del seu llançament l'any 2007 [298], s'ha emprat per facilitar la recerca en la metabolòmica en general, amb especial atenció per la química clínica i pel descobriment de nous biomarcadors. Aquesta base de dades s'ha millorat significativament en els darrers anys. La darrera versió del 2018 (versió 4.0) consta de 114.100 entrades de metabòlits [307]. La HMDB conté informació tant de metabòlits que es troben comunament en mostres humanes de forma natural (per exemple, fluids biològics) com d'altres menys freqüents (fàrmacs, metabòlits procedents de la ingesta de fàrmacs i aliments). Aquesta base de dades posa a l'abast informació clínica, química, espectral, bioquímica i enzimàtica dels metabòlits, la qual

es pot descarregar en format XML. A més, aquesta informació sovint es troba vinculada a altres bases de dades com KEGG, PubChem, ChemSpider, MetaCyc, ChEBI [308], Swiss-Prot [309] i GenBank [310, 311]. De manera paral·lela, el repositori de la HMDB reuneix fins a 22.198 espectres experimentals de MS/MS, tant de baixa com d'alta resolució.

Yeast Metabolome Database (YMDB)

La YMDB és també una base de dades gratuïta (http://www.ymdb.ca) que descriu detalladament el metaboloma de *Saccharomyces cerevisiae*. Aquesta base de dades va sorgir de la idea del repositori del metabolisme humà HMDB i, per tant, comparteix moltes de les seves opcions. Actualment, la YMDB (versió 2.0) recull informació química, física i biològica de més de 16.000 metabòlits del llevat cobrint la majoria del seu metaboloma. Conté els metabòlits del llevat ja descrits en llibres, articles científics i altres bases de dades electròniques. A més, un gran nombre d'entrades (aproximadament 13.000 compostos) contenen també espectres de RMN o de MS/MS [312].

METLIN metabolite database

METLIN és una base de dades que es va desenvolupar l'any 2004 de forma electrònica i completament gratuïta (http://metlin.scripps.edu/) per a facilitar l'assignació de metabòlits detectats per espectrometria de masses [296, 313]. METLIN conté actualment més de 240.000 compostos que inclouen metabòlits endògens de diferents organismes (per exemple, plantes, bacteris i humans) i compostos exògens, com ara fàrmacs i altres substàncies orgàniques sintètiques. Aquesta base de dades ofereix els espectres de MS/MS d'alta resolució de més de 22.000 patrons autèntics a tres energies de col·lisió diferents (10V, 20V i 40V) adquirides amb un espectròmetre de masses Q-TOF d'Agilent Technologies i amb ionització ESI positiva i negativa. En total, reuneix un total de més de 68.000 espectres MS/MS d'alta resolució. Això fa que METLIN sigui un dels recursos espectrals més grans per a la metabolòmica basada en MS. La cerca de metabòlits es pot realitzar mitjançant l'ús de la relació *m/z* de l'ió o a partir de la fórmula molecular, del nom del compost, del nombre CAS o del codi KEGG.

Actualment, la base de dades METLIN també permet carregar espectres de MS/MS dels usuaris per així poder comparar-los automàticament amb els espectres de MS/MS emmagatzemats en la base de dades. La cerca MS/MS retorna una llista de coincidències obtingudes per similitud espectral la qual cosa facilita la posterior identificació dels compostos.

MassBank

MassBank és un dipòsit d'espectres de masses d'accés obert (http://www.massbank.jp) dissenyat per compartir públicament espectres de masses de referència de patrons purs per facilitar l'assignació de metabòlits als usuaris [314]. Aquest repositori conté espectres de masses d'adquirits en diferents condicions depenent de la configuració de l'espectròmetre de masses. Inclou espectres d'ionització ESI (60% del conjunt de dades), d'impacte electrònic (EI), d'ionització química (CI), d'ionització química a pressió atmosfèrica (APCI) i d'ionització-desorció amb làser assistida per una matriu (MALDI) i de diferents analitzadors de masses de baixa (QqQ, trampa iònica) i alta (QTOF, Orbitrap) resolució. Actualment, la base de dades MassBank conté gairebé uns 19.000 espectres de MS (principalment d'EI) i 28.400 espectres de MS/MS (principalment d'ESI-QTOF i ESI-Orbitrap). Una de les característica distintives de MassBank és que proporciona 'espectres combinats' (un 2% del total d'espectres de la base de dades). Aquests espectres combinats són espectres obtinguts a partir de la combinació de tots els espectres de dissociació induïda per col·lisió (CID) generats per a un mateix compost [297]. Aquesta opció té com a objectiu fer que la identificació de metabòlits sigui independent de la configuració de l'instrument emprat i del fabricant de l'espectròmetre de masses.

Cada entrada en la base de dades MassBank descriu el nom del compost, l'estructura química i les condicions experimentals (plataforma MS, mètodes cromatogràfic, temps de retenció de l'ió precursor). Molts d'aquests camps es troben vinculats a altres bases de dades per a permetre una major caracterització biològica. MassBank permet cercar els compostos d'interès a través del seu nom, de la massa exacta, la fórmula molecular i l'estructura química. A més, ofereix als usuaris la possibilitat de comparar els seus espectres MS^n amb espectres experimentals MS/MS de la base de dades mostrant el percentatge de semblança entre espectres i el nombre d'ions producte coincidents.

National Institute of Science and Technology (NIST) database

La base de dades NIST es va desenvolupar originàriament per aplicacions de GC-MS amb ionització per impacte electrònic (*electron ionization*, EI). Però, avui en dia, aquesta base de dades comercial disposa de més de 234.000 espectres ESI-MS/MS de diferents molècules petites, com metabòlits, lípids i pèptids biològicament actius. Un tret característic de la NIST és que ofereix espectres de MS/MS d'un gran ventall d'ions comunament formats durant la ionització ESI. A part d'espectres MS/MS de l'ió $[M+H]^+$ en mode positiu i $[M-H]^-$ en mode negatiu, inclou espectres MS/MS d'ions com $[M+H-H_2O]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M+H-NH_3]^+$, $[2M+H]^+$, $[M-H-H_2O]^-$, $[2M-H]^-$, $[M-2H]^{2-}$. Això, és particularment útil ja que els adductes predominants en ESI varien segons el metabòlit o les fases mòbils o solvents utilitzats. Tots aquests espectres de MS/MS estan adquirits amb diferents instruments d'alta i baixa resolució i a un ampli rang d'energies tant en ESI positiu com negatiu [305].

La base de dades NIST s'ha convertit en un repositori d'espectres de masses molt potent, el qual es troba enllaçat al programari de la majoria de proveïdors instrumentals.

2.4.2. Interpretació biològica de les dades òmiques

En el context dels estudis metabolòmics d'aquesta Tesi, una vegada s'ha identificat aquell conjunt de metabòlits que han canviat significativament degut a l'estímul estudiat (per exemple, contaminant químic), el següent repte és intentar entendre la resposta dinàmica del sistema biològic a aquest estímul. La interpretació de la informació biològica de les dades òmiques és essencial per a la comprensió completa del sistema biològic, dels mecanismes bioquímics subjacents i dels possibles efectes a nivell de fenotip. Els gens, els metabòlits i les proteïnes solen associar-se a vies i a processos biològics específics indicatius de l'estat de l'organisme. Per exemple, la majoria dels metabòlits mesurats en els estudis metabolòmics estan associats a determinats processos metabòlics de les cèl·lules vives com, per exemple, la glicòlisi o el metabolisme dels lípids. Els canvis de concentració de metabòlits poden indicar efectes en la viabilitat cel·lular (apoptosi), en els nivells d'oxigenació (anòxia o isquèmia) o l'homeòstasi en general, entre d'altres. Hi ha també metabòlits que estan associats específicament a determinats danys tissulars, atrofia muscular o estrès oxidatiu. És a dir,

cadascun dels metabòlits que conformen el sistema pot donar informació del seu estat [315]. De manera anàloga, qualsevol canvi en l'abundància relativa dels gens (genòmica), dels transcrits de mRNA (transcriptòmica) i de les proteïnes (proteòmica) també reflecteix el comportament de l'organisme. L'estudi dels canvis en els gens, les proteïnes i els metabòlits per separat ofereix una visió parcial del comportament i/o funcionament de l'organisme. Moltes vegades és necessària la interpretació de la informació biològica de tots ells de forma conjunta a través d'estratègies de fusió de dades per a una millor comprensió global de l'estructura i la dinàmica del sistema. En el marc dels estudis òmics, es pot obtenir una millor interpretació biològica dels canvis en aquests tipus de molècules a partir de l'ús de bases de dades o eines de visualització de les rutes metabòliques. Les bases de dades metabòliques són clau per desxifrar les rutes bioquímiques o les interaccions metabòlit-gen-proteïna. Aquestes eines proporcionen una visió més general del metabolisme del sistema (metabòlits, gens, proteïnes i enzims, i les vies que formen) i ajuden a detectar les rutes metabòliques alterades per assolir una correcta interpretació biològica. Entre les bases de dades de rutes metabòliques més comunes es troben KEGG [299], Reactome [300], Wikipathways [301] o Biocyc [302, 316]. A continuació, es farà una breu descripció de les característiques de la base de dades emprada en aquesta Tesi, la base de dades KEGG.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

La base de dades KEGG és probablement l'eina més emprada i completa per a l'exploració de les rutes metabòliques [317, 318]. És una base de dades totalment gratuïta via web (http://www.genome.jp/kegg/) que es va desenvolupar l'any 1995 al laboratori Kanehisa de l'Institut d'Investigacions Químiques de Kyoto. KEGG és un recurs bioinformàtic que permet comprendre les diferents funcions dels sistemes biològics a nivell molecular. Aquest recurs atorga una visió general del metabolisme d'un ampli ventall d'organismes, conferint informació genòmica, química i de les rutes metabòliques de milers d'organismes [319].

En realitat, KEGG és una base de dades integrada que esta constituïda per 16 bases de dades més petites. Aquestes es classifiquen segons si ofereixen informació dels sistemes (PATHWAY, BRITE i MODULE), informació genòmica (ORTHOLOGY, GENOME i GENES, SSDB), informació química

(LIGAND, COMPOUND, GLYCAN, REACTION i ENZYME) i informació biomèdica (DISEASE, DRUG, ENVIRON i MEDICUS). La base de dades PATHWAY és especialment útil per a la visualització dels conjunts de dades moleculars, especialment conjunts de dades a gran escala de genòmica, transcriptòmica, proteòmica i metabolòmica, en mapes i així, facilitar la interpretació biològica de les dades òmiques. KEGG PATHWAY conté rutes metabòliques que conformen diagrames metabòlics de senyalització de metabòlits, gens i proteïnes. La **Figura 2.28** mostra un exemple de representació gràfica de les rutes metabòliques a través de la base de dades KEGG on es poden visualitzar de forma detallada les rutes afectades amb els corresponents metabòlits i gens que hi estan involucrats.

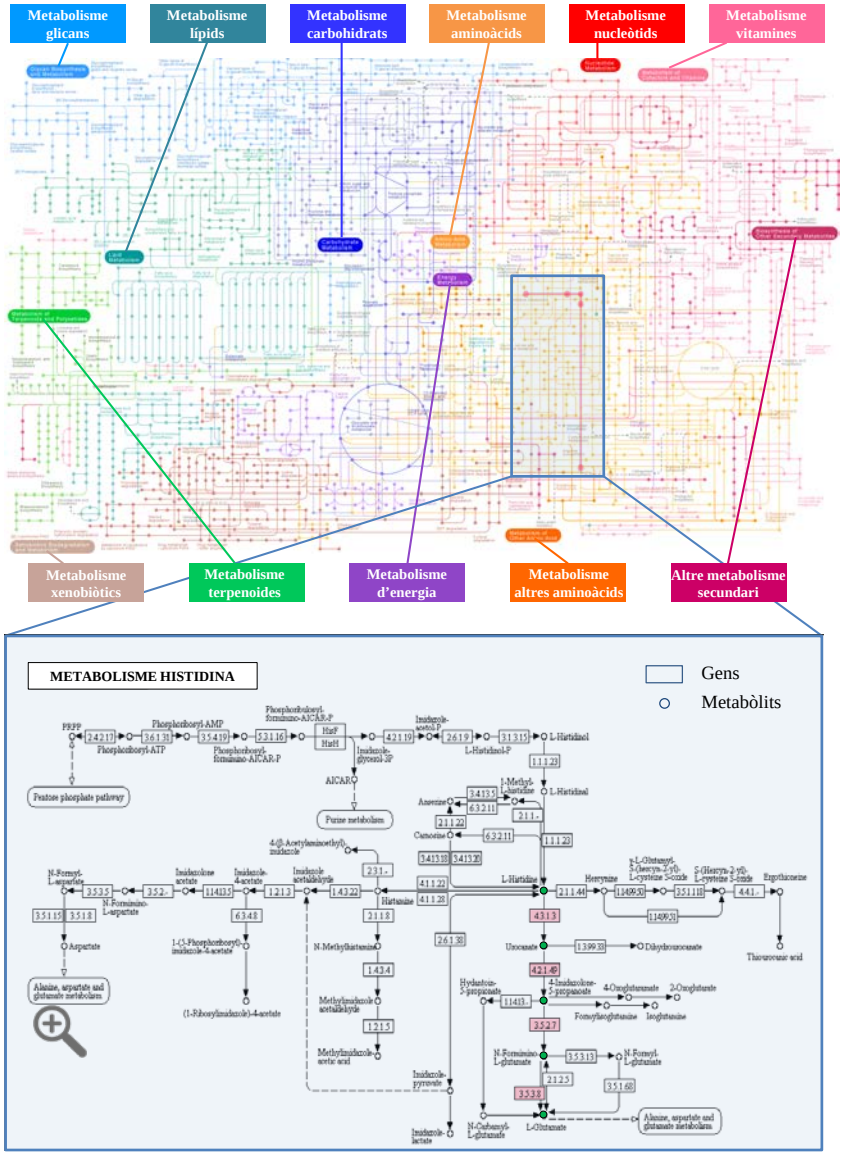


Figura 2.28. Mapa cartogràfic KEGG per a l'anàlisi global de les rutes metabòliques.

2.5. REFERÈNCIES

1. Harrison, R. J., Understanding genetic variation and function- the applications of next generation sequencing, *Seminars in Cell & Developmental Biology*. 2012, *23*, 230-236.
2. Stahl, P. L., Lundeberg, J., Toward the single-hour high-quality genome, *Annual Review of Biochemistry*. 2012, *81*, 359-378.
3. Bartel, D. P., MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*. 2004, *116*, 281-297.
4. Morozova, O., Marra, M. A., Applications of next-generation sequencing technologies in functional genomics, *Genomics*. 2008, *92*, 255-264.
5. Slonim, D. K., Yanai, I., Getting Started in Gene Expression Microarray Analysis, *PLoS Computational Biology*. 2009, *5*, e1000543.
6. Van Vliet, A. H., Next generation sequencing of microbial transcriptomes: challenges and opportunities, *FEMS Microbiology Letters.* 2010, *302*, 1-7.
7. Angel, T. E., Aryal, U. K., Hengel, S. M., Baker, E. S., Kelly, R. T., Robinson, E. W., Smith, R. D., Mass spectrometry-based proteomics: existing capabilities and future directions, *Chemical Society Reviews*. 2012, *41*, 3912-3928.
8. Thelen, J. J., Miernyk, J. A., The proteomic future: where mass spectrometry should be taking us, *Biochemical Journal*. 2012, *444*, 169-181.
9. Bundy, J. G., Davey, M. P., Viant, M. R., Environmental metabolomics: a critical review and future perspectives, *Metabolomics*. 2008, *5*, 3.
10. Viant, M. R., Sommer, U., Mass spectrometry based environmental metabolomics: A primer and review, *Metabolomics*. 2013, *9*, 144-158.
11. Spies, D., Ciaudo, C., Dynamics in Transcriptomics: Advancements in RNA-Seq Time Course and Downstream Analysis, *Computational and Structural Biotechnology Journal*. 2015, *13*, 469-477.
12. Aebersold, R., Mann, M., Mass spectrometry-based proteomics, *Nature*. 2003, *422*, 198-207.
13. Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., van Dam, K., Oliver, S. G., A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, *Nature Biotechnology*. 2001, *19*, 45-50.
14. Mahner, M., Kary, M., What exactly are genomes, genotypes and phenotypes? And what about phenomes?, *Journal of Theoretical Biology*. 1997, *186*, 55-63.
15. Lindon, J. C., Nicholson, J. K., Holmes, E. (Eds.), *The Handbook of Metabonomics and Metabolomics*,1st Edition, Elsevier, 2011.
16. Longnecker, K., Futrelle, J., Coburn, E., Kido Soule, M. C., Kujawinski, E. B., Environmental metabolomics: Databases and tools for data analysis, *Marine Chemistry*. 2015, *177*, 366-373.
17. Horgan, R. P., Kenny, L. C., 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics, *The Obstetrician & Gynaecologist*. 2011, *13*, 189-195.
18. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., *et al.*, Complementary DNA sequencing: expressed sequence tags and human genome project, *Science*. 1991, *252*, 1651-1656.
19. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M.C., *et al.*, Initial sequencing and analysis of the human genome, *Nature*. 2001, *409*, 860-921.
20. Venter, J. C., Adams, M. D., Myers, E.W., Li, P.W., Mural, R.J., *et al.*, The sequence of the human genome, *Science*. 2001, *291*, 1304-1351.
21. Schacherer, J., Ruderfer, D. M., Gresham, D., Dolinski, K., Botstein, D., Kruglyak, L., Genome-Wide Analysis of Nucleotide-Level Variation in Commonly Used Saccharomyces cerevisiae Strains, *PLoS One*. 2007, *2*, e322.
22. Dong, Z., Chen, Y., Transcriptomics: advances and approaches, *Science China Life Sciences*. 2013, *56*, 960-967.
23. Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., Transcriptomics technologies, *PLoS Computational Biology*. 2017, *13*, e1005457.
24. Lockhart, D. J., Winzeler, E. A., Genomics, gene expression and DNA arrays, *Nature*. 2000, *405*, 827-836.
25. 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.
26. Patti, G. J., Yanes, O., Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy, *Nature Reviews Molecular Cell Biology*. 2012, *13*, 263-269.

27. 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*. 1999, *29*, 1181-1189.
28. 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.
29. Griffiths, W. J., *Metabolomics, Metabonomics and Metabolite Profiling*, 1st Edition. Royal Society of Chemistry, 2007.
30. Robertson, D. G., Watkins, P. B., Reily, M. D., Metabolomics in toxicology: preclinical and clinical applications, *Toxicological Sciences.* 2011, *120 Suppl 1*, S146-170.
31. Hu, C., Xu, G., Mass-spectrometry-based metabolomics analysis for foodomics, *TrAC Trends in Analytical Chemistry*. 2013, *52*, 36-46.
32. Ishihara, A., Matsuda, F., Miyagawa, H., Wakasa, K., Metabolomics for metabolically manipulated plants: effects of tryptophan overproduction, *Metabolomics*. 2007, *3*, 319-334.
33. Puchades-Carrasco, L., Pineda-Lucena, A., Metabolomics Applications in Precision Medicine: An Oncological Perspective, *Current Topics in Medicinal Chemistry*. 2017, *17*, 2740-2751.
34. Glassbrook, N., Beecher, C., Ryals, J., Metabolic profiling on the right path, *Nature Biotechnology*. 2000, *18*, 1142-1143.
35. Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., Kell, D. B., Metabolomics by numbers: acquiring and understanding global metabolite data, *Trends Biotechnology*. 2004, *22*, 245-252.
36. Dudley, E., Yousef, M., Wang, Y., Griffiths, W. J., Targeted metabolomics and mass spectrometry, *Advances in Protein Chemistry and Structural Biology*. 2010, *80*, 45-83.
37. Gutala, R. V., Reddy, P. H., The use of real-time PCR analysis in a gene expression study of Alzheimer's disease post-mortem brains, *Journal of Neuroscience Methods*. 2004, *132*, 101-107.
38. Mastrokolias, A., Pool, R., Mina, E., Hettne, K. M., van Duijn, E., *et al*., Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples, *Metabolomics.* 2016, *12*, 137.
39. Lohmann, S., Herold, A., Bergauer, T., Belousov, A., Betzl, G., *et al*., Gene expression analysis in biomarker research and early drug development using function tested reverse transcription quantitative real-time PCR assays, *Methods.* 2013, *59*, 10-19.
40. Wang, F., Liu, X., Liu, C., Liu, Z., Sun, L., Effects of antibiotic antitumor drugs on nucleotide levels in cultured tumor cells: an exploratory method to distinguish the mechanisms of antitumor drug action based on targeted metabolomics, *Acta Pharmaceutica Sinica B.* 2015, *5*, 223-230.
41. Huang, S. S., Benskin, J. P., Chandramouli, B., Butler, H., Helbing, C. C., Cosgrove, J. R., Xenobiotics Produce Distinct Metabolomic Responses in Zebrafish Larvae (Danio rerio), *Environmental Science & Technology*. 2016, *50*, 6526-6535.
42. Santos, D., Matos, M., Coimbra, A. M., Developmental toxicity of endocrine disruptors in early life stages of zebrafish, a genetic and embryogenesis study, *Neurotoxicology and Teratology.* 2014, *46*, 18-25.
43. Menni, C., Zierer, J., Valdes, A. M., Spector, T. D., Mixing omics: combining genetics and metabolomics to study rheumatic diseases, *Nature Reviews Rheumatology*. 2017, *13*, 174-181.
44. Hu, Z. Z., Huang, H., Wu, C. H., Jung, M., Dritschilo, A., Riegel, A. T., Wellstein, A., Omics-based molecular target and biomarker identification, *Methods in Molecular Biology*. 2011, *719*, 547-571.
45. Robertson, D. G., Metabonomics in Toxicology: A Review, *Toxicological Sciences*. 2005, *85*, 809-822.
46. Yu, S. Y., Paul, S., Hwang, S. Y., Application of the emerging technologies in toxicogenomics: An overview, *BioChip Journal*. 2016, *10*, 288-296.
47. Zhou, B., Xiao, J. F., Tuli, L., Ressom, H. W., LC-MS-based metabolomics, *Molecular BioSystems*. 2012, *8*, 470-481.
48. Dunn, W. B., Wilson, I. D., Nicholls, A. W., Broadhurst, D., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, *Bioanalysis*. 2012, *4*, 2249-2264.
49. Engskog, M. K. R., Haglöf, J., Arvidsson, T., Pettersson, C., LC–MS based global metabolite profiling: the necessity of high data quality, *Metabolomics*. 2016, *12*, 114.
50. Gika, H. G., Zisi, C., Theodoridis, G., Wilson, I. D., Protocol for quality control in metabolic profiling of biological fluids by U(H)PLC-MS, *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences.* 2016, *1008*, 15-25.

51. Want, E. J., Wilson, I. D., Gika, H., Theodoridis, G., Plumb, R. S., Shockcor, J., Holmes, E., Nicholson, J. K., Global metabolic profiling procedures for urine using UPLC-MS, *Nature Protocols*. 2010, *5*, 1005-1018.
52. Ge, Y., Wang, D. Z., Chiu, J. F., Cristobal, S., Sheehan, D., Silvestre, F., Peng, X., Li, H., Gong, Z., Lam, S. H., Environmental omics: Current status and future directions, *Journal of Integrated Omics*. 2013, *3*, 75-87.
53. Morrison, N., Wood, A. J., et al., Standard Annotation of Environmental OMICS Data: Application to the Transcriptomics Domain, *OMICS: A Journal of Integrative Biology*. 2006, *10*, 172-178.
54. Edison, A., Hall, R., Junot, C., Karp, P., Kurland, I., Mistrik, R., *et al.*, The Time Is Right to Focus on Model Organism Metabolomes, *Metabolites*. 2016, *6*, 8.
55. Fields, S., Johnston, M., Whither Model Organism Research?, *Science*. 2005, *307*, 1885-1886.
56. Griffiths, E., *What is a model?*, Sheffield University. 2010.
57. Karathia, H., Vilaprinyo, E., Sorribas, A., Alves, R., Saccharomyces cerevisiae as a Model Organism: A Comparative Study, *PLoS One*. 2011, *6*, e16015.
58. Saha, R., Chowdhury, A., Maranas, C. D., Recent advances in the reconstruction of metabolic models and integration of omics data, *Current Opinion in Biotechnology*. 2014, *29*, 39-45.
59. Goldstein, B., King, N., The Future of Cell Biology: Emerging Model Organisms, *Trends Cell Biology*. 2016, *26*, 818-824.
60. Richards, O. W., The analysis of growth as illustrated by yeast, *Cold Spring Harbor Symposia on Quantitative Biology*. 1934, *2*, 157-166.
61. Richards, O. W., Haynes, F. W., Oxygen consumption and carbon dioxide production during the growth of yeast, *Plant Physiology*. 1932, *7*, 139-144.
62. Botstein, D., Fink, G. R., Yeast: an experimental organism for 21st Century biology, *Genetics*. 2011, *189*.
63. Broach, J. R., Nutritional Control of Growth and Development in Yeast, *Genetics*. 2012, *192*, 73-105.
64. Arroyo-Lopez, F. N., Orlic, S., Querol, A., Barrio, E., Effects of temperature, pH and sugar concentration on the growth parameters of Saccharomyces cerevisiae, S. kudriavzevii and their interspecific hybrid, *International Journal of Food Microbiology*. 2009, *131*, 120-127.
65. Torija, M. J., Rozès, N., Poblet, M., Guillamón, J. M., Mas, A., Effects of fermentation temperature on the strain population of Saccharomyces cerevisiae, *International Journal of Food Microbiology*. 2003, *80*, 47-53.
66. Bergman, L. W., Growth and maintenance of yeast, *Methods in Molecular Biology*. 2001, *177*, 9-14.
67. Streisinger, G., Walker, C., Dower, N., Knauber, D., Singer, F., Production of clones of homozygous diploid zebra fish (Brachydanio rerio), *Nature*. 1981, *291*, 293-296.
68. Mayden, R. L., Tang, K. L., Conway, K. W., Freyhof, J., Chamberlain, S., Haskins, M., Schneider, L., Sudkamp, M., Wood, R. M., Agnew, M., Bufalino, A., Sulaiman, Z., Miya, M., Saitoh, K., He, S., Phylogenetic relationships of Danio within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species, *Journal of Experimental Zoology-Part B: Molecular and Developmental Evolution*. 2007, *308B*, 642-654.
69. Spence, R., Gerlach, G., Lawrence, C., Smith, C., The behaviour and ecology of the zebrafish, Danio rerio, *Biological Reviews of the Cambridge Philosophical Society*. 2008, *83*, 13-34.
70. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., Schilling, T. F., Stages of embryonic development of the zebrafish, *Developmental Dynamics*. 1995, *203*, 253-310.
71. Peterson, R. T., MacRae, C. A., Systematic Approaches to Toxicology in the Zebrafish, *Annual Review of Pharmacology and Toxicology*. 2012, *52*, 433-453.
72. Hung, M. W., Zhang, Z. J., Li, S., Lei, B., Yuan, S., Cui, G. Z., Man Hoi, P., Chan, K., Lee, S. M. Y., From Omics to Drug Metabolism and High Content Screen of Natural Product in Zebrafish: A New Model for Discovery of Neuroactive Compound, *Evidence-Based Complementary and Alternative Medicine*. 2012, *2012*, 20.
73. Willemsen, R., van't Padje, S., van Swieten, J. C., Oostra, B. A., in: De Deyn, P.P., Van Dam, D., (Eds.), *Animal Models of Dementia*, Humana Press, Totowa, NJ 2011, pp. 255-269.
74. Berghmans, S., Butler, P., Goldsmith, P., Waldron, G., Gardner, I., *et al*., Zebrafish based assays for the assessment of cardiac, visual and gut function-potential safety screens for early drug discovery, *Journal of Pharmacological and Toxicological Methods*. 2008, *58*, 59-68.
75. Lieschke, G. J., Currie, P. D., Animal models of human disease: zebrafish swim into view, *Nature Reviews Genetics*. 2007, *8*, 353-367.

76. Sukardi, H., Chng, H. T., Chan, E. C., Gong, Z., Lam, S. H., Zebrafish for drug toxicity screening: bridging the in vitro cell-based models and in vivo mammalian models, *Expert Opinion on Drug Metabolism & Toxicology*. 2011, *7*, 579-589.
77. Directive2010/63/EU. Directive 2010/63/EU Of The European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes, *Official Journal of the European Union*. 2010, *L276*, 33-79.
78. Zhang, C., Willett, C., Fremgen, T., Zebrafish: an animal model for toxicological studies, *Current Protocols in Toxicology*. 2003, *17*, 1.7.1-1.7.18.
79. Garcia, G. R., Noyes, P. D., Tanguay, R. L., Advancements in zebrafish applications for 21st century toxicology, *Pharmacology & Therapeutics.* 2016, *161*, 11-21.
80. Howe, D. G., Bradford, Y. M., et al., ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics, *Nucleic Acids Research*. 2013, *41*, D854-860.
81. Parng, C., Seng, W. L., Semino, C., McGrath, P., Zebrafish: a preclinical model for drug screening, *Assay and Drug Development Technologies.* 2002, *1*, 41-48.
82. Peterson, R. T., Link, B. A., Dowling, J. E., Schreiber, S. L., Small molecule developmental screens reveal the logic and timing of vertebrate development, *Proceedings of the National Academy of Sciences of the United States of America*. 2000, *97*, 12965-12969.
83. Hill, A. J., Teraoka, H., Heideman, W., Peterson, R. E., Zebrafish as a model vertebrate for investigating chemical toxicity, *Toxicological Sciences.* 2005, *86*, 6-19.
84. Carvan, M. J., Heiden, T. K., Tomasiewicz, H., The utility of zebrafish as a model for toxicological research, *Biochemistry and Molecular Biology of Fishes*. 2005, *6*, 3-41.
85. Dave, G., Andersson, K., Berglind, R., Hasselrot, B., Toxicity of eight solvent extraction chemicals and of cadmium to water fleas, Daphnia magna, rainbow trout, Salmo gairdneri, and zebrafish, Brachydanio rerio, *Comparative Biochemistry and Physiology - Part C: Toxicology*. 1981, *69c*, 83-98.
86. Nagel, R., Isberner, K., in: Braunbeck T, Hinton DE, Streit B (Eds.), Testing of chemicals with fish-a critical evaluation of tests with special regard to zebrafish, *Fish Ecotoxicology*, Birkhäuser Basel, Basel. 1998, pp. 337-352.
87. Braunbeck, T., Boettcher, M., Hollert, H., Kosmehl, T., Lammer, E., Leist, E., Rudolf, M., Seitz, N., Towards an alternative for the acute fish LC(50) test in chemical assessment: the fish embryo toxicity test goes multi-species-an update, *Altex*. 2005, *22*, 87-102.
88. Annamalai, J., Namasivayam, V., Endocrine disrupting chemicals in the atmosphere: Their effects on humans and wildlife, *Environment International*. 2015, *76*, 78-97.
89. Jordao, R., Garreta, E., Campos, B., Lemos, M. F., Soares, A. M., Tauler, R., Barata, C., Compounds altering fat storage in Daphnia magna, *Science of Total Environment*. 2016, *545-546*, 127-136.
90. Nagato, E. G., Simpson, A. J., Simpson, M. J., Metabolomics reveals energetic impairments in Daphnia magna exposed to diazinon, malathion and bisphenol-A, *Aquatic toxicology*. 2016, *170*, 175-186.
91. Zhang, C., Wang, J., Zhang, S., Zhu, L., Du, Z., Wang, J., Acute and subchronic toxicity of pyraclostrobin in zebrafish (Danio rerio), Chemosphere. 2017, *188*, 510-516.
92. Lam, S. H., Hlaing, M. M., Zhang, X., Yan, C., Duan, Z., Zhu, L., Ung, C. Y., Mathavan, S., Ong, C. N., Gong, Z., Toxicogenomic and phenotypic analyses of bisphenol-A early-life exposure toxicity in zebrafish, *PLoS One*. 2011, *6*, e28273.
93. McGonnell, I. M., Fowkes, R. C., Fishing for gene function – endocrine modelling in the zebrafish, *Journal of Endocrinology*. 2006, *189*, 425-439.
94. Segner, H., Zebrafish (Danio rerio) as a model organism for investigating endocrine disruption, *Comp Comparative Biochemistry and Physiology - Part C: Toxicology & Pharmacology*. 2009, *149*, 187-195.
95. Giulivo, M., Lopez de Alda, M., Capri, E., Barceló, D., Human exposure to endocrine disrupting compounds: Their role in reproductive systems, metabolic syndrome and breast cancer. A review, *Environmental Research*. 2016, *151*, 251-264.
96. Diamanti-Kandarakis, E., Bourguignon, J. P., Giudice, L. C., Hauser, R., Prins, G. S., Soto, A. M., Zoeller, R. T., Gore, A. C., Endocrine-disrupting chemicals: an Endocrine Society scientific statement, *Endocrine Reviews.* 2009, *30*, 293-342.
97. Legler, J., Fletcher, T., Govarts, E., Porta, M., Blumberg, B., Heindel, J. J., Trasande, L., Obesity, diabetes, and associated costs of exposure to endocrine-disrupting chemicals in the European Union, *Journal of Clinical Endocrinology & Metabolism*. 2015, *100*, 1278-1288.

98. Casals-Casas, C., Desvergne, B., Endocrine disruptors: from endocrine to metabolic disruption, *Annual Review of Physiology*. 2011, *73*, 135-162.
99. Bradbury, J., Small Fish, Big Science, *PLoS Biology*. 2004, *2*, e148.
100. Oliveira, E., Barata, C., Piña, B., Endocrine disruption in the omics era: New views, new hazards, new approaches, *Open Biotechnology Journal*. 2016, *10*, 20-35.
101. Noyes, P. D., Garcia, G. R., Tanguay, R. L., Zebrafish as an in vivo model for sustainable chemical design, *Green Chemistry*. 2016, *18*, 6410-6430.
102. Deepak, S. A., Kottapalli, K. R., Rakwal, R., Oros, G., Rangappa, K. S., Iwahashi, H., Masuo, Y., Agrawal, G. K., Real-Time PCR: Revolutionizing Detection and Expression Analysis of Genes, *Current Genomics*. 2007, *8*, 234-251.
103. Auburn, R. P., Kreil, D. P., Meadows, L. A., Fischer, B., Matilla, S. S., Russell, S., Robotic spotting of cDNA and oligonucleotide microarrays, *Trends Biotechnology*. 2005, *23*, 374-379.
104. George, I. L., Maria, B., Emmanouil, S., Apostolos, Z., in: Mehdi, K.P. (Eds.), Microarrays, *Encyclopedia of Information Science and Technology,* 3rd Edition, IGI Global, Hershey, PA, USA 2015, pp. 5593-5606.
105. Wang, Z., Gerstein, M., Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics.* 2009, *10*, 57-63.
106. Ozsolak, F., Milos, P. M., RNA sequencing: advances, challenges and opportunities, *Nature Reviews Genetics*. 2011, *12*, 87-98.
107. Szabo, D. T., in: Gupta, R. C. (Eds.), Transcriptomic biomarkers in safety and risk assessment of chemicals, *Biomarkers in Toxicology*, 1st Edition, Academic Press, Boston. 2014, pp. 1033-1038.
108. Kussmann, M., Raymond, F., Affolter, M., OMICS-driven biomarker discovery in nutrition and health, *Journal of Biotechnology*. 2006, *124*, 758-787.
109. Charpe, A. M., in: Ravi, I., Baunthiyal, M., Saxena, J. (Eds.), DNA Microarray, *Advances in Biotechnology*, Springer, India, New Delhi. 2014, pp. 71-104.
110. Tarca, A. L., Romero, R., Draghici, S., Analysis of microarray experiments of gene expression profiling, *American journal of obstetrics and gynecology*. 2006, *195*, 373-388.
111. Draghici, S., Khatri, P., Shah, A., Tainsky, M., Assessing the functional bias of commercial microarrays using the Onto-Compare database, *Biotechniques.* 2003,*34*, 55-61.
112. Mochida, K., Shinozaki, K., Advances in omics and bioinformatics tools for systems analyses of plant functions, *Plant Cell Physiology*. 2011, *52*, 2017-2038.
113. Lister, R., Gregory, B. D., Ecker, J. R., Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond, *Current Opinion in Plant Biology.* 2009, *12*, 107-118.
114. Bryant, D. W., Jr., Priest, H. D., Mockler, T. C., Detection and quantification of alternative splicing variants using RNA-Seq, *Methods in Molecular Biology*. 2012, *883*, 97-110.
115. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., Liu, X., Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells, *PLoS One*. 2014, *9*, e78644.
116. Qian, X., Ba, Y., Zhuang, Q., Zhong, G., RNA-Seq Technology and Its Application in Fish Transcriptomics, *OMICS : a Journal of Integrative Biology*. 2014, *18*, 98-110.
117. Core, L. J., Waterfall, J. J., Lis, J. T., Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, *Science*. 2008, *322*, 1845-1848.
118. Marguerat, S., Bähler, J., RNA-Seq: from technology to biology, *Cellular and Molecular Life Sciences*. 2010, *67*, 569-579.
119. Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., Chinnaiyan, A. M., Transcriptome sequencing to detect gene fusions in cancer, *Nature*. 2009, *458*, 97-101.
120. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*. 2008, *320*, 1344-1349.
121. Strickler, S. R., Bombarely, A., Mueller, L. A., Designing a transcriptome next-generation sequencing project for a nonmodel plant species, *American Journal of Botany*. 2012, *99*, 257-266.
122. Nicholson, J. K., Wilson, I. D., Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism, *Nature Reviews Drug Discovery*. 2003, *2*, 668-676.
123. Harrigan, G. G., LaPlante, R. H., Cosma, G. N., Cockerell, G., Goodacre, R., Maddox, J. F., Luyendyk, J. P., Ganey, P. E., Roth, R. A., Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity, *Toxicology Letters*. 2004, *146*, 197-205.

124. Johnson, H. E., Broadhurst, D., Kell, D. B., Theodorou, M. K., Merry, R. J., Griffith, G. W., High-Throughput Metabolic Fingerprinting of Legume Silage Fermentations via Fourier Transform Infrared Spectroscopy and Chemometrics, *Applied and Environmental Microbiology*. 2004, *70*, 1583-1592.
125. Howlett, R. M., Davey, M. P., Kelly, D. J., Metabolomic Analysis of Campylobacter jejuni by Direct-Injection Electrospray Ionization Mass Spectrometry, *Methods in Molecular Biology*. 2017, *1512*, 189-197.
126. Dunn, W. B., Ellis, D. I., Metabolomics: Current analytical platforms and methodologies, *TrAC Trends in Analytical Chemistry*. 2005, *24*, 285-294.
127. De Raad, M., Fischer, C. R., Northen, T. R., High-throughput platforms for metabolomics, *Current Opinion in Chemical Biology*. 2016, *30*, 7-13.
128. Bjerrum, J. T., Metabonomics: analytical techniques and associated chemometrics at a glance, *Methods in Molecular Biology*. 2015,*1277*, 1-14.
129. Halket, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K. P., Fraser, P. D., Bramley, P. M., Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS, *Journal of Experimental Botany*. 2005, *56*, 219-243.
130. Liu, X., Locasale, J. W., Metabolomics: A Primer, *Trends in Biochemical Sciences*. 2017, *42*, 274-284.
131. Simon-Manso, Y., Lowenthal, M. S., Kilpatrick, L. E., Sampson, M. L., Telu, K. H., *et al.*, Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources, *Analytical Chemistry*. 2013, *85*, 11725-11731.
132. Hummel, J., Selbig, J., Walther, D., Kopka, J., in: Nielsen J, Jewett MC (Eds.), The Golm Metabolome Database: a database for GC-MS based metabolite profiling, *Metabolomics: A Powerful Tool in Systems Biology*, Springer Berlin Heidelberg, Berlin, Heidelberg. 2007, pp.75-95.
133. Skogerson, K., Wohlgemuth, G., Barupal, D. K., Fiehn, O., The volatile compound BinBase mass spectral database, *BMC bioinformatics*. 2011, *12*, 321.
134. Ramautar, R., Mayboroda, O. A., Somsen, G. W., de Jong, G. J., CE-MS for metabolomics: Developments and applications in the period 2008-2010, *Electrophoresis*. 2011, *32*, 52-65.
135. Heiger, D., *High Performance Capillary Electrophoresis: A Primer*, Agilent Technologies Inc. 2010.
136. Schmitt-Kopplin, P., Frommberger, M., Capillary electrophoresis-mass spectrometry: 15 years of developments and applications, *Electrophoresis*. 2003, *24*, 3837-3867.
137. Schmitt-Kopplin, P., Fekete, A., The CE-Way of Thinking: "All Is Relative!", *Methods in Molecular Biology*. 2016, *1483*, 3-19.
138. Ramautar, R., Berger, R., van der Greef, J., Hankemeier, T., Human metabolomics: strategies to understand biology, *Current Opinion in Chemical Biology*. 2013, *17*, 841-846.
139. Kuehnbaum, N. L., Britz-McKibbin, P., New Advances in Separation Science for Metabolomics: Resolving Chemical Diversity in a Post-Genomic Era, *Chemical Reviews*. 2013, *113*, 2437-2468.
140. Ramautar, R., Nevedomskaya, E., Mayboroda, O. A., Deelder, A. M., Wilson, I. D., Gika, H. G., Theodoridis, G. A., Somsen, G. W., de Jong, G. J., Metabolic profiling of human urine by CE-MS using a positively charged capillary coating and comparison with UPLC-MS, *Molecular BioSystems*. 2011, *7*, 194-199.
141. Mukhopadhyay, R., DNA sequencers: the next generation, *Analytical Chemistry*. 2009, *81*, 1736-1740.
142. Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M., Nishioka, T., Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry, *Analytical Chemistry*. 2002, *74*, 2233-2239.
143. Hirayama, A., Wakayama, M., Soga, T., Metabolome analysis based on capillary electrophoresis-mass spectrometry, *TrAC Trends in Analytical Chemistry*. 2014, *61*, 215-222.
144. Kok, M. G. M., Somsen, G. W., de Jong, G. J., The role of capillary electrophoresis in metabolic profiling studies employing multiple analytical techniques, *TrAC Trends in Analytical Chemistry*. 2014, *61*, 223-235.
145. Wakayama, M., Hirayama, A., Soga, T., in: Bjerrum JT (Eds.), Capillary Electrophoresis-Mass Spectrometry, *Metabonomics: Methods and Protocols*, Springer New York, New York, 2015, pp. 113-122.
146. Mischak, H., Vlahou, A., Ioannidis, J. P. A., Technical aspects and inter-laboratory variability in native peptide profiling: The CE–MS experience, *Clinical Biochemistry*. 2013, *46*, 432-443.

147. Mischak, H., How to get proteomics to the clinic? Issues in clinical proteomics, exemplified by CE-MS, *Proteomics-Clinical Applications*. 2012, *6*, 437-442.
148. Daimon, M., Soga, T., Hozawa, A., Oizumi, T., Kaino, W., *et al*., Serum Glycerophosphate Levels are Increased in Japanese Men with Type 2 Diabetes, *Internal Medicine*. 2012, *51*, 545-551.
149. Pont, L., Benavente, F., Jaumot, J., Tauler, R., Alberch, J., Ginés, S., Barbosa, J., Sanz-Nebot, V., Metabolic profiling for the identification of Huntington biomarkers by on-line solid-phase extraction capillary electrophoresis mass spectrometry combined with advanced data analysis tools, *Electrophoresis*. 2016, *37*, 795-808.
150. Takahashi, N., Washio, J., Mayanagi, G., Metabolomics of supragingival plaque and oral bacteria, *Journal of Dental Research*. 2010, *89*, 1383-1388.
151. Kitajima, T., Jigami, Y., Chiba, Y., Cytotoxic mechanism of selenomethionine in yeast, *Journal of Biological Chemistry*. 2012, *287*, 10032-10038.
152. Sato, S., Soga, T., Nishioka, T., Tomita, M., Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection, *The Plant Journal*. 2004, *40*, 151-163.
153. Delatte, T. L., Schluepmann, H., Smeekens, S. C. M., De Jong, G. J., Somsen, G. W., Capillary electrophoresis-mass spectrometry analysis of trehalose-6-phosphate in Arabidopsis thaliana seedlings, *Analytical and Bioanalytical Chemistry*. 2011, *400*, 1137-1144.
154. Sugimoto, M., Kaneko, M., Onuma, H., Sakaguchi, Y., Mori, M., Abe, S., Soga, T., Tomita, M., Changes in the Charged Metabolite and Sugar Profiles of Pasteurized and Unpasteurized Japanese Sake with Storage, *Journal of Agricultural and Food Chemistry*. 2012, *60*, 2586-2593.
155. Çelebier, M., Ibáñez, C., Simó, C., Cifuentes, A., in: Kurien BT, Scofield RH (Eds.), *Protein Electrophoresis: Methods and Protocols*, Humana Press, Totowa, NJ 2012, pp. 185-195.
156. Ramautar, R., Somsen, G., de Jong, G., CE-MS in metabolomics, *Electrophoresis*. 2009, *30*, 276-291.
157. 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.
158. Swartz, M. E., UPLC™: An Introduction and Review, *Journal of Liquid Chromatography & Related Technologies*. 2005, *28*, 1253-1263.
159. Fanali, S., Haddad, P., Poole, C., Schoenmakers, P., Lloyd, D., *Liquid chromatography: fundamentals and instrumentation*, 2nd Edition, Elsevier. 2013.
160. Harder, U., Koletzko, B., Peissner, W., Quantification of 22 plasma amino acids combining derivatization and ion-pair LC-MS/MS, *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*. 2011, *879*, 495-504.
161. Bidlingmeyer, B. A., Deming, S. N., Price, W. P., Sachok, B., Petrusek, M., Retention mechamism for reversed-phase ion-pair liquid chromatography, *Journal of Chromatography A*. 1979, *186*, 419-434.
162. Hemstrom, P., Irgum, K., Hydrophilic interaction chromatography, *Journal of Separation Science*. 2006, *29*, 1784-1821.
163. Alpert, A. J., Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds, *Journal Chromatography*. 1990, *499*, 177-196.
164. Wernisch, S., Pennathur, S., Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics, *Analytical and Bioanalytical Chemistry*. 2016, *408*, 6079-6091.
165. 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.
166. T'Kindt, R., Storme, M., Deforce, D., Van Bocxlaer, J., Evaluation of hydrophilic interaction chromatography versus reversed-phase chromatography in a plant metabolomics perspective, *Journal of Separation Science*. 2008, *31*, 1609-1614.
167. Guo, Y., Gaiki, S., Retention and selectivity of stationary phases for hydrophilic interaction chromatography, *Journal of Chromatography A*. 2011, *1218*, 5920-5938.
168. Greco, G., Letzel, T., Main interactions and influences of the chromatographic parameters in HILIC separations, *Journal of Chromatographic Science*. 2013, *51*, 684-693.
169. Gritti, F., dos Santos Pereira, A., Sandra, P., Guiochon, G., Efficiency of the same neat silica column in hydrophilic interaction chromatography and per aqueous liquid chromatography, *Journal of Chromatography A*. 2010, *1217*, 683-688.

170. Jandera, P., Stationary and mobile phases in hydrophilic interaction chromatography: a review, *Analytica Chimica Acta*. 2011, *692*, 1-25.
171. Buszewski, B., Noga, S., Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique, *Analytical and Bioanalytical Chemistry*. 2012, *402*, 231-247.
172. Shou, W. Z., Naidong, W., Simple means to alleviate sensitivity loss by trifluoroacetic acid (TFA) mobile phases in the hydrophilic interaction chromatography–electrospray tandem mass spectrometric (HILIC–ESI/MS/MS) bioanalysis of basic compounds, *Journal of Chromatography B*. 2005, *825*, 186-192.
173. Ikegami, T., Tomomatsu, K., Takubo, H., Horie, K., Tanaka, N., Separation efficiencies in hydrophilic interaction chromatography, *Journal of Chromatography A*. 2008, *1184*, 474-503.
174. McCalley, D. V., Study of the selectivity, retention mechanisms and performance of alternative silica-based stationary phases for separation of ionised solutes in hydrophilic interaction chromatography, *Journal of Chromatography A*. 2010, *1217*, 3408-3417.
175. Choi, M. Y., Chai, C., Park, J. H., Lim, J., Lee, J., Kwon, S. W., Effects of storage period and heat treatment on phenolic compound composition in dried Citrus peels (Chenpi) and discrimination of Chenpi with different storage periods through targeted metabolomic study using HPLC-DAD analysis, *Journal of Pharmaceutical and Biomedical Analysis*. 2011, *54*, 638-645.
176. Cevallos-Cevallos, J. M., Rouseff, R., Reyes-De-Corcuera, J. I., Untargeted metabolite analysis of healthy and Huanglongbing-infected orange leaves by CE-DAD, *Electrophoresis*. 2009, *30*, 1240-1247.
177. Lei, Z., Huhman, D. V., Sumner, L. W., Mass Spectrometry Strategies in Metabolomics, *The Journal of Biological Chemistry*. 2011, *286*, 25435-25442.
178. Gowda, G. A. N., Djukovic, D., Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges, *Methods in Molecular Biology*. 2014, *1198*, 3-12.
179. Kai, H., Kinoshita, K., Harada, H., Uesawa, Y., Maeda, A., Suzuki, R., Okada, Y., Takahashi, K., Matsuno, K., Establishment of a direct-injection electron ionization-mass spectrometry metabolomics method and its application to Lichen profiling, *Analytical Chemistry*. 2017, *89*, 6408-6414.
180. 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.
181. McLafferty, F. W., Tandem mass spectrometry, *Science*. 1981, *214*, 280.
182. Fenn, J., Electrospray ionization mass spectrometry: How it all began, *Journal of Biomolecular Techniques:JBT*. 2002, *13*, 101-118.
183. 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.
184. Hilton, G. R., Benesch, J. L., Two decades of studying non-covalent biomolecular assemblies by means of electrospray ionization mass spectrometry, *Journal of The Royal Society Interface*. 2012, *9*, 801-816.
185. Cotter, R. J., Time-of-flight mass spectrometry for the structural analysis of biological molecules, *Analytical Chemistry*. 1992, *64*, 1027a-1039a.
186. Marshall, A. G., Hendrickson, C. L., High-resolution mass spectrometers, *Annual Review of Analytical Chemistry*. 2008, *1*, 579-599.
187. Guilhaus, M., Selby, D., Mlynski, V., Orthogonal acceleration time-of-flight mass spectrometry, *Mass Spectrometry Reviews*. 2000, *19*, 65-107.
188. Dawson, J. H. J., Guilhaus, M., Orthogonal-acceleration time-of-flight mass spectrometer, *Rapid Communications in Mass Spectrometry*. 1989, *3*, 155-159.
189. Romero-González, R., Frenich, A. G., *Applications in High Resolution Mass Spectrometry: Food Safety and Pesticide Residue Analysis*, 1st Edition, Elsevier, 2017.
190. Eliuk, S., Makarov, A., Evolution of Orbitrap Mass Spectrometry Instrumentation, *Annual Review of Analytical Chemistry*. 2015, *8*, 61-80.
191. Wang, Y., Gu, M., The Concept of Spectral Accuracy for MS, *Analytical Chemistry*. 2010, *82*, 7055-7062.
192. Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., ProteoWizard: open source software for rapid proteomics tools development, *Bioinformatics*. 2008, *24*, 2534-2536.
193. Chambers, M. C., Maclean, B., et al., A cross-platform toolkit for mass spectrometry and proteomics, *Nature Biotechnology*. 2012, *30*, 918-920.
194. Tautenhahn, R., Bottcher, C., Neumann, S., Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*. 2008, *9*, 504.

195. Farrés, M., Piña, B., Tauler, R., Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS, *Metabolomics*. 2015, *11*, 210-224.
196. 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.
197. 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.
198. 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.
199. 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.
200. Gorrochategui, E., Jaumot, J., Tauler, R., A protocol for LC-MS metabolomic data processing using chemometric tools, *Nature Protocol Exchange.* 2015. doi:10.1038/protex.2015.102.
201. Strutz, T., (Eds.), *Data Fitting and Uncertainty-A practical introduction to weighted least squares and beyond*, 2nd Edition, Springer Vieweg, 2016.
202. Eilers, P. H. C., Boelens, H. F. M., *Baseline Correction with Asymmetric Least Squares Smoothing*, 2005.
203. Eilers, P. H. C., A Perfect Smoother, *Analytical chemistry*. 2003, *75*, 3631-3636.
204. Nielsen, N. P. V., Carstensen, J. M., Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *Journal of Chromatography A*. 1998, *A*, 17-35.
205. Pierce, K. M., Hope, J. L., Johnson, K. J., Wright, B. W., Synovec, R. E., Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, *Journal of Chromatography A*. 2005, *1096*, 101-110.
206. Christin, C., Smilde, A. K., Hoefsloot, H. C., Suits, F., Bischoff, R., Horvatovich, P. L., Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms, *Analytical Chemistry*. 2008, *80*, 7012-7021.
207. Skov, T., van den Berg, F., Tomasi, G., Bro, R., Automated alignment of chromatographic data, *Journal of Chemometrics*. 2006, *20*, 484-497.
208. 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.
209. Trygg, J., Gabrielsson, J., Lundstedt, T., in: Tauler R, Walczak B (Eds.), *Comprehensive Chemometrics*, 1st Edition, Elsevier, Oxford, 2009, pp. 1-8.
210. Garcia-Reiriz, A. G., Olivieri, A. C., Teixido, E., Ginebreda, A., Tauler, R., Chemometric modeling of organic contaminant sources in surface waters of a mediterranean river basin, *Environmental Science: Processes & Impacts*. 2014, *16*, 124-134.
211. Tauler, R., Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*. 1995, *30*, 133-146.
212. Ruckebusch, C., Duponchel, L., Sombret, B., Huvenne, J. P., Saurina, J., Time-Resolved Step-Scan FT-IR Spectroscopy: Focus on Multivariate Curve Resolution, *Journal of Chemical Information and Computer Sciences*. 2003, *43*, 1966-1973.
213. Piqueras, S., Duponchel, L., Tauler, R., de Juan, A., Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares, *Anal Chim Acta*. 2011, *705*, 182-192.
214. Zhang, X., Tauler, R., Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging, *Analytica Chimica Acta*. 2013, *762*, 25-38.
215. Puig-Castellví, F., Alfonso, I., Pinã, B., Tauler, R., 1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis, *Scientific Reports*. 2016, *6*. 30982.
216. Wold, S., Esbensen, K., Geladi, P., Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*. 1987, *2*, 37-52.
217. Golub, G. H., Reinsch, C., Singular value decomposition and least squares solutions, *Numerische Mathematik*. 1970, *14*, 403-420.

218. Windig, W., Gallaguer, N. B., Shaver, J. M., Wise, B. M., *A new approach for interactive self-modeling mixture analysis*, Elsevier, Amsterdam, The Netherlands, 2005.
219. 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.
220. De Juan, A., Jaumot, J., Tauler, R., Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, *Analytical Methods*. 2014, *6*, 4964-4976.
221. Abdollahi, H., Tauler, R., Uniqueness and rotation ambiguities in multivariate curve resolution methods, *Chemometrics and Intelligent Laboratory Systems*. 2011, *108*, 100-111.
222. De Juan, A., Vander Heyden, Y., Tauler, R., Massart, D. L., Assessment of new constraints applied to the alternating least squares method, *Analytica Chimica Acta*. 1997, *346*, 307-318.
223. Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., Zhu, F., Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis, *Scientific Reports*. 2016, *6*, 38881.
224. Hugelier, S., Piqueras, S., Bedia, C., de Juan, A., Ruckebusch, C., Application of a sparseness constraint in multivariate curve resolution-Alternating least squares, *Analytica Chimica Acta*. 2017.
225. 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.
226. Wishart, D. S., Current progress in computational metabolomics, *Briefings in Bioinformatics*. 2007, *8*, 279-293.
227. Kim, T. K., T test as a parametric statistic, *Korean Journal of Anesthesiology*. 2015, *68*, 540-546.
228. Bin, Z., Yuqing, Z., Mann-Whitney U test and Kruskal-Wallis test should be used for comparisons of differences in medians, not means: Comment on the article by van der Helm-van Mil et al, *Arthritis & Rheumatism*. 2009, *60*, 1565-1565.
229. Kim, H. Y., Analysis of variance (ANOVA) comparing means of more than two groups, *Restorative Dentistry & Endodontics*. 2014, *39*, 74-77.
230. 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.
231. Massart, D. L., Buydens, L. M. C., Vandeginste, B. G. M., *Handbook of Chemometrics and Qualimetrics*, 1st Edition, Elsevier, Amsterdam, The Netherlands, 1997.
232. Miller, J. N., Miller, J. C., *Statistics and chemometrics for analytical chemistry*, Pearson Education, England, 2010.
233. Alonso, A., Marsal, S., Julià, A., Analytical Methods in Untargeted Metabolomics: State of the Art in 2015, *Frontiers in Bioengineering and Biotechnology*. 2015, *3*, 23.
234. Joliffe, I. T., Morgan, B. J., Principal component analysis and exploratory factor analysis, *Statistical Methods in Medical Research*. 1992, *1*, 69-95.
235. Benjamini, Y., Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society Series B (Methodological)*. 1995, *57*, 289-300.
236. 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.
237. Wold, S., Analysis of similarities and dissimilarities between chromatographic liquid phases by means of pattern cognition, *Journal of Chromatographic Science*. 1975, *13*, 525-532.
238. Köhn, H.-F., Hubert, L. J., (Eds.), *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2014.
239. Madala, N. E., Piater, L. A., Steenkamp, P. A., Dubery, I. A., Multivariate statistical models of metabolomic data reveals different metabolite distribution patterns in isonitrosoacetophenone-elicited Nicotiana tabacum and Sorghum bicolor cells, *SpringerPlus*. 2014, *3*, 254.
240. Bartel, J., Krumsiek, J., Theis, F. J., Statistical methods for the analysis of high-throughput metabolomics data, *Computational and Structural Biotechnology Journal*. 2013, *4*, e201301009.
241. Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., Veselkov, K. A., Daviglus, M. L., Kesteloot, H., Ueshima, H., Zhao, L., Nicholson, J. K., Elliott, P., Human metabolic phenotype diversity and its association with diet and blood pressure, *Nature*. 2008, *453*, 396-400.
242. Gadjev, I., Vanderauwera, S., Gechev, T. S., Laloi, C., Minkov, I. N., Shulaev, V., Apel, K., Inzé, D., Mittler, R., Van Breusegem, F., Transcriptomic Footprints Disclose Specificity of Reactive Oxygen Species Signaling in Arabidopsis, *Plant Physiology*. 2006, *141*, 436-445.

243. Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., Gavin, A. C., Visualization of omics data for systems biology, *Nature Methods*. 2010, *7*, S56-68.
244. Jolliffe, I. T., *Principal Component Analysis*, Springer New York, New York, 1986.
245. Terrado, M., Kuster, M., Raldúa, D., Lopez de Alda, M., Barceló, D., Tauler, R., Use of chemometric and geostatistical methods to evaluate pesticide pollution in the irrigation and drainage channels of the Ebro river delta during the rice-growing season, *Analytical and Bioanalytical Chemistry*. 2007, *387*, 1479-1488.
246. Bengraïne, K., Marhaba, T. F., Using principal component analysis to monitor spatial and temporal changes in water quality, *Journal of Hazardous Materials*. 2003, *100*, 179-195.
247. Yetukuri, L. R., *Bioinformatics approaches for the analysis of lipidomics data*, VTT Publications, 2010.
248. Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J., Wold, S., Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm), *Analytical and Bioanalytical Chemistry*. 2004, *380*, 419-429.
249. Trygg, J., Holmes, E., Lundstedt, T., Chemometrics in metabonomics, *Journal of Proteome Research*. 2007, *6*, 469-479.
250. Wold, S., Ruhe, A., Wold, H., Dunn, I., WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing*. 1984, *5*, 735-743.
251. Martens, H., Naes, T., *Multivariate calibration*, John Wiley & Sons,1992.
252. Jonsson, P., Bruce, S. J., Moritz, T., Trygg, J., Sjostrom, M., Plumb, R., Granger, J., Maibaum, E., Nicholson, J. K., Holmes, E., Antti, H., Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets, *Analyst*. 2005, *130*, 701-707.
253. Perez-Enciso, M., Tenenhaus, M., Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach, *Human Genetics*. 2003, *112*, 581-592.
254. Geladi, P., Kowalski, B. R., Partial least-squares regression: a tutorial, *Analytica Chimica Acta*. 1986, *185*, 1-17.
255. Barker, M., Rayens, W., Partial least squares for discrimination, *Journal of Chemometrics*. 2003, *17*, 166-173.
256. Rajalahti, T., Arneberg, R., Kroksveen, A. C., Berle, M., Myhr, K. M., Kvalheim, O. M., Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, *Analytical Chemistry*. 2009, *81*, 2581-2590.
257. Wold, S., Johansson, E., Cocchi, M., PLS-partial least squares projections to latent structures, *3D QSAR in Drug Design*. 1993, *1*, 523-550.
258. 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.
259. Ståhle, L., Wold, S., Analysis of variance (ANOVA), *Chemometrics and Intelligent Laboratory System*. 1989, 259-272.
260. Mardia, K. V., Kent, J. T., Bibby, J. M., *Multivariate Analysis*, 1st Edition, Academic Press, 1979.
261. Scheiner, S. M., Gurevitch, J., Multiple response variables and multi-species interactions, *Design and Analysis of Ecological Experiments* 1993, 94-112.
262. Ståhle, L., Wold, S., Multivariate analysis of variance (MANOVA), *Chemometrics and Intelligent Laboratory System*.1990, 127-141.
263. 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.
264. 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, *Analytica Chimica Acta*. 2015, *899*, 1-12.
265. Vis, D. J., Westerhuis, J. A., Smilde, A. K., van der Greef, J., Statistical validation of megavariate effects in ASCA, *BMC Bioinformatics*. 2007, *8*, 322.
266. Hoefsloot, H. C. J., Vis, D. J., Westerhuis, J. A., Smilde, A. K., Jansen, J. J., in: Tauler R, Walczak B (Eds.), Multiset Data Analysis: ANOVA Simultaneous Component Analysis and Related Methods, *Comprehensive Chemometrics*, Elsevier, Oxford, 2009, pp.453-472.

267. Marini, F., de Beer, D., Walters, N. A., de Villiers, A., Joubert, E., Walczak, B., Multivariate analysis of variance of designed chromatographic data. A case study involving fermentation of rooibos tea, *Journal of Chromatography A*. 2017, *1489*, 115-125.
268. Ledoit, O., Wolf, M., A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*. 2004, *88*, 365-411.
269. Richards, S. E., Dumas, M. E., Fonville, J. M., Ebbels, T. M. D., Holmes, E., Nicholson, J. K., Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework, *Chemometrics and Intelligent Laboratory Systems*. 2010, *104*, 121-131.
270. Gligorijević, V., Pržulj, N., Methods for biological data integration: perspectives and challenges, *Journal of the Royal Society Interface*. 2015, *12*, 20150571.
271. Ebbels, T. M. D., Cavill, R., Bioinformatic methods in NMR-based metabolic profiling, *Progress in Nuclear Magnetic Resonance Spectroscopy*. 2009, *55*, 361-374.
272. Cavill, R., Jennen, D., Kleinjans, J., Briede, J. J., Transcriptomic and metabolomic data integration, *Briefings in Bioinformatics*. 2016, *17*, 891-901.
273. Boccard, J., Rudaz, S., Harnessing the complexity of metabolomic data with chemometrics, *Journal of Chemometrics*. 2014, *28*, 1-9.
274. Schouteden, M., Van Deun, K., Pattyn, S., Van Mechelen, I., SCA with rotation to distinguish common and distinctive information in linked data, *Behavior Research Methods*. 2013, *45*, 822-833.
275. Acar, E., Rasmussen, M. A., Savorani, F., Næs, T., Bro, R., Understanding data fusion within the framework of coupled matrix and tensor factorizations, *Chemometrics and Intelligent Laboratory Systems*. 2013, *129*, 53-63.
276. Alter, O., Brown, P. O., Botstein, D., Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, *Proceedings of the National Academy of Sciences of the United States of America*. 2003, *100*, 3351-3356.
277. Bouhaddani, S. e., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., Uh, H.-W., Evaluation of O2PLS in Omics data integration, *BMC Bioinformatics*. 2016, *17*, S11.
278. Löfstedt, T., Trygg, J., OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation, *Journal of Chemometrics*. 2011, *25*, 441-455.
279. Kuligowski, J., Perez-Guaita, D., Sanchez-Illana, A., Leon-Gonzalez, Z., de la Guardia, M., Vento, M., Lock, E. F., Quintas, G., Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE), *Analyst*. 2015, *140*, 4521-4529.
280. Acar, E., Bro, R., Smilde, A. K., Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations, *Proceedings of the IEEE*. 2015, *103*, 1602-1620.
281. Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L. W., Trengove, R., Wolfender, J.-L., Metabolite identification: are you sure? And how do your peers gauge your confidence?, *Metabolomics*. 2014, *10*, 350-353.
282. Sumner, L. W., Lei, Z., Nikolau, B. J., Saito, K., Roessner, U., Trengove, R., Proposed quantitative and alphanumeric metabolite identification metrics, *Metabolomics*. 2014, *10*, 1047-1049.
283. Salek, R. M., Arita, M., Dayalan, S., Ebbels, T., Jones, A. R., Neumann, S., Rocca-Serra, P., Viant, M. R., Vizcaíno, J.-A., Embedding standards in metabolomics: the Metabolomics Society data standards task group, *Metabolomics*. 2015, *11*, 782-783.
284. Sumner, L. W., Amberg, A., et al., Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), *Metabolomics: Official journal of the Metabolomic Society*. 2007, *3*, 211-221.
285. Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., Viant, M. R., Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics, *Metabolomics*. 2013, *9*, 44-66.
286. Heinonen, M., Shen, H., Zamboni, N., Rousu, J., Metabolite identification and molecular fingerprint prediction through machine learning, *Bioinformatics*. 2012, *28*, 2333-2341.
287. Gerlich, M., Neumann, S., MetFusion: integration of compound identification strategies, *Journal of Mass Spectrometry*. 2013, *48*, 291-298.
288. Li, L., Li, R., Zhou, J., Zuniga, A., Stanislaus, A. E., Wu, Y., Huan, T., Zheng, J., Shi, Y., Wishart, D. S., Lin, G., MyCompoundID: using an evidence-based metabolome library for metabolite identification, *Analytical Chemistry*. 2013, *85*, 3401-3408.
289. Allen, F., Pon, A., Wilson, M., Greiner, R., Wishart, D., CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra, *Nucleic Acids Research*. 2014, *42*, W94-W99.

290. Allen, F., Greiner, R., Wishart, D., Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, *Metabolomics*. 2015, *11*, 98-110.
291. Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., van Schaik, R., Vervoort, J., Substructure-based annotation of high-resolution multistage MS^n spectral trees, *Rapid Communications in Mass Spectrometry*. 2012, *26*, 2461-2471.
292. Ridder, L., van der Hooft, J. J., Verhoeven, S., de Vos, R. C., Bino, R. J., Vervoort, J., Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea, *Analytical Chemistry*. 2013, *85*, 6033-6040.
293. Fiehn, O., Barupal, D. K., Kind, T., Extending biochemical databases by metabolomic surveys, *Journal of Biological Chemistry*. 2011, *286*, 23637-23643.
294. Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules, *Nucleic Acids Research*. 2009, *37*, W623-W633.
295. Pence, H. E., Williams, A., ChemSpider: An online chemical information resource, *Journal of Chemical Education*. 2010, *87*, 1123-1124.
296. Zhu, Z. J., Schultz, A. W., Wang, J., Johnson, C. H., Yannone, S. M., Patti, G. J., Siuzdak, G., Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database, *Nature Protocols*. 2013, *8*, 451-460.
297. Horai, H., Arita, M., *et al*., MassBank: a public repository for sharing mass spectral data for life sciences, *Journal of Mass Spectrometry*. 2010, *45*, 703-714.
298. Wishart, D. S., Tzur, D., *et al*., HMDB: the human metabolome database, *Nucleic Acids Research*. 2007, *35*, D521-D526.
299. Kanehisa, M., Goto, S., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*. 2000, *28*, 27-30.
300. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., Stein, L., Reactome: a knowledgebase of biological pathways, *Nucleic Acids Research*. 2005, *33*, D428-432.
301. Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., Pico, A. R., WikiPathways: building research communities on biological pathways, *Nucleic Acids Research*. 2012, *40*, D1301-D1307.
302. Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N., Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nucleic Acids Research*. 2005, *33*, 6083-6089.
303. Smith, C. A., Maille, G. O., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., Siuzdak, G., METLIN: A metabolite mass spectral database, *Therapeutic Drug Monitoring*. 2005, *27*, 747-751.
304. Sana, T. R., Roark, J. C., Li, X., Waddell, K., Fischer, S. M., Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS, *Journal of Biomolecular Techniques*. 2008, *19*, 258-266.
305. Yang, X., Neta, P., Stein, S. E., Quality control for building libraries from electrospray ionization tandem mass spectra, *Analytical Chemistry*. 2014, *86*, 6393-6400.
306. Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A. C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., Wilson, M., Wishart, D. S., YMDB: the yeast metabolome database, *Nucleic Acids Research*. 2012, *40*, D815-D820.
307. Wishart, D. S., Feunang, Y. D., et al., HMDB 4.0: the human metabolome database for 2018, *Nucleic Acids Research*. 2018, *46*, D608-D617.
308. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M., ChEBI: a database and ontology for chemical entities of biological interest, *Nucleic Acids Research*. 2008, *36*, D344-D350.
309. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., *et al*., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Research*. 2003, *31*, 365-370.
310. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., Wheeler, D. L., GenBank, *Nucleic Acids Research*. 2000, *28*, 15-18.
311. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., GenBank, *Nucleic Acids Research*. 2013, *41*, D36-D42.

312. Ramirez-Gaona, M., Marcu, A., Pon, A., Guo, A. C., Sajed, T., Wishart, N. A., Karu, N., Djoumbou Feunang, Y., Arndt, D., Wishart, D. S., YMDB 2.0: a significantly expanded version of the yeast metabolome database, *Nucleic Acids Research*. 2017, *45*, D440-D445.
313. Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., Siuzdak, G., METLIN: A technology platform for identifying knowns and unknowns, *Analytical Chemistry*. 2018, *90*, 3156-3164.
314. Nishioka, T., MassBank: Database of mass spectra and its applications to metabolomics, *CICSJ Bulletin*. 2014, *32*, 62.
315. Wishart, D. S., in: Matthiesen R (Eds.), *Bioinformatics Methods in Clinical Research*, Humana Press, Totowa. NJ 2010, pp. 283-313.
316. Caspi, R., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic Acids Research*. 2008, *36*, D623-D631.
317. Kanehisa, M., KEGG for linking genomes to life and the environment, *Nucleic Acids Research*. 2008, *36*, D480-D484.
318. Okuda, S., KEGG atlas mapping for global analysis of metabolic pathways, *Nucleic Acids Research*. 2008, *36*, W423-W426.
319. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., KEGG: new perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Research*. 2017, *45*, D353-D361.

CAPÍTOL 3.

Desenvolupament i optimització de mètodes analítics i quimiomètrics en estudis de metabolòmica no dirigida

3.1. INTRODUCCIÓ

L'anàlisi no dirigida del metaboloma dels sistemes biològics és un problema analític complex, principalment a causa de la gran diversitat de concentracions i propietats fisicoquímiques dels metabòlits endògens. Actualment, un dels grans reptes de la metabolòmica no dirigida és el desenvolupament d'estratègies analítiques que permetin cobrir el màxim nombre de metabòlits de l'organisme ja que no hi ha cap tècnica analítica que permeti estudiar el metaboloma íntegre. En aquesta Tesi, s'han proposat diferents metodologies analítiques no dirigides adequades per a la separació, detecció i caracterització de biomarcadors metabòlics mitjançant la cromatografia de líquids (LC) i l'electroforesi capil·lar (CE) acoblades a l'espectrometria de masses (MS).

D'una banda, la LC és avui en dia la tècnica de separació més emprada en els estudis de metabolòmica. No obstant, l'elevada polaritat i hidrofilicitat dels metabòlits fa que presentin una baixa retenció en les fases estacionàries convencionals de fase invertida raó per la qual han aparegut nous modes de separació, com la cromatografia de líquids d'interacció hidrofílica (HILIC) [1]. Així, en aquesta Tesi s'ha posat especial èmfasi en la cromatografia HILIC ja que és una bona alternativa per a l'anàlisi de compostos de naturalesa polar [2, 3]. Per entendre el seu funcionament, s'han avaluat cinc fases estacionàries HILIC diferents per a la seva aplicació en estudis de metabolòmica no dirigida. A més, també s'ha investigat la influència de les condicions experimentals (modificador orgànic, contingut d'additiu de la fase mòbil i el pH) en la interacció dels metabòlits amb els grups funcionals de la fase estacionària i així, proposar finalment una metodologia HILIC-MS (article I). D'altra banda, la CE és encara un tècnica poc estesa en el camp de la metabolòmica. Malgrat això, aquesta plataforma ha demostrat repetidament ser una estratègia poderosa i eficaç per a la separació de metabòlits polars relacionats amb el metabolisme primari [4]. Per aquest motiu, en aquesta Tesi es proposa una metodologia analítica basada en CE-MS per dur a terme estudis metabolòmics no dirigits (article II).

Actualment, l'extrema complexitat lligada a l'anàlisi de les dades de metabolòmica no dirigida segueix sent un dels reptes en el camp de la quimiometria. Les dades megavariants generades en

aquests estudis amb milions de senyals per mostra analitzada requereixen de mètodes quimiomètrics avançats per a la seva modelització [5, 6]. Degut a les enormes dimensions dels arxius, en aquest capítol també es revisen diferents estratègies de tractament de dades metabolòmiques no dirigides de MS basades en una etapa de compressió de les dades seguida de la resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS). Així doncs, l'extracció quimiomètrica de la informació present en els conjunts de dades obtinguts experimentalment per LC-MS i CE-MS ha permès la detecció i identificació dels metabòlits clau del comportament dels sistemes biològics investigats en cada cas.

Ara bé, un dels grans reptes pendents en la biologia dels sistemes i en l'anàlisi de les dades òmiques és el desenvolupament de nous enfocaments que permetin millorar la comprensió dels processos biològics [7, 8]. Per aquest motiu, els mètodes de fusió de dades són avui en dia un dels principals objectes d'investigació en el camp de la quimiometria i de la bioinformàtica per tal d'integrar el coneixement de les diferents ciències òmiques. Les estratègies de fusió permeten l'estudi simultani de conjunts de dades que provenen de diferents plataformes analítiques, nivells omics, organismes o tipus de mostra (diferents teixits, fluids biològics, etc.). En conseqüència, en aquest capítol es presenta la integració de dades metabolòmiques no dirigides obtingudes mitjançant diferents plataformes analítiques (CE-MS i HILIC-MS) fent servir l'estratègia combinada de la compressió mitjançant la selecció de les regions d'interès, ROI, i de la resolució per MCR-ALS (mètode ROIMCR) (article III). D'aquesta manera, s'obté informació més precisa sobre els canvis en els perfils metabòlics dels organismes biològics i es genera una interpretació biològica més completa i veraç.

Per demostrar la idoneïtat de les metodologies analítiques no dirigides (HILIC-MS i CE-MS) i de tractament de dades proposades es van avaluar els perfils metabòlics de mostres de llevat (*Saccharomyces cerevisiae*) en condicions de creixement òptimes i estressants. *S. cerevisiae* és un organisme eucariota unicel·lular àmpliament utilitzat com a model en investigacions metabolòmiques fonamentals i aplicades [9]. El llevat requereix de condicions específiques per a un creixement òptim. En condicions d'estrès ambiental, com l'estrès tèrmic o fonts de carboni pobres, sofreix alteracions en el seu metabolisme que poden afectar al seu cicle vital.

3.2. PUBLICACIONS

- **Article científic I.** Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites. **E. Ortiz-Villanueva**, M. Navarro-Reig, J. Jaumot, R. Tauler. *Analytical Methods* 9 (2017) 774-785.

En aquest article s'avalua el comportament de diferents fases estacionàries HILIC per a l'anàlisi d'una mescla model de metabòlits mitjançant LC-DAD. Amb aquesta finalitat, es va emprar un disseny experimental factorial complet tenint en compte diferents paràmetres cromatogràfics que influeixen en la separació en les columnes HILIC, com el tipus de modificador orgànic, la força iònica o el pH de la fase aquosa. La capacitat de retenció de cadascuna de les fases estacionàries a les diferents condicions experimentals i la millor configuració cromatogràfica per dur a terme estudis de metabolòmica es va determinar mitjançant l'anàlisi dels cromatogrames amb mètodes quimiomètrics i la funció resposta cromatogràfica (CRF) de Berridge.

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

En aquest article es presenten les metodologies analítiques i quimiomètriques posades a punt en els estudis de metabolòmica no dirigida mitjançant CE-MS presentats en aquesta Tesi. L'anàlisi quimiomètrica de les dades es basa en l'aplicació del mètode MCR-ALS per a la resolució dels perfils electroforètics i els espectres dels metabòlits presents en les mostres analitzades en *full scan* i en mode *profile*. Aquesta estratègia va permetre resoldre diversos problemes electroforètics, com són les contribucions del soroll de fons, les relacions de senyal/soroll pobres, la asimetria dels pics i el desplaçament dels pics electroforètics. A més, es demostra la utilitat de la metodologia proposada en l'estudi dels canvis dels perfils metabòlics de mostres de llevat (*Saccharomyces cerevisiae*) cultivades a dues temperatures diferents, 30 °C i 37 °C.

- **Article científic III.** Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data. **E. Ortiz-Villanueva**, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler. J. Jaumot. *Analytica Chimica Acta* 978 (2017) 10-23.

En aquest tercer treball es presenten dues estratègies de fusió o integració de dades metabolòmiques obtingudes a partir de l'aplicació de metodologies no dirigides de CE-MS i LC-MS i de mètodes quimiomètrics. Aquestes estratègies es basen en la selecció de les regions d'interès, ROI, seguida del procés de resolució MCR-ALS (ROIMCR). En primer lloc, es van integrar i interpretar conjuntament els metabòlits obtinguts a partir de les anàlisis per separat de les dades de CE-MS i LC-MS. En segon lloc, es va proposar fer la integració directa (fusió de nivell baix) de les dades experimentals de CE-MS i LC-MS abans de ser analitzades per MCR-ALS. Aquests resultats, permeten generar la seva interpretació conjunta. La idoneïtat i els beneficis d'aquestes dues estratègies d'integració es va avaluar en l'estudi del metaboloma de mostres de llevat (*S.cerevisiae*) cultivades amb dues fonts de carboni diferents: acetat (fermentable) i glucosa (no fermentable).

3.2.1. Article científic I.

Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites.

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

Analytical Methods 9 (2017) 774-785.

Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2017, **9**, 774

Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites†

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

Different hydrophilic interaction liquid chromatography (HILIC) stationary phases have been evaluated using different chemometric methods with the aim of their application in metabolomics studies. Experimental factors, such as the type of HILIC stationary phase (*i.e.* amide, amine, zwitterionic and diol) and the mobile phase conditions (organic co-solvent, pH and ionic strength) were assessed using a full factorial experimental design. A test sample mixture of metabolites with diverse physicochemical properties (amino acids, nucleotides, nucleosides, and sugars among others) was analyzed by liquid chromatography with a diode array detector (LC-DAD) using five different HILIC columns. Application of multivariate curve resolution alternating least squares (MCR-ALS) method, allowed the full chromatographic peak resolution of all mixture constituents. This approach was particularly helpful in the case of methanol samples where the quality of the chromatographic separation (resolution) was lower in consequence of the co-solvent perturbation on the water layer formation at the surface of the stationary phase. Then, Berridge chromatographic response function (CRF), based on peak resolution, retention times and number of peaks, was used for the investigation of the best HILIC column configuration for future metabolomics studies. The best chromatographic configuration resulted in being the amide and zwitterionic HILIC stationary phases in combination with acetonitrile as organic co-solvent of the mobile phase.

Received 31st October 2016
Accepted 3rd January 2017

DOI: 10.1039/c6ay02976k

www.rsc.org/methods

1. Introduction

Metabolomics studies aim to characterize the complete endogenous low-molecular weight compounds (metabolites) present in biological systems.[1,2] Metabolites have diverse physicochemical properties and are usually found at a broad range of concentrations in living organisms.[3,4] In the last years, new approaches have gained importance to increase coverage capability of these compounds.

Liquid chromatography (LC) is an attractive platform for *-omic* analyses because of its versatility, precision and high concentration sensitivity. In the metabolomics field, reversed-phase liquid chromatography (RPLC) coupled to C18 or C8 stationary phases present very low retention for the highly polar and hydrophilic compounds which are usually found in metabolomic samples. The development of analytical alternatives and technologies, such as ion-pair chromatography (IPC)[5] or relatively novel hydrophilic interaction liquid chromatography (HILIC),[6] had overcome this major drawback of more traditional approaches. Nowadays, the use of HILIC is widely recognized as an alternative for carrying out metabolomics studies,[3,7] solving polar selectivity issues associated with the use of RPLC columns. HILIC stationary phases can be considered to be a combination of a normal stationary phase (NP) and an RP mobile phase, containing a high percentage of organic solvent. Under such conditions, HILIC stationary phases provide enhanced retention for strong polar compounds on their surface active groups.

In the last decade, HILIC stationary phase supports and surface chemistry advancements have provided solutions to specific separation problems such as those related to the determination of short chain carboxylic acids,[8] carbohydrates,[9,10] amino acids,[11] nucleosides and nucleotides,[12] and peptides.[13] Consequently, several types of HILIC columns have been described, including plain silica, neutral polar chemically bonded, ion-exchange and zwitterionic stationary phases.[14–17] However, according to Alpert's,[18,19] the retention mechanism of HILIC chromatographic systems is more complex than for RPLC systems.[20] This is due to the different retention patterns depending on the specific stationary phase considered. HILIC retention is a function of various molecular interactions, such as adsorption, ion exchange, and analyte partitioning between the mobile phase and the water-rich layer at the surface of the stationary phase.[21] Accordingly, the chemistry occurring at the stationary phase and the composition of the mobile phase (pH,

Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain. E-mail: roma.tauler@idaea.csic.es; Tel: +34934006140

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6ay02976k

This journal is © The Royal Society of Chemistry 2017

View Article Online
Paper Analytical Methods

ionic strength, and organic co-solvent composition) play a crucial role in HILIC selectivity and retention mechanisms.[22,23] Therefore, the selection of the most suitable pair of stationary and mobile phases is a critical decision when a comprehensive analysis of metabolites is desired.

Until now few studies have attempted a statistical experimental design approach and optimization of the best separation conditions using HILIC stationary phases with the goal of reducing experimental efforts and increasing the reliability of the results.[19] In most of the previous studies about HILIC columns, data analysis involved the application of multivariate exploratory and classification methods, such as principal component analysis (PCA)[19,24,25] or hierarchical cluster analysis (HCA)[19] to investigate the different behavior of HILIC columns. Some of the previously published articles have also reported the use of desirability functions for the automated comparison of HILIC stationary phases.[7,13] However, these kinds of approaches could miss relevant information from the data due to the not completely resolved peaks in the chromatographic separations of a large number of analytes. In this point, the application of chemometric tools for dealing with these complex datasets is highly recommended. These methods allow obtaining the resolution of peaks (overlapped, embedded) that could not be resolved using a single chromatographic separation. There are several chemometric methods that can be used to resolved these chromatographic data as detailed in the reviews of de Juan[26] or Amigo.[27] From all these families of methods, multivariate curve resolution alternating least squares (MCR-ALS) can be presented as a powerful method to get into the detail of this HILIC chromatographic data.[28] To the best of our knowledge, most works have been only focused on the comparison of less number of chromatographic conditions,[19] or using a narrow range of compounds or only a specific family of them,[13,25] without using the advantages of advanced chemometric methods, such as those from MCR-ALS to solve complex chromatographic separations.

In this work, the capability of different HILIC stationary phases (amine, amide, zwitterionic and diol) was evaluated for metabolomics studies by the combination of two strategies. On the one side, a full factorial experimental design was applied where the nature of the HILIC stationary phase and of the mobile phase (organic co-solvent, pH and ionic strength) were considered as experimental factors. This experimental design allowed the screening of the influence of these factors on the chromatographic behaviour of a mixture of preselected metabolites from different families analyzed by LC-DAD. On the other side, MCR-ALS approach was used to explore the effects caused by these chromatographic parameters in the total resolution of the elution profiles and pure UV spectra of all targeted metabolites. Finally, the use of the Berridge chromatographic response function (CRF)[29] is proposed for the evaluation of the different chromatographic conditions to gather the best configuration for the analysis of the considered metabolite mixture after MCR-ALS based chemometric metabolite resolution.

2. Materials and methods

2.1. Chemicals and reagents

All chemicals used in the preparation of solutions were of analytical reagent grade. Acetic acid (glacial), ammonia (25%), methanol (HPLC grade) and acetonitrile (HPLC grade) were purchased from Merck (Darmstadt, Germany). Ammonium acetate was provided by Sigma-Aldrich (St. Louis, MO, USA). Water with conductivity lower than 0.05 μS cm^{-1} was obtained using a Milli-Q water purification system (Millipore, Molsheim, France).

2.2. Standards mixture and working solutions

A test mixture of 12 chemical compounds, including reduced glutathione, D-fructose 1,6-bisphosphate sodium salt hydrate (F2,6BP), β-nicotinamide adenine dinucleotide (NADH), adenosine 5′-monophosphate disodium salt (AMP), hypoxanthine, uridine, inosine, cytidine, L-phenylalanine, L-tryptophan, L-tyrosine and L-citrulline was used as a model metabolite mixture. All standards were purchased from Sigma-Aldrich (St. Louis, MO, USA).

For each metabolite, one standard solution (1000 μg mL^{-1}) was prepared and stored in the freezer at −20 °C until its use. Working standard solutions were obtained by dilution of the metabolite stock solutions with water. A sample mixture of 12 metabolites, from different families (*i.e.* amino acids, nucleotides, nucleosides and sugars among others (purine, tripeptide)) and with various structures and physicochemical properties, was prepared and used as a test sample. Diluted standard solutions (concentrations of 20 μg mL^{-1}) were employed to evaluate the distinct chromatographic conditions used in this work.

2.3. Instrumentation and procedures

2.3.1. Stationary phases. In this work, five different HILIC stationary phases (BEH amide, amide, amine, zwitterionic and diol, see Table 1 for their main specifications) were tested to evaluate their suitability and performance.

2.3.2. LC-DAD. LC-DAD experiments were performed in an Agilent Infinity 1200 series system coupled to a diode array detector (DAD) (Agilent Technologies, Waldbronn, Germany). LC control and separation data acquisition were performed using ChemStation Software (Agilent Technologies).

Chromatographic separations were carried out using chromatographic conditions detailed in Table 1 (solvents, flow rate, and elution gradient). Ammonium acetate was used to prepare all aqueous mobile phases. Sample injection was performed with an autosampler at 4 °C, and the injection volume was 5 μL. Mobile phases were degassed for 15 min by sonication before using them. DAD detection was carried out considering a spectral range from 190 to 500 nm.

The effect of the four different factors on the chromatographic behavior was statistically assessed using a full factorial experimental design. The experimental factors were the HILIC stationary phase, and the other three factors related to the mobile phase composition: organic co-solvent, pH, and ionic

Table 1 Specifications of the HILIC columns and experimental conditions employed for the chromatographic separations

Column					Separation conditions	
Code	Name	Manufacturer	Stationary phase and surface chemistry	Dimensions	Flow (mL min^{-1})	Elution gradient (A: organic phase; B: aqueous phase)
A	XBridge™ Amide	Waters (Milford, MA, USA)	BEH amide –Linker–$CONH_2$, trifunctional amide	150 mm × 4.6 mm id, 5 μm	0.15	0–4 min, 5% B; 4–34 min, 5 to 70% B; 34–42 min, at 70% B; and 42–44 min, 70 to 5% B
B	TSK Gel Amide-80	Tosoh Bioscience (Tokyo, Japan)	Amide –$CONH_2$, non-ionic carbamoyl	250 mm × 2.1 mm id, 5 μm	0.15	0–3 min, 5% B; 3–27 min, 5 to 70% B; 27–30 min, at 70% B; and 30–32 min, 70 to 5% B
C	Luna-NH_2	Phenomenex (Torrance, CA, USA)	Amine ~NH_2	250 mm × 2.1 mm id, 5 μm	0.15	0–3 min, 5% B; 3–27 min, 5 to 70% B; 27–30 min, at 70% B; and 30–32 min, 70 to 5% B
D	ZIC-HILIC	SeQuant (Umeå, Sweden)	Zwitterionic ~$CH_2N^+(CH_3)_2$–CH_2–CH_2–CH_2–SO_3^-	250 mm × 2.1 mm id, 5 μm	0.15	0–3 min, 5% B; 3–27 min, 5 to 70% B; 27–30 min, at 70% B; and 30–32 min, 70 to 5% B
E	Acclaim™ Mixed-Mode HILIC-1	Thermo Scientific (Sunnyvale, CA, USA)	Diol ~$Si(CH_3)_2C_3H_{18}CH(OH)CH_2$–OH	150 mm × 2.1 mm id, 5 μm	0.15	0–2 min, 5% B; 2–16 min, 5 to 70% B; 16–20 min, at 70% B; and 20–22 min, 70 to 5% B

strength. Five HILIC stationary phases (columns) were evaluated (BEH amide, amide, amine, zwitterionic and diol, see Table 1). In the case of the mobile phase, two commonly used organic phase co-solvents were tested: methanol and acetonitrile; three pH values: acid (pH of the buffer was 3.0), moderately acid (pH of the buffer was 5.5) and neutral (pH of the buffer was 7.0); and three ionic strengths in the aqueous phase: low (5.0 mM), medium (25.0 mM), and high (50.0 mM). Preliminary studies showed that in all the considered cases (different pH values and organic co-solvent contents[30]) the pH at the highest organic content (95%) varied approximately up to 2 units in the case of acetonitrile, and up to 1.5 in the case of methanol. However, this shifting was considered not to dwarf the effect of pH significantly on the separation of the investigated metabolites.

To sum up, the experimental design considering all the factor levels (full factorial design) gave a total number of 90 experiments which were randomly injected twice (see ESI Table S1†). In addition, pure metabolite samples were also individually injected.

2.4. Data analysis

LC-DAD data from the set of metabolite mixture samples previously described were analyzed by chemometric methods with the goal of evaluating the separation capacity of the five HILIC stationary phases tested under the full factorial experimental conditions previously detailed. MCR-ALS was applied to resolve the chromatographic peaks of all metabolites in the analyzed test mixture. This complete resolution would not be possible without the help of MCR-ALS, due to the strong coelutions present in the used chromatographic conditions among the 12 metabolites considered in the mixture. This problem would be even more challenging in the case of the analysis of biological samples where hundreds of metabolites would be simultaneously present. From MCR-ALS resolved elution profiles, individual peak parameters for every metabolite (such as peak width and retention time) were obtained and used to investigate the performance of the different chromatographic combinations of stationary and mobile phases by means of Berridge chromatographic response function (CRF) defined for this purpose.

2.4.1. Data preparation and preprocessing. ChemStation .csv files were imported into the MATLAB® environment (The Mathworks Inc. Natick, MA, USA). In the case of LC-DAD, each chromatographic run provided a single data matrix $\mathbf{D}$ ($I \times J$) in which the rows were the spectra recorded at every retention time ($i = 1,\ldots I$), and the columns were the elution profiles obtained at every wavelength ($j = 1,\ldots J$). For instance, the number of rows ranged from 4800 to 9000 depending on the chromatographic run whereas the number of measured wavelengths was 156.

Since the experimental design considered 90 different experiment conditions, a total number of 90 data matrices were finally obtained, each one corresponding to the analysis of the metabolite mixture test sample under a particular chromatographic condition. Chromatograms from every data matrix were baseline/background corrected using the AsLS (asymmetric least squares) algorithm.[31] Additionally, in order to reduce the

This journal is © The Royal Society of Chemistry 2017

size of the data matrices to be analyzed, retention times (rows) and wavelengths (columns) where no signal was detected were removed. The final number of considered wavelengths (columns) was 49 ranging from 215 to 280 nm for all chromatographic experiments, and the number of retention times (rows) varied between 300 and 3000 seconds.

2.4.2. MCR-ALS resolution of chromatographic peaks. MCR-ALS is a chemometric method particularly useful to analyze complex multicomponent mixture systems, in particular chromatographic systems with strongly coeluted contributions.[32,33] In the case of LC-DAD data, MCR-ALS decomposes the original data matrix $\mathbf{D}$ ($I \times J$) according to a bilinear model into two factor matrices, $\mathbf{C}$ and $\mathbf{S}^T$, as is shown in eqn (1):

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

where $\mathbf{C}$ ($I \times M$) matrix contains the elution profiles of the M resolved contributions in the considered sample, $\mathbf{S}^T$ ($M \times J$) has their pure UV spectra and $\mathbf{E}$ ($I \times J$) is the residuals matrix with the background absorption unexplained by the model. Pure spectra allow the identification of the metabolites in each resolved component, whereas the corresponding resolved elution profiles permit retrieving their chromatographic peak parameters.[34,35]

Data sets obtained in the chromatographic analyses of the metabolite test mixture using different stationary phases and under the same mobile phase conditions (pH, ionic strength, and organic co-solvent) were analyzed simultaneously by MCR-ALS using the column-wise augmented data matrix strategy shown in eqn (2) (18 column-wise augmented data matrices). $\mathbf{D}_{\text{BEH amide}}$, $\mathbf{D}_{\text{amide}}$, $\mathbf{D}_{\text{amine}}$, $\mathbf{D}_{\text{zwitterionic}}$ and $\mathbf{D}_{\text{diol}}$ are the data matrices obtained in the different chromatographic runs using every HILIC stationary phase in all combinations of mobile phase conditions (combination of pH, ionic strength, and organic co-solvent). In addition, to facilitate the resolution and identification of all metabolites, chromatographic data matrices obtained in the analysis of every one of the 12 metabolites, $\mathbf{D}_{\text{standard},1}$ to $\mathbf{D}_{\text{standard},n}$, $n = 1,\ldots12$, using the HILIC amide stationary phase and mobile phase conditions of pH 5.5, ionic strength 5 mM and acetonitrile as organic co-solvent, were also added to each one of the 18 augmented data matrices.

$$\mathbf{D}_{\text{aug}} = \begin{bmatrix} \mathbf{D}_{\text{BEH amide}} \\ \mathbf{D}_{\text{amide}} \\ \mathbf{D}_{\text{amine}} \\ \mathbf{D}_{\text{zwitterionic}} \\ \mathbf{D}_{\text{diol}} \\ \mathbf{D}_{\text{standard},1} \\ \cdots \\ \mathbf{D}_{\text{standard},n} \end{bmatrix} = \mathbf{C}_{\text{aug}}\mathbf{S}^T + \mathbf{E}_{\text{aug}}$$
$$= \begin{bmatrix} \mathbf{C}_{\text{BEH amide}} \\ \mathbf{C}_{\text{amide}} \\ \mathbf{C}_{\text{amine}} \\ \mathbf{C}_{\text{zwitterionic}} \\ \mathbf{C}_{\text{diol}} \\ \mathbf{C}_{\text{standard},1} \\ \cdots \\ \mathbf{C}_{\text{standard},n} \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_{\text{BEH amide}} \\ \mathbf{E}_{\text{amide}} \\ \mathbf{E}_{\text{amine}} \\ \mathbf{E}_{\text{zwitterionic}} \\ \mathbf{E}_{\text{diol}} \\ \mathbf{E}_{\text{standard},1} \\ \cdots \\ \mathbf{E}_{\text{standard},n} \end{bmatrix} \quad (2)$$

Eqn (2) summarizes the extension of the bilinear model used by MCR-ALS when it is applied to the multirun chromatographic analyses of the metabolite mixture with the different tested HILIC columns at one particular experimental condition. $\mathbf{C}_{\text{aug}}$ in eqn (2) gives the resolved elution profiles of the 12 metabolites for every chromatographic run in the different HILIC stationary phases, $\mathbf{C}_{\text{BEH amide}}$, $\mathbf{C}_{\text{amide}}$, $\mathbf{C}_{\text{amine}}$, $\mathbf{C}_{\text{zwitterionic}}$ and $\mathbf{C}_{\text{diol}}$, and $\mathbf{C}_{\text{standard},1}$ to $\mathbf{C}_{\text{standard},n}$ have the chromatographic profiles of the pure metabolites. $\mathbf{S}^T$ has the MCR-ALS resolved pure spectra of all the components (12 metabolites) in all simultaneously analyzed datasets.[34] The ALS constrained optimization iteratively resolved eqn (2) for the elution and spectra profiles[35–37] under non-negativity constraints for both elution and spectra profiles, unimodality for elution profiles, and equal height normalization for spectra profiles.[35,37,38] The incorporation of the data matrices coming from the LC-DAD analysis of the pure metabolites (standards) favored the resolution of the complete system since unique spectra were obtained for them and facilitate the proper resolution of some the metabolites with more highly overlapped elution profiles. Pure metabolites were only analyzed under three different conditions (number 19, 25 and 31 in Table S1†) taking into account each studied pH with 5.0 mM of ionic strength and acetonitrile as co-solvent using TSK Gel Amide-80 column. The correspondence between standards and species in the mixtures was also implemented during the resolution. Correct application of the proposed model and resolution of the 12 metabolites was finally checked by comparison of the 12 MCR-ALS resolved spectra profiles with the spectra of the known metabolites. In all cases (except for F2,6BP) a perfect agreement was achieved. More detailed explanations about the MCR-ALS procedure and constraints are given in previous publications.[33,38] In addition, details of the strategy used for resolution of the 18 augmented chromatographic systems, one for each combination of the three pH values, three ionic strength values, and of the two organic co-solvents are given in the ESI, Table S2.†

2.4.3. Evaluation of chromatographic performance. Resolved elution profiles obtained by MCR-ALS permitted to retrieve peak retention times and peak widths for the 11 completely resolved metabolites in each HILIC stationary phase and for each chromatographic run at different experimental conditions. Using this information, an investigation about the performance of the different stationary phases and of the behavior of the various chromatographic conditions can be achieved. The use of Berridge chromatographic response function (CRF) is proposed to summarize these results (see below).

2.4.3.1. Investigation of HILIC stationary phases behavior in LC-DAD. Differences in retention time values obtained for every metabolite in each chromatographic run were investigated. The data matrix, $\mathbf{D}_{\text{tr}}$, was built up with the retention times for each metabolite at each chromatographic condition. This matrix had a number of rows equal to the number of chromatographic conditions (90 runs) and a number of columns equal to the number of resolved metabolites (11 compounds). Before the analysis, all retention times were normalized using the median retention time as a reference for all compounds in a particular chromatographic run. This retention times data matrix was

subjected to statistical evaluation using two different methods: ANOVA-simultaneous component analysis (ASCA)[39] and principal component analysis (PCA).[40] The ASCA method combines the advantages of ANOVA and simultaneous component analysis (SCA, a method similar to principal component analysis). In ASCA, the retention times data matrix ($\mathbf{D}_{tr}$) was split into the different effect data matrices containing the level averages for each factor, and the interaction data matrices describing the possible interactions between the considered factors. In this work, an ANOVA model considering four factors and their two-way interactions was considered as is shown in eqn (3):

$$\mathbf{D} = \overline{\mathbf{D}} + \mathbf{D}_{SP} + \mathbf{D}_{pH} + \mathbf{D}_{I} + \mathbf{D}_{org} + \mathbf{D}_{SP-pH} + \mathbf{D}_{SP-I} + \mathbf{D}_{SP-org} + \mathbf{D}_{pH-I} + \mathbf{D}_{pH-org} + \mathbf{D}_{I-org} + \mathbf{D}_{residual} \quad (3)$$

In this equation, $\mathbf{D}$ is the retention times raw data matrix, $\overline{\mathbf{D}}$ is the grand mean data matrix, $\mathbf{D}_{SP}$ is the matrix of effects of the different stationary phases, $\mathbf{D}_{pH}$ is the matrix of effects of pH values, $\mathbf{D}_{I}$ is the matrix of effects of ionic strength, $\mathbf{D}_{org}$ is the matrix of effects of organic co-solvent. $\mathbf{D}_{SP-pH}$, $\mathbf{D}_{SP-I}$ and $\mathbf{D}_{SP-org}$ are the interaction matrices of the different stationary phases with pH, ionic strength and organic co-solvent, respectively. $\mathbf{D}_{pH-I}$ and $\mathbf{D}_{pH-org}$ corresponds to the interactions between pH and ionic strength and the organic co-solvent, whereas $\mathbf{D}_{I-org}$ is the interaction between ionic strength and the organic co-solvent. Finally, $\mathbf{D}_{residual}$ represents residuals not explained by the ANOVA model.

One of the advantages of ASCA[39] is that allows estimating the statistical significance of each individual factor and of their interactions by using a permutation test.[41] Therefore, the null hypothesis H_0 of no experimental effect can be tested against the alternative hypothesis of the presence of experimental effects due to the factors and interactions with a confidence level of the *p*-value. In this work, the number of permutations used in ASCA had been set to 10 000. More details about the ASCA method can be found elsewhere.[42,43]

In addition, apart from the application of the ASCA method, PCA was also performed on the whole data matrix $\mathbf{D}_{tr}$, in order to identify relationships between the various chromatographic conditions. This method compresses the information of the variables into a smaller number of non-correlated variables known as principal components. These principal components are built as linear combinations of the original variables, retaining most of the valuable information about the experimental data variance without overlapping information among them by the application of orthogonality constraints.[40]

2.4.3.2. Evaluation of the optimal chromatographic conditions (HILIC stationary phases). Finally, the Berridge CRF was used to evaluate the quality of the chromatographic separation of the investigated metabolites at each one of the considered experimental conditions. As stated above, MCR-ALS analysis of the 18 different chromatographic data sets enabled the resolution of the elution profiles for every metabolite in each stationary phase at the diverse experimental conditions (90 different conditions). Therefore, chromatographic parameters, such as retention time and peak widths were calculated for every case. Taking account all this information, the overall quality of the individual chromatographic runs (in one particular HILIC stationary phase with one particular mobile phase composition) could be assessed using the Berridge CRF defined as (eqn (4)):[29]

$$\text{Berridge CRF} = \sum_{i=1}^{N-1} R_i + N^{\propto} - \beta|t_M - t_L| - \gamma|t_0 - t_1| \quad (4)$$

where R_i is the resolution among a pair of consecutive peaks, N is the total number of metabolites detected in the chromatogram, t_M, t_L, t_1 and t_0 are the maximum acceptable time, the experimental retention time of the last and first peak, and the minimum retention time of the first peak, respectively. Finally, α, β, and γ are arbitrary weighting parameters (usually set to values between 0 and 3).

Berridge CRF values for the considered chromatographic conditions were evaluated to detect the settings that provided better separation conditions of investigated metabolites. In addition, effects of the experimental factors used in the experimental design (stationary phase, pH, ionic strength and organic co-solvent) were also statistically assessed by an ANOVA analysis of the obtained Berridge CRF values also using the model given in eqn (3).

2.4.4. Software. Experimental design was performed using ChemStation B.04.01 LC 1200 series (Agilent Technologies, Waldbronn, Germany). Most of the calculations and data analysis were performed under MATLAB R2014a (The Mathworks Inc. Natick, MA, USA). PLS Toolbox 8.1 (Eigenvector Research Inc., Wenatchee, WA, USA) was used for generation of the full factorial experimental design, PCA and ASCA calculations. MCR-ALS toolbox[32] was used for MCR-ALS analysis of the chromatographic experiments under different experimental conditions and the resolution of the chromatographic elution and spectra profiles of the 12 metabolites in each case.

3. Results and discussion

3.1. Preliminary analysis and MCR-ALS resolution of LC-DAD chromatograms

In Fig. 1, the chromatograms at 270 nm of the mixture of the 12 metabolites at the different stationary phases and experimental conditions (organic co-solvents (acetonitrile and methanol), pH values (acid, moderately acid, and neutral), and ionic strengths (low, medium, and high)) are given. The whole set of different investigated conditions following a full factorial statistical design with the four factors at their various levels is detailed in the ESI Table S1.† Fig. 1A and B give the chromatograms at 270 nm using acetonitrile and methanol as organic co-solvents, respectively.

Chromatograms in Fig. 1A and B differ significantly upon variations of the organic co-solvent because the retention mechanism was drastically modified using these two mobile phases (methanol and acetonitrile present very different elution strengths). In some cases, peaks were distorted, especially in the case of using methanol as organic co-solvent. According to the HILIC separation mechanism, methanol could appear not to be appropriated in HILIC separations. In this work, the use of methanol was further investigated taking the advantage of

 This journal is © The Royal Society of Chemistry 2017

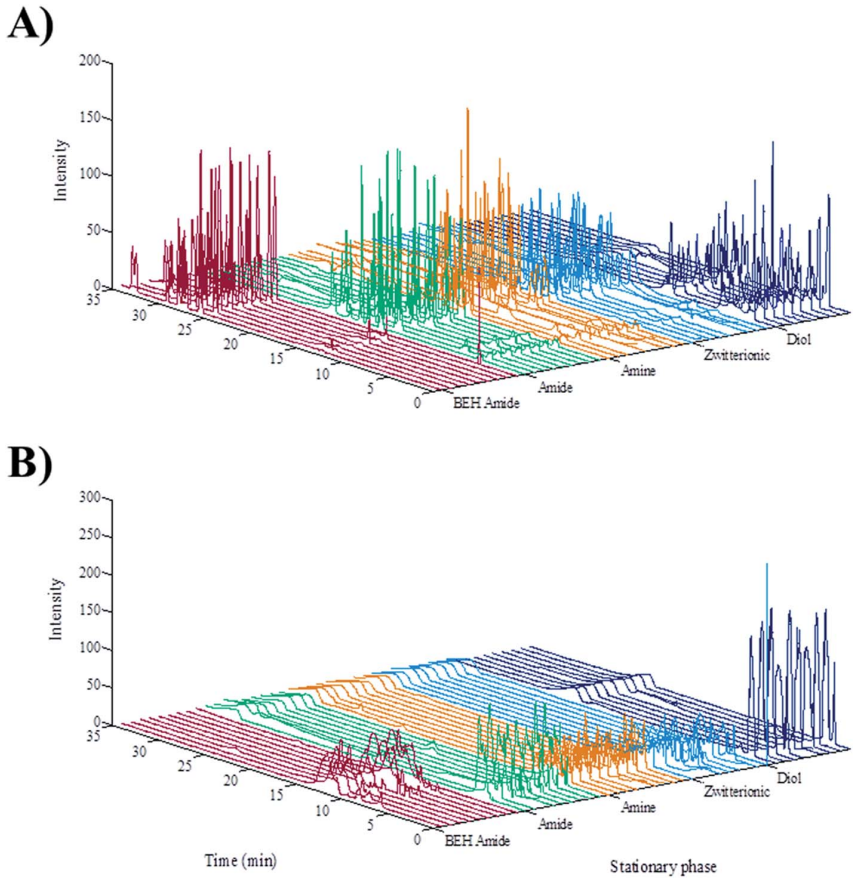


Fig. 1 Chromatograms of model mixture using: (A) acetonitrile and (B) methanol as organic co-solvent in the studied LC-DAD separations. All chromatograms collected in the different experimental conditions are shown grouped according to the HILIC stationary phase: BEH amide (red), amide (green), amine (orange), zwitterionic (cyan) and diol (blue). Within group variation is caused by the different pH and ionic strength conditions used in the different chromatographic runs.

using chemometric enhanced resolution to improve the obtained chromatographic resolution. However, results showed that acetonitrile provides better chromatographic separation both given chromatographic and chemometric resolution.

When the chromatographic results were compared considering the distinct HILIC stationary phases, the chromatographic behavior of the diol linked surface stationary phase was clearly different from the behavior of the other stationary phases. Peaks eluted at much lower retention times in both organic co-solvents. In contrast, the other stationary phases (amide, amine and zwitterionic) presented more similar profiles and were slightly affected by pH and ionic strength factors, giving only small peak intensity changes. The visual observation of these chromatographic differences already gave a preliminary insight of the effects of the different studied factors.

A deeper study of the performance of the various chromatographic conditions would require the evaluation of the behavior of the elution profiles of each metabolite in each stationary phase and at the diverse experimental conditions. Due to the overlapped metabolite signals in both chromatographic and spectral modes, chemometric resolution of the pure elution and spectral profiles of each metabolite in the different chromatographic runs at the different experimental conditions provided a means for the evaluation of the investigated experimental factors.

As an example, Fig. 2 shows the MCR-ALS results obtained in the case of the mobile phase conditions at pH 5.5, ionic strength 50 mM and acetonitrile as organic co-solvent (number 11 in ESI Table S2†) for the five investigated HILIC stationary phases. Again, the different behavior of the diol stationary phase is evident (Fig. 2D). As it is shown in Fig. 1, diol stationary phase appeared to be the less selective stationary phase giving two very separate peak regions, most of them with rather low signal intensities. The other stationary phases behaved more similarly and with better peak separation. Chromatograms resolved for BEH amide (Fig. 2A), zwitterionic (Fig. 2B) and amide (Fig. 2E) stationary phases were rather similar whereas the amine stationary phase (Fig. 2C) suggested a different elution behavior, in agreement with previous reports.[25] In the case of BEH amide column, analyses could have been performed using a higher flow rate to avoid extensively long metabolite retention. In this way, the difference of internal diameter between this column (4.6 mm) and the other four columns (2.1 mm) studied would be better to have identical

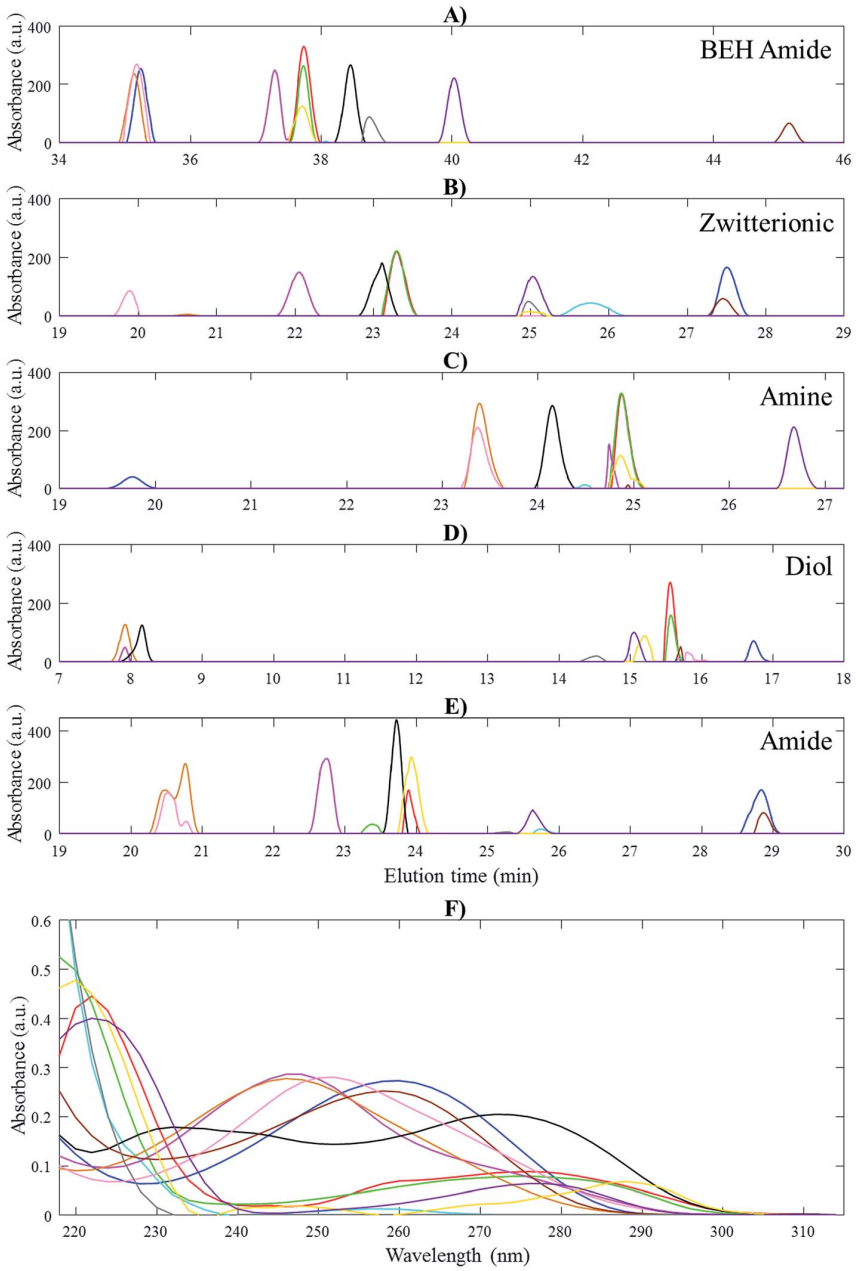


Fig. 2 MCR-ALS resolved profiles for the different chromatographic runs on the five different stationary phases (A) BEH amide, (B) zwitterionic, (C) amine, (D) diol, and (E) amide; at pH 5.5, 50 mM ionic strength and acetonitrile as organic co-solvent chromatographic condition (number 11 in Table S2†). (A–E) Elution profiles obtained for the different stationary phases and (F) MCR-ALS resolved UV spectra. Different metabolites are identified by the following lines legend: uridine (blue), L-tryptophan (red), L-tyrosine (green), L-phenylalanine (cyan), inosine (magenta), hypoxanthine (orange), cytidine (black), NADH (brown), AMP (pink), L-citrulline (gray), F2,6BP (yellow) and reduced glutathione (violet).

linear velocity in all experimental conditions. When chromatograms obtained at the other experimental conditions were analyzed, similar conclusions were drawn. Fig. 2F shows the MCR-ALS UV spectra profiles resolved for the 11 detected metabolites in this particular case (acetonitrile as organic co-solvent of the mobile phase conditions and pH 5.5 and ionic strength of 50 mM). Again, their strong overlapping compels the possibility of selection of a unique wavelength for their separation and analysis, and therefore it requires the use of multivariate (multiwavelength DAD detection) chemometric resolution methods, such as MCR-ALS.

Using the information obtained by MCR-ALS analysis of the 18 augmented data matrices (at the different analytical conditions of the experimental design), the resolved elution and

 This journal is © The Royal Society of Chemistry 2017

spectra profiles obtained in each case were compared. Elution profiles of the 11 metabolites (all except F2,6BP) were identified by their corresponding MCR-ALS resolved spectra. F2,6BP could not be well resolved because its UV absorption is at very low wavelengths (below 220 nm), significantly overlapped with the mobile phase and background absorption (acetonitrile and methanol absorb at the same wavelengths (210–220 nm)), which hindered its proper resolution. In the case of NADH metabolite, the resolution of its elution profile was not possible at some of the experimental conditions due to its low intensity signals and, therefore, its retention time values could not be calculated. Finally, the data matrix with all retention times (Table S3†) containing information about the position and width of the elution profiles of the 11 metabolites for the 90 studied samples was obtained to be further investigated using chemometric methods.

3.2. Study of the behavior of HILIC stationary phases

As already said above, the performance of the different chromatographic conditions was assessed using the table of retention times of the peak maxima of the MCR-ALS resolved elution profiles, for each metabolite at every condition of the experimental design. Peak maxima retention times of the considered metabolites were normalized using the median retention time as a reference for all metabolites in each individual chromatographic condition. In this way, the effect of elution gradient variations was eliminated.

Statistical significance of the four considered experimental factors (stationary phase, organic co-solvent, pH and ionic strength) as well as two-way interactions (stationary phase–pH, stationary phase–ionic strength, stationary phase–organic co-solvent, pH–ionic strength, pH–organic co-solvent, ionic strength–organic co-solvent) was assessed using the ASCA method[39] and permutation test[42] described in the method section. Results showed that only the effects of the type of stationary phase and organic co-solvent were statistically significant (p-value of 0.0001), whereas pH and ionic strength factors returned p-values equal to 0.7 and 0.2, respectively. Moreover, only the interaction between the type of HILIC stationary phase and the organic co-solvent was statistically significant (p-value of 0.0001), whereas all the other two-way interaction combinations gave p-values higher than 0.15. Hence, from ASCA statistical test, the two factors that seemed to be relevant to differentiate between the chromatographic behavior of the considered metabolites were the stationary phase and the organic co-solvent.

Subsequent interpretation of SCA scores of the matrix related with the organic co-solvent (at mean levels) (Fig. 3A) shows differences between samples analyzed with acetonitrile and methanol co-solvents. In addition, the principal component scores (at mean levels) for each considered HILIC stationary phase (Fig. 3B) revealed some trends. PC1 clearly distinguished the diol linked surface (Acclaim™ Mixed-Mode HILIC-1 column) stationary phase with a very large and negative PC1 score value. In contrast, the two amides and the zwitterionic stationary phases had similar positive PC1 score values, whereas amine surface (Luna-NH_2 column) stationary phase presented an intermediate score value. PC2 confirmed the differentiation of the diol surface stationary phase but, in this case, amine surface behaved more similarly to zwitterionic and amide stationary phases.

On the other side, SCA scores of the matrix of interactions between stationary phases and organic co-solvents confirmed the information described above. PC2 *vs.* PC1 scores scatter plot, when the organic co-solvent was considered as a factor (Fig. 3C), differentiates between methanol (colored red) and acetonitrile organic co-solvent mobile phase samples (colored blue). In contrast, when considering the stationary phase as a factor (Fig. 3D), PC1 scores shows some trends for amine, zwitterionic and amide (both columns) stationary phase samples, from left to right. However, only diol HILIC stationary phase samples were clearly distinguished from the other HILIC stationary phases.

In order to confirm and complete these results, a complementary PCA exploration of the whole chromatographic data set was carried out. Results confirmed that the most important factor was the organic co-solvent. Scores plot in Fig. S1A† shows a clear differentiation between acetonitrile and methanol samples. Again, the strange behavior of the set of samples corresponding to the diol stationary phase is clearly distinguished.

When samples measured with the amine stationary phase at neutral pH were excluded, and PCA was applied to the rest of samples analyzed with methanol, samples did not show any cluster (Fig. S1B†). In contrast, in the case of acetonitrile samples, diol stationary phase samples were forming a cluster clearly separated from the rest of samples (Fig. S1C†).

Furthermore, PCA was also applied to all samples except diol stationary phase samples. Fig. S1D† reveals the presence of three clear groups of samples related respectively with zwitterionic, amine and BEH amide (XBridge™ amide) HILIC stationary phases. In contrast, samples measured using amide (TSK Gel Amide-80) stationary phase showed an average behavior between the one obtained for samples measured using amine (basic) and zwitterionic (acid) stationary phases.

3.3. Chromatographic response function results evaluation

The quality of the separations achieved using the different chromatographic conditions was assessed using the CRF based on the Berridge model.[29] Calculation of the Berridge CRF values for each condition was carried out setting the α, β and γ parameters to 1, 0.5 and 0.1, respectively, as recommended in the literature.[44] An ANOVA model was built on the logarithm (base 10) of the CRF values obtained in the evaluation of every experimental chromatographic condition. HILIC stationary phase and organic co-solvent were confirmed again to be statistically significant factors (p-values lower than 0.0001 in both cases). Ionic strength had a large p-value (0.4), and it was considered not having a significant effect. On the other hand, the effect of pH resulted in being more important in this case with a p-value lower than 0.1. Finally, the effect of the interactions between stationary phase–ionic strength (p-value < 0.0001)

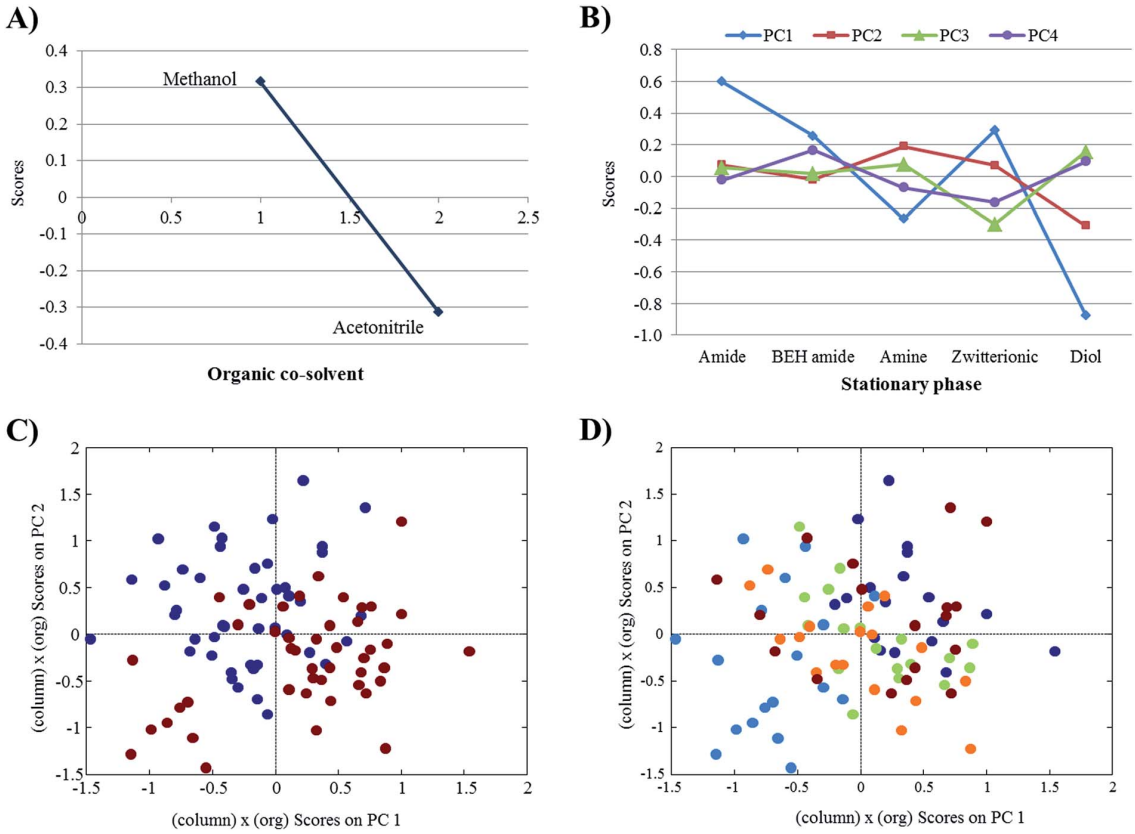


Fig. 3 Summary of ASCA results on the matrix of the retention times of the peak maxima of the elution profiles of all 11 metabolites at all 90 experimental conditions resolved by MCR-ALS. (A) PC1 scores plot for the organic co-solvent (mobile phase) factor matrix; (B) first four principal components score plot for each HILIC stationary phase at the different experimental conditions. SCA scores of the interaction matrix: (C) PC1 *vs.* PC2 ASCA scores plot for HILIC stationary phase factor colored according solvent co-solvent (acetonitrile samples in blue and methanol samples in red); and (D) PC2 *vs.* PC1 ASCA plot for HILIC stationary phase factor colored according to the HILIC stationary phase: BEH amide (red), amide (blue), amine (orange), diol (cyan) and zwitterionic (green).

and the stationary phase–organic co-solvent (p-value < 0.0001) resulted in being statistically significant (see Table 2).

Fig. 4 summarizes the Berridge CRF values obtained for the different chromatographic conditions, considering only the stationary phase and the organic phase co-solvent as a factor. According to the organic co-solvent, it can be seen that, in general, CRF values obtained using acetonitrile were slightly higher than those obtained with methanol (Fig. 4A and B), which confirmed again that acetonitrile seemed the best choice for the chromatographic analysis of metabolites using the selected HILIC columns.[19] In the case of the selection of the different stationary phases (Fig. 4C and D), the trends were not so clear, but in general, they were in agreement to those previously reported in the literature.[13,25,45,46] These works showed that amide-bonded and zwitterionic stationary phases gave a better retention of the metabolites compared with those provided by amino and diol stationary phases. Amide and zwitterionic stationary phases provided the highest CRF values, which imply that they would be the best options for metabolomics studies. In general, the best results were obtained when using acetonitrile as organic co-solvent whereas pH and ionic strength did not provide a clear and noticeable trend about the best experimental conditions to work with, probably because they are very much specific metabolite-dependent and not general. In the case of the two amide stationary phases, differences in the conditions were observed linked to the different surface chemistry of these two HILIC stationary phases. For instance, in

Table 2 ANOVA results for the analysis of CRF values

Factor	F	p-Value	Significant
Global model	10.30	<0.0001	Yes
Stationary phase	12.78	<0.0001	Yes
pH	2.51	0.0909	No
Ionic strength	0.87	0.4251	No
Organic co-solvent	22.78	<0.0001	Yes
Stationary phase × pH	0.49	0.8581	No
Stationary phase × ionic strength	5.30	<0.0001	Yes
Stationary phase × organic co-solvent	62.03	<0.0001	Yes
pH × ionic strength	0.44	0.7800	No
pH × organic co-solvent	0.28	0.7555	No
Ionic strength × organic co-solvent	1.81	0.1744	No

This journal is © The Royal Society of Chemistry 2017

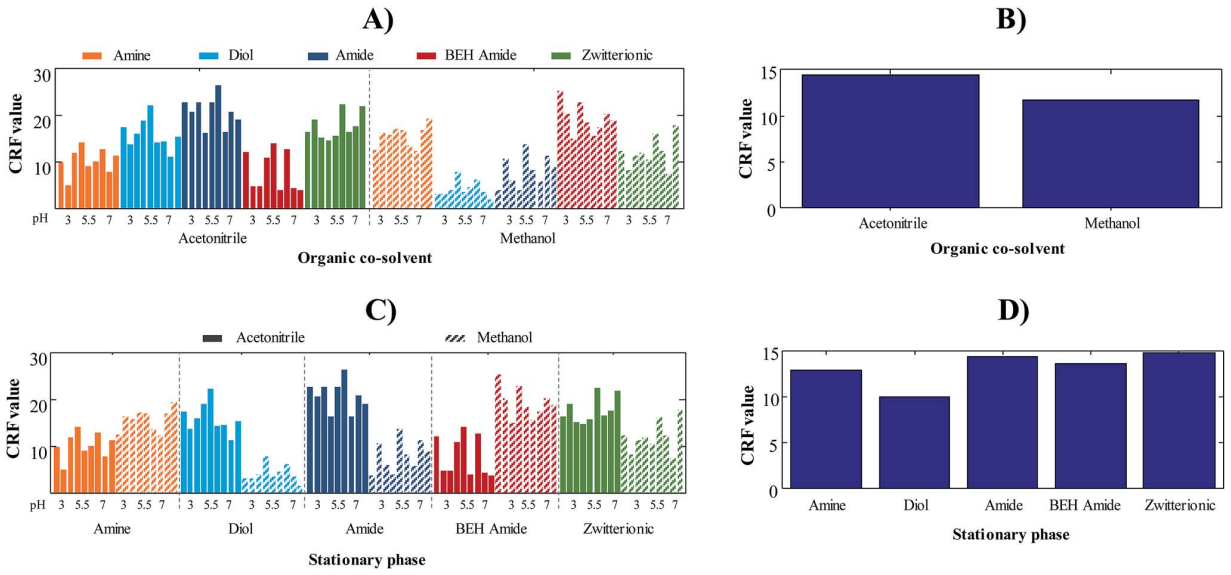


Fig. 4 Graphical summarized representation of CRF results for the 90 chromatographic conditions investigated in this work considering the experiments according to: the organic co-solvent using (A) the individual CRF values and (B) the average value for each solvent; the different stationary phases using (C) the CRF values and (D) the average value. In (A) and (C) the stationary phases are represented in different colors and organic co-solvent using different color shape: acetonitrile (plain color) and methanol (striped color). Within each group, samples are ordered according to increasing pH value (three experiments for each pH). The three conditions in the same pH are arranged agree with ionic strength (from 5.0 to 50.0 mM).

the case of TSK Gel Amide-80, the largest CRF values were obtained for the experiments using acetonitrile as organic co-solvent. In contrast, when XBridge™ Amide column was used, this trend was not so clear, and methanol as organic co-solvent provided similar results to those using acetonitrile. Amine and, especially, diol stationary phases gave lower CRF values which hinder their general application in metabolomics studies.

Results obtained using the proposed CRF approach were in agreement with those previously obtained by PCA. Similar behavior was detected for amide and zwitterionic HILIC stationary phases (samples analyzed in these two stationary phases were clustered in dense groups) whereas diol and amine stationary phases showed much larger variability.

4. Conclusions

In this work, the application of the MCR-ALS chemometric method to the analysis of highly complex HILIC chromatographic data have helped to get a deeper insight in the obtained separation.

ASCA results demonstrated that the main factors affecting chromatographic separation and determination of metabolite mixtures were the HILIC stationary phase and the organic phase co-solvent. In addition, the interaction between these two factors was also statistically significant. The optimization of Berridge chromatographic response function, CRF, allowed the evaluation of the separation performance achieved in the different experimental chromatographic conditions after MCR-ALS peak profiles resolution. As a conclusion, in the analysis of the investigated metabolite mixture, the two amide and zwitterionic HILIC stationary phases gave the highest (the best) CRF values. Amine stationary phase showed acceptable intermediate results, whereas diol stationary phase provided worse separation of the metabolites (lower CRF values) (especially when methanol was used as organic co-solvent of the mobile phase). Additionally, acetonitrile mobile phase gave greater results from a chromatographic separation point of view. Although, MCR-ALS resolution resulted in being especially useful in the case of methanol samples where the quality of the chromatographic separation (resolution) was lower giving improved CRF values, closer to these ones obtained using acetonitrile. CRF criteria took into account not only the chromatographic resolution also the chemometrics enhanced resolution, which was precisely the goal of the study.

From the results obtained in this work, amide and zwitterionic columns showed a similar behavior and they are confirmed to be the best choice for metabolomics studies regarding retention and hydrophilic selectivity using HILIC chromatographic columns, especially when only polar metabolites were considered. In contrast, studied diol stationary phase showed a fair performance (poor retention) when dealing with the set of selected strongly polar compounds because it is more hydrophobic than the other HILIC stationary phases. However, diol stationary phase could be considered an option when both polar and non-polar metabolites need to be simultaneously analyzed (for instance, in lipidomics studies). Acetonitrile remained the logical option to be used as the organic co-solvent, since in all cases provided better separations than methanol from both chromatographic and chemometric criteria. In the case of pH and ionic strength, both factors

This journal is © The Royal Society of Chemistry 2017

appeared to have a minor effect in the chromatographic analysis when compared with the effects of the stationary phase and the organic co-solvent.

Further studies are needed to confirm and generalize the robustness of the obtained results, the validity of the used chemometric approach and the conclusions derived from them. In these new studies, designed to gather more relevant information in the metabolomic field, the main priority will be given to select optimal chromatographic conditions for each one of the considered stationary phases, select a larger set of metabolites and to employ mass spectrometry as detection technique.

Conflict of interest

The authors declare that they have no competing interests.

Abbreviations

ASCA	ANOVA-simultaneous component analysis
CRF	Chromatographic response function
HILIC	Hydrophilic interaction liquid chromatography
IPC	Ion-pair chromatography
MCR	Multivariate curve resolution
ALS	Alternating least squares
NP	Normal phase
PCA	Principal component analysis

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. Research funding from MINECO Spain Grant No. CTQ2015-66254-C2-1-P is also acknowledged. The authors are grateful to Ms Marta Fontal from IDAEA-CSIC (Barcelona, Spain) for technical assistance in the use of the LC-DAD system.

References

1 J. K. Nicholson and J. C. Lindon, *Systems Biology: Metabonomics Nature*, 2008, **455**, 1054–1056.
2 S. G. Villas-Bôas, S. Mas, M. Åkesson, J. Smedsgaard and J. Nielsen, Mass spectrometry in metabolome analysis, *Mass Spectrom. Rev.*, 2005, **24**, 613–646.
3 S. Wernisch and S. Pennathur, Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics, *Anal. Bioanal. Chem.*, 2016, 1–13.
4 M. Cao, K. Fraser, J. Huege, T. Featonby, S. Rasmussen and C. Jones, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, *Metabolomics*, 2015, **11**, 696–706.
5 U. Harder, B. Koletzko and W. Peissner, Quantification of 22 plasma amino acids combining derivatization and ion-pair LC-MS/MS, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2011, **879**, 495–504.
6 P. Hemström and K. Irgum, Hydrophilic interaction chromatography, *J. Sep. Sci.*, 2006, **29**, 1784–1821.
7 I. Sampsonidis, M. Witting, W. Koch, C. Virgiliou, H. G. Gika, P. Schmitt-Kopplin and G. A. Theodoridis, Computational analysis and ratiometric comparison approaches aimed to assist column selection in hydrophilic interaction liquid chromatography-tandem mass spectrometry targeted metabolomics, *J. Chromatogr. A*, 2015, **1406**, 145–155.
8 G. Marrubini, A. Pedrali, P. Hemström, T. Jonsson, P. Appelblad and G. Massolini, Column comparison and method development for the analysis of short-chain carboxylic acids by zwitterionic hydrophilic interaction liquid chromatography with UV detection, *J. Sep. Sci.*, 2013, **36**, 3493–3502.
9 Q. Fu, T. Liang, Z. Li, X. Xu, Y. Ke, Y. Jin and X. Liang, Separation of carbohydrates using hydrophilic interaction liquid chromatography, *Carbohydr. Res.*, 2013, **379**, 13–17.
10 A. J. Alpert, M. Shukla, A. K. Shukla, L. R. Zieske, S. W. Yuen, M. A. J. Ferguson, A. Mehlert, M. Pauly and R. Orlando, Hydrophilic-interaction chromatography of complex carbohydrates, *J. Chromatogr. A*, 1994, **676**, 191–202.
11 M. De Person, P. Chaimbault and C. Elfakir, Analysis of native amino acids by liquid chromatography/electrospray ionization mass spectrometry: comparative study between two sources and interfaces, *J. Mass Spectrom.*, 2008, **43**, 204–215.
12 D. García-Gómez, E. Rodríguez-Gonzalo and R. Carabias-Martínez, Stationary phases for separation of nucleosides and nucleotides by hydrophilic interaction liquid chromatography, *TrAC, Trends Anal. Chem.*, 2013, **47**, 111–128.
13 S. Van Dorpe, V. Vergote, A. Pezeshki, C. Burvenich, K. Peremans and B. De Spiegeleer, Hydrophilic interaction LC of peptides: columns comparison and clustering, *J. Sep. Sci.*, 2010, **33**, 728–739.
14 B. Buszewski and S. Noga, Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique, *Anal. Bioanal. Chem.*, 2011, **402**, 231–247.
15 P. Jandera, Stationary and mobile phases in hydrophilic interaction chromatography: a review, *Anal. Chim. Acta*, 2011, **692**, 1–25.
16 R.-I. Chirita, C. West, A.-L. Finaru and C. Elfakir, Approach to hydrophilic interaction chromatography column selection: application to neurotransmitters analysis, *J. Chromatogr. A*, 2010, **1217**, 3091–3104.
17 T. Ikegami, K. Tomomatsu, H. Takubo, K. Horie and N. Tanaka, Separation efficiencies in hydrophilic interaction chromatography, *J. Chromatogr. A*, 2008, **1184**, 474–503.
18 H. Gika, G. Theodoridis, F. Mattivi, U. Vrhovsek and A. Pappa-Louisi, Hydrophilic interaction ultra performance liquid chromatography retention prediction under gradient elution, *Anal. Bioanal. Chem.*, 2012, **404**, 701–709.
19 A. Periat, B. Debrus, S. Rudaz and D. Guillarme, Screening of the most relevant parameters for method development in

 This journal is © The Royal Society of Chemistry 2017

ultra-high performance hydrophilic interaction chromatography, *J. Chromatogr. A*, 2013, **1282**, 72–83.

20 E. Tyteca, A. Périat, S. Rudaz, G. Desmet and D. Guillarme, Retention modeling and method development in hydrophilic interaction chromatography, *J. Chromatogr. A*, 2014, **1337**, 116–127.

21 N. P. Dinh, T. Jonsson and K. Irgum, Probing the interaction mode in hydrophilic interaction chromatography, *J. Chromatogr. A*, 2011, **1218**, 5880–5891.

22 G. Greco and T. Letzel, Main interactions and influences of the chromatographic parameters in HILIC separations, *J. Chromatogr. Sci.*, 2013, **51**, 684–693.

23 T. Rakić, B. Jančić Stojanović, A. Malenović, D. Ivanović and M. Medenica, Improved chromatographic response function in HILIC analysis: application to mixture of antidepressants, *Talanta*, 2012, **98**, 54–61.

24 S. Noga, S. Bocian and B. Buszewski, Hydrophilic interaction liquid chromatography columns classification by effect of solvation and chemometric methods, *J. Chromatogr. A*, 2013, **1278**, 89–97.

25 Y. Kawachi, T. Ikegami, H. Takubo, Y. Ikegami, M. Miyamoto and N. Tanaka, Chromatographic characterization of hydrophilic interaction liquid chromatography stationary phases: hydrophilicity, charge effects, structural selectivity, and separation efficiency, *J. Chromatogr. A*, 2011, **1218**, 5903–5919.

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 J. M. Amigo, T. Skov and R. Bro, ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics, *Chem. Rev.*, 2010, **110**, 4582–4605.

28 E. Gorrochategui, J. Jaumot, S. Lacorte and R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow, *TrAC, Trends Anal. Chem.*, 2016, **82**, 425–442.

29 J. C. Berridge, Unattended optimisation of reversed-phase high-performance liquid chromatographic separations using the modified simplex algorithm, *J. Chromatogr. A*, 1982, **244**, 1–14.

30 D. V. McCalley, Is hydrophilic interaction chromatography with silica columns a viable alternative to reversed-phase liquid chromatography for the analysis of ionisable compounds?, *J. Chromatogr. A*, 2007, **1171**, 46–55.

31 H. F. M. Boelens, R. J. Dijkstra, P. H. C. Eilers, F. Fitzpatrick and J. A. Westerhuis, New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection, *J. Chromatogr. A*, 2004, **1057**, 21–30.

32 J. Jaumot, R. Gargallo, A. de Juan and R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, *Chemom. Intell. Lab. Syst.*, 2005, **76**, 101–110.

33 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.

34 S. Mas, G. Fonrodona, R. Tauler and J. Barbosa, Determination of phenolic acids in strawberry samples by means of fast liquid chromatography and multivariate curve resolution methods, *Talanta*, 2007, **71**, 1455–1463.

35 R. Tauler and D. Barceló, Multivariate curve resolution applied to liquid chromatography-diode array detection, *TrAC, Trends Anal. Chem.*, 1993, **12**, 319–327.

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

37 R. Tauler, A. Smilde and B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *J. Chemom.*, 1995, **9**, 31–58.

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

39 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.

40 S. Wold, K. Esbensen and P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37–52.

41 G. Zwanenburg, H. C. Hoefsloot, J. A. Westerhuis, J. J. Jansen and A. K. Smilde, ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison, *J. Chemom.*, 2011, **25**, 561–567.

42 A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R. J. A. N. Lamers, J. van der Greef and M. E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, *Bioinformatics*, 2005, **21**, 3043–3048.

43 M. Farrés, M. Villagrasa, E. Eljarrat, D. Barceló and R. Tauler, Chemometric evaluation of different experimental conditions on wheat (*Triticum aestivum* L.) development using liquid chromatography mass spectrometry (LC-MS) profiles of benzoxazinone derivatives, *Anal. Chim. Acta*, 2012, **731**, 24–31.

44 R. M. B. O. Duarte and A. C. Duarte, A new chromatographic response function for use in size-exclusion chromatography optimization strategies: application to complex organic mixtures, *J. Chromatogr. A*, 2010, **1217**, 7556–7563.

45 A. Kumar, J. C. Heaton and D. V. McCalley, Practical investigation of the factors that affect the selectivity in hydrophilic interaction chromatography, *J. Chromatogr. A*, 2013, **1276**, 33–46.

46 G. Schuster and W. Lindner, Comparative characterization of hydrophilic interaction liquid chromatography columns by linear solvation energy relationships, *J. Chromatogr. A*, 2013, **1273**, 73–94.

Informació suplementària de l'article científic I.

Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites.

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

Analytical Methods 9 (2017) 774-785.

Table S1. Full factorial experimental design carried out in this work.

Experimental design				
Exp. ID	**Column**	**pH**	**Ionic strength (mM)**	**Organic co-solvent**
1	A	3	5	Acetonitrile
2	A	3	25	Acetonitrile
3	A	3	50	Acetonitrile
4	A	3	5	Methanol
5	A	3	25	Methanol
6	A	3	50	Methanol
7	A	5.5	5	Acetonitrile
8	A	5.5	25	Acetonitrile
9	A	5.5	50	Acetonitrile
10	A	5.5	5	Methanol
11	A	5.5	25	Methanol
12	A	5.5	50	Methanol
13	A	7	5	Acetonitrile
14	A	7	25	Acetonitrile
15	A	7	50	Acetonitrile
16	A	7	5	Methanol
17	A	7	25	Methanol
18	A	7	50	Methanol
19	B	3	5	Acetonitrile
20	B	3	25	Acetonitrile
21	B	3	50	Acetonitrile
22	B	3	5	Methanol
23	B	3	25	Methanol
24	B	3	50	Methanol
25	B	5.5	5	Acetonitrile
26	B	5.5	25	Acetonitrile
27	B	5.5	50	Acetonitrile
28	B	5.5	5	Methanol
29	B	5.5	25	Methanol
30	B	5.5	50	Methanol
31	B	7	5	Acetonitrile
32	B	7	25	Acetonitrile
33	B	7	50	Acetonitrile
34	B	7	5	Methanol
35	B	7	25	Methanol
36	B	7	50	Methanol
37	C	3	5	Acetonitrile
38	C	3	25	Acetonitrile
39	C	3	50	Acetonitrile
40	C	3	5	Methanol
41	C	3	25	Methanol
42	C	3	50	Methanol
43	C	5.5	5	Acetonitrile
44	C	5.5	25	Acetonitrile
45	C	5.5	50	Acetonitrile
46	C	5.5	5	Methanol
47	C	5.5	25	Methanol
48	C	5.5	50	Methanol
49	C	7	5	Acetonitrile
50	C	7	25	Acetonitrile
51	C	7	50	Acetonitrile
52	C	7	5	Methanol
53	C	7	25	Methanol
54	C	7	50	Methanol
55	D	3	5	Acetonitrile
56	D	3	25	Acetonitrile
57	D	3	50	Acetonitrile
58	D	3	5	Methanol
59	D	3	25	Methanol
60	D	3	50	Methanol
61	D	5.5	5	Acetonitrile
62	D	5.5	25	Acetonitrile
63	D	5.5	50	Acetonitrile
64	D	5.5	5	Methanol
65	D	5.5	25	Methanol
66	D	5.5	50	Methanol
67	D	7	5	Acetonitrile
68	D	7	25	Acetonitrile
69	D	7	50	Acetonitrile
70	D	7	5	Methanol
71	D	7	25	Methanol
72	D	7	50	Methanol
73	E	3	5	Acetonitrile
74	E	3	25	Acetonitrile
75	E	3	50	Acetonitrile
76	E	3	5	Methanol
77	E	3	25	Methanol
78	E	3	50	Methanol
79	E	5.5	5	Acetonitrile
80	E	5.5	25	Acetonitrile
81	E	5.5	50	Acetonitrile
82	E	5.5	5	Methanol
83	E	5.5	25	Methanol
84	E	5.5	50	Methanol
85	E	7	5	Acetonitrile
86	E	7	25	Acetonitrile
87	E	7	50	Acetonitrile
88	E	7	5	Methanol
89	E	7	25	Methanol
90	E	7	50	Methanol

Column code as in Table 1 (A: XBridge™ Amide; B: TSK Gel Amide-80; C: Luna-NH_2; D: ZIC-HILIC; E: Acclaim™ Mixed-Mode HILIC-1). pH values: 3 - acid, 5.5 - moderately acid, 7 - neutral.

Table S2. MCR-ALS augmented matrices covering the different chromatographic conditions (same pH, ionic strength and organic co-solvent of the mobile phase) for the different stationary phases considered.

MCR-ALS analysis number*	Column		
	pH	Ionic strength (mM)	Organic co-solvent
1	3	5	Acetonitrile
2	3	5	Methanol
3	3	25	Acetonitrile
4	3	25	Methanol
5	3	50	Acetonitrile
6	3	50	Methanol
7	5.5	5	Acetonitrile
8	5.5	5	Methanol
9	5.5	25	Acetonitrile
10	5.5	25	Methanol
11	5.5	50	Acetonitrile
12	5.5	50	Methanol
13	7	5	Acetonitrile
14	7	5	Methanol
15	7	25	Acetonitrile
16	7	25	Methanol
17	7	50	Acetonitrile
18	7	50	Methanol

*Every MCR-ALS augmented data matrix contains experiments carried out using the five different HILIC columns in the same mobile phase conditions. For instance, MCR-ALS analysis number 1 contained experiments number 1 (XBridge™ Amide), 19 (TSK Gel Amide-80), 37 (Luna-NH_2), 55 (ZIC-HILIC) and 73 (Acclaim™ Mixed-Mode HILIC-1) of Table S1. All these runs were carried out using a mobile phase at pH 3.0 with an ionic strength of 5 mM and acetonitrile as organic co-solvent.

Table S3. Retention times of the MCR-ALS resolved elution profiles of the 11 metabolites for the 90 studied experimental conditions.

	Retention time (min)										
Exp. ID	**Uridine**	**L-tryptophan**	**L-tyrosine**	**Phenylalanine**	**Inosine**	**Hypoxanthine**	**Cytidine**	**NADH**	**AMP**	**L-citrulline**	**Red. Glutathione**
1	35.07	37.74	37.73	39.87	37.1	34.94	39.25	45.1	42.97	39.82	39.87
2	35.46	37.75	37.75	40.25	37.53	35.3	39.17	46.04	35.36	40.29	40.22
3	35.25	39.46	37.61	39.32	37.42	35.09	38.99	37.47	34.99	39.32	39.44
4	17.69	17.13	17.78	9.17	18.35	16.46	19.64	12.91	15.86	18.07	11.95
5	17.56	14.46	16.69	16.52	16.5	16.34	17.58	13.23	19.77	19.91	12.47
6	17.4	13.65	16.6	16.78	13.7	16.25	17.26	15.23	18.17	16.7	13.52
7	35.42	37.87	37.86	42.44	37.37	35.29	38.57	42.45	40.81	36.77	41.85
8	35.13	37.84	37.83	39.15	37.22	35.05	38.4	44.7	43.92	37.58	37.23
9	35.4	37.92	37.92	38.94	37.45	35.33	38.63	45.38	38.85	38.93	40.23
10	8.94	16.9	17.04	15.92	12.6	16.78	15.28	8.42	12.31	12.35	17.02
11	12.68	13.84	16.63	16.71	12.52	13.67	12.62	10.32	12.67	16.71	16.62
12	14.39	16.37	12.59	12.42	16.42	16.19	13.09	14.48	14.43	12.42	14.73
13	34.61	37.69	37.69	37.02	36.77	34.53	38	41.63	39.21	38.28	41.63
14	34.87	37.72	37.71	38.81	37.02	34.82	38.2	44.31	37.07	37.88	44.31
15	35.03	37.86	37.85	39.32	37.25	35.01	38.4	45.25	44.76	39.38	40.14
16	12.68	13.86	16.63	12.58	12.53	13.68	12.59	12.71	14.61	16.66	17.1
17	11.94	16.92	17.09	14.47	15.68	15.32	16.76	14.67	11.93	17.15	9.55
18	16.54	7.3	7.91	7.96	11.22	16.57	11.27	16.37	16.76	16.74	7.95
19	20.39	23.98	23.99	24.09	22.59	20.63	25.5	28.51	28.12	27.92	28.52
20	28.29	23.92	23.92	23.79	22.53	20.26	23.57	28.29	27.58	23.82	28.29
21	29.7	25.25	23.67	24.94	22.76	20.66	24.33	29.71	24.83	24.52	25.02
22	6.75	6.92	6.17	6.24	6.99	6.32	6.63	-	6.37	6.06	9.13
23	9.51	6.49	9.51	7.03	9.62	9.02	9.37	8.46	6.98	8.52	8.46
24	6.75	4.64	6.53	4.66	6.3	6.26	6.66	5.9	6.63	4.66	4.62
25	20.36	23.87	23.87	24.03	22.57	20.6	24.26	27.25	25.66	22.08	25.31
26	21.18	23.89	23.89	23.85	23.09	21.23	24.04	28.33	27.88	22.75	23.05
27	28.98	24.01	23.51	24.05	22.87	20.87	23.86	29.01	28.9	25.4	25.77
28	6.75	6.61	6.75	6	6.28	6.34	6.82	-	6.28	6.28	6.08

29	6.76	4.61	4.69	3.22	5.86	6.3	4.76	4.74	4.71	3.22	4.77
30	6.75	4.59	6.51	4.37	4.8	6.3	6.61	5.19	5.1	4.4	4.8
31	19.59	23.76	23.77	23.63	22.16	19.95	23.88	26.65	25.53	23.89	26.65
32	28.29	23.91	23.93	23.81	22.53	20.39	23.57	28.29	27.58	23.83	28.29
33	21.23	23.91	23.91	24.28	23.22	21.39	24.16	29.03	28.94	23.91	25.6
34	6.63	6.81	6.71	6.1	6.18	6.28	6.71	-	5.59	6.71	7.66
35	6.67	4.62	6.71	4.64	4.62	6.29	6.75	-	6.32	4.66	4.66
36	6.89	4.6	6.76	4.37	4.61	6.39	6.73	5.22	6.3	7.06	6.89
37	22.92	25.41	25.41	24.94	24.4	22.92	24.34	-	24.6	25.94	28.94
38	21.69	43.02	25.43	24.55	24.62	22.15	25.08	21.94	24.52	24.47	24.58
39	20.84	40.72	24.78	31.22	23.59	21.71	24.64	21.73	23.65	24.4	31.22
40	6.6	11.63	11.67	10.8	11.01	9.97	7.65	-	11.35	8.5	10.49
41	6.58	11.26	11.28	10.58	10.34	9.44	7.77	5.35	8.29	5.76	10.34
42	6.43	5.48	10.56	9.88	5.62	8.31	7.73	4.69	6.28	9.83	5.05
43	21.26	25.62	25.62	25.45	26.06	24.79	24.71	22.52	25.22	24.88	24.08
44	23.71	25.01	25.01	24.65	24.95	23.72	24.52	23.23	23.31	27.87	24.96
45	23.48	25	24.99	24.6	25.49	23.5	24.27	22.87	23.6	24.87	26.81
46	6.6	11.91	12.56	9.96	12.4	9.17	7.32	6.05	9.98	12.48	12.56
47	6.55	11.52	11.51	8.44	5.08	9.06	8.01	5.89	8.52	4.53	5.04
48	6.49	9.93	10.88	8.11	10.06	9.15	7.58	5.8	6.03	5.47	5.05
49	22.51	25.28	25.29	24.88	24.17	22.52	24.17	21.48	24.86	24.98	28.7
50	24.93	25.33	25.33	25.47	24.58	24.93	24.05	25.1	24.6	25.26	25.12
51	25.17	25.09	25.09	24.74	22.98	24.03	24.25	24.11	24.06	23.62	26.97
52	7.49	9.09	10.16	7.31	8.33	7.49	6.41	8.54	6.88	5.17	6.39
53	6.51	5.51	11.32	8.47	10.26	9.26	5.56	5.63	6.02	8.57	11.64
54	6.62	9.69	11.06	3.95	5.09	7.54	8.64	5.55	7.58	7.5	3.95
55	19.87	23.68	23.68	23.1	21.75	19.6	26.41	26.95	26.79	29.09	26.97
56	20.21	25.1	23.36	25.49	22.11	20.08	23.54	27.93	25.37	25.51	26.23
57	26.53	25.1	23.2	25.03	21.8	24.91	23.98	27.99	24.94	24.99	24.74
58	7.12	9.36	9.31	8.98	6.99	7.16	9.39	-	8.13	8.93	8.99
59	5.25	5.65	8.27	5.57	5.65	5.72	5.77	5.82	5.38	5.39	5.57
60	9.07	5.49	7.95	5.47	5.54	6.94	5.21	8.17	6.23	5.4	5.5
61	24.92	23.79	23.8	22.46	22.66	20.88	23.64	25.25	24.92	24	25.25

62	26.77	23.26	23.31	23.23	21.84	21.83	22.81	26.91	26.8	27.33	21.83
63	27.64	23.42	23.42	25.1	22.16	19.99	23.24	27.6	27.72	25.1	25.16
64	3.42	5.58	7.84	5.46	5.62	6.91	5.57	4.28	3.43	7.81	7.85
65	4.17	5.28	7.82	6.76	6.93	7.04	8.59	5.27	5.3	5.32	5.3
66	5.37	7.9	7.85	5.95	6.08	7.04	8.54	11.87	5.35	6.16	5.23
67	24.27	23.31	24.2	23.3	21.86	19.66	22.89	24.36	22.76	23.23	24.56
68	26.75	23.26	23.28	23.14	21.81	21.84	22.77	26.8	26.71	23.16	26.83
69	27.87	23.69	23.69	25.34	22.49	22.51	23.52	27.76	20.53	25.31	25.44
70	3.34	5.65	7.73	3.34	5.67	6.88	5.23	4.45	3.31	8.48	4.44
71	5.31	5.64	7.75	5.32	5.7	5.74	5.74	5.27	5.34	7.54	5.17
72	5.42	6.34	7.9	8.24	11.16	7.14	8.87	-	5.43	8.03	8.11
73	15.36	15.26	15.4	15.25	5.82	7.61	7.95	15.94	15.42	14.86	15.26
74	16.14	15.48	15.48	15.6	7.77	7.71	15.48	5.79	7.85	15.54	15.58
75	15.89	15.09	15.14	15.27	15.48	15.99	8.05	14.04	6.86	15.16	15.56
76	3.87	3.26	3.44	3.69	4.22	3.91	3.91	3.87	3.5	3.38	3.24
77	3.58	3.95	3.95	3.46	4.24	3.95	3.95	3.08	3.64	3.65	3.27
78	3.24	3.95	3.34	3.4	3.95	3.34	3.93	4.35	3.64	3.6	3.96
79	5.32	15.65	15.68	15.07	5.77	5.31	8.8	14.61	15.26	13.28	12.07
80	16.42	15.66	15.21	3.36	3.79	5.56	8.3	15.77	16.39	16.31	3.78
81	16.82	15.64	15.65	15.22	7.95	7.95	8.19	8	15.81	14.6	15.14
82	2.39	3.24	3.98	2.41	4.28	3.99	2.45	2.43	2.35	4.03	3.98
83	2.65	2.88	3.93	2.65	2.89	3.94	2.79	2.77	3.83	4	2.71
84	3.18	3.4	3.91	2.67	4.23	3.41	3.91	3.59	3.14	3.99	3.43
85	5.21	15.52	15.54	6.22	5.91	7.57	7.92	15.46	15.14	7.36	8.76
86	16.5	15.54	15.58	15	7.81	7.71	8.02	7.77	16.5	15.73	7.77
87	15.68	14.97	15.07	14.93	7.86	7.84	8.06	7.85	15.64	15.08	14.94
88	2.3	3.28	3.46	3.62	3.91	3.86	3.84	2.34	3.53	3.47	2.33
89	2.55	3.32	3.9	2.7	3.65	3.49	3.44	2.81	3.09	3.44	3.36
90	3.14	3.38	3.69	3.41	3.64	3.72	3.31	-	3.5	2.86	3.56

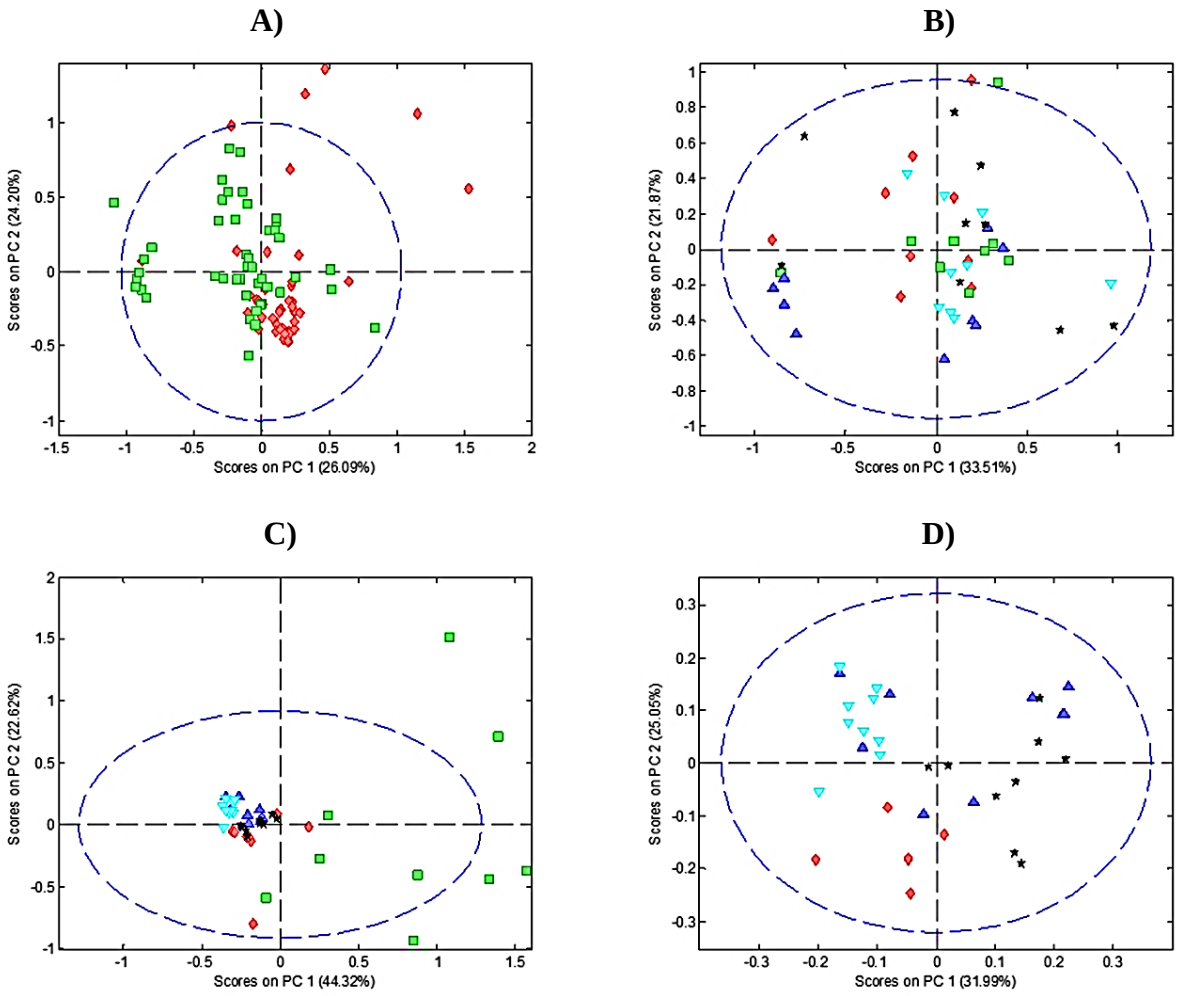


Figure S1. PCA models results. PC2 *vs.* PC1 scores plot considering: (A) all samples classified by organic co-solvent: acetonitrile samples (red diamonds) and methanol samples (green squares); (B) methanol samples classified by stationary phase: (BEH Amide (cyan triangles,▼), Amide (blue triangles,▲), Amine (red diamonds), Diol (green squares) and Zwitterionic (black stars); (C) acetonitrile samples classified by stationary phase (colors as in (B)); (D) acetonitrile samples after diol samples removal classified by stationary phase (colors as in (B)).

3.2.2. Article científic II.

Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling.

E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler.

Electrophoresis 36 (2015) 2324-2335.

Electrophoresis 2015, *36*, 2324–2335

Elena Ortiz-Villanueva[1]
Joaquim Jaumot[1]
Fernando Benavente[2]
Benjamín Piña[1]
Victoria Sanz-Nebot[2]
Romà Tauler[1]

[1]Department of Environmental Chemistry, IDAEA-CSIC, Barcelona, Spain
[2]Department of Analytical Chemistry, University of Barcelona, Barcelona, Spain

Received January 21, 2015
Revised March 16, 2015
Accepted March 20, 2015

Research Article

Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling

In this work, an untargeted approach based on capillary electrophoresis-mass spectrometry (CE-MS) in combination with multivariate data analyses is proposed as a high-throughput general methodology for metabolomic studies. First, total ion electropherograms (TIEs) were considered for exploratory and classification purposes by means of principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). Then, multivariate curve resolution alternating least squares (MCR-ALS) was applied to the multiple full scan MS data sets. This strategy permitted the resolution of a large number of metabolites being characterized by their electrophoretic peaks and their corresponding mass spectra. The proposed approach allowed solving additional electrophoretic issues, such as background noise contributions, low signal-to-noise ratios, asymmetric peaks and migration time shifts. The usefulness of the proposed methodology is demonstrated in a comparative study of the metabolic profiles from baker's yeast (*Saccharomyces cerevisiae*) samples cultured at two temperatures, 30°C and 37°C. A total number of 80 metabolites were relevant to yeast samples differentiation at the two temperatures and almost 50 of them were tentatively identified based on their accurate experimental molecular mass. The results show that changes in amino acid, nucleotide and lipid metabolic pathways participated in the acclimatization of yeast cells to grow at 37°C.

Keywords:
Capillary electrophoresis-mass spectrometry / Metabolic profiling / Multivariate data analyses / *Saccharomyces cerevisiae* / Untargeted analysis
DOI 10.1002/elps.201500027

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Metabolomics is a relatively modern field that aims to obtain a comprehensive coverage of low molecular weight compounds from biological systems [1, 2]. In the last few years, metabolic responses to external stimuli have been extensively investigated according to the changes in their concentration levels detected by means of different high-throughput analytical platforms.

Among all these platforms, capillary electrophoresis-mass spectrometry (CE-MS) analysis for targeted and untargeted metabolomics studies has been proposed as a complementary tool to other most commonly used techniques, such as GC-MS, LC-MS and NMR. CE-MS has the advantage of providing relevant information about charged and highly polar metabolites, in a relatively fast and simple way [2, 3]. CE is a promising analytical alternative due to its highly efficient separations, instrumental simplicity, full automation, low reagent and sample consumption, minor sample treatment and high versatility due to its multiple available modes. However, it is also true that CE has poor concentration sensitivity due to the small volume of sample loaded into the capillary and has limited reproducibility due to the inner wall changes induced by sample matrix in bare fused silica

Correspondence: Dr. Joaquim Jaumot, Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18–26, 08034 Barcelona, Spain
E-mail: joaquim.jaumot@idaea.csic.es
Fax: +34934006100

Abbreviations: **COW**, Correlation optimized warping; **IS**, Internal standard; **MCR-ALS**, Multivariate curve resolution by alternating least squares; **MWCO**, Molecular weight cut-off; **PCA**, Principal component analysis; **PLS-DA**, Partial least squares discriminant analysis; **TIE**, Total ion electropherograms; **VIP**, Variable importance in projection; **YPD**, Yeast extract peptone dextrose medium

Colour Online: See the article online to view Figs. 1–4 in colour

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

capillaries. Low sensitivity can be improved by coupling CE with MS, which also provides structural information about the separated compounds. Problems of reproducibility causing high variability in migration times or peak areas that hinder compound identifications and quantification can be adequately corrected by using coated capillaries [4], appropriate sample pretreatments or proper alignment algorithms [5]. The use of CE-MS is widely recognized nowadays as a potential analytical platform to perform small-scale and high-throughput metabolic profiling of biological systems [6, 7] and to assess changes in metabolic pathways on food consumption [8], drug administration [9], and disease disorders [10, 11]. Investigation of metabolic alterations caused by such stressing situations might be of great interest in biomarker discovery.

Data analysis tools play a crucial role to achieve biological interpretation and knowledge in the metabolomics field. Therefore, the need to analyze huge sets of very complex data generated in -omic studies using high-throughput instrumental analytical techniques have encouraged researchers to develop and apply advanced data analysis tools and novel workflows [12].

Regarding CE-MS in the -omic sciences, there is a current trend towards the application of chemometric tools in data processing. Once CE-MS data fulfill the required format, different data processing procedures can be applied before data analysis, such as PCA and partial least square discriminant analysis (PLS-DA) for exploratory and classification purposes [11, 13–15], respectively. However, relevant information could be missed by these kinds of approaches due to the presence of overlapped and embedded peaks in the raw CE-MS data. Accordingly, MCR-ALS can be proposed as a very powerful tool to allow a deeper and more extensive analysis of the CE-MS data. MCR-ALS allows overcoming different problems, such as retention time shifts among different injections, background noise contributions, and signal-to-noise ratios. Moreover, MCR-ALS method is able to resolve overlapped electrophoretic peaks and to provide the electrophoretic and mass spectra profiles of the constituents in the analyzed samples. Several published articles focus on the application of MCR-ALS to solve similar problems in LC-MS [12, 16, 17] and GC-MS [18]; but to the best of our knowledge, only few studies have been previously reported for the use of CE with MCR-ALS in metabolomic applications. For instance, MCR-ALS was used to improve feature detection in CE-UV [19], which was a challenging task because of the large migration time shifts and the heterogeneity of peak shapes.

The aim of this work is to propose a high-throughput general untargeted metabolomic approach based on CE-MS combined with advanced multivariate data analysis tools to obtain reliable information about collected data. The suitability of the presented methodology is demonstrated in the study of baker's yeast (*Saccharomyces cerevisiae*) metabolic profiles under mild heat stress conditions. *S. cerevisiae* is a well-known unicellular eukaryote organism widely used as a model for fundamental and applied metabolomic studies. Baker's yeast requires specific internal conditions for optimal growth, but environmental stressors, such as a moderate change of temperature, can perturb and disrupt its normal metabolic pathways. This interesting feature converts baker's yeast in a suitable candidate to demonstrate the potential of the proposed methodology based on the combination of CE-MS and chemometric methods.

2 Materials and methods

2.1 Chemicals and reagents

All chemicals used in the preparation of buffers and solutions were analytical reagent grade. Formic acid (98-100%), acetic acid (glacial), ammonia (25%), hydrochloric acid (25%), sodium hydroxide, methanol (HPLC grade) and 2-propanol (HPLC and MS grade) were purchased from Merck (Darmstadt, Germany). Bacteriological peptone, yeast extract, D-glucose, sodium hydrogen phosphate, sodium chloride, potassium dihydrogen phosphate, potassium chloride and bovine serum albumin (BSA) were provided by Sigma-Aldrich (St. Louis, MO, USA). Water with conductivity lower than 0.05 $\mu S \cdot cm^{-1}$ was obtained using a Milli-Q water purification system (Millipore, Molsheim, France).

Citric acid, succinic acid, malic acid, fumaric acid, lactic acid, phosphoenolpyruvic acid, palmitic acid, reduced and oxidized glutathione, L-carnitine hydrochloride, D-glucose 6-phosphate sodium salt, D-fructose 1,6-bisphosphate sodium salt hydrate, trehalose, D-mannitol, β-nicotinamide adenine dinucleotide, adenosine 5'-monophosphate disodium salt, adenosine 5'-diphosphate sodium salt, adenosine 5'-triphosphate disodium salt hydrate, guanosine 5'-triphosphate sodium salt hydrate, acetyl coenzyme A sodium salt and nucleoside (47310-U) and amino acid (A9906) mixtures, used as test standards, were purchased from Sigma-Aldrich (St. Louis, MO, USA). L-Methionine sulfone, used as the IS, was also supplied by Sigma-Aldrich (St. Louis, MO, USA).

2.2 Electrolytes, sheath liquids and standard solutions

An aqueous standard solution (1000 $\mu g \cdot mL^{-1}$) of each standard was prepared and stored in the freezer at -20°C until their use. Working standard solutions were obtained by diluting the stock solutions with water. Diluted standard solutions were used to optimize CE-MS methods and to spike yeast extract samples. The background electrolyte (BGE) contained 1 M of acetic acid (pH 2.3) for the detection in positive mode (ESI+) and 25 mM of ammonium acetate adjusted to pH 8.5 with ammonia in negative mode (ESI−). The sheath liquid solutions for CE-MS experiments consisted of a hydro-organic mixture of 60:40 (v/v) 2-propanol:water containing 0.05% (v/v) of formic acid in ESI+ and 0.5% (v/v) of ammonia in ESI− [7]. All solutions were passed through a 0.22 μm nylon filter (MSI, Westboro, MA, USA) before analysis and were stored at 4°C. The sheath liquid solutions were degassed for 15 min by sonication before using them.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

A solution of 1% (m/v) of BSA in PBS (0.01 M sodium hydrogen phosphate, 0.0015 M potassium dihydrogen phospate, 0.14 M sodium chloride, 0.0027 M potassium chloride, pH 7.2) was used to passivate the 3 kDa MWCO cellulose acetate filters (Amicon® Ultra-0.5 filters, Millipore) [20].

2.3 Sample preparation

2.3.1 Yeast strains and culture conditions

A single colony of baker's yeast (*Saccharomyces cerevisiae* strain BY4741) cells was first grown overnight at 30°C and 150 rpm in 50 mL of non-selective medium (yeast extract peptone dextrose medium, YPD / 20 $g{\cdot}L^{-1}$ bacteriological peptone, 10 $g{\cdot}L^{-1}$ yeast extract, 20 $g{\cdot}L^{-1}$ glucose) to obtain the initial culture (pre-culture). Once yeast pre-culture grew, a larger culture was prepared in a 1 L flask with 800 mL of YPD medium inoculated with 20 mL of yeast pre-culture. Thereafter, 12 fractions of the stock culture were individually incubated at 150 rpm for seven hours at 30°C (six fractions) and 37°C (six fractions), respectively. Figure S1 (Supporting Information) summarizes this workflow graphically. During all the process, yeast growth was monitored by measuring absorbance at 660 nm.

Separately, a new set of six yeast samples (three for each temperature condition) was grown using the same procedure as a validation set for the proposed methodology.

2.3.2 Quenching and yeast metabolite extraction

The sample treatment consisted of cleaning-up and quenching procedures followed by metabolite extraction. Culture samples were poured first into 50 mL tubes, and the metabolism was rapidly inactivated on ice. Samples were centrifuged at 3300 × *g* for 5 min at 4°C, and the supernatant was discarded. Then the pellets were washed twice with 1 mL of PBS at 3300 × *g* for 5 min at 4°C. The clean pellets were snap-frozen in liquid nitrogen and freeze-dried overnight.

An amount of 20 mg of freeze-dried yeast pellet were weighted into 2 mL plastic tubes, and metabolites were extracted with a mixture of 250 μL of cold methanol and 250 μL of cold water. After vortexing 30 s, the mixture was centrifuged at 11 000 × *g* for 15 min at 4°C to isolate the supernatant, and 350 μL of cold chloroform was added. Samples were vortexed 30 s, placed on ice for 10 min and centrifuged again at 11 000 × *g* for 15 min at 4°C. The aqueous fractions were filtered using 3 kDa MWCO filters previously passivated with BSA to avoid metabolite loss through adsorption on the inner walls of the plastic sample reservoir [20]. Filtrates were finally evaporated to dryness under nitrogen gas and reconstituted with 100 μL of water containing L-Methionine sulfone (IS) at a concentration of 5 $\mu g{\cdot}mL^{-1}$. The yeast extracts were stored at −80°C until the analysis.

2.4 Apparatus and procedures

pH measurements were performed with a Crison 2002 potentiometer and a Crison electrode 52–03 (Crison Instruments, Barcelona, Spain). Centrifugation was carried out in a Serie Digicen 21 centrifuge (Ortoalresa, Madrid, Spain). Sample incubation was performed with an Innova 40/40R incubator shaker (New Brunswick Scientific, Edison, USA). All bare fused-silica capillaries were supplied by Polymicro Technologies (Phoenix, AZ, USA).

2.4.1 CE-MS

CE-MS experiments were performed in an HP^{3D} CE system coupled with an orthogonal G1603A sheath-flow interface to a 6220 oa-TOF LC/MS mass spectrometer (Agilent Technologies, Waldbronn, Germany). The sheath liquid was delivered at a flow rate of 3.3 $\mu L{\cdot}min^{-1}$ by a KD Scientific 100 series infusion pump (Holliston, MA, USA). CE control and separation data acquisition were performed using ChemStation Software (Agilent Technologies) that was running in combination with the MassHunter workstation software (Agilent Technologies) for control and data acquisition of the TOF mass spectrometer.

A 72 cm total length (L_T) × 75 μm id × 360 μm od bare fused-silica capillary was used for the electrophoretic separations at 25°C, and all capillary rinses were performed at 930 mbar. New capillaries were flushed with 1 M NaOH (20 min), water (15 min) and BGE (30 min). The system was finally equilibrated by applying for 15 min the separation voltage (17 kV, normal polarity, i.e. anode in the inlet). Between workdays, the capillary was conditioned by rising successively with 0.1 M NaOH (5 min), water (10 min), and BGE (15 min). Both activation and conditioning procedures were performed off-line in order to avoid the unnecessary entrance of NaOH into the MS system. All samples were hydrodynamically injected at 50 mbar for 5 s. Between runs, the capillary was rinsed for 3 min with BGE.

The oa-TOF mass spectrometer was operated both in positive and negative mode using the following parameters: capillary voltage 4000 V, drying gas temperature 200°C, drying gas flow rate 4 $L{\cdot}min^{-1}$, nebulizer gas 7 psig, fragmentor voltage 215 V, skimmer voltage 60 V and OCT 1 RF Vpp voltage 300 V. Data were collected in profile mode at 1 spectrum/s (approximately 10 000 transients/spectrum) with a m/z range of 85–1100 working in the extended dynamic range mode (2 GHz) with the mass range set to standard (m/z 1700).

2.5 Data analysis

CE-MS data was analyzed by a combination of advanced chemometric tools to evaluate the most significant metabolic changes induced by heat stress conditions in yeast. Figure 1 shows a summary of the data analysis workflow, which is explained in detail in this section.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

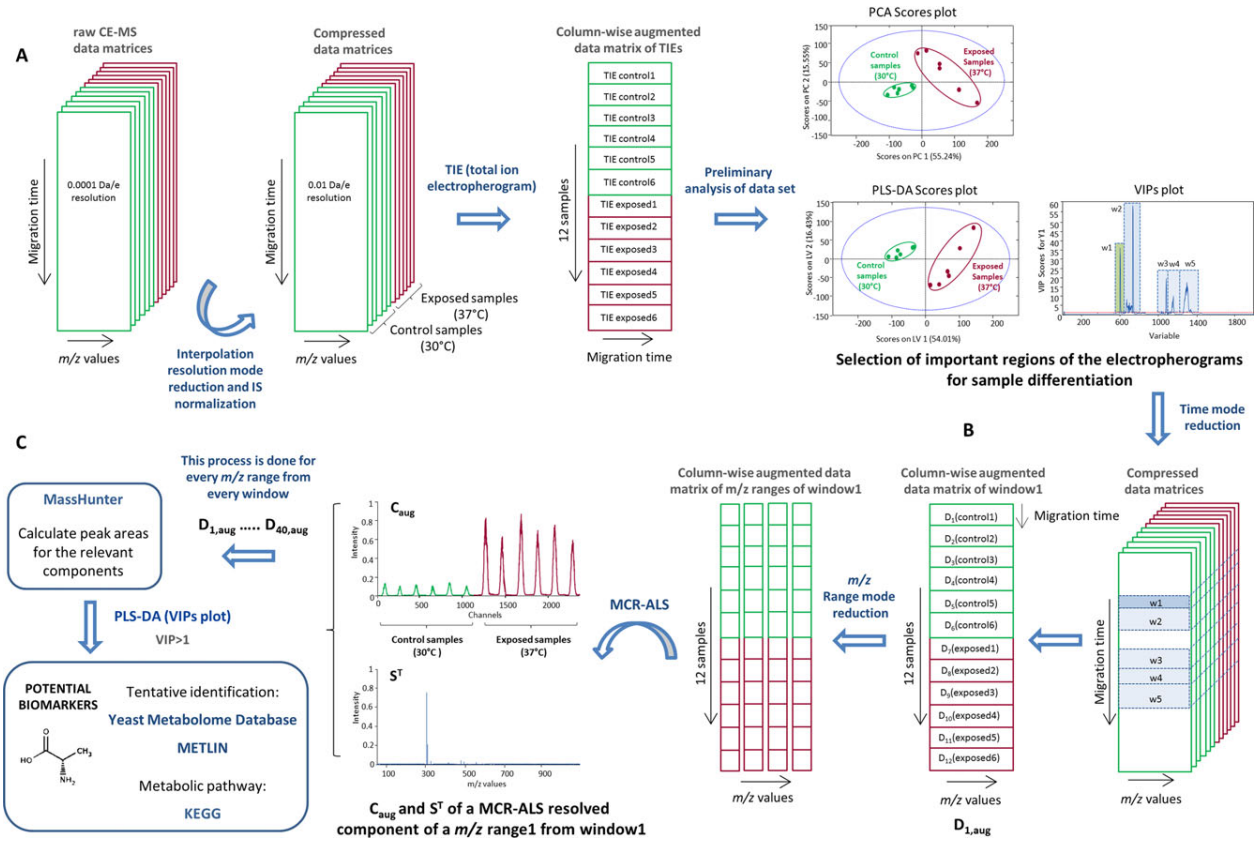


Figure 1. Workflow of CE-MS data analysis: (A) Pre-processing and preliminary data analysis, (B) data arrangement and MCR-ALS analysis in order to detect metabolites, and (C) tentative identification of relevant metabolites and metabolic pathways.

2.5.1 Data pre-processing and preliminary analysis of data set

In the first stage of the data processing, the accurate mass full scan raw MassHunter electropherograms were converted to .txt data files by ProteoWizard software [21] and imported to MATLAB (The Mathworks Inc. Natick, MA, USA). Due to the vast size of the full MS scan files in profile mode, input full resolution (0.0001 Da/e) were interpolated to a compressed data matrix at 0.01 Da/e resolution. In this way, every CE-MS sample provided a data matrix with 2126 rows (migration times, 25 mins of electrophoretic run) and 101 501 columns (*m/z* values, from 85 to 1100) (see Fig. 1A). Finally, the intensity scale of every compressed data matrix was normalized taking into account the area of the IS (L-Methionine sulfone) to correct the instrumental intensity drifts among injections and to scale the data internally.

After that, an exploratory study of the TIE of the 12 yeast samples was performed. A data matrix of dimensions 12 × 2126 (number of samples × migration times) was obtained by arranging the individual TIE data vectors one on top of each other. Electrophoretic peaks were aligned in the time domain to correct the moderate variations in migration times of metabolites among runs using the COW method [22]. TIE vectors were baseline corrected by automatic weighted least squares and mean-centered pretreatment methods. Data from the two different types of yeast cultures were subjected to preliminary analysis exploration by PCA and PLS-DA [23,24]. PCA allows an initial exploration of CE-MS data and the differentiation of yeast samples according to culture temperature conditions and the detection of possible outlier samples. Furthermore, PLS-DA is used for the discrimination of yeast samples on the basis of their metabolic profile changes according to growth temperature. In addition to sample discrimination, PLS-DA loadings can provide information about the most significant electrophoretic regions (windows, w) for differentiation between control (30°C), and mildly high temperature exposed (37°C) yeast samples. In this work, these critical regions have been selected using variable importance in the projection (VIP) scores [25–27]. Figure 1A shows the results of data pre-processing and PCA and PLS-DA analyses of the TIE in positive mode (ESI+ data).

2.5.2 Full scan MS data arrangement and MCR-ALS analysis

MCR-ALS is a chemometric method especially useful to analyse multicomponent systems with strongly overlapping contributions, such as those present in CE separations where

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

the electrophoretic behaviour of metabolites is rather similar [28]. In the case of CE-MS, full scan MS data matrix **D** contains the experimental mass spectra at all retention times in their rows and the electropherograms at all *m/z* values in their columns. MCR-ALS decomposes the original data matrix **D** using a bilinear model which produces the electrophoretic and MS spectral of the resolved contributions. MCR-ALS analysis of the data matrix **D** gives two factor matrices, **C** and $\mathbf{S}^T$, as in eq. (1):

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

where matrix **C** has the electrophoretic profiles of the resolved contributions (components), $\mathbf{S}^T$ has their mass spectra and **E** has the residuals unexplained by the model.

In this work, MCR-ALS was applied to resolve the metabolite profiles from the acquired CE-MS data [29, 30]. When different samples are simultaneously analysed and compared by MCR-ALS, the full scan MS data matrices obtained in the analysis of each sample are arranged in a column-wise augmented data matrix configuration (see Fig. 1B). For instance, a column-wise augmented data matrix $\mathbf{D}_{aug}$ (eq. (2)) for a particular *m/z* range was built up by arranging data matrices ($\mathbf{D}_k$, corresponding to the 12 yeast samples) one on top of each other.

$$\mathbf{D}_{aug} = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_{12} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_{12} \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{12} \end{bmatrix} = \mathbf{C}_{aug}\mathbf{S}^T + \mathbf{E}_{aug} \quad (2)$$

In this case, MCR-ALS provided a common matrix of the mass spectra of the resolved components ($\mathbf{S}^T$) for all samples, and a set of matrices of describing the resolved electrophoretic profiles ($\mathbf{C}_{aug}$) in every sample. Electrophoretic peaks resolved in matrix $\mathbf{C}_{aug}$ are allowed to vary in position (shifts) and shape among samples [29, 31]. This aspect is especially useful in the case of CE data where migration shifts among samples occurs and, so, the alignment of electrophoretic peaks before analysis is not needed.

In this study, the initially compressed CE-MS data matrices were further reduced in their time mode using VIP scores obtained in the preliminary PLS-DA analysis of TIE. In accordance with the VIP scores, the electropherograms were partitioned in five different time windows for both ESI+ and ESI− ionization modes (see Fig. 1B). Then, these time windows were further reduced in their *m/z* mode dimension in four different *m/z* ranges of approximately 300 Da. This time and *m/z* windowing approach permitted MCR-ALS analysis of the huge sized original CE-MS raw data within computer processing limitations and reasonable data processing times. Finally, $\mathbf{D}_{aug}$ was normalized by using the adaptation of the MinMax transformation [32] which allows the resolution of low concentrated metabolites, as well as the major metabolites during MCR-ALS analysis. MinMax procedure rescales each column of the raw data matrix by subtracting the minimum value to each element of the column and dividing the result by the range of the column. In order to avoid unwanted noise effects, only values clearly above the noise level of each data matrix were considered.

In MCR-ALS data analysis, an initial estimation of the number of independent contributions to the observed data variance (components) is required. In order to do this, the singular value decomposition (SVD) method [33] is applied. Subsequently, initial estimations of either $\mathbf{C}_{aug}$ and $\mathbf{S}^T$ are given (for instance using pure variable detection methods [34]). Finally, an ALS optimization under constraints is performed to solve eq. (2) for $\mathbf{C}_{aug}$ and $\mathbf{S}^T$ factor matrices. Constraints can be applied to ALS optimization in order to provide chemical meaning to the resolved electrophoretic and mass spectra profiles. Also, the effect of rotation ambiguities can be decreased using constraints properly. [28, 35, 36]. Constraints, such as non-negativity for electrophoretic ($\mathbf{C}_{aug}$) and spectra ($\mathbf{S}^T$) profiles, and spectral normalization (equal height) were applied [29, 35–37]. The quality of MCR-ALS models was evaluated by the percent of variance explained (R^2) and lack of fit (LOF) values [37].

2.5.3 Detection and identification of potential metabolites

For every resolved MCR-ALS component, detection and evaluation of the metabolites causing differences between control and exposed samples was carried out following the procedure shown in Fig. 2 that details the last part of Fig. 1. First, electropherogram (peak) profiles of control and exposed samples (on the columns of $\mathbf{C}_{aug}$ matrix) were compared using a non-parametric Mann–Whitney *U* test on their areas to determine whether there was a significant difference between the means of the two groups of profile areas (first statistical test, t-1, in Fig. 2). Only resolved components of $\mathbf{C}_{aug}$ that showed a significant difference between the means of the groups (*p*-value lower than 0.05) were selected. In this case, *p*-values were not corrected using a control method to correct for multiple comparisons because the aim of this stage is to obtain the maximum number of metabolites. If false positive metabolites were selected in this step, they would be reevaluated, and very likely discarded, in the following steps of the analysis. Next, their corresponding mass spectra profiles ($\mathbf{S}^T$) were used to identify the *m/z* values causing the differentiation between control and exposed samples. Finally, peak areas of these candidate *m/z* values were recovered from the full scan CE-MS data using the MassHunter workstation software, taking as a reference the *m/z* value and the migration time of the MCR-ALS resolved components. Despite MCR-ALS resolved electropherograms in $\mathbf{C}_{aug}$ contained information about the areas of these features, due to the applied MinMax normalization it was preferred returning to the original raw data to evaluate the contribution of each feature in each sample (areas were also finally normalized considering the IS).

These preselected peak areas for the candidate metabolites were autoscaled, and PLS-DA was then applied to

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

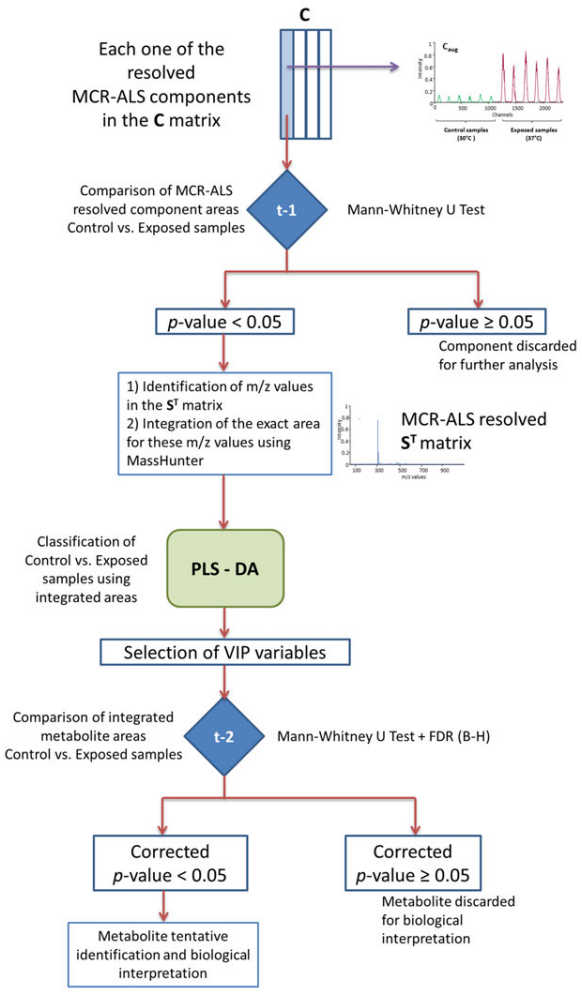


Figure 2. Scheme of the workflow for the detection of the potential metabolites from the MCR-ALS resolved components.

identify the most important metabolites responsible for the sample discrimination according to the VIP scores of the PLS-DA model. Statistical significance of the variables selected considering the VIP scores of the PLS-DA model was assessed by a Mann–Whitney *U* test controlled by the False Discovery Rate (FDR) using the Benjamini–Hochberg procedure on the calculated peak areas of control and exposed samples (second statistical test, t-2, in Fig. 2). In this case, the multiple hypothesis testing was carried out to assure that the finally selected metabolites were not false positives. Thereafter, the accurate exact *m/z* values of the finally VIP selected metabolites were searched in on-line databases resources, such as METLIN Metabolite Database [38] and Yeast Metabolome Database (YMDB) [39]. Finally, the list of the tentatively identified metabolites was used to investigate the possible metabolic pathways and mechanisms affected by the heat stress according to the KEGG database [40].

Validation of the obtained results (discrimination between control and exposed samples and identification of the potential metabolites) was additionally performed using an external set of six samples (three control and three exposed samples).

2.5.4 Software

Most of the calculations and data analysis were performed under MATLAB R2013a (The Mathworks Inc. Natick, MA, USA). PLS Toolbox 7.3.1 (Eigenvector Research Inc., Wenatchee, WA, USA) was used for PCA, PLS-DA and VIP calculations and MCR-ALS toolbox [31] was used for resolution of electropherogram and mass spectral metabolite profiles from full MS scan augmented data matrices.

3 Results and discussion

3.1 CE-MS method development and baker's yeast analysis

In the preliminary part of this work, various experimental CE-MS methodologies were tested for an adequate yeast metabolome analysis. Particular attention was given to the optimization of separation and detection conditions in both ESI+ and ESI− ionization modes. Several test mixtures containing model chemical compounds, covering a broad range of metabolite families with different physicochemical properties and concentrations from 10 to 100 $\mu g \cdot mL^{-1}$ (i.e. organic acids, sugars, nucleotides, nucleosides and amino acids), were used to optimize the experimental conditions. CE-MS conditions were specially selected in terms of sensitivity and reproducibility in order to ensure a comprehensive and reliable metabolite profiling.

Acetic and formic acid were investigated as acidic BGEs for CE separation in ESI+ mode at concentrations ranging from 0.5 to 1.0 M or in a mixture of 50 mM of acetic acid and 50 mM of formic acid, as it has been previously proposed in preceding metabolomic works [6, 7]. Under these conditions, some organic acids, nucleotides and sugars could not be detected because they were poorly ionized in ESI+ in accordance with their electrical charge in the BGE. For the detected compounds (i.e. the rest of organic acids, nucleosides and amino acids), the mixture of acetic and formic acid provided lower signal-to-noise ratios than pure acetic or formic acid solutions. BGEs of 1.0 M acetic or formic acid showed the best results in terms of sensitivity. Acetic acid was finally preferred due to the lower CE currents (<50 μA) that are recommended to prevent mass spectrometer electronics damage due to arc discharge in the interface. In addition, 2-propanol and methanol were tested as organic modifiers in the sheath liquid using different percentages ranging from 50 to 100% (v/v) with 0.05% (v/v) of formic acid. A sheath liquid composition of 60:40 (v/v) for a 2-propanol: water ratio resulted in optimum detection sensitivity in ESI+ mode. Sensitivity could not be further improved by changing the amount of

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

formic acid in the sheath liquid in the range from 0.05% to 0.5% (v/v).

ESI− mode is preferred in the cases in which ESI+ sensitivity is poor. Despite the sheath liquid allows to adjust pH value of the sprayed solution, it is recommended to use a weakly basic volatile BGE of ammonium formate or acetate, better than ammonium bicarbonate or more basic BGEs with ammonia [4,41–43]. A 25 mM of ammonium acetate (pH 8.5) was used as a BGE to analyse the different test mixtures. At this pH value, anions migrated towards the cathode pushed by the EOF and sensitivity in ESI− was good for most of the test compounds that could not be analysed by ESI+. As in ESI+ mode, 2-propanol and methanol mixtures were investigated as organic modifiers of the sheath liquid at different percentages ranging from 50 to 100% (v/v). Again a 60:40 (v/v) of 2-propanol: water ratio also resulted in optimum detection sensitivity. Then, the addition of ammonia ranging from 0.05 to 0.5% (v/v) was tested, and the best results were obtained with 0.5% (v/v) of ammonia.

3.2 Total ion electropherograms preliminary exploration

The optimized CE-MS conditions in ESI+ and ESI− were applied to the analysis of changes in the metabolome of yeast samples cultured at optimal growth temperature (30°C) and under heat stress conditions (37°C) using an untargeted metabolomic approach. Examples of CE-MS electropherograms belonging to the different yeast extracts are given in Fig. 3A and B. Both electropherograms showed a complex profile with compounds comigrating at different concentrations in less than 25 minutes. Separation was faster and profiles significantly different than in a previous work by LC-MS [12].

In the case of ESI+ data, the TIE data matrix with 12 samples and 2126 migration times was firstly analyzed by PCA. Three principal components already explained 78% of the data variance. Figure 3C shows the scores plot of the first two components, where PC1 clearly separated samples in relation to the temperature culture. PLS-DA applied to the TIE data matrix allowed perfect sample discrimination (see Supporting Information Fig. S2A), and the detection of the most relevant variables (i.e., electrophoretic time regions) for the classification of the samples (Supporting Information Fig. S2C). Cross-validated (leave-one-out) PLS-DA model with a single latent variable (LV) explained 54% of the total TIE data matrix (**X**) variance and the 81% of the total sample class (**Y**) variance. Most important electrophoretic regions were selected according to VIP scores (with values higher than 1). Finally, only five electrophoretic regions (time windows) were selected for further MCR-ALS analysis (normalized time windows: 7.2–7.6, 7.5–9.4, 11.5–13.2, 13.1–14.5, and 14.4–16.8 minutes).

TIE data matrix acquired in ESI− mode was also examined by PCA and PLS-DA. Initial exploratory analysis performed by PCA revealed differences in samples according to the culture temperature (see Fig. 3D), but in contrast to the ESI+ scores plot, the two groups were partially overlapped. In this case, four principal components were needed to explain the 78% of the total data variance. Leave-one-out cross-validated PLS-DA model with three latent variables accounted for a 68% of the TIE **X**-variance and a 97% of the sample class **Y**-variance, with a partial separation between samples (see Supporting Information Fig. S2B). Once again, VIP scores allowed selecting five electrophoretic regions

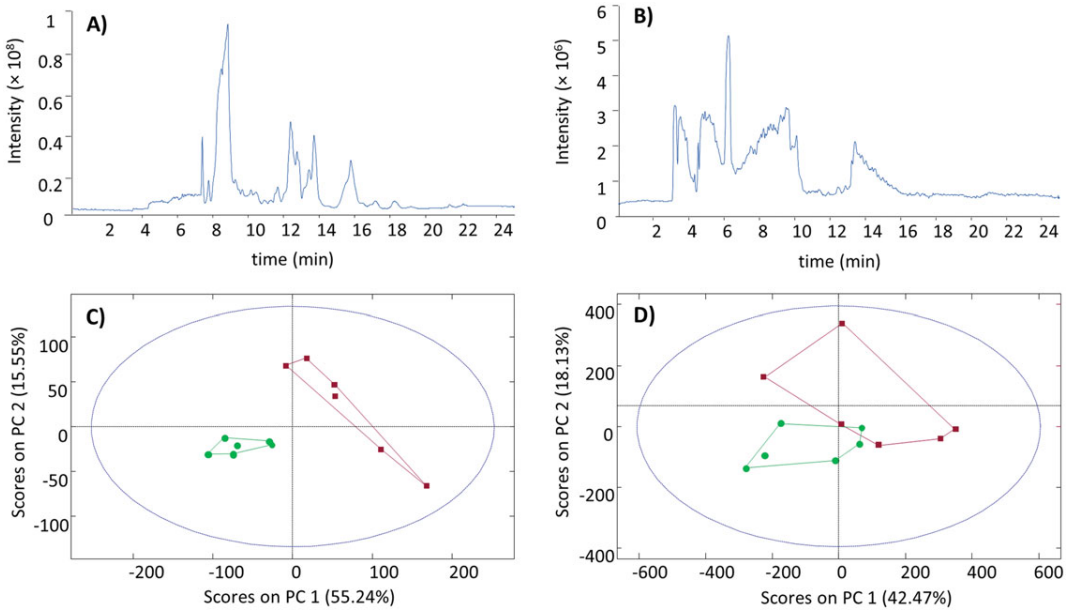


Figure 3. TIEs of yeast extracts analyzed by CE-MS: (A) ESI+ and (B) ESI–. PCA scores plots for the 12 TIEs of the yeast extracts: (C) ESI+ and (D) ESI–. Green circles are control samples (cultured at 30°C) and red squares are exposed yeast samples (cultured at 37°C).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

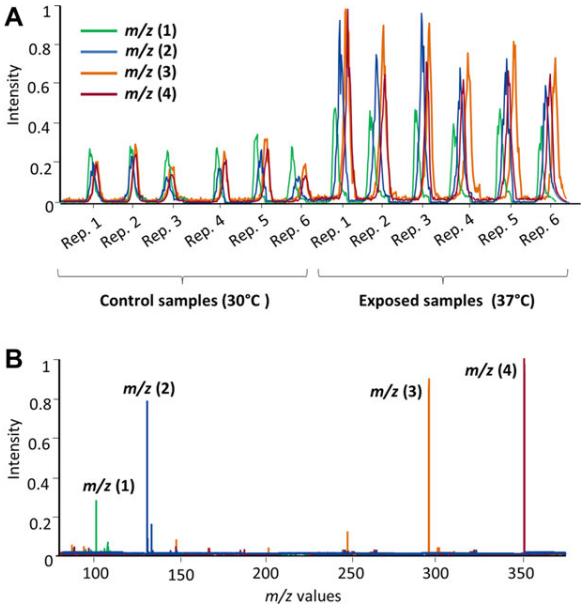


Figure 4. Example of MCR-ALS resolution simultaneously applied to a column-wise augmented data matrix of *m/z* range 85–385 Da corresponding to window 1 (w1, 7.2–7.6 minutes) in ESI+. (A) Resolved comigrated electrophoretic peaks for the different samples and (B) resolved mass spectra.

for further MCR-ALS analysis (Supporting Information Fig. S2D), that were different than in ESI+ mode (normalized time windows: 2.8–4.5, 4.4–6, 5.9–6.7, 6.6–10.6, and 12.7–16.5 minutes).

3.3 Results of MCR-ALS analysis and detection of the most relevant metabolites

As mentioned in Section 2.5.2, the direct untargeted analysis of full scan CE-MS raw data is rather complicated due to the multiple comigrated electrophoretic peaks. For this reason, MCR-ALS was employed to facilitate this task. MCR-ALS was applied using a column-wise augmented data matrix containing the information about the 12 samples (control and exposed samples) simultaneously and allowed the resolution of the electropherogram profiles and corresponding mass spectra of the yeast metabolites.

MCR-ALS analysis in each ESI mode was performed separately on column-wise augmented data matrices of different *m/z* ranges (at the resolution of 0.01 Da/e) (five for each ESI mode), corresponding to the previously PLS-DA VIP selected time windows. A total number of 40 column-wise augmented matrices (10 time windows × 4 *m/z* intervals) were separately analysed. The number of components selected were related to the number of electrophoretic peaks, despite the fact that some of these resolved components could be assigned to contributions such as solvent background or instrumental noise. In most of the cases, MCR-ALS showed an explained variance (R^2) higher than 97%, which is rather satisfactory considering the noise level and the normalization of the data set.

Figure 4 shows MCR results of the first time window (w1) augmented data matrix (migration times from 7.2-7.6) corresponding to the *m/z* range of 85–385 Da. As it can be observed, in w1 there was a single electrophoretic peak and MCR-ALS successfully resolved four comigrated components ($\mathbf{C}_{aug}$) that were more concentrated in stressed samples and could be identified by their corresponding mass spectra ($\mathbf{S}^T$). In this MCR-ALS analysis, other components were resolved that could correspond to non-metabolite contributions, such as noise interferences. A statistical analysis of the profile areas of the resolved electropherograms was performed to differentiate between control and exposed samples. Only components with a *p*-value lower than 0.05 in a Mann–Whitney *U* test comparing the means of the areas of control and exposed samples were finally considered as we consider that components with a *p*-value higher than 0.05 could be related to non-changing metabolites between control and exposed samples or to non-metabolite contributions. The mass spectra of these components (from $\mathbf{S}^T$) allow their tentative identification. After the resolution and analysis of the 40 augmented data matrices, the total number of statistically significant metabolite profiles detected in ESI+ and ESI− modes were 94 and 30, respectively.

Furthermore, the peak areas of these features (*m/z* values) were directly integrated with MassHunter for each sample in the CE-MS raw data. The final selection of the most relevant features (metabolites) for temperature effect discrimination was performed by PLS-DA from their peak areas in each sample. One latent variable allowed the perfect discrimination between the two groups of samples in both ionization modes, explaining most of X-variance (93% in ESI+ and 95% in ESI−) and Y-variance (99% in both cases). Once again, VIP scores (values higher than 1) of the PLS-DA model were used as a feature selection tool in order to choose the most relevant candidate metabolites. Finally, the number of metabolites was reduced to 59 for ESI+ and 21 for ESI−. Statistical significance of the finally selected metabolites was assessed by means of a Mann–Whitney *U* test with *p*-values corrected by FDR using the Benjamini–Hochberg procedure. Results showed that the larger corrected *p*-value was 0.0027 which confirmed the differences between the control and exposed samples considering the selected metabolites.

This list of candidate metabolites was validated using a set of six external samples (three control samples and three exposed samples). Areas of the finally preselected metabolites were determined in these validation samples, and PLS-DA was performed to examine whether sample discrimination could be achieved in both ESI modes. Furthermore, the validation set were correctly predicted obtaining sensitivity and specificity values of 1.0.

3.4 Tentative metabolite identification and biological interpretation

The most contributing metabolites to sample discrimination (59 for ESI+ and 21 for ESI−) were tentatively identified.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Tentative identification of the relevant metabolites detected by the combination of CE-MS and MCR-ALS (mass accuracy ≤ 15 ppm)

Met. ID	Compound	Molecular formula	Adduct	Calculated mass (Da)	Measured mass (Da)	Error (ppm)	Fold-change	Trend
ESI+ mode								
1	Choline	$C_5H_{14}NO$	$[M+H-H_2O]^+$	86.0970	86.0973	3.5	5.4	UP
	Choline	$C_5H_{14}NO$	M	104.1075	104.1077	1.6	7.9	UP
2	Pyruvic acid	$C_3H_4O_3$	$[M+Na]^+$	111.0053	111.0054	1.2	2.7	UP
3	Pipecolic acid	$C_6H_{11}NO_2$	$[M+H-H_2O]^+$	112.0751	112.0763	10.6	1.9	UP
4	Uracil	$C_4H_4N_2O_2$	$[M+H]^+$	113.0346	113.0343	2.3	2.1	DOWN
5	L-Valine	$C_5H_{11}NO_2$	$[M+H]^+$	118.0863	118.0866	2.9	2.0	UP
6	Pantetheine	$C_{11}H_{22}N_2O_4S$	$[M+2K+H]^{3+}$	119.0212	119.0204	6.8	2.1	UP
7	Gamma-Butyrolactone	$C_4H_6O_2$	$[M+K]^+$	124.9999	125.0002	2.1	7.0	UP
8	4-Hydroxy-L-proline	$C_5H_9NO_3$	$[M+H]^+$	132.0655	132.0658	2.1	4.2	UP
9	Adenine	$C_5H_5N_5$	$[M+H]^+$	136.0618	136.0621	2.4	5.0	DOWN
10	4-Deoxypyridoxine	$C_8H_{11}NO_2$	$[M+H-H_2O]^+$	136.0751	136.0755	2.8	1.4	UP
11	Pyrimidine	$C_4H_4N_2$	$[M+isoprop+H]^+$	141.1028	141.1028	0.1	2.1	UP
12	LysoPE(14:1(9Z)/0:0)	$C_{19}H_{38}NO_7P$	$[M+3H]^{3+}$	142.0868	142.0871	2.1	2.2	UP
13	Sorbitol-6-phosphate	$C_6H_{15}O_9P$	$[M+H+Na]^{2+}$	143.0209	143.0200	6.5	1.9	DOWN
14	Urea	CH_4N_2O	$[2M+Na]^+$	143.0539	143.0537	1.7	1.5	DOWN
15	Propionylcholine	$C_8H_{18}NO_2$	M	160.1338	160.1349	7.2	3.2	UP
16	Phosphocholine	$C_5H_{14}NO_4P$	$[M+H-H_2O]^+$	166.0633	166.0628	3.0	6.7	UP
17	Phosphorylcholine	$C_5H_{15}NO_4P$	M	184.0739	184.0736	1.5	2.5	UP
18	Thymine	$C_5H_6N_2O_2$	$[M+isoprop+H]^+$	187.1083	187.1088	2.9	2.2	UP
19	Pentadecanoic acid	$C_{15}H_{30}O_2$	$[M+H-2H_2O]^+$	207.2096	207.2078	8.6	31.3	UP
20	L-Dihydroorotic acid	$C_5H_6N_2O_4$	$[M+isoprop+H]^+$	219.0981	219.0980	0.4	1.6	UP
21	L-3-Hydroxykynurenine	$C_{10}H_{12}N_2O_4$	$[M+H]^+$	225.087	225.0883	5.9	15.1	UP
22	Biotin	$C_{10}H_{16}N_2O_3S$	$[M+H-H_2O]^+$	227.0843	227.0861	7.9	4.1	DOWN
23	Adenosine	$C_{10}H_{13}N_5O_4$	$[M+H-H_2O]^+$	250.0929	250.0934	2.0	3.5	UP
24	Glycerophosphocholine	$C_8H_{21}NO_6P^+$	$[M+H]^+$	258.1101	258.1112	4.3	8.1	UP
25	L-Threonine	$C_4H_9NO_3$	$[2M+Na]^+$	261.1057	261.1085	10.7	3.5	UP
26	Homocysteine	$C_4H_9NO_2S$	$[2M+3H_2O+2H]^+$	298.0939	298.0958	6.3	3.6	UP
27	Deoxyadenosine monophosphate	$C_{10}H_{14}N_5O_6P$	$[M+H]^+$	332.0754	332.0754	0.1	1.8	DOWN
28	Nicotinamide ribotide	$C_{11}H_{15}N_2O_8P$	$[M+H]^+$	335.0639	335.0642	1.0	3.9	DOWN
29	S-Adenosylmethioninamine	$C_{14}H_{23}N_6O_3S^+$	M	355.1552	355.156	2.2	4.9	UP
30	Biocytin	$C_{16}H_{28}N_4O_4S$	$[M+H]^+$	373.1904	373.1923	5.1	2.3	DOWN
31	Cytidine monophosphate	$C_9H_{14}N_3O_8P$	$[M+isoprop+H]^+$	384.1172	384.1179	1.8	2.6	UP
32	Oxitriptan	$C_{11}H_{12}N_2O_3$	$[2M+H]^+$	441.1769	441.1785	3.7	30.9	UP
33	Citicoline	$C_{14}H_{26}N_4O_{11}P_2$	$[M+H]^+$	489.1146	489.1155	1.8	1.9	DOWN
	Glycerophosphocholine	$C_8H_{21}NO_6P^+$	$[2M+H]^+$	515.2129	515.2145	3.1	276.6	UP
	Glycerophosphocholine	$C_8H_{21}NO_6P^+$	$[2M+K]^+$	553.1688	553.1711	4.2	34.8	UP
ESI- mode								
34	L-Alanine	$C_3H_7NO_2$	$[M-H]^-$	88.0404	88.0412	9.1	1.7	UP
35	Benzoic acid	$C_7H_6O_2$	$[M-H]^-$	121.0295	121.0305	8.2	5.4	UP
36	Itaconic acid	$C_5H_6O_4$	$[M-H]^-$	129.0193	129.0206	9.8	2.5	UP
37	D-Mannitol	$C_6H_{14}O_6$	$[M-H]^-$	181.0718	181.0733	8.5	2.2	UP
38	L-Tryptophan	$C_{11}H_{12}N_2O_2$	$[M-H]^-$	203.0826	203.0846	9.8	1.8	DOWN
39	Cytidine	$C_9H_{13}N_3O_5$	$[M-H]^-$	242.0782	242.0815	13.5	11.2	UP
40	S(8)succinyldihydrolipoamide	$C_{12}H_{21}NO_4S_2$	$[M-H]^-$	306.0839	306.0833	2.0	1.9	DOWN
	Glycerophosphocholine	$C_8H_{21}NO_6P^+$	$[M+HAc-H]^-$	316.1167	316.1204	11.8	13.3	UP
41	4-Hydroxy-L-threonine	$C_4H_9NO_4$	$[2M+HAc-H]^-$	329.1202	329.1181	6.3	3.6	UP
42	Trehalose	$C_{12}H_{22}O_{11}$	$[M-H]^-$	341.1089	341.1104	4.3	2.8	UP
43	L-Tyrosine	$C_9H_{11}NO_3$	$[2M-H]^-$	361.1405	361.1429	6.6	7.1	DOWN
	Biocytin	$C_{16}H_{28}N_4O_4S$	$[M-H]^-$	371.1759	371.1748	2.8	2.7	DOWN
44	Leukotriene B4	$C_{20}H_{32}O_4$	$[M+K-2H]^-$	373.1787	373.1832	12.2	7.7	UP
45	L-Phenylalanine	$C_9H_{11}NO_2$	$[2M+FA-H]^-$	375.1562	375.1556	1.5	8.6	UP

Statistical significance of the selected metabolites was assessed by FDR using Benjamini–Hochberg corrected Mann–Whitney *U* test (larger corrected *p*-value was 0.0027).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. Pathway analysis (KEGG) of affected metabolites by temperature

	Total number of affected metabolites	Metabolites[a)]
sce01100 Metabolic pathways	29	1,2,3,4,5,8,9,14,16,18,20,21,22,23,25,27,28,29,30,31,32,33,34,35,39,42,43,44,45
sce01110 Biosynthesis of secondary metabolites	9	2,3,5,6,25,35,42,43,45
sce02010 ABC transporters	9	1,5,14,22,25,34,37,42,45
sce01230 Biosynthesis of amino acids	6	2,5,25,34,43,45
sce00240 Pyrimidine metabolism	6	5,25,31,34,43,45
sce00970 Aminoacyl-tRNA biosynthesis	5	5,25,34,43,45
sce01210 2-Oxocarboxylic acid metabolism	4	2,5,43,45
sce00330 Arginine and proline metabolism	4	2,8,14,29
sce00770 Pantothenate and CoA biosynthesis	4	2,4,5,6
sce00230 Purine metabolism	4	9,14,23,27
sce00564 Glycerophospholipid metabolism	4	1,16,24,33
sce00360 Phenylalanine metabolism	4	2,35,43,45
sce00260 Glycine, serine and threonine metabolism	3	1,2,25
sce00270 Cysteine and methionine metabolism	3	2,5,25

a) According to metabolites identification field of Table 1.

Taking advantage of the highly accurate experimental molecular mass values provided by CE-MS with an oa-TOF analyser, a small deviation from the calculated (theoretical) molecular mass was used to evaluate the accuracy of possible molecular formulas (≤15 ppm). Tentative identities of metabolites as well as their folding trends (up and down relative to the control samples), are summarized in Table 1. Each candidate and the involved metabolic pathways were searched in different on-line databases.

The change in temperature affected three main metabolic pathways (see Table 2 and Supporting Information Fig. S3). Most of the identified metabolites belonged to amino acid or nucleotide/nucleoside metabolism pathways (some of them also appeared as "secondary metabolism" in Table 2), probably reflecting an acclimatization of the protein and nucleic acid synthesis to mildly high temperatures. The analysis also revealed alterations in the concentrations of the metabolites implicated in the choline/phosphoglycerocholine metabolism pathway (see Table 2 and Supporting Information Fig. S3). These results were consistent with the expected changes on lipid composition upon a shift of temperature, as this particular pathway has been already related to acclimatization to temperature [44], and its disruption results in a poor performance at suboptimal temperatures [45]. Other metabolites affected by the temperature and identified were also related to lipid metabolism, like LysoPE(14:1(9Z)/0:0) or leukotriene B4 (Table 1). Finally, an increase on trehalose at the higher temperature, consequent with its role as temperature and stress-protector of the yeast cell [46], plus other key metabolites, like pyruvate or urea (Supporting Information Fig. S3).

The response of baker's yeast to heat stress has been profusely analyzed by transcriptomic analyses [47–50], and more recently, by metabolomic studies [12,51]. Despite the different genetic backgrounds and experimental strategies, all these analyses identified glucose/energy metabolism, respiration and amino acid synthesis and metabolism as the metabolic pathways more affected by mild heat stress, together with the already mentioned accumulation of trehalose [12, 46–51]. Whereas transcriptomic analyses usually failed to detect any temperature-related effect on lipid metabolism, metabolomic studies consistently showed changes in lipid and phospholipid metabolism [12, 51]. Changes in lipid composition have been genetically identified as crucial on the acclimatization of the yeast cells to temperature changes [45, 52]. It is likely than other metabolic changes equally important for yeast growth could have been overlooked by transcriptomic or other mRNA quantitation-based analyses, whereas they can be adequately described by metabolomic studies. Therefore, the development of fast and reliable methodologies for metabolome analyses is therefore of paramount importance in the study of the behavior of living cells to the ever-changing environment.

4 Concluding remarks

The combination of CE-MS and chemometric data analysis strategy has demonstrated to be suitable for high-throughput untargeted metabolomics of yeast samples stressed by temperature.

Two CE-MS separation and detection methodologies were established to ensure a reliable metabolite profiling of yeast samples. The combination of MCR-ALS with other chemometric tools, such as PLS-DA, allowed the comprehensive analysis of the CE-MS metabolomic data, resolving electrophoretic peaks and mass spectra of a large number of metabolites. This approach overcomes typical electrophoretic problems, such as migration time shifts among injections and background noise contributions that are detrimental to metabolite identification. Finally, a list of metabolites whose concentrations change with temperature in control and

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

exposed yeast samples was provided, most of them were tentatively identified and the corresponding metabolic pathways affected by temperature were discussed. Acclimatization to mildly high temperature (37°C) seemed to influence the metabolism of the building blocks of proteins and amino acids, although at this point it is difficult to elaborate a model linking these alterations on metabolite concentrations and specific phenotypes. Temperature also seemed to alter the lipid metabolism, consistent with the proposed importance of changes in membrane fluidity to cope with temperature changes.

To conclude, the proposed strategy can be extended to CE untargeted metabolomic studies of other biological systems for biomarker discovery and metabolite pathways investigation, which are typically investigated by LC-MS, GC-MS or NMR.

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. Some part of this study was also supported by a grant from the Spain's Ministry of Education and Science (CTQ2011-27130). Also, recognition from the Catalan government (grant 2014 SGR 1106) is acknowledged. JJ acknowledges a CSIC JAE-Doc contract cofounded by the FSE.

The authors declare that they have no competing interests.

5. References

[1] Villas-Bôas, S. G., Mas, S., Åkesson, M., Smedsgaard, J., Nielsen, J., *Mass Spectrom. Rev*. 2005, *24*, 613–646.

[2] Hirayama, A., Wakayama, M., Soga, T., *TrAC Trend Anal. Chem*. 2014, *61*, 215–222.

[3] Kok, M. G. M., Somsen, G. W., de Jong, G. J., *TrAC Trend Anal. Chem*. 2014, *61*, 223–235.

[4] Tanaka, Y., Higashi, T., Rakwal, R., Wakida, S.-i., Iwahashi, H., *Electrophoresis*. 2008, *29*, 2016–2023.

[5] Nevedomskaya, E., Derks, R., Deelder, A., Mayboroda, O., Palmblad, M., *Anal. Bioanal. Chem*. 2009, *395*, 2527–2533.

[6] Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., Nishioka, T., *J. Proteome Res*. 2003, *2*, 488–494.

[7] Benavente, F., van der Heijden, R., Tjaden, U. R., van der Greef, J., Hankemeier, T., *Electrophoresis*. 2006, *27*, 4570–4584.

[8] Ibáñez, C., Simó, C., García-Cañas, V., Gómez-Martínez, Á., Ferragut, J. A., Cifuentes, A., *Electrophoresis*. 2012, *33*, 2328–2336.

[9] Canuto, G. B., Castilho-Martins, E., Tavares, M. M., Rivas, L., Barbas, C., López-Gonzálvez, Á., *Anal. Bioanal. Chem*. 2014, *406*, 3459–3476.

[10] Ibáñez, C., Simó, C., Martín-Álvarez, P. J., Kivipelto, M., Winblad, B., Cedazo-Mínguez, A., Cifuentes, A., *Anal. Chem*. 2012, *84*, 8532–8540.

[11] González-Domínguez, R., García, A., García-Barrera, T., Barbas, C., Gómez-Ariza, J. L., *Electrophoresis*. 2014, *35*, 3321–3330.

[12] Farrés, M., Piña, B., Tauler, R., *Metabolomics*. 2015, *11*, 210–224.

[13] Kim, J., Choi, J. N., John, K. M. M., Kusano, M., Oikawa, A., Saito, K., Lee, C. H., *J. Agric. Food Chem*. 2012, *60*, 9746–9753.

[14] Alberice, J. V., Amaral, A. F. S., Armitage, E. G., Lorente, J. A., Algaba, F., Carrilho, E., Márquez, M., García, A., Malats, N., Barbas, C., *J. Chromatogr. A* 2013, *1318*, 163–170.

[15] Tseng, Y. J., Kuo, C.-T., Wang, S.-Y., Liao, H.-W., Chen, G.-Y., Ku, Y.-L., Shao, W.-C., Kuo, C.-H., *Electrophoresis*. 2013, *34*, 2918–2927.

[16] Sánchez Pérez, I., Culzoni, M. J., Siano, G. G., Gil García, M. D., Goicoechea, H. C., Martínez Galera, M., *Anal. Chem*. 2009, *81*, 8335–8346.

[17] Siano, G. G., Sánchez Pérez, I., Gil García, M. D., Martínez Galera, M., Goicoechea, H. C., *Talanta*. 2011, *85*, 264–275.

[18] Parastar, H., Jalali-Heravi, M., Sereshti, H., Mani-Varnosfaderani, A., *J. Chromatogr. A* 2012, *1251*, 176–187.

[19] Szymańska, E., Markuszewski, M. J., Van der Heyden, Y., Kaliszan, R., *Electrophoresis*. 2009, *30*, 3573–3581.

[20] Pont, L., Benavente, F., Barbosa, J., Sanz-Nebot, V., *J. Sep. Sci*. 2013, *36*, 3896–3902.

[21] Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., *Bioinformatics*. 2008, *24*, 2534–2536.

[22] Tomasi, G., van den Berg, F., Andersson, C., *J. Chemometr*. 2004, *18*, 231–241.

[23] Joliffe, I., Morgan, B., *Stat. Methods Med. Res*. 1992, *1*, 69–95.

[24] Barker, M., Rayens, W., *J. Chemometr*. 2003, *17*, 166–173.

[25] Wold, S., Sjöström, M., Eriksson, L., *Chemometr. Intell. Lab*. 2001, *58*, 109–130.

[26] Wold, S., Johansson, A., Cocchi, M., in: Kubinyi, H. (Eds.), *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM Science Publishers, Leiden 1993, pp. 523–550.

[27] Wold, S., in: vande Waterbeemd, H. (Eds.), *Chemometric Methods in Molecular Design*, Verlag Chemie, Weinheim 1995, pp. 195–218.

[28] Saurina, J., in: Hanrahan, G., Gomez, F. A. (Eds.), *Chemometric Methods in Capillary Electrophoresis*, John Wiley & Sons, Inc., New Jersey 2009, pp. 199–226.

[29] Tauler, R., *Chemometr. Intell. Lab*. 1995, *30*, 133–146.

[30] Benavente, F., Andón, B., Giménez, E., Olivieri, A. C., Barbosa, J., Sanz-Nebot, V., *Electrophoresis*. 2008, *29*, 4355–4367.

[31] Jaumot, J., Gargallo, R., de Juan, A., Tauler, R., *Chemometr. Intell. Lab*. 2005, *76*, 101–110.

[32] Garcia-Reiriz, A. G., Olivieri, A. C., Teixido, E., Ginebreda, A., Tauler, R., *Environ. Sci. Process Impacts*. 2014, *16*, 124–134.

[33] Golub, G., Solna, K., Dooren, P. V., *SIAM J. Matrix Anal. Appl*. 2000, *22*, 1–19.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[34] Windig, W., Guilment, J., *Anal. Chem.* 1991, *63*, 1425–1432.

[35] Tauler, R., Barceló, D., *TrAC Trend Anal. Chem.* 1993, *12*, 319–327.

[36] Tauler, R., Smilde, A., Kowalski, B., *J. Chemometr.* 1995, *9*, 31–58.

[37] de Juan, A., Jaumot, J., Tauler, R., *Anal. Methods* 2014, *6*, 4964–4976.

[38] Smith, C. A., Maille, G. O., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., Siuzdak, G., *Ther. Drug Monit.* 2005, *27*, 747–751.

[39] Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A. C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., Wilson, M., Wishart, D. S., *Nucleic Acids Res.* 2012, *40*, D815-D820.

[40] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., *Nucleic Acids Res.* 2012, *40*, 109–114.

[41] Shamsi, S. A., Miller, B. E., *Electrophoresis.* 2004, *25*, 3927–3961.

[42] Stutz, H., *Electrophoresis.* 2005, *26*, 1254–1290.

[43] Benavente, F., Sanz-Nebot, V., Barbosa, J., van der Heijden, R., van der Greef, J., Hankemeier, T., *Electrophoresis.* 2007, *28*, 944–949.

[44] Dowd, S. R., Bier, M. E., Patton-Vogt, J. L., *J. Biol. Chem.* 2001, *276*, 3756–3763.

[45] Redón, M., Borrull, A., López, M., Salvadó, Z., Cordero, R., Mas, A., Guillamón, J. M., Rozès, N., *Yeast.* 2012, *29*, 443–452.

[46] Estruch, F., *Fems Microbiol. Rev.* 2000, *24*, 469–486.

[47] Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O., *Mol. Biol. Cell* 2000, *11*, 4241–4257.

[48] Sakaki, K., Tashiro, K., Kuhara, S., Mihara, K., *J. Biochem.* 2003, *134*, 373–384.

[49] Becerra, M., Lombardía, L. J., González-Siso, M. I., Rodríguez-Belmonte, E., Hauser, N. C., Cerdán, M. E., *Compar. Funct. Genom.* 2003, *4*, 366–375.

[50] Mensonides, F. I. C., Hellingwerf, K. J., de Mattos, M. J. T., Brul, S., *Food Res. Int.* 2013, *54*, 1103–1112.

[51] Strassburg, K., Walther, D., Takahashi, H., Kanaya, S., Kopka, J., *OMICS* 2010, *14*, 249–259.

[52] Jarolim, S., Ayer, A., Pillay, B., Gee, A. C., Phrakaysone, A., Perrone, G. G., Breitenbach, M., Dawes, I. W., *G3-Genes Genomes Genet.* 2013, *3*, 2321–2333.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Informació suplementària de l'article científic II.

Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling.

E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler.

Electrophoresis 36 (2015) 2324-2335.

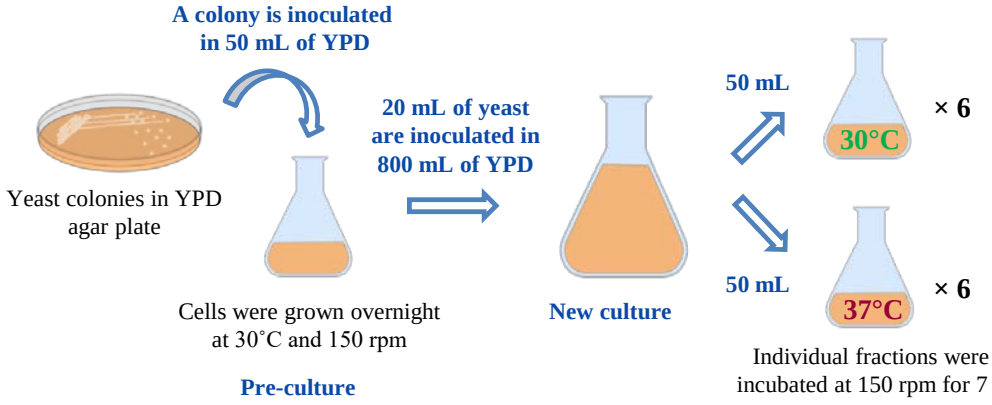


Figure S1. Scheme of the steps of the yeast culture protocol.

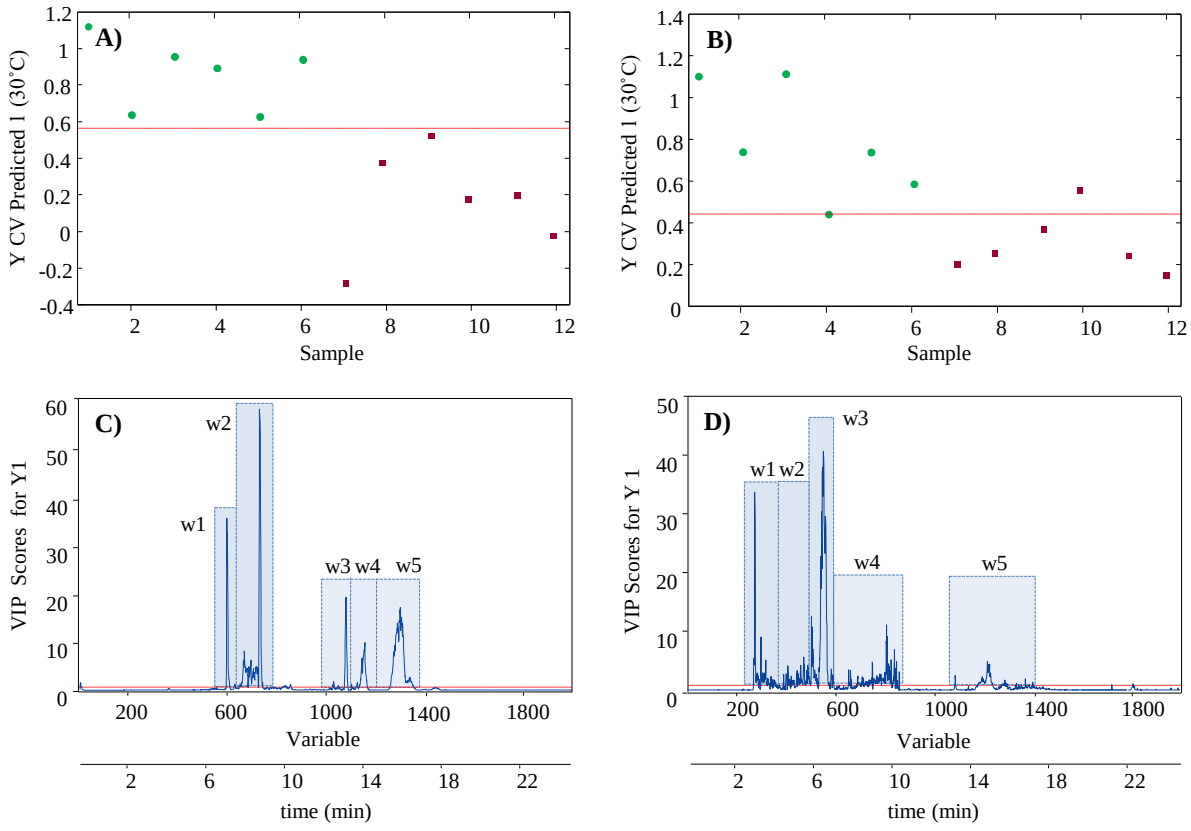


Figure S2. PLS-DA scores plots for the 12 TIE yeast extracts: (A) ESI+, and (B) ESI-. Green circles are control samples (30˚C) and red squares are exposed yeast samples (37˚C). VIP scores plots for selection the regions of interest of the electropherograms: (C) ESI+, and (D) ESI-.

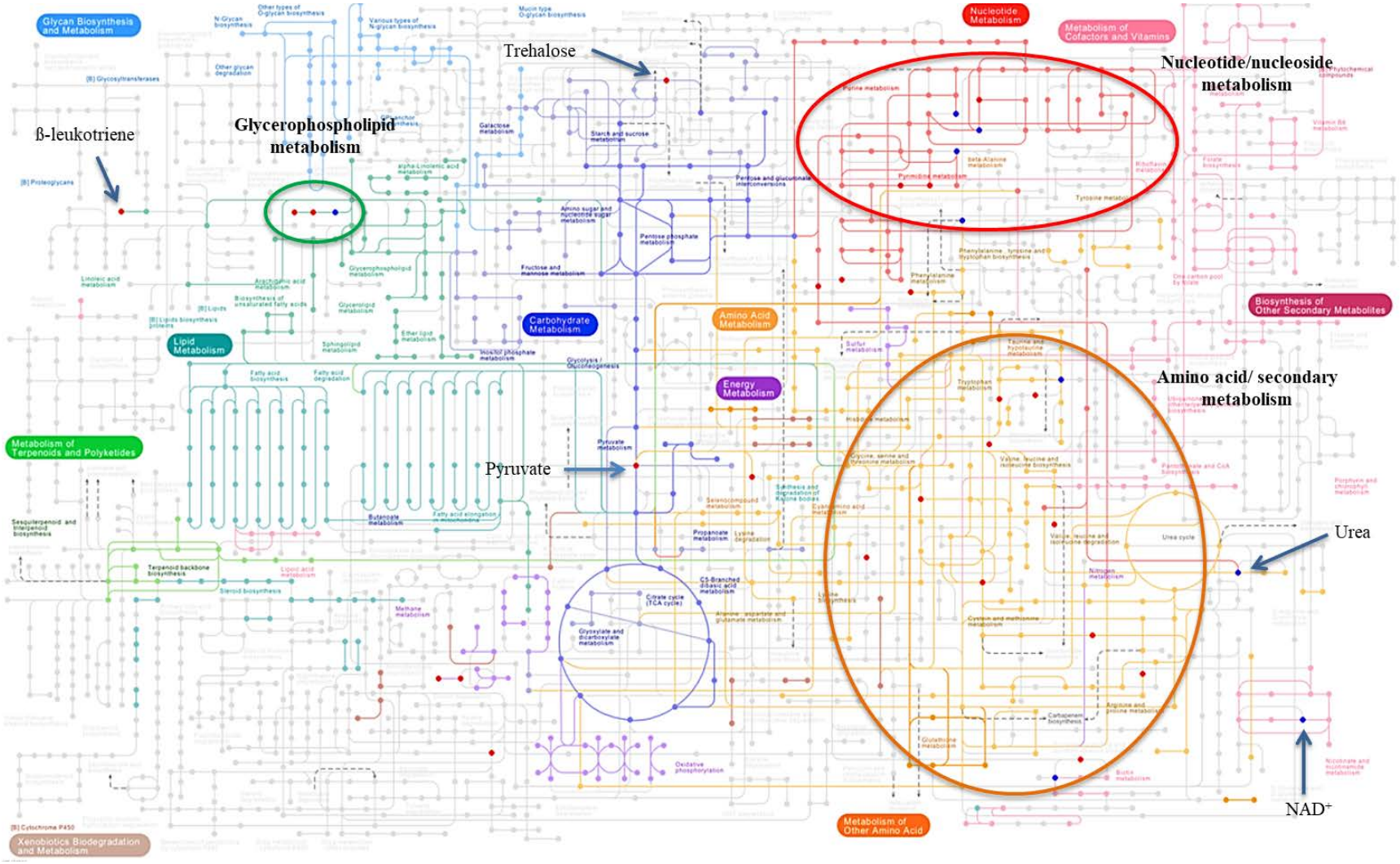


Figure S3. Metabolic pathways and mechanisms of *Saccharomyces cerevisiae* affected by temperature. Red dots show the metabolites up-regulated compared to control samples, whereas the blue dots show the down-regulated ones.

3.2.3. Article científic III.

Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data.

E. Ortiz-Villanueva, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler. J. Jaumot.

Analytica Chimica Acta 978 (2017) 10-23.

Analytica Chimica Acta 978 (2017) 10–23

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data

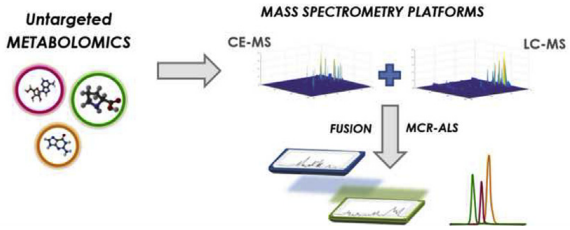
Elena Ortiz-Villanueva [a], Fernando Benavente [b], Benjamín Piña [a], Victoria Sanz-Nebot [b], Romà Tauler [a], Joaquim Jaumot [a, *]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 645, 08028 Barcelona, Spain

HIGHLIGHTS

- Two data fusion strategies were proposed for untargeted metabolomics studies.
- Data fusion and results integration approaches were based on MCR-ALS.
- Goodness of proposed strategies was proven in a metabolomic study of yeast growth.
- Proposed chemometric approaches allowed the joint analysis of CE-MS and LC-MS data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 September 2016
Received in revised form
7 March 2017
Accepted 25 April 2017
Available online 11 May 2017

Keywords:
Knowledge integration
Data fusion
Capillary electrophoresis-mass spectrometry
Liquid chromatography-mass spectrometry
MCR-ALS
Untargeted metabolomics

ABSTRACT

In this work, two knowledge integration strategies based on multivariate curve resolution alternating least squares (MCR-ALS) were used for the simultaneous analysis of data from two metabolomic platforms. The benefits and the suitability of these integration strategies were demonstrated in a comparative study of the metabolite profiles from yeast (*Saccharomyces cerevisiae*) samples grown in non-fermentable (acetate) and fermentable (glucose) carbon source. Untargeted metabolomics data acquired by capillary electrophoresis-mass spectrometry (CE-MS) and liquid chromatography-mass spectrometry (LC-MS) were jointly analysed. On the one hand, features obtained by independent MCR-ALS analysis of each dataset were joined to obtain a biological interpretation based on the combined metabolic network visualization. On the other hand, taking advantage of the common spectral mode, a low-level data fusion strategy was proposed merging CE-MS and LC-MS data before the MCR-ALS analysis to extract the most relevant features for further biological interpretation. Then, results obtained by the two presented methods were compared. Overall, the study highlights the ability of MCR-ALS to be used in any of both knowledge integration strategies for untargeted metabolomics. Furthermore, enhanced metabolite identification and differential carbon source response detection were achieved when considering a combination of LC-MS and CE-MS based platforms.

© 2017 Elsevier B.V. All rights reserved.

Abbreviations: BGE, Background electrolyte; BSA, Bovine serum albumin; CMTF, Coupled matrix and tensor factorization; COW, Correlation optimized warping; DISCO-SCA, Distinctive and common components with simultaneous component analysis; GSVD, Generalized singular value decomposition; JIVE, Joint and individual variation explained; MCR-ALS, Multivariate curve resolution alternating least squares; MWCO, Molecular weight cut-off; O2PLS, Two-way orthogonal projections to latent structures; OnPLS, Multiblock orthogonal projections to latent structures; PBS, Phosphate buffered saline; PIPES, 2,2'-(1,4-Piperazinediyl)diethanesulfonic acid; PLS-DA, Partial least squares discriminant analysis; YPD, Yeast extract peptone dextrose; YEPA, Yeast extract peptone acetate.

* Corresponding author.
E-mail address: joaquim.jaumot@idaea.csic.es (J. Jaumot).

http://dx.doi.org/10.1016/j.aca.2017.04.049
0003-2670/© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In the framework of *omic* sciences, there is a need towards the application of chemometric methods to analyse the large amount of data generated in chemical and biological studies. In the last few years, advanced data analysis tools and novel approaches have been developed to enhance the acquisition of new knowledge and improve the understanding of biological processes [1,2]. Data fusion and integration methods are nowadays the subject of active research in computational statistics and chemometrics [3–7]. These new methods allow the simultaneous study of datasets considering diverse viewpoints, such as datasets coming from different analytical platforms, *omic* levels, organisms or sample types [8,9]. For instance, in an inter-platform data fusion procedure for metabolomics, the same samples are jointly analysed using diverse methods or techniques [5,10]. In this way, the strengths of a particular analytical platform can be used to compensate the weaknesses of the rest, in an attempt to gather better metabolic information with greater accuracy and lower uncertainty. So, these strategies combining different sources of information are promising tools for fundamental *omic* studies as they allow an improved biomarker detection, hence a better characterization of biological responses.

From a chemometric point of view, extensive work has been done in the field of data fusion (that can also be known as multiset analysis). More recently, these data fusion strategies have been applied in diverse research fields and a classification of these possible different approaches has been proposed distinguishing high-level, mid-level and low-level data fusion [5,11,12]. High-level fusion implies optimal preprocessing and modelling procedures for each data block separately. Different models outputs are then jointly evaluated to provide a global overview, which is often hardly interpretable. Similarly to this high-level fusion, integration of the results obtained in these individual analyses of the blocks can yield an enhanced biological interpretation. For instance, identified features can be visualized at the same time in a metabolic or gene network to gain a deeper knowledge of the underlying biological processes, *i.e.* pathway based integration [10,13]. In contrast, low-level and mid-level fusion strategies aim to combine first data blocks to obtain later a model with an improved joint interpretation. Low-level fusion generates vast size fused data with a large amount of variables; whereas mid-level fusion is based on a previous dimensionality compression of data blocks where only a reduced number of pre-selected variables (usually the most relevant or combinations of them) from each data block are fused and jointly interpreted. Regarding the chemometric methods used for these data fusion studies, several proposals have been done considering these data fusion levels. In the case of high-level data fusion, each dataset is analysed individually using traditional chemometric methods for sample classification and feature detection [9,14–17]. More interestingly, additional efforts have been done for the development of mid-level data fusion methods. The methods found in the literature try to identify the common and specific variance coming from each one of the analysed blocks after a feature selection to reduce the size of the dataset. Various methods should be highlighted such as GSVD (generalized singular value decomposition) [18], O2PLS (two-way orthogonal projections to latent structures) [19], OnPLS (multiblock orthogonal projections to latent structures) [20], DISCO-SCA (distinctive and common components with simultaneous component analysis) [9], JIVE (joint and individual variation explained) [21], and CMTF (coupled matrix and tensor factorization) [14]. Some of these mid-level data fusion methods can also be used for low-level data fusion depending on the raw data characteristics (usually an appropriate block scaling is required). However, other methods could also be an option for this purpose, such as multivariate curve resolution alternating least squares (MCR-ALS). MCR-ALS allows the joint analysis of multiple datasets from different samples (experiments) or techniques or from different samples and techniques simultaneously. The benefits of MCR-ALS in diverse research fields have been extensively discussed in previous literature not related to the *omic* sciences for the join analysis of different data sources (*i.e.* spectroscopies, physical or chemical parameters) or different samples (*i.e.* biological processes, chromatographic runs) [22–26]. In addition, to the best of our knowledge, this is the first study investigating this low-level data fusion strategy in CE-MS and LC-MS inter-platforms for untargeted metabolomics.

There is a broad variety of instrumental techniques that can be used for *omic* studies. Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the most widely used techniques in metabolomics [27,28]. NMR is accepted as the most advisable platform to obtain reproducible results with negligible coefficients of variation, but its sensitivity is low. In contrast, hyphenated MS-based platforms are highly recommended for the great selectivity and sensitivity. Gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) provide easy and reliable metabolite separation, detection, identification and quantification. These platforms often produce data multisets containing partly complementary information (in terms of selectivity of the separations, polarity of the compounds detected and concentration ranges) that jointly analysed may reveal underlying pathways in highly complex samples, difficult to extract otherwise. Particularly, CE-MS presents some properties that complement the more commonly used techniques (reversed-phase LC-MS and GC-MS). Thus, CE-MS provides information about charged and highly polar compounds in a fast and simple way (without the need of chemical derivatization) and using an extremely small volume of sample. On particular occasions, these multiple analytical platforms can be useful to cover all the changes induced in an organism by an external stimulus. However, the datasets from these diverse analytical sources can be heterogeneous, and the analysis is still challenging. Several articles have proved the importance of data fusion models in untargeted *omic* studies, using an inter-platform fusion of ^{1}H NMR and LC-MS [29], ^{1}H NMR and GC-MS [30] or LC-MS and GC-MS [31,32]. However, to the best of our knowledge, only a few of these studies have been previously reported using CE (*e.g.* CE and NMR) [33].

The main aim of this work is the proposal of two different inter-platform knowledge integration strategies based on the application of MCR-ALS for the joint analysis of untargeted metabolomics data. The suitability of these two data analysis strategies presented in this work is demonstrated in a comparative study of the metabolic changes induced in yeast (*Saccharomyces cerevisiae*) growth by a non-fermentable carbon source (acetate) with regard to the typical fermentable carbon source (glucose). The selection of this biological system is due to the advantages of yeast as an optimal model organism. First, yeast metabolism is well known and it is essentially as complex as any other eukaryotic organisms. This knowledge facilitates the identification of the detected metabolites, the interpretation of the observed metabolic changes and the extrapolation of the biological interpretation to any other eukaryote. The changes in the yeast metabolome derived from growing on fermentable or non-fermentable carbon sources are of major interest for biotechnology and food industries. These industries rely on the unique yeast metabolic properties for a vast number of applications, from bakery, brewery, and wine-making to the production of different recombinant proteins or its use as single-cell catalysts in fine chemistry [34–36].

2. Materials and methods

2.1. Chemicals and reagents

All chemicals used in the preparation of buffers, mobile phases and solutions were analytical reagent grade or better. Formic acid, glacial acetic acid, ammonium acetate, potassium acetate, acetonitrile and 2-propanol (HPLC and MS grade) were purchased from Merck (Darmstadt, Germany). Chloroform was supplied by Carlo Erba (Peypin, France). Bacteriological peptone, yeast extract, D-glucose, sodium chloride, disodium hydrogen phosphate, potassium dihydrogen phosphate, potassium chloride, sodium hydroxide (NaOH) and bovine serum albumin (BSA) were provided by Sigma-Aldrich (St. Louis, MO, US). Water with conductivity lower than 0.05 μS cm^{-1} was obtained using a Milli-Q water purification system (Millipore, Molsheim, France).

L-methionine sulfone and 2,2'-(1,4-piperazinediyl)diethanesulfonic acid (PIPES), used as internal standards (IS), were also supplied by Sigma-Aldrich.

2.2. Apparatus

pH measurements were performed using a Crison 2002 potentiometer and a Crison electrode 52–03 (Crison Instruments, Barcelona, Spain). Centrifugation was carried out in a Serie Digicen 21 centrifuge (Ortoalresa, Madrid, Spain). Sample incubation was performed with an Innova 40/40R incubator shaker (New Brunswick Scientific, Edison, NJ, US).

2.3. Procedures

2.3.1. Sample preparation

2.3.1.1. Yeast strains. A colony of *S. cerevisiae* BY4741 was grown overnight at 30 °C and 150 rpm in 60 mL of non-selective medium (yeast extract peptone dextrose medium, YPD/20 g L^{-1} bacteriological peptone, 10 g L^{-1} yeast extract and 20 g L^{-1} glucose) so as to obtain the presporulation media (pre-culture). Then, two larger sporulation media (cultures) were prepared in 1 L flasks: one containing 950 mL of YPD medium and the other with 950 mL of yeast extract peptone acetate, YEPA (20 g L^{-1} bacteriological peptone, 10 g L^{-1} yeast extract, and 20 g L^{-1} potassium acetate). Both flasks were inoculated with 28 mL of yeast pre-culture to an optical density of 0.1 at 600 nm (OD600). Thereafter, six fractions of 150 mL from each stock culture were individually incubated at 150 rpm at 30 °C until 0.9 of OD600, and the cultures were then placed on ice. Potassium acetate as a carbon source resulted in a decrease in the growth rate (see Supplementary Material Fig. S1).

2.3.1.2. Yeast metabolite extraction. Yeast cells harvest was performed by centrifugation of each culture at 3300*g* for 5 min at 4 °C discarding the supernatant. Then, pellets were washed three times with 3 mL of phosphate buffered saline (PBS) at 3300 g for 5 min at 4 °C. The clean pellets were snap-frozen in liquid nitrogen and freeze-dried overnight.

From each freeze-dried yeast pellet, 20 mg and 8 mg were separately weight into Eppendorf tubes for CE-MS and LC-MS samples, respectively. Next, metabolites were extracted with a mixture of 250 μL of cold methanol and 250 μL of cold water containing the internal standard (IS1) (L-methionine sulfone) at a final concentration of 5 μg mL^{-1}. After vortexing 30 s, the mixture was centrifuged at 11000*g* for 15 min at 4 °C to isolate the supernatant, and 350 μL of cold chloroform were added. The mixtures were vortexed 30 s, placed on ice for 10 min and centrifuged again at 11000*g* for 15 min at 4 °C. Aqueous fractions were separated and finally evaporated to dryness under nitrogen.

Before the analysis, samples were reconstituted with 100 μL of water (CE-MS) or 100 μL of 1:1 acetonitrile:water (v/v) (LC-MS) containing PIPES (IS2) at a concentration of 5 μg mL^{-1}. For CE-MS analysis, the yeast extracts were filtered using 3 kDa molecular weight cut-off (MWCO) cellulose acetate filters (Amicon® Ultra-0.5 filters, Millipore, Bedford, MA, US) previously passivated with a solution of 1% (m/v) of BSA in PBS (10 mM disodium hydrogen phosphate, 1.5 mM potassium dihydrogen phosphate, 140 mM sodium chloride and 2.7 mM potassium chloride, pH 7.2) to avoid metabolite loss through adsorption on the inner walls of the plastic sample reservoir. For LC-MS analysis, the extracts were only filtered through 0.22 μm filters (Ultrafree®-MC, Millipore Bedford, MA, US) at 11000 g for 4 min at 4 °C.

For each internal standard, an aqueous solution (1000 μg mL^{-1}) was prepared and stored in the freezer at −20 °C until its use. Working standard solutions were obtained by diluting the stock solutions with water. Diluted standard solutions were used to spike yeast extract samples. Quality control (QC) samples were generated by pooling 15 μL of all the studied samples (extracts) and taking it into 2 mL vials.

2.3.2. CE-MS

All CE-MS experiments were performed in an HP^{3D} CE system coupled with an orthogonal G1603A sheath-flow interface to a 6220 oa-TOF LC/MS mass spectrometer (Agilent Technologies, Waldbronn, Germany). The sheath liquid consisted of a hydro-organic mixture of 60:40 (v/v) 2-propanol:water with 0.05% (v/v) of formic acid and was delivered at a flow rate of 3.3 μL min^{-1} by a KD Scientific 100 series infusion pump (Holliston, MA, US). CE control and separation data acquisition were performed using ChemStation Software (Agilent Technologies) that was running in combination with the MassHunter workstation software (Agilent Technologies) for control and data acquisition of the TOF mass spectrometer.

A 72 cm total length (L_T), 75 μm inner diameter and 360 μm outer diameter bare fused-silica capillary was used for the electrophoretic separations at 25 °C, and all capillary rinses were performed at 930 mbar. All bare fused-silica capillaries were supplied by Polymicro Technologies (Phoenix, AZ, US). New capillaries were flushed with 1 M NaOH (20 min), water (15 min) and background electrolyte (BGE) (30 min, acetic acid 1.0 M). The system was finally equilibrated by applying for 15 min the separation voltage (25 kV, normal polarity, *i.e.* anode in the inlet). Between workdays, the capillary was conditioned by rising successively with 0.1 M NaOH (5 min), water (10 min), and BGE (15 min). Both activation and conditioning procedures were performed off-line in order to avoid the unnecessary entrance of NaOH into the mass spectrometer. All samples (six glucose- and six acetate-grown) were randomly injected at 50 mbar for 5 s. During the whole electrophoretic experiment, six quality control and six blank samples were injected to control batch quality. Between runs, the capillary was rinsed for 3 min with BGE. All solutions were passed through a 0.22 μm nylon filter (MSI, Westboro, MA, US) before analysis and were stored at 4 °C. BGE and sheath liquid solutions were degassed for 15 min by sonication before use.

The TOF mass spectrometer was operated in positive mode using the following parameters: capillary voltage 4000 V, drying gas temperature 250 °C, drying gas flow rate 4 L min^{-1}, nebulizer gas 7 psig, fragmentor voltage 150 V, skimmer voltage 65 V and OCT 1 RF Vpp voltage 300 V. Data were collected in profile mode at 1 spectrum/s (approximately 10000 transients/spectrum) in the *m/z* range of 85–1000, working in the extended dynamic range mode (2 GHz) with the mass range set to standard (*m/z* 1700).

2.3.3. LC-MS

LC-MS experiments were carried out in an Agilent Infinity 1200

series system coupled with an orthogonal G1385-44300 interface (Agilent Technologies) to the same TOF mass spectrometer. LC and MS control, separation and data acquisition were performed using MassHunter workstation software. The TOF mass spectrometer measurement parameters were as described for CE-MS, except for the drying gas temperature (350 °C), drying gas flow rate (8 L min^{-1}) and nebulizer gas (30 psig).

For the chromatographic separations, a TSK Gel Amide-80 column (250 mm length, 2.1 mm inner diameter and 5 μm particle size) from Tosoh Bioscience (Tokyo, Japan) was used at room temperature with gradient elution at a flow rate of 0.15 mL min^{-1}. HILIC chromatographic mode was employed, and elution solvents were acetonitrile (**A**) and 5 mM of ammonium acetate adjusted to pH 5.5 with acetic acid (**B**). The elution gradient was as follows: 0–8 min, linear gradient from 25 to 30% **B**; 8–12 min, from 30 to 60% **B;** 12–17 min, 60% **B;** 17–20 min, back from 60% to 25% **B**; and from 20 to 27 min, 25% **B**. Solvents were degassed for 15 min by sonication before use. Sample injection was performed with an autosampler at 4 °C, and the injection volume was 5 μL. All samples (six glucose-grown and six acetate-grown) were randomly injected. Quality control and blank samples were also injected at regular intervals during the whole chromatographic batch.

2.4. Data analysis

MCR-ALS was applied to resolve the metabolite profiles from the electrophoretic and chromatographic runs for an untargeted metabolomic analysis. Fig. 1 shows a summary of the datasets and MCR-ALS analysis workflow, which is detailed in this section.

2.4.1. Data pre-processing: conversion, ROI compression and data arrangement

In the first stage of the data processing, the accurate mass full scan raw MassHunter chromatograms and electropherograms were converted to mzXML data files by ProteoWizard software [37] and imported to MATLAB (The Mathworks Inc. Natick, MA, US). Due to the vast size of the full MS scan files in profile mode and the storage requirements, the input full resolution data was compressed through "regions of interest" (ROI) search (Fig. 1A) [38–40]. The usefulness of ROI compression has been previously demonstrated elsewhere [39,41]. Using this compression, for every sample (chromatographic or electrophoretic run) ROI values are searched among every migration or elution time of the considered electrophoretic or chromatographic run. These vectors are reorganized into a matrix grouping the common ROIs among the different migration or elution times and the final *m/z* values of each ROI are calculated as the mean of the *m/z* values obtained for that specific ROI (see Supplementary Material Fig. S2A, example of an LC-MS sample). This strategy allowed obtaining for every considered sample a data matrix with a number of rows equals to the total number of elution (or migration times) of the considered run and a number of columns equals to the total number of identified ROIs.

ROI approach requires the input of a signal-to-noise threshold value, mass accuracy (*i.e. m/z* error) and the minimum number of migration or elution times to be considered as a peak for each ROI. In our case, these parameters were respectively set at 0.1% of maximum MS intensity signal (threshold) independently observed for the entire set of LC-MS and CE-MS samples. Mass accuracy was set to 0.05 Da/e in both LC-MS and CE-MS batches because the same MS instrument was used, and the minimum number of elution or migration times in a peak was set to 25. When this approach is applied, a relatively low number of *m/z* values were considered to be further investigated. Then, the intensity scale of each MS-ROI data matrix (CE or LC compressed matrix, Fig. 1A) was normalized taking into account the area of the internal standards; to correct the instrumental intensity drifts among injections and to adjust the data scale. After normalization, individual MS-ROI matrices were arranged in three different column-wise augmented data matrices to find common and uncommon ROI values among all considered samples in each case. These column-wise configurations were obtained by a pairwise search of ROI among the previous generated individual ROI data matrices (corresponding to each chromatographic or electrophoretic run). In the end, this MS-ROI augmentation strategy provided data matrices with an equal number of *m/z* values (column of the final data matrix) (see Supplementary Material Fig. S2B for an example of an LC-MS-ROI augmentation strategy). The dimensions of the ROI augmented data matrices were the total number of migration or elution times considered in the whole set of samples (rows) and the total number of considered *m/z* ROI values (columns). This augmentation strategy was based on a first step of a pairwise combination of glucose- and acetate-grown samples for each metabolomic platform. Next, all the samples obtained using one technique were combined to obtain the data matrix to be analysed. The first MS-ROI augmented data matrix (Multiset CE-MS, $\mathbf{D_{augCE}}$) was built with ROI search between the six glucose- and the six acetate-grown samples analysed by CE-MS (Fig. 1A, left). The second MS-ROI augmented data matrix (Multiset LC-MS, $\mathbf{D_{augLC}}$) was obtained with the six glucose- and the six acetate-grown samples from different individual LC-MS analyses (Fig. 1A, right). Finally, the third MS-ROI augmented data matrix (JOINT dataset, $\mathbf{D_{augTOTAL}}$) was obtained after a pairwise search of ROI among CE-MS-ROI and LC-MS-ROI multisets, as is shown in Fig. 1C. These augmented data matrices were considered for the evaluation of the different integration strategies. First, results obtained in the independent analysis of $\mathbf{D_{augCE}}$ and $\mathbf{D_{augLC}}$ were merged for further biological interpretation (Fig. 1B). Second, low-level data fusion was evaluated by the analysis of $\mathbf{D_{augTOTAL}}$ which contained information from both CE-MS and LC-MS platforms (Fig. 1C).

2.4.2. MCR-ALS analysis for metabolite detection and identification

MCR-ALS is a powerful chemometric method especially useful to analyse multicomponent systems with strongly overlapped contributions [42,43], such as those present in chromatograms and electropherograms of complex mixtures in metabolomics studies [44,45]. In this study, each platform provided an MS data matrix **D** where the rows were the experimental mass spectra at all migration or elution times, and the columns were the electropherograms (CE) or chromatograms (LC) at every *m/z* value. After ROI compression, MCR-ALS decomposed an individual MS-ROI compressed data matrix **D** using a bilinear model that produced the concentration profiles and MS spectra of the resolved contributions. MCR-ALS analysis of the data matrix **D** gave two factor matrices, **C** and $\mathbf{S^T}$, as in Eq. (1):

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

where **C** is the matrix of electrophoretic or chromatographic profiles of the resolved contributions (components), $\mathbf{S^T}$ contains their mass spectra, and **E** is the matrix containing the residuals unexplained by the model. Peaks resolved in matrix **C** are allowed to vary in position (shifts) and shape among samples. This aspect is especially useful in the case of CE data where migration shifts among samples can be significant and, therefore, the alignment of electrophoretic peaks before MCR-ALS analysis is not needed [44].

The singular value decomposition (SVD) method was used to

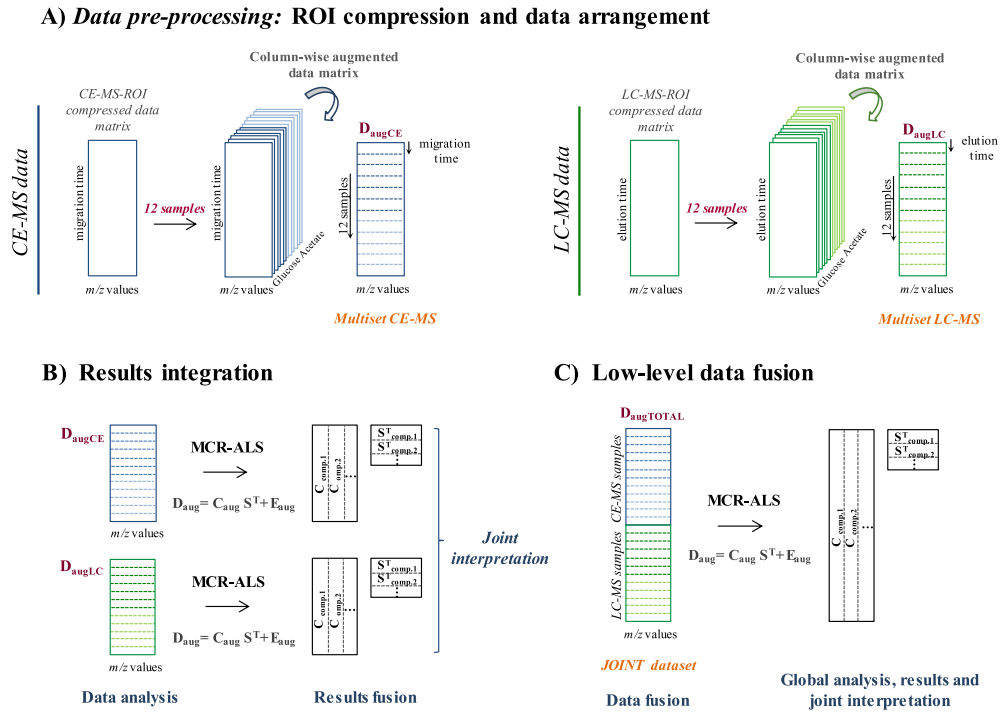


Fig. 1. Data analysis workflow: (A) Pre-processing and data arrangement, (B) Independent analysis for results integration and (C) Low-level data fusion before MCR-ALS analysis to detect and tentatively identify relevant metabolites.

establish the initial guess of the number of components (independent contributions) to the observed data variance [46]. After this initial guess, several MCR-ALS models were carried out changing the number of components (around ±10% of the initial number of components) to evaluate the on the models as a function of the different number of components considered. Subsequently, initial estimations of either elution or migration profiles (**C** matrix) or mass spectra ($\mathbf{S^T}$ matrix) were determined to start the ALS optimization. An adaptation of the SIMPLISMA method for the detection of purest variables was used considering the preliminary estimated number of components and the noise level threshold of 10% [47]. Finally, an ALS optimization was performed under non-negativity constraints for electrophoretic or chromatographic (**C**) and spectra ($\mathbf{S^T}$) profiles, as well as spectral normalization (equal height) to provide chemical meaning to the resolved elution and mass spectra profiles and minimize possible intensity or rotational ambiguities [22,48,49]. The fit quality was evaluated from the explained data variance (R^2) and the percentage of lack of fit (*LOF*) [22].

2.4.2.1. Independent MCR-ALS analysis of LC-MS and CE-MS data. The first strategy consists of the combination of the results obtained considering independently the augmented MS-ROI data matrices for CE-MS and LC-MS multisets to obtain a global visualization of the affected pathways (Fig. 1B). $\mathbf{D_{augCE}}$ and $\mathbf{D_{augLC}}$, which contained 12 samples (six glucose- and six acetate-grown samples), were separately analysed by MCR-ALS.

For every resolved MCR-ALS component, peak areas of glucose- and acetate-grown samples (on the columns of $\mathbf{C_{aug}}$ matrix, Eq. (1)) were statistically assessed using a nonparametric Mann–Whitney *U* test. A False Discovery Rate (FDR) for multiple comparison testing using the Benjamini–Hochberg procedure [50] in their areas was then used to control whether there was a significant difference between the means of the two groups of profile areas. Only resolved components that showed a statistically significant difference among groups (p-value<0.05) after FDR assessment were selected. Next, their corresponding mass spectra profiles ($\mathbf{S^T}$, Eq. (1)) were used to identify the *m/z* values causing the differentiation between both groups of samples. Finally, the accurate *m/z* values of these differential metabolites were searched in on-line databases, such as Yeast Metabolome Database (YMDB) [51] and METLIN metabolite database [52], to be tentatively identified. The lists of the tentatively identified metabolites by CE-MS and LC-MS were jointly analysed to obtain a global interpretation. Assignation of the different metabolites to standard yeast metabolic pathways was performed using the KEGG Pathway analysis tool [53].

2.4.2.2. MCR-ALS for low-level data fusion. A novel strategy for the low-level data fusion of CE-MS and LC-MS data was proposed based on the joint analysis by MCR-ALS of the multiset containing information from both platforms. CE-MS and LC-MS multisets were further merged in a column-wise augmented data matrix containing the 24 samples after ROI search (Fig. 1C). Prior to multiset merging, a scaling factor (*i.e.* CE-MS data was divided by 3 taking into account the ratio of the first singular value calculated for each multiset) was applied to balance the intensity differences among CE-MS and LC-MS intensity signals. This scaling forced that the total variance explained by each one of the blocks was similar to give equal weight during the MCR-ALS resolution (in the case of the independent analysis of the different blocks, this scaling was not required).

This multiset structure also hold an extended bilinear model, as applied for the individual multisets in the previous section in accordance with Eq. (1):

$$\mathbf{D_{augTOTAL}} = \begin{bmatrix} \boldsymbol{D_{CE\_G,1}} \\ \boldsymbol{D_{CE\_G,2}} \\ \vdots \\ \boldsymbol{D_{CE\_A,5}} \\ \boldsymbol{D_{CE\_A,6}} \\ \boldsymbol{D_{LC\_G,1}} \\ \boldsymbol{D_{LC\_G,2}} \\ \vdots \\ \boldsymbol{D_{LC\_A,5}} \\ \boldsymbol{D_{LC\_A,6}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C_{CE\_G,1}} \\ \boldsymbol{C_{CE\_G,2}} \\ \vdots \\ \boldsymbol{C_{CE\_A,5}} \\ \boldsymbol{C_{CE\_A,6}} \\ \boldsymbol{C_{LC\_G,1}} \\ \boldsymbol{C_{LC\_G,2}} \\ \vdots \\ \boldsymbol{C_{LC\_A,5}} \\ \boldsymbol{C_{LC\_A,6}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \boldsymbol{E_{CE\_G,1}} \\ \boldsymbol{E_{CE\_G,2}} \\ \vdots \\ \boldsymbol{E_{CE\_A,5}} \\ \boldsymbol{E_{CE\_A,6}} \\ \boldsymbol{E_{LC\_G,1}} \\ \boldsymbol{E_{LC\_G,2}} \\ \vdots \\ \boldsymbol{E_{LC\_A,5}} \\ \boldsymbol{E_{LC\_A,6}} \end{bmatrix}$$
$$= \mathbf{C_{aug}S^T} + \mathbf{E_{aug}} \quad (2)$$

where subindexes *CE* and *LC* indicate the analytical platform, *G,n* (n = 1, …,6) the glucose-grown yeast sample and *A,n* (n = 1, …,6) the potassium acetate-grown sample.

$\mathbf{D_{augTOTAL}}$ contains data from all the experiments (CE-MS and LC-MS runs); $\mathbf{C_{augTOTAL}}$ the resolved electrophoretic and chromatographic profiles in each sample and $\mathbf{S^T}$ their corresponding mass spectra. This approach takes advantage of the fact that the MS spectra obtained for both platforms (LC-MS and CE-MS) should have common features. Therefore, metabolites that showed contributions for the two considered platforms should be reduced in a single component characterized by its mass spectrum and with a set of different CE and LC profiles in $\mathbf{C_{augTOTAL}}$ giving quantitative information related to CE-MS or LC-MS runs. Finally, peak areas of the CE and LC migration or elution profiles of all the resolved MCR-ALS components were statistically assessed and interpreted as described in the previous section, to identify the potential metabolites for glucose- and acetate-grown yeast samples differentiation and the affected metabolic pathways.

2.4.3. Software

Calculations and data analyses were performed using MATLAB R2016a (The Mathworks Inc. Natick, MA, USA) and MCR-ALS toolbox [16].

3. Results and discussion

This section describes the information that can be gathered using MCR-ALS in an inter-platform fusion procedure of untargeted metabolomics data acquired using CE-MS and LC-MS. As a study case, the effects on the yeast metabolome caused by a non-fermentable carbon source (potassium acetate) were compared with those observed with a fermentable carbon source (glucose).

Before the different integration strategies, the use of ROI approach allowed a valuable compression of each sample data matrix without loss of spectral accuracy as mentioned in section 2.4.1 [39]. Each one of the CE-MS and LC-MS samples were compressed to a single data matrix of size 1500 rows (total number of migration or elution times) and, approximately, 1200 (CE-MS) or 1600 (LC-MS) columns (number of *m/z* values providing relevant information, from 85 to 1000 Da/e). After ROI compression and pairwise evaluation, two ROI column-wise augmented data matrices were obtained ($\mathbf{D_{augCE}}$ and $\mathbf{D_{augLC}}$). On one hand, $\mathbf{D_{augCE}}$ contained 18000 rows (migration times) and 1699 columns (*m/z* values). On the other hand, $\mathbf{D_{augLC}}$ had 18000 rows (elution times) and 1686 columns (Fig. 1B). An example of the full scan CE-MS and LC-MS data of the different yeast extracts after ROI compression is given in Fig. 2. Both CE-MS electropherograms and LC-MS chromatograms presented complex profiles with comigrating or coeluting compounds at a broad range of concentrations (see zoomed samples). The visual inspection of the electropherograms and chromatograms already showed distinct features that permitted the differentiation of the samples according to the carbon source used in the yeast growth. However, in these cases, due to the complexity of the profiles, advanced chemometric methods are necessary to interpret the information in detail. Quality of the chromatographic and electrophoretic runs was assessed by the analysis of QC samples [54,55]. Instrumental conditions during the entire LC-MS and CE-MS batches were stable and, no further corrections were applied. Preliminary analysis of QC samples also allowed discarding the presence of outliers as all glucose- and acetate-grown samples had a similar within group behavior, both for LC-MS and CE-MS platforms.

3.1. Results integration: CE-MS and LC-MS individual analysis

MCR-ALS was independently performed on $\mathbf{D_{augCE}}$ and $\mathbf{D_{augLC}}$. Each data multiset had the migration or elution and spectral information for the 12 yeast samples (six glucose-grown and six acetate-grown yeast samples) (Fig. 1B). MCR-ALS analysis allowed the resolution of migration or elution and spectra profiles of the yeast metabolites with the different carbon sources. The MCR-ALS

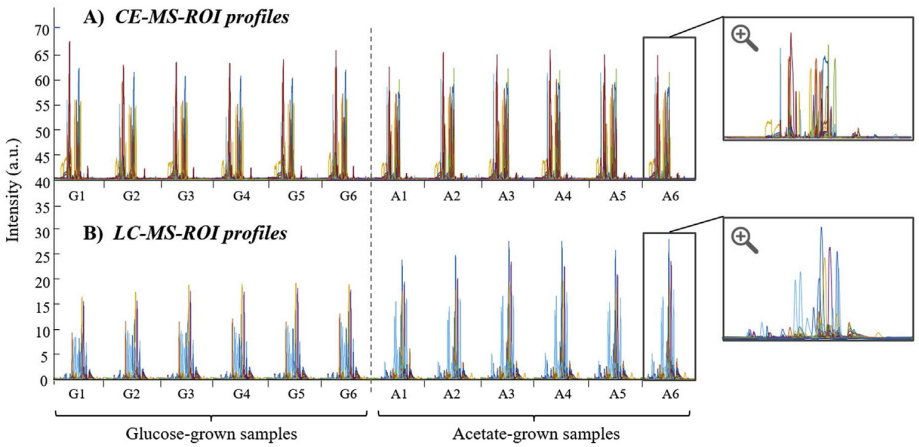


Fig. 2. CE-MS electropherograms and LC-MS chromatograms of multiple samples after ROI compression. (A) Electropherograms for the six glucose-grown samples are on the left (G1-G6) whereas electropherograms for the six acetate-grown samples are on the right (A1-A6). (B) Chromatograms for the six glucose-grown samples are on the left (G1-G6) whereas chromatograms for the six acetate-grown samples are on the right (A1-A6). As an example a zoomed view of samples A6 are depicted for both CE-MS and LC-MS.

results for each augmented data matrix ($\mathbf{D_{augCE}}$ and $\mathbf{D_{augLC}}$) (R^2, *LOF*, number of resolved components, number of differential components between glucose- and acetate-grown samples and number of differential metabolites identified) are shown in Table 1.

The R^2 values in both cases were higher than 99%, and the LOF values were lower than 3% considering a large number of components (approximately, 100 components in both cases). Due to the noise filtering properties of the ROI compression procedure which removes meaningless signals below an intensity threshold, both R^2 and *LOF* parameters presented excellent values. MCR-ALS resolved components according to their migration/elution profiles and mass spectra, grouping the information of several features related to the same compound together in a single MCR-ALS component. In the positive ion mode, these features usually correspond to isotopic peaks of the nominal *m/z* value with hydrogen cations or other adducts (*e.g.* with Na^+ or K^+) of the same compound. In addition, a few number of MCR-ALS resolved components could be associated with solvent or background contributions. A statistical analysis of the resolved MCR-ALS profile areas was performed to find the components that showed significant differences among glucose- and potassium acetate-grown samples by CE-MS and LC-MS (similar results were obtained if multivariate PLS-DA model were

Table 1
MCR-ALS data fitting results for the analysis of the MS-ROI augmented data matrices.

MCR-ALS results						
Dataset	No. of ROI (*m/z* values)	R^2 (%)	*LOF* (%)	No. of components	No. of differential components	No. of differential metabolites identified
$\mathbf{D_{augCE}}$	1699	99.9	1.9	103	47	40
$\mathbf{D_{augLC}}$	1686	99.9	2.7	98	36	32
$\mathbf{D_{augTOTAL}}$	2665	98.7	7.1	160	71	60

Table 2
List of identified metabolites that show statistically significance to differentiate between the two carbon-source samples in CE-MS (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] ≤10 ppm).

N	Metabolite	Molecular formula	Ion assignment	Measured mass (Da)	Error (ppm)	Fold-change	Trend[b]	t_m (min)	*p*-value
1	4-Aminobutanoate (GABA)	C4H9NO2	$[M+H]^+$	104.0715	8.6	3.8	UP	10.7	0.005
2	**L-Serine**	C3H7NO3	$[M+H]^+$	106.0507	7.5	2.7	DOWN	11.3	0.005
3	**L-Proline**	C5H9NO2	$[M+H]^+$	116.0716	8.6	3.0	UP	12.6	0.005
4	**L-Valine**	C5H11NO2	$[M+H]^+$	118.0853	8.5	4.5	UP	13.1	0.005
5	L-Threonine	C4H9NO3	$[M+H]^+$	120.0664	7.5	1.5	UP	11.7	0.005
6	gamma-Amino-gamma-cyanobutanoate	C5H8N2O2	$[M+H]^+$	129.0669	8.5	75.8	UP	9.5	0.005
7	**L-Leucine**	C6H13NO2	$[M+H]^+$	132.1028	6.8	1.3	UP	11.1	0.005
8	L-Ornithine	C5H12N2O2	$[M+H]^+$	133.0983	8.3	1.7	UP	7.0	0.005
9	L-Aspartate	C4H7NO4	$[M+H]^+$	134.0455	5.2	7.5	DOWN	13.3	0.005
10	2-deoxy-D-ribose	C5H10O4	$[M+H]^+$	135.0665	9.6	3.8	DOWN	13.3	0.005
11	5-(2-Hydroxyethyl)-4-methylthiazole	C6H9NOS	$[M+H]^+$	144.0492	9.7	3.8	DOWN	6.4	0.005
12	4-guanidinobutanoic acid	C5H11N3O2	$[M+H]^+$	146.0908	8.2	7.1	UP	7.4	0.005
13	L-Glutamine	C5H10N2O3	$[M+H]^+$	147.0778	9.5	1.3	UP	11.8	0.049
14	L-Lysine	C6H14N2O2	$[M+H]^+$	147.1140	8.2	6.7	DOWN	7.2	0.005
15	**L-Glutamate**	C5H9NO4	$[M+H]^+$	148.0617	8.8	1.4	DOWN	12.3	0.009
16	**L-Methionine**	C5H11NO2S	$[M+H]^+$	150.0591	5.3	1.4	DOWN	11.8	0.031
17	**Guanine**	C5H5N5O	$[M+H]^+$	152.0581	9.3	1.2	DOWN	8.4	0.028
18	**L-Carnitine**	C7H15NO3	$[M+H]^+$	162.1135	6.2	1.9	UP	8.1	0.005
19	**L-Phenylalanine**	C9H11NO2	$[M+H]^+$	166.0879	9.6	2.5	DOWN	12.1	0.005
20	L-Arginine	C6H14N4O2	$[M+H]^+$	175.1207	9.7	1.3	DOWN	7.5	0.005
21	**Acetyllysine**	C8H16N2O3	$[M+H]^+$	189.1249	7.9	1.7	DOWN	12.0	0.005
22	**N-acetyl-L-glutamate**	C7H11NO5	$[M+H]^+$	190.0701	4.7	1.3	UP	18.6	0.005
23	S-Adenosyl-L-homocysteine	C14H20N6O5S	$[M+2H]^+$	193.0700	9.8	2.1	DOWN	9.1	0.005
24	L-Tryptophan	C11H12N2O2	$[M+H]^+$	205.0961	5.4	1.4	DOWN	9.6	0.009
25	L-Cystathionine	C7H14N2O4S	$[M+H]^+$	223.0768	9.4	2.8	UP	11.5	0.005
26	D-Erythro-imidazole-glycerol-phosphate	C6H11N2O6P	$[M+H]^+$	239.0443	6.7	2.7	DOWN	16.7	0.009
27	**Glycerophosphocholine**	C8H20NO6P	$[M+H]^+$	258.1122	8.1	1.5	DOWN	17.3	0.005
28	trans-4-Hydroxy-L-proline	C5H9NO3	$[2M+H]^+$	263.1263	9.5	1.9	UP	5.9	0.005
29	L-Asparagine	C4H8N2O3	$[2M+H]^+$	265.1153	3.8	3.1	DOWN	6.4	0.005
30	**Adenosine**[a] or	C10H13N5O4	$[M+H]^+$	268.1062	8.2	2.7	DOWN	9.3	0.005
	2′-Deoxyguanosine[a]	C10H16N5O4	$[M+H]^+$	268.1062	8.2	2.7	DOWN	9.3	0.005
31	**N6-(L-1,3-Dicarboxypropyl)-L-lysine**	C11H20N2O6	$[M+H]^+$	277.1416	7.9	11.9	UP	12.6	0.005
32	**N-(L-Arginino)succinate**	C10H18N4O6	$[M+H]^+$	291.1324	8.6	13.9	UP	10.5	0.005
33	**5′-Methylthioadenosine**	C11H15N5O3S	$[M+H]^+$	298.0982	4.7	6.2	UP	9.4	0.005
34	**Glutathione**	C10H17N3O6S	$[M+H]^+$	308.0940	9.4	1.2	DOWN	14.0	0.005
35	Cytidine monophosphate (CMP)	C9H14N3O8P	$[M+H]^+$	324.0604	4.0	3.9	DOWN	17.4	0.005
36	**Trehalose**	C12H22O11	$[M+H]^+$	343.1212	6.7	5.0	UP	17.2	0.005
37	**Adenosine monophosphate (AMP)**[a] **or**	C10H14N5O7P	$[M+H]^+$	348.0735	8.9	1.3	DOWN	17.5	0.005
	Deoxyguanosine monophosphate (dGMP)[a]	C10H14N5O7P	$[M+H]^+$	348.0735	8.9	1.3	DOWN	17.5	0.005
38	5-Amino-6-(5'-phosphoribosylamino)uracil	C9H15N4O9P	$[M+H]^+$	355.0621	7.9	2.1	UP	11.9	0.005
39	Adenylsuccinic acid	C14H18N5O11P	$[M+H]^+$	464.0839	5.6	4.8	UP	26.7	0.005
40	**Nicotinamide adenine dinucleotide (NAD^+)**	C21H27N7O14P2	$[M+H]^+$	664.1189	3.8	1.2	DOWN	17.8	0.018

Metabolites in bold letters were also detected by LC-MS (Table 3).
[a] These metabolites provided an identification conflict.
[b] Trend was determined taking as a reference the response of control samples.

generated, see Supplementary Material). Only the components that showed a statistically significant difference were selected as relevant (47 and 36 components in CE-MS and LC-MS, respectively, Table 1). Taking advantage of the accurate and high-resolution mass measurements and since no loss of performance was expected after ROI peak compression, most of these differential components were tentatively identified as metabolites from the molecular mass values measured in the MCR-ALS resolved pure mass spectra ($\mathbf{S}^{\mathbf{T}}$), searching in YMDB and METLIN databases (40 and 32 metabolites in CE-MS and LC-MS, respectively, Table 1). Tables 2 and 3 show the tentatively identified metabolites that led to differentiate among glucose- and acetate-grown samples by CE-MS and LC-MS, respectively. Metabolite identity, detected ion, measured mass, *m/z* relative error (≤10 ppm), migration or elution time, adjusted *p*-value (<0.05), fold change and their folding trends are given. It is also described if metabolites were detected by only one or both analytical platforms.

Then, features independently obtained for each platform were integrated for the joint evaluation and interpretation of the results. Fig. 3A shows a Venn diagram with the relationships among the total number of metabolites identified, highlighting the common metabolites and those specific for each platform (see Tables 2 and 3 for the identities). Only 20 out of 40 (CE-MS) and 20 out of 32 (LC-MS) metabolites were common between both techniques. Therefore, for a comprehensive differentiation it was also necessary to consider the specific metabolites for CE-MS (20) and LC-MS (12). In Tables 2 and 3, the identities of these metabolites are given, suggesting that both platforms provided partly complementary information that allows a more reliable detection and identification of metabolites to differentiate between glucose- and acetate-grown samples. As it can be seen, both techniques retrieved a large number of amino acids and nucleotides. However, CE-MS appeared to provide higher selectivity for amino acids and their intermediates, whereas LC-MS analyses also retrieved some lipidomic information (*e.g.* LysoPC(16:1(9Z)) and LysoPC(18:1(9Z))).

3.2. Low-level data fusion: CE-MS and LC-MS joint analysis

CE-MS and LC-MS raw data were also jointly analysed building an MS-ROI column-wise augmented data matrix containing the 24 CE-MS and LC-MS samples (Fig. 1C). After ROI compression and pairwise evaluation, $\mathbf{D}_{\mathbf{augTOTAL}}$ contained 36000 rows (number of migration and elution times) and 2665 columns (*m/z* values, from 85 to 1000 Da/e). In this case, a total number of 160 MCR-ALS components were resolved with an R^2 higher than 98% and a *LOF* value lower than 8% (see Table 1). Fig. 4 shows an example of four of the resolved MCR-ALS components when samples from both platforms are considered (in the case of the independent analysis discussed in the previous section a similar figure could be obtained but with samples of a single metabolomics platform, either CE-MS or LC-MS). The electrophoretic and chromatographic profiles ($\mathbf{C}_{\mathbf{aug}}$) of these four components (Fig. 4A) and their corresponding pure spectra ($\mathbf{S}^{\mathbf{T}}$) (Fig. 4B) are illustrated. As it can be observed, two of the features (MCR-ALS components) were specific for CE-MS (blue line, 223 *m/z*) and LC-MS (green line, 245 *m/z*), whereas the other two features were common (red, 166 *m/z* and orange, 277 *m/z*).

Table 3
List of identified metabolites that show statistically significance to differentiate between the two carbon-source samples for LC-MS data (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] ≤10 ppm).

N	Metabolite	Molecular formula	Ion assignment	Measured mass (Da)	Error (ppm)	Fold-change	Trend[b]	t_r (min)	*p*-value
1	**L-Serine**	C3H7NO3	$[M+H]^+$	106.0495	3.8	4.7	DOWN	14.8	0.004
2	Uracil	C4H4N2O2	$[M+H]^+$	113.0352	5.3	3.9	DOWN	6.1	0.013
3	**L-Proline**	C5H9NO2	$[M+H]^+$	116.0712	5.2	1.6	UP	13.0	0.004
4	**L-Valine**	C5H11NO2	$[M+H]^+$	118.087	5.9	1.4	UP	12.7	0.004
5	Pyroglutamic acid	C5H7NO3	$[M+H]^+$	130.0497	1.5	1.6	UP	14.6	0.004
6	**L-Leucine**	C6H13NO2	$[M+H]^+$	132.1030	8.3	1.2	UP	10.2	0.004
7	**L-Glutamate**	C5H9NO4	$[M+H]^+$	148.0595	6.1	1.3	DOWN	12.7	0.004
8	**L-Methionine**	C5H11NO2S	$[M+H]^+$	150.0576	4.7	1.6	DOWN	10.6	0.004
9	**Guanine**	C5H5N5O	$[M+H]^+$	152.0553	9.1	1.5	DOWN	8.7	0.004
10	**L-Carnitine**	C7H15NO3	$[M+H]^+$	162.1139	8.6	1.4	UP	15.2	0.004
11	**L-Phenylalanine**	C9H11NO2	$[M+H]^+$	166.0871	4.8	2.9	DOWN	8.8	0.004
12	L-Tyrosine	C9H11NO3	$[M+H]^+$	182.0798	7.7	1.7	DOWN	10.7	0.004
13	**Acetyllysine**	C8H16N2O3	$[M+H]^+$	189.1226	4.2	1.8	DOWN	13.8	0.022
14	**N-acetyl-L-glutamate**	C7H11NO5	$[M+H]^+$	190.0699	5.8	2.0	UP	8.0	0.004
15	L-Acetylcarnitine	C9H17NO4	$[M+H]^+$	204.1248	8.8	1.8	UP	14.0	0.004
16	4-phospho-L-aspartate	C4H5NO7P	$[M+H]^+$	210.9889	6.2	3.0	UP	9.1	0.004
17	2-deoxy-D-ribose 1-phosphate	C5H11O7P	$[M+H]^+$	215.0309	3.3	6.3	DOWN	5.6	0.004
18	O-Phospho-L-homoserine	C4H10NO6P	$[M+Na]^+$	222.0150	5.4	2.4	UP	7.9	0.004
19	lipoamide	C8H15NOS2	$[M+Na]^+$	228.0508	9.2	2.8	UP	6.1	0.004
20	Biotin	C10H16N2O3S	$[M+H]^+$	245.0943	4.5	1.6	UP	6.0	0.022
21	**Glycerophosphocholine**	C8H20NO6P	$[M+H]^+$	258.1105	1.5	1.7	DOWN	15.3	0.004
22	**Adenosinea[a] or**	C10H13N5O4	$[M+H]^+$	268.1018	8.2	3.1	DOWN	7.3	0.007
	2′-Deoxyguanosine[a]	C10H16N5O4	$[M+H]^+$	268.1018	8.2	3.1	DOWN	7.3	0.007
23	**N6-(L-1,3-Dicarboxypropyl)-L-lysine**	C11H20N2O6	$[M+H]^+$	277.1400	2.2	13.8	UP	14.5	0.004
24	**N-(L-Arginino)succinate**	C10H18N4O6	$[M+H]^+$	291.1305	2.1	13.7	UP	14.7	0.004
25	**5′-Methylthioadenosine**	C11H15N5O4S	$[M+H]^+$	298.0941	9.1	5.5	UP	5.1	0.004
26	**Glutathione**	C10H17N3O6S	$[M+H]^+$	308.0895	5.2	1.3	DOWN	12.7	0.004
27	Uridine 5'-monophosphate (UMP)	C9H13N2O9P	$[M+H]^+$	325.0411	6.2	1.3	UP	10.8	0.004
28	Trehalose[a]	C12H22O11	$[M+H]^+$	343.1226	2.5	6.4	UP	15.5	0.004
29	**Adenosine monophosphate (AMP)[a] or**	C10H14N5O7P	$[M+H]^+$	348.0687	4.9	1.4	DOWN	10.1	0.004
	Deoxyguanosine monophosphate (dGMP)[a]	C10H14N5O7P	$[M+H]^+$	348.0687	4.9	1.4	DOWN	10.1	0.004
30	LysoPC(16:1(9Z))	C24H48NO7P	$[M+K]^+$	532.2769	5.8	1.8	UP	4.1	0.004
31	LysoPC(18:1(9Z))	C26H52NO7P	$[M+K]^+$	560.3145	5.7	3.9	UP	4.0	0.004
32	**Nicotinamide adenine dinucleotide (NAD^+)**	C21H27N7O14P2	$[M+H]^+$	664.1137	4.1	1.5	DOWN	13.7	0.004

Metabolites in bold letters were also detected by CE-MS (Table 2).
[a] These metabolites provide an identification conflict.
[b] Trend was determined taking as a reference the response of control samples.

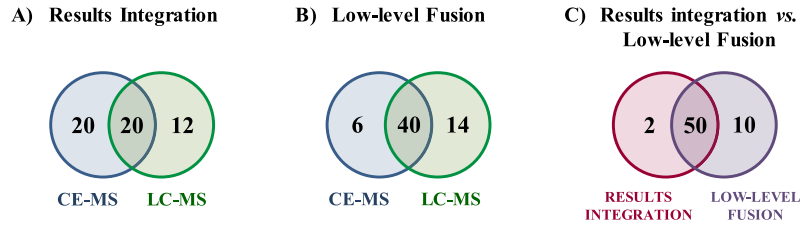


Fig. 3. Venn diagram of the identified metabolites that differentiate between both carbon sources. (A) Results integration: CE-MS and LC-MS independent analysis. (B) Low-level data fusion strategy: CE-MS and LC-MS were merged before the analysis. (C) Global comparison between independent analysis results integration and low-level data fusion strategies.

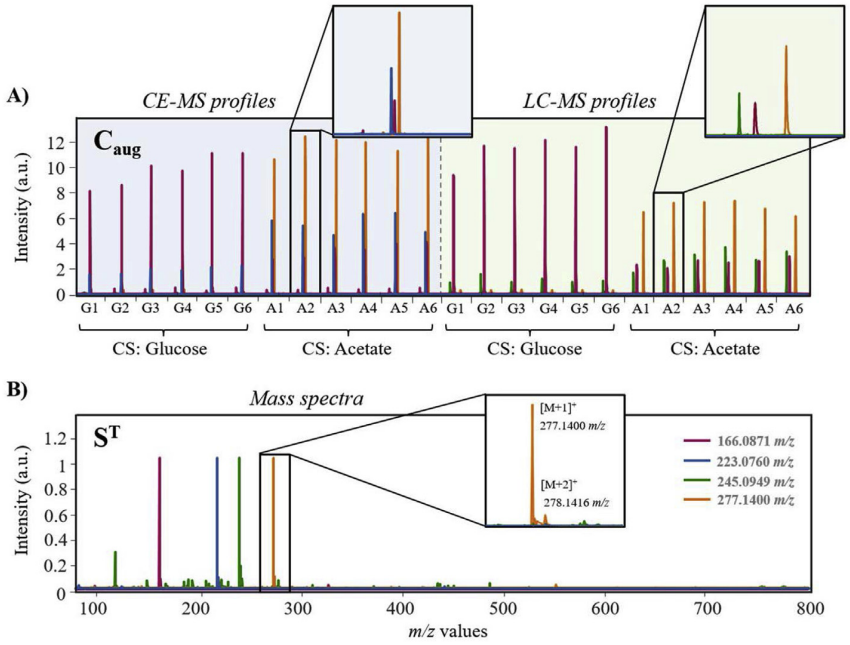


Fig. 4. An example of four resolved components by MCR-ALS after low-level data fusion: (A) Migration and elution profiles for the different samples. Left side: CE-MS samples considering the two carbon source (CS) fonts: glucose (G1-G6) and acetate (A1-A6). Right side: LC-MS samples considering the two carbon source fonts: glucose (G1-G6) and acetate (A1-A6). Insets show a detailed view of the four depicted components for an acetate-grown sample (A2) measured by both platforms (CE-MS and LC-MS). (B) Mass spectra for the four components. The inset shows the various ion contribution due to the H isotopes resolved for the singly-charged molecular ion of the metabolite with nominal monoisotopic *m/z* value 277.

Statistically significant metabolites to differentiate between glucose- and acetate-grown samples were determined and identified using the same procedures described above. A total number of 71 differential components were statistically significant, and 60 were tentatively identified as metabolites (Table 1). Fig. 3B summarises in a Venn diagram the relationships among the metabolites identified by the low-level data fusion results. 40 metabolites were common for both platforms, and 6 and 14 metabolites were specific for CE-MS and LC-MS, respectively. Both techniques appeared to retrieve complementary information as almost a 20% of the identified metabolites were detected by a single technique. These results also suggested that the low-level data fusion strategy was more efficient than the independent analysis in the metabolome coverage from the LC-MS and CE-MS multisets. As is shown in Fig. 3C, the low-level data fusion allowed to identify a higher number of metabolites (60 *vs.* 52), suggesting that despite most of the metabolites are common for the different approaches, this difference in the metabolome coverage could provide a deeper insight of the affected pathways. Only a few number of the resolved potential metabolites (2 metabolites) in the independent analysis strategy were not resolved using the low data fusion, *e.g.* 5-Amino-6-(5'-phosphoribosylamino)uracil, lysoPC(18:1(9Z)). In contrast, MCR-ALS model for low-level fused data led to discover 10 metabolites that had not previously been detected using the other strategy.

Table 4 shows the candidate metabolites identified by this low-level data fusion strategy for each one of the platforms. According to our previous comment of Fig. 3C, some of the metabolites were considered to be specific for CE-MS or LC-MS according to previously obtained results for the independent analysis (Tables 2 and 3) were now categorized as common metabolites by the low-level fusion approach (Table 4). This fact shows that low-level data fusion strategy appears to be a powerful strategy for the resolution of the metabolites of interest, especially low abundant (or poorly detected) metabolites. Since low intensity signals could be overlapped with other background or compound contributions, they

Table 4
Differential metabolites obtained by a low-level fusion CE-MS and LC-MS analysed by MCR-ALS (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] ≤10 ppm).

N	Metabolite	Molecular formula	Ion assignment	Measured mass (Da)	Error (ppm)	Fold-change	Trend[e]	t (min)	*p*-value
CE-MS									
1	**4-Aminobutanoate (GABA)**	C4H9NO2	$[M+H]^+$	104.0715	8.6	3.8	UP	10.7	0.005
2	**L-Serine**	C3H7NO3	$[M+H]^+$	106.0507	7.5	2.7	DOWN	11.3	0.005
3	**L-Proline**	C5H9NO2	$[M+H]^+$	116.0716	8.6	3.0	UP	12.6	0.005
4	**L-Valine**	C5H11NO2	$[M+H]^+$	118.0853	8.5	4.5	UP	13.1	0.005
5	**L-Threonine**	C4H9NO3	$[M+H]^+$	120.0664	7.5	1.5	UP	11.7	0.005
6	**gamma-Amino-gamma-cyanobutanoate**	C5H8N2O2	$[M+H]^+$	129.0669	8.5	75.8	UP	9.5	0.005
7	**Pyroglutamic acid**[a]	C5H7NO3	$[M+H]^+$	130.0511	9.2	1.5	UP	11.9	0.011
8	**L-Leucine**	C6H13NO2	$[M+H]^+$	132.1028	6.8	1.3	UP	11.1	0.005
9	**L-Ornithine**	C5H12N2O2	$[M+H]^+$	133.0983	8.3	1.7	UP	7.0	0.005
10	**L-Aspartate**	C4H7NO4	$[M+H]^+$	134.0455	5.2	7.5	DOWN	13.3	0.005
11	2-deoxy-D-ribose	C5H10O4	$[M+H]^+$	135.0665	9.6	3.8	DOWN	13.3	0.005
12	**5-(2-Hydroxyethyl)-4-methylthiazole**	C6H9NOS	$[M+H]^+$	144.0492	9.7	3.8	DOWN	6.4	0.005
13	**4-guanidinobutanoic acid**	C5H11N3O2	$[M+H]^+$	146.0908	8.2	7.1	UP	7.4	0.005
14	**L-Glutamine**	C5H10N2O3	$[M+H]^+$	147.0778	9.5	1.3	UP	11.8	0.049
15	L-Lysine	C6H14N2O2	$[M+H]^+$	147.1140	8.2	6.7	DOWN	7.2	0.005
16	**L-Glutamate**	C5H9NO4	$[M+H]^+$	148.0617	8.8	1.4	DOWN	12.3	0.009
17	**L-Methionine**	C5H11NO2S	$[M+H]^+$	150.0591	5.3	1.4	DOWN	11.8	0.031
18	**Guanine**	C5H5N5O	$[M+H]^+$	152.0581	9.3	1.2	DOWN	8.4	0.028
19	**L-Carnitine**	C7H15NO3	$[M+H]^+$	162.1135	6.2	1.9	UP	8.1	0.005
20	**L-Phenylalanine**	C9H11NO2	$[M+H]^+$	166.0879	9.6	2.5	DOWN	12.1	0.005
21	**L-Arginine**	C6H14N4O2	$[M+H]^+$	175.1207	9.7	1.3	DOWN	7.5	0.005
22	**Acetyllysine**	C8H16N2O3	$[M+H]^+$	189.1249	7.9	1.7	DOWN	12	0.005
23	**N-acetyl-L-glutamate**	C7H11NO5	$[M+H]^+$	190.0701	4.7	1.3	UP	18.6	0.005
24	**S-Adenosyl-L-homocysteine**	C14H20N6O5S	$[M+2H]^+$	193.0700	9.8	2.1	DOWN	9.1	0.005
25	**O-Phospho-L-homoserine**[a]	C4H10NO6P	$[M+H]^+$	200.0313	3.0	2.4	UP	18.5	0.011
26	3-(indol-3-yl)pyruvate	C11H9NO3	$[M+H]^+$	203.0589	5.9	3.1	UP	5.7	0.018
27	**L-Acetylcarnitine**[a]	C9H17NO4	$[M+H]^+$	204.1248	8.8	1.6	UP	8.6	0.011
28	**L-Tryptophan**	C11H12N2O2	$[M+H]^+$	205.0961	5.4	1.4	DOWN	9.6	0.009
29	**4-phospho-L-aspartate**[a]	C4H8NO7P	$[M+H]^+$	210.9892	7.6	3.3	UP	5.7	0.011
30	L-Cystathionine	C7H14N2O4S	$[M+H]^+$	223.0768	9.4	2.8	UP	11.5	0.005
31	**D-Erythro-imidazole-glycerol-phosphate**	C6H11N2O6P	$[M+H]^+$	239.0443	6.7	2.7	DOWN	16.7	0.009
32	**Glycerophosphocholine**	C8H20NO6P	$[M+H]^+$	258.1122	8.1	1.5	DOWN	17.3	0.005
33	**trans-4-Hydroxy-L-proline**	C5H9NO3	$[2M+H]^+$	263.1263	9.5	1.9	UP	5.9	0.005
34	L-Asparagine	C4H8N2O3	$[2M+H]^+$	265.1153	3.8	3.1	DOWN	6.4	0.005
35	**Adenosine**[d] **or**	C10H13N5O4	$[M+H]^+$	268.1062	8.2	2.7	DOWN	9.3	0.005
	2′-Deoxyguanosine[d]	C10H16N5O4	$[M+H]^+$	268.1062	8.2	2.7	DOWN	9.3	0.005
36	**N6-(L-1,3-Dicarboxypropyl)-L-lysine**	C11H20N2O6	$[M+H]^+$	277.1416	7.9	11.9	UP	12.6	0.005
37	**N-(L-Arginino)succinate**	C10H18N4O6	$[M+H]^+$	291.1324	8.6	13.9	UP	10.5	0.005
38	**Aminoimidazole ribonucleotide**[c]	C8H14N3O7P	$[M+H]^+$	296.0671	9.8	2.8	DOWN	17.2	0.011
39	**5′-Methylthioadenosine**	C11H15N5O3S	$[M+H]^+$	298.0982	4.7	6.2	UP	9.4	0.005
40	**Glutathione**	C10H17N3O6S	$[M+H]^+$	308.0940	9.4	1.2	DOWN	14.0	0.005
41	Cytidine monophosphate (CMP)	C9H14N3O8P	$[M+H]^+$	324.0604	4.0	3.9	DOWN	17.4	0.005
42	**Trehalose**	C12H22O11	$[M+H]^+$	343.1212	6.7	5.0	UP	17.2	0.005
43	**Adenosine monophosphate (AMP)**[d] **or**	C10H14N5O7P	$[M+H]^+$	348.0735	8.9	1.3	DOWN	17.5	0.005
	Deoxyguanosine monophosphate (dGMP)[d]	C10H14N5O7P	$[M+H]^+$	348.0735	8.9	1.3	DOWN	17.5	0.005
44	**Guanosine monophosphate (GMP)**[c]	C10H14N5O8P	$[M+H]^+$	364.0673	5.5	1.3	DOWN	23.6	0.045
45	**Adenylsuccinic acid**	C14H18N5O11P	$[M+H]^+$	464.0839	5.6	4.8	UP	26.7	0.005
46	**Nicotinamide adenine dinucleotide (NAD^+)**	C21H27N7O14P2	$[M+H]^+$	664.1189	3.8	1.2	DOWN	17.8	0.018
LC-MS									
1	**4-Aminobutanoate (GABA)**[b]	C4H9NO2	$[M+H]^+$	104.0702	3.8	3.0	UP	14.8	0.045
2	**L-Serine**	C3H7NO3	$[M+H]^+$	106.0495	3.8	4.7	DOWN	14.8	0.004
3	Uracil	C4H4N2O2	$[M+H]^+$	113.0352	5.3	3.9	DOWN	6.1	0.013
4	**L-Proline**	C5H9NO2	$[M+H]^+$	116.0712	5.2	1.6	UP	13.0	0.004
5	**L-Valine**	C5H11NO2	$[M+H]^+$	118.0870	5.9	1.4	UP	12.7	0.004
6	**L-Threonine**[b]	C4H9NO3	$[M+H]^+$	120.0656	0.8	1.7	UP	14.3	0.016
7	**gamma-Amino-gamma-cyanobutanoate**[b]	C5H8N2O2	$[M+H]^+$	129.0652	4.6	43.0	UP	14.6	0.005
8	**Pyroglutamic acid**	C5H7NO3	$[M+H]^+$	130.0497	1.5	1.6	UP	14.6	0.004
9	**L-Leucine**	C6H13NO2	$[M+H]^+$	132.1030	8.3	1.2	UP	10.2	0.004
10	**L-Ornithine**[b]	C5H12N2O2	$[M+H]^+$	133.0971	0.4	1.7	UP	15.0	0.005
11	**L-Aspartate**[b]	C4H7NO4	$[M+H]^+$	134.0442	4.5	9.3	DOWN	13.6	0.005
12	**5-(2-Hydroxyethyl)-4-methylthiazole**[b]	C6H9NOS	$[M+H]^+$	144.0472	4.2	3.4	DOWN	4.6	0.005
13	**4-guanidinobutanoic acid**[b]	C5H11N3O2	$[M+H]^+$	146.0921	0.7	5.9	UP	14.4	0.005
14	**L-Glutamine**[b]	C5H10N2O3	$[M+H]^+$	147.0770	4.1	1.8	UP	14.6	0.045
15	**L-Glutamate**	C5H9NO4	$[M+H]^+$	148.0595	6.1	1.3	DOWN	12.7	0.004
16	**L-Methionine**	C5H11NO2S	$[M+H]^+$	150.0576	4.7	1.6	DOWN	10.6	0.004
17	**Guanine**	C5H5N5O	$[M+H]^+$	152.0553	9.1	1.5	DOWN	8.7	0.004
18	Phosphoglycolic acid	C2H5O6P	$[M+H]^+$	156.9887	6.4	9.1	UP	6.7	0.005
19	**L-Carnitine**	C7H15NO3	$[M+H]^+$	162.1139	8.6	1.4	UP	15.2	0.004
20	**L-Phenylalanine**	C9H11NO2	$[M+H]^+$	166.0871	4.8	2.9	DOWN	8.8	0.004
21	**L-Arginine**[b]	C6H14N4O2	$[M+H]^+$	175.1200	5.7	1.6	DOWN	10.9	0.005
22	L-cysteinylglycine	C5H10N2O3S	$[M+H]^+$	179.0497	6.7	1.2	DOWN	12.8	0.005
23	L-Tyrosine	C9H11NO3	$[M+H]^+$	182.0798	7.7	1.7	DOWN	10.7	0.004
24	**Acetyllysine**	C8H16N2O3	$[M+H]^+$	189.1226	4.2	1.8	DOWN	13.8	0.022
25	**N-acetyl-L-glutamate**	C7H11NO5	$[M+H]^+$	190.0699	5.8	2.0	UP	8.0	0.004

(*continued on next page*)

Table 4 (*continued*)

N	Metabolite	Molecular formula	Ion assignment	Measured mass (Da)	Error (ppm)	Fold-change	Trend[e]	t (min)	*p*-value
26	**S-Adenosyl-L-homocysteine**[b]	C14H20N6O5S	$[M+2H]^+$	193.0699	9.3	1.7	DOWN	3.7	0.009
27	**L-Acetylcarnitine**	C9H17NO4	$[M+H]^+$	204.1248	8.8	1.8	UP	14	0.004
28	**L-Tryptophan**[b]	C11H12N2O2	$[M+H]^+$	205.0961	5.4	1.4	DOWN	8.3	0.016
29	**4-phospho-L-aspartate**	C4H5NO7P	$[M+H]^+$	210.9889	6.2	3.0	UP	9.1	0.004
30	2-deoxy-D-ribose 1-phosphate	C5H11O7P	$[M+H]^+$	215.0309	3.3	6.3	DOWN	5.6	0.004
31	**O-Phospho-L-homoserine**	C4H10NO6P	$[M+Na]^+$	222.0150	5.4	2.4	UP	7.9	0.004
32	lipoamide	C8H15NOS2	$[M+Na]^+$	228.0508	9.2	2.8	UP	6.1	0.004
33	Biotin	C10H16N2O3S	$[M+H]^+$	245.0943	4.5	1.6	UP	6.0	0.022
34	gamma-L-Glutamyl-L-cysteine	C8H14N2O5S	$[M+H]^+$	251.0695	0.4	2.4	DOWN	17.9	0.005
35	**D-Erythro-imidazole-glycerol-phosphate**[b]	C6H11N2O6P	$[M+NH_4]^+$	256.0710	6.6	4.0	DOWN	12.9	0.027
36	**Glycerophosphocholine**	C8H20NO6P	$[M+H]^+$	258.1105	1.5	1.7	DOWN	15.3	0.004
37	**trans-4-Hydroxy-L-proline**[b]	C5H9NO3	$[2M+H]^+$	263.1226	4.6	1.8	UP	13.1	0.005
38	**Adenosine**[d] **or**	C10H13N5O4	$[M+H]^+$	268.1018	8.2	3.1	DOWN	7.3	0.007
	2′-Deoxyguanosine[d]	C10H16N5O4	$[M+H]^+$	268.1018	8.2	3.1	DOWN	7.3	0.007
39	**N6-(L-1,3-Dicarboxypropyl)-L-lysine**	C11H20N2O6	$[M+H]^+$	277.1400	2.2	13.8	UP	14.5	0.004
40	Cytidine	C9H13N3O5	$[M+K]^+$	282.0491	1.4	2.8	DOWN	9.5	0.005
41	Guanosine	C10H13N5O5	$[M+H]^+$	284.0980	3.2	2.4	DOWN	9.6	0.005
42	**N-(L-Arginino)succinate**	C10H18N4O6	$[M+H]^+$	291.1305	2.1	13.7	UP	14.7	0.004
43	**Aminoimidazole ribonucleotide**[c]	C8H14N3O7P	$[M+H]^+$	296.0658	5.4	2.0	DOWN	15.3	0.005
44	**5′-Methylthioadenosine**	C11H15N5O4S	$[M+H]^+$	298.0941	9.1	5.5	UP	5.1	0.004
45	**Glutathione**	C10H17N3O6S	$[M+H]^+$	308.0895	5.2	1.3	DOWN	12.7	0.004
46	Uridine 5'-monophosphate (UMP)	C9H13N2O9P	$[M+H]^+$	325.0411	6.2	1.3	UP	10.8	0.004
47	**Trehalose**	C12H22O11	$[M+H]^+$	343.1226	2.5	6.4	UP	15.5	0.004
48	**Adenosine monophosphate (AMP)**[d] **or**	C10H14N5O7P	$[M+H]^+$	348.0687	4.9	1.4	DOWN	10.1	0.004
	Deoxyguanosine monophosphate (dGMP)[d]	C10H14N5O7P	$[M+H]^+$	348.0687	4.9	1.4	DOWN	10.1	0.004
49	**Guanosine monophosphate (GMP)**[c]	C10H14N5O8P	$[M+H]^+$	364.0629	6.6	1.6	DOWN	13.6	0.005
50	**Adenylsuccinic acid**[b]	C14H18N5O11P	$[M+H]^+$	464.0789	5.2	3.7	UP	13.6	0.005
51	LysoPC(16:1(9Z))	C24H48NO7P	$[M+K]^+$	532.2769	5.8	1.8	UP	4.1	0.004
52	Oxidized glutathione	C20H32N6O12S2	$[M+H]^+$	613.1568	3.9	1.4	DOWN	14.5	0.005
53	Uridine diphosphate-N-acetylglucosamine	C17H27N3O17P2	$[M+K]^+$	646.0423	3.7	6.4	UP	13.1	0.005
54	**Nicotinamide adenine dinucleotide (NAD^+)**	C21H27N7O14P2	$[M+H]^+$	664.1137	4.1	1.5	DOWN	13.7	0.004

Metabolites in bold letters were detected by both analytical platforms using the low-level data fusion.
[a] New metabolite previously detected in the individual analysis of LC-MS data (Table 3).
[b] New metabolite previously detected in the individual analysis of CE-MS data (Table 2).
[c] New differential metabolite detected by both platforms: CE-MS and LC-MS.
[d] These metabolites provided an identification conflict.
[e] Trend was determined taking as a reference the response of control samples.

could be only resolved when the joint analysis of CE-MS and LC-MS was performed using the advantage of the unique MS spectra ($\mathbf{S}^{\mathbf{T}}$). In this way, low abundant (or poorly detected) metabolites that were well resolved by one of the techniques could be better detected and resolved by the other one if experimental conditions are less favourable (*e.g.* ion suppression promoted by poor separation resolution from other contributions or low concentrated samples). For instance, L-Aspartate was poorly detected by LC-MS, whereas in CE-MS ion signal intensity was high (see Supplementary Material Fig. S3). However, now it could also be detected and resolved in both cases thanks to the low-level data fusion.

Briefly, advantages of the individual techniques could be combined and the drawbacks minimized to improve the quality of the results. For instance, it is known that CE-MS required a minor amount of sample in comparison with LC-MS but at the cost of preparing samples at a higher concentration. In this work, we have shown that these limitations could be overcome by the combination of CE-MS with LC-MS, allowing the detection of extremely low abundant compounds using the main benefits of each technique. Complementary information provided by CE-MS and LC-MS is in part shown in the results obtained by these two platforms. However, in this case, most of the identified metabolites were detected by both platforms, since an HILIC column was used in LC-MS experiments. Nowadays HILIC is widely applied as an alternative to reversed phase chromatography for determination of polar compounds by LC-MS, as an alternative to CE-MS. Our results confirmed the suitability of the proposed low-level data fusion strategy to obtain reliable information from the fused data. Furthermore, they open new possibilities to exploit MS data fusion potential in more orthogonal situations and conditions (like for instance, in reversed-phase LC-MS and CE-MS inter-platforms), which undoubtedly should increase the analytical metabolome coverage.

3.3. Biological interpretation

KEGG analysis assigned a total of 50 metabolites to known yeast metabolic pathways (or modules, in KEGG terminology) taking into account the 60 identified metabolites by low-level data fusion. Forty out of these 50 metabolites (all except 4-aminobutanoate (GABA), 4-guanidinobutanoic acid, 5-(2-hydroxyethyl)-4-methylthiazole, 5'-methylthioadenosine, biotin, nicotinamide adenine dinucleotide (NAD^+), phosphoglycolic acid, trans-4-hydroxy-L-proline, trehalose and uridine diphosphate-N-acetylglucosamine) were associated with three basic metabolic pathways (Table 5, a metabolic map showing the position of affected metabolites by the carbon source is presented in Supplementary Material Fig. S4): amino acid metabolism (24 metabolites), nucleotide metabolism (13 metabolites, one of them was also in the previous group), and glutathione metabolism (6 metabolites, two of them were also in the first group). The data showed a general decrease in metabolite levels for nucleotide and glutathione metabolism in acetate-grown samples; in contrast, intermediate amino acid metabolites showed a general increase, whereas concentrations of the precursor amino acids were generally reduced, except for L-Glutamine, L-Ornithine, L-Leucine, L-Proline, L-Valine and L-Threonine (Table 5). These changes can be explained as a response to the poor carbon source such as acetate,

Table 5
Functional annotation (KEGG) of yeast metabolites with significant differences between glucose- and acetate-grown samples detected by Low-level data fusion strategy.

KEGG module (#hits)	KEGG code	Metabolite	Trend[b]
sce01230 Biosynthesis of amino acids (24)	C00021	S-Adenosyl-L-homocysteine	DOWN
	C00025	L-Glutamate[a]	DOWN
	C00047	L-Lysine	DOWN
	C00049	L-Aspartate	DOWN
	C00062	L-Arginine	DOWN
	C00064	L-Glutamine[a]	UP
	C00065	L-Serine	DOWN
	C00073	L-Methionine	DOWN
	C00077	L-Ornithine[a]	UP
	C00078	L-Tryptophan	DOWN
	C00079	L-Phenylalanine	DOWN
	C00082	L-Tyrosine	DOWN
	C00123	L-Leucine	UP
	C00148	L-Proline	UP
	C00152	L-Asparagine	DOWN
	C00183	L-Valine	UP
	C00188	L-Threonine	UP
	C00449	N6-(L-1,3-Dicarboxypropyl)-L-lysine	UP
	C00624	N-Acetyl-L-glutamate	UP
	C01102	O-Phospho-L-homoserine	UP
	C02291	L-Cystathionine	UP
	C03082	4-Phospho-L-aspartate	UP
	C03406	N-(L-Arginino)succinate	UP
	C04666	D-Erythro-imidazole-glycerol-phosphate	DOWN
sce00230 Purine metabolism/sce00240 Pyrimidine metabolism (13)	C00020/C00362	AMP/dGMP	DOWN
	C00055	CMP	DOWN
	C00064	L-Glutamine[a]	UP
	C00105	UMP	UP
	C00106	Uracil	DOWN
	C00144	GMP	DOWN
	C00212/C00330	Adenosine/2′-Deoxyguanosine	DOWN
	C00242	Guanine	DOWN
	C00387	Guanosine	DOWN
	C00475	Cytidine	DOWN
	C00672	2-Deoxy-D-ribose 1-phosphate	DOWN
	C03373	Aminoimidazole ribotide	DOWN
	C03794	Adenylsuccinic acid	UP
sce00480 Glutathione metabolism (6)	C00025	L-Glutamate[a]	DOWN
	C00051	Glutathione	DOWN
	C00127	Oxidized glutathione	DOWN
	C00669	gamma-L-Glutamyl-L-cysteine	DOWN
	C01879	Pyroglutamic acid	UP
	C00077	L-Ornithine[a]	UP

[a] Metabolites involved in different metabolic pathways.
[b] Trend was determined taking as a reference the response of control samples.

and the use of internal pools of amino acids and other metabolites to feed essential secondary metabolism. A general reduction in transcription of genes from the pentose phosphate pathway and other nucleotide biosynthesis-related genes has been previously associated to such culture media by microarray analysis [56], but to our knowledge this is the first report indicating an actual decrease in nucleotide and nucleoside metabolite levels as a consequence of the use of C2 carbon sources (acetate) in yeast.

4. Concluding remarks

The application of different knowledge integration strategies based on MCR-ALS allowed the joint, untargeted, comprehensive, reliable and high-throughput analysis of the complex CE-MS and LC-MS metabolomics data from a comparative study of yeast samples grown in a non-fermentable (acetate) and a fermentable carbon source (glucose).

Overall, the number of identified metabolites from the resolved MCR-ALS components that were significant to differentiate between glucose- and acetate-grown yeast samples using both strategies was similar despite the low-level data fusion approach provided a higher metabolome coverage. For this reason, when possible, the use of low-level data fusion is recommended because it was more efficient in resolving low abundant (or poorly detected) metabolites with one of the techniques, if signal intensity or separation with the other technique was good enough. Furthermore, obtained results clearly demonstrated that the joint analysis of the CE-MS and LC-MS platform results using these integration strategies enabled an improved metabolite coverage than without combining the information from both techniques. Hence, these knowledge integration approaches allowed better understanding of the underlying biochemical changes of the studied biological system due to different culture conditions, which were mainly associated with amino acid, nucleotide, and glutathione metabolism.

In this study, a CE-MS and LC-MS inter-platform was used to test the proposed methodology for low-level data fusion. However, due to the flexibility of the MCR-ALS method for multiset data analysis, this methodology can be easily extended to other analytical platforms for metabolomics research to facilitate more accurate and unequivocal biological information and interpretation. Other matrix augmentation strategies can be attempted in the generation

and analysis of multiset data. Individual independent analysis of every dataset and final combined interpretation of these individual results will always be an option. Depending on the data structures to be simultaneously analysed (*i.e.* different *omic* levels such as metabolomics and transcriptomics for the same samples, or different metabolomic instrumental analysis platforms such as LC-MS and NMR also for the same samples), the proposed low-level approach (or mid-level after a preliminary feature selection) can be also attempted by the analysis of row-wise concatenated (fused) data.

Acknowledgements

This research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 320737. Some part of this study was also supported by a grant from the Spanish Ministry of Economy and Competitiveness (CTQ2014-56777-R and CTQ2015-66254). Also, recognition from the Catalan government (grant 2014 SGR 1106) is acknowledged.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.aca.2017.04.049.

References

[1] M.M.W.B. Hendriks, F.A. Eeuwijk, R.H. Jellema, J.A. Westerhuis, T.H. Reijmers, H.C.J. Hoefsloot, A.K. Smilde, Data-processing strategies for metabolomics studies, TrAC - Trends Anal. Chem. 30 (2011) 1685–1698.
[2] C.H. Johnson, J. Ivanisevic, H.P. Benton, G. Siuzdak, Bioinformatics: the next frontier of metabolomics, Anal. Chem. 87 (2015) 147–156.
[3] L. Blanchet, A. Smolinska, Data fusion in metabolomics and proteomics for biomarker discovery, Methods Mol. Biol. 1362 (2016) 209–223.
[4] J. Boccard, S. Rudaz, Harnessing the complexity of metabolomic data with chemometrics, J. Chemom. 28 (2014) 1–9.
[5] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (2013) 28–44.
[6] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (2015) 1449–1477.
[7] A.K. Smilde, J.A. Westerhuis, R. Boqué, Multiway multiblock component and covariates regression models, J. Chemom. 14 (2000) 301–331.
[8] E. Acar, A.J. Lawaetz, M.A. Rasmussen, R. Bro, Structure-revealing data fusion model with applications in metabolomics, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS,, 2013, pp. 6023–6026.
[9] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (2013) 822–833.
[10] S.E. Richards, M.E. Dumas, J.M. Fonville, T.M.D. Ebbels, E. Holmes, J.K. Nicholson, Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework, Chemom. Intell. Lab. Syst. 104 (2010) 121–131.
[11] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinforma. 11 (2010) 395.
[12] A.K. Smilde, M.J. Van Der Werf, S. Bijlsma, B.J.C. Van Der Werff-Van Der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Anal. Chem. 77 (2005) 6729–6736.
[13] P. Vernocchi, L. Vannini, D. Gottardi, F. Del Chierico, D.I. Serrazanetti, M. Ndagijimana, M.E. Guerzoni, Integration of datasets from different analytical techniques to assess the impact of nutrition on human metabolome, Front. Cell. Infect. Microbiol. 2 (2012) 156.
[14] E. Acar, R. Bro, A.K. Smilde, Data fusion in metabolomics using coupled matrix and tensor factorizations, Proc. IEEE 103 (2015) 1602–1620.
[15] E. Acar, M.A. Rasmussen, F. Savorani, T. Næs, R. Bro, Understanding data fusion within the framework of coupled matrix and tensor factorizations, Chemom. Intell. Lab. Syst. 129 (2013) 53–63.
[16] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12.
[17] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787.
[18] O. Alter, P.O. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3351–3356.
[19] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, J. Trygg, Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data, Plant J. 52 (2007) 1181–1191.
[20] T. Löfstedt, J. Trygg, OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemom. 25 (2011) 441–455.
[21] J. Kuligowski, D. Pérez-Guaita, A. Sánchez-Illana, Z. León-González, M. De La Guardia, M. Vento, E.F. Lock, G. Quintás, Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE), Analyst 140 (2015) 4521–4529.
[22] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976.
[23] J. Jaumot, R. Eritja, R. Tauler, R. Gargallo, Resolution of a structural competition involving dimeric G-quadruplex and its C-rich complementary strand, Nucleic Acids Res. 34 (2006) 206–216.
[24] J. Jaumot, B. Piña, R. Tauler, Application of multivariate curve resolution to the analysis of yeast genome-wide screens, Chemom. Intell. Lab. Syst. 104 (2010) 53–64.
[25] S. Mas, R. Tauler, A. de Juan, Chromatographic and spectroscopic data fusion analysis for interpretation of photodegradation processes, J. Chromatogr. A 1218 (2011) 9260–9268.
[26] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36.
[27] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: a review, Anal. Chim. Acta 914 (2016) 17–34.
[28] A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Modern analytical techniques in metabolomics analysis, Analyst 137 (2012) 293–300.
[29] J. Forshed, H. Idborg, S.P. Jacobsson, Evaluation of different techniques for data fusion of LC/MS and 1H-NMR, Chemom. Intell. Lab. Syst. 85 (2007) 102–109.
[30] B. Biais, J.W. Allwood, C. Deborde, Y. Xu, M. Maucourt, B. Beauvoit, W.B. Dunn, D. Jacob, R. Goodacre, D. Rolin, A. Moing, 1H NMR, GC-EI-TOFMS, and data set correlation for fruit metabolomics: application to spatial metabolite analysis in melon, Anal. Chem. 81 (2009) 2884–2894.
[31] R.A. van den Berg, C.M. Rubingh, J.A. Westerhuis, M.J. van der Werf, A.K. Smilde, Metabolomics data exploration guided by prior knowledge, Anal. Chim. Acta 651 (2009) 173–181.
[32] W. Yao, M. He, Y. Jiang, L. Zhang, A. Ding, Y. Hu, Integrated LC/MS and GC/MS metabolomics data for the evaluation of protection function of fructus ligustri lucidi on mouse liver, Chromatographia 76 (2013) 1171–1179.
[33] I. Garcia-Perez, A. Couto Alves, S. Angulo, J.V. Ll, J. Utzinger, T.M.D. Ebbels, C. Legido-Quigley, J.K. Nicholson, E. Holmes, C. Barbas, Bidirectional correlation of NMR and capillary electrophoresis fingerprints: a new approach to investigating Schistosoma mansoni infection in a mouse model, Anal. Chem. 82 (2010) 203–210.
[34] P.V. Attfield, Stress tolerance: the key to effective strains of industrial baker's yeast, Nat. Biotechnol. 15 (1997) 1351–1357.
[35] I. Borodina, J. Nielsen, Advances in metabolic engineering of yeast Saccharomyces cerevisiae for production of chemicals, Biotechnol. J. 9 (2014) 609–620.
[36] J. Nielsen, C. Larsson, A. van Maris, J. Pronk, Metabolic engineering of yeast for production of fuels and chemicals, Curr. Opin. Biotechnol. 24 (2013) 398–404.
[37] D. Kessner, M. Chambers, R. Burke, D. Agus, P. Mallick, ProteoWizard: open source software for rapid proteomics tools development, Bioinformatics 24 (2008) 2534–2536.
[38] M. Farrés, B. Piña, R. Tauler, LC-MS based metabolomics and chemometrics study of the toxic effects of copper on Saccharomyces cerevisiae, Metallomics 8 (2016) 790–798.
[39] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow, TrAC - Trends Anal. Chem. 82 (2016) 425–442.
[40] A.S. Marques, C. Bedia, K.M.G. Lima, R. Tauler, Assessment of the effects of As(III) treatment on cyanobacteria lipidomic profiles by LC-MS and MCR-ALS, Anal. Bioanal. Chem. 408 (2016) 5829–5841.
[41] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinforma. 9 (2008) 504.
[42] 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.
[43] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[44] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, 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. 407 (2015) 8835–8847.
[45] E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler, Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling, Electrophoresis 36 (2015) 2324–2335.
[46] G. Golub, K. Sølna, P. Van Dooren, Computing the SVD of a general matrix product/quotient, SIAM J. Matrix Anal. Appl. 22 (2000) 1–19.
[47] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432.
[48] R. Tauler, D. Barceló, Multivariate curve resolution applied to liquid chromatography-diode array detection, Trends Anal. Chem. 12 (1993) 319–327.
[49] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data

analysis and ambiguity in multivariate curve resolution, J. Chemom. 9 (1995) 31–58.
[50] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1995) 289–300.
[51] T. Jewison, C. Knox, V. Neveu, Y. Djoumbou, A.C. Guo, J. Lee, P. Liu, R. Mandal, R. Krishnamurthy, I. Sinelnikov, M. Wilson, D.S. Wishart, YMDB: the yeast metabolome database, Nucleic Acids Res. 40 (2012).
[52] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: a metabolite mass spectral database, Ther. Drug Monit. 27 (2005) 747–751.
[53] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res. 40 (2012).
[54] J.A. Kirwan, D.I. Broadhurst, R.L. Davidson, M.R. Viant, Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow, Anal. Bioanal. Chem. 405 (2013) 5147–5157.
[55] J. Kuligowski, Á. Sánchez-Illana, D. Sanjuán-Herráez, M. Vento, G. Quintás, Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC), Analyst 140 (2015) 7810–7817.
[56] P. Daran-Lapujade, M.L.A. Jansen, J.M. Daran, W. Van Gulik, J.H. De Winde, J.T. Pronk, Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae: a chemostat culture study, J. Biol. Chem. 279 (2004) 9125–9138.

Informació suplementària a l'article científic III.

Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data.

E. Ortiz-Villanueva, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler. J. Jaumot.

Analytica Chimica Acta 978 (2017) 10-23.

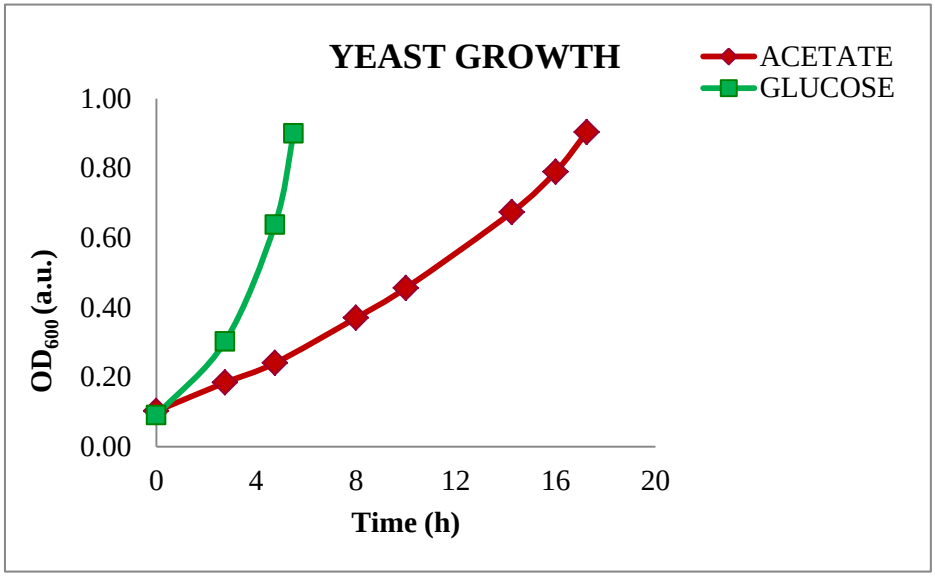


Figure S1. Growth rate of *S. cerevisiae* strain (BY4741) in YPD (glucose) and YEPA (acetate) media.

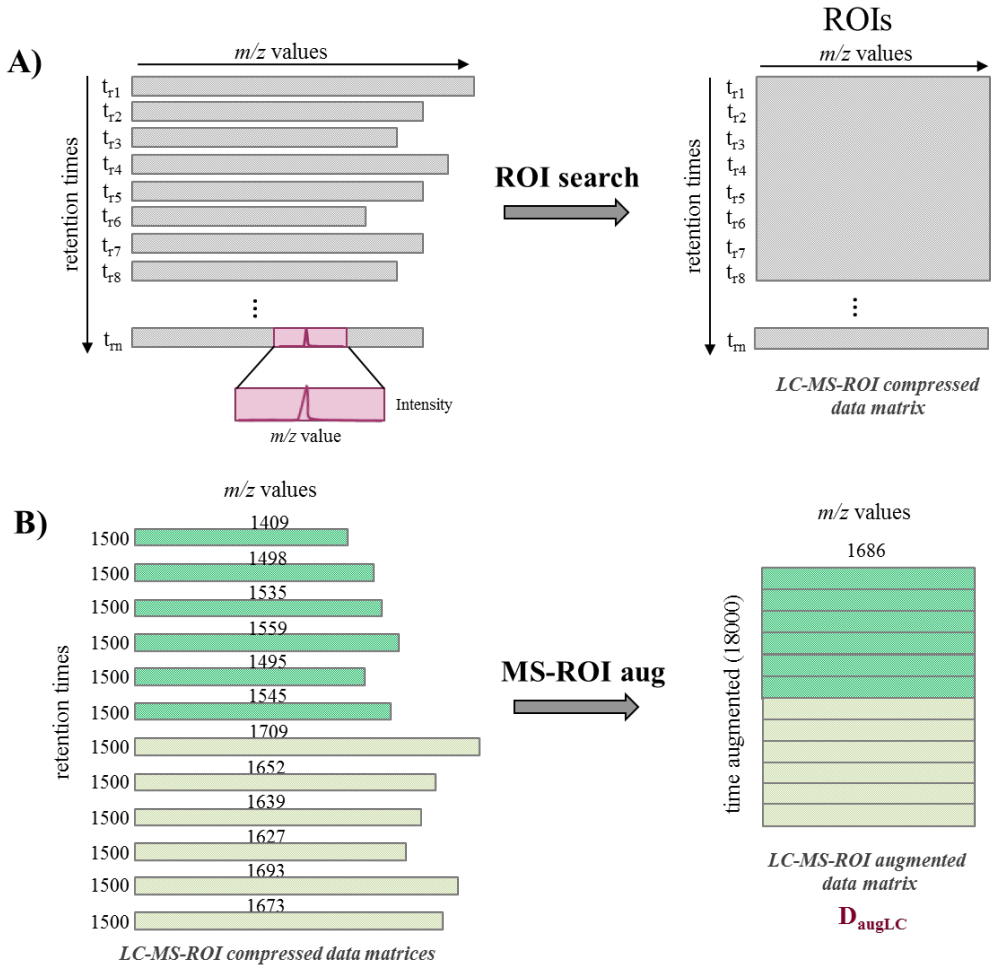


Figure S2. (A) Steps involved in the compression of one LC-MS data matrix by the ROI compression method: chromatographic information at each retention time is reorganized into a matrix grouping the common ROIs among the different retention times fulfilling the search parameters. (B) MS-ROI augmentation strategy among all considered data matrices: A pairwise search of ROI among the previous generated individual ROI data matrices (each rectangle on the left corresponds to a single chromatographic run) provides a data matrix with an equal number of *m/z* values for all the considered samples.

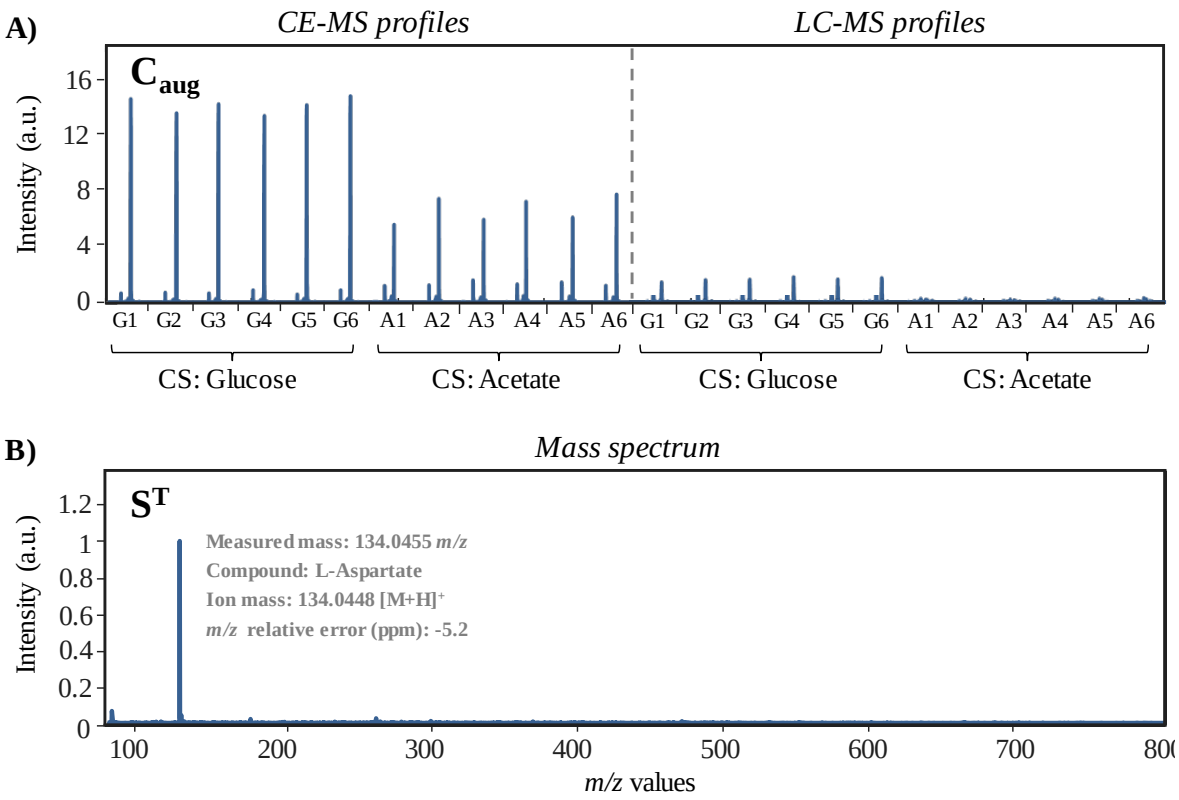


Figure S3. Example of a metabolite that was resolved by CE-MS and could also be detected in LC-MS data by MCR-ALS after low-level data fusion: (A) migration and elution profiles for the different samples. Left side: CE-MS samples considering the two carbon source (CS) fonts: glucose (G1-G6) and acetate (A1-A6). Right side: LC-MS samples considering the two carbon source fonts: glucose (G1-G6) and acetate (A1-A6). (B) Mass spectrum of the metabolite. (MCR-ALS could only resolve L-Aspartate in CE-MS data when it was independently analysed before results integration, Table 2 and 3).

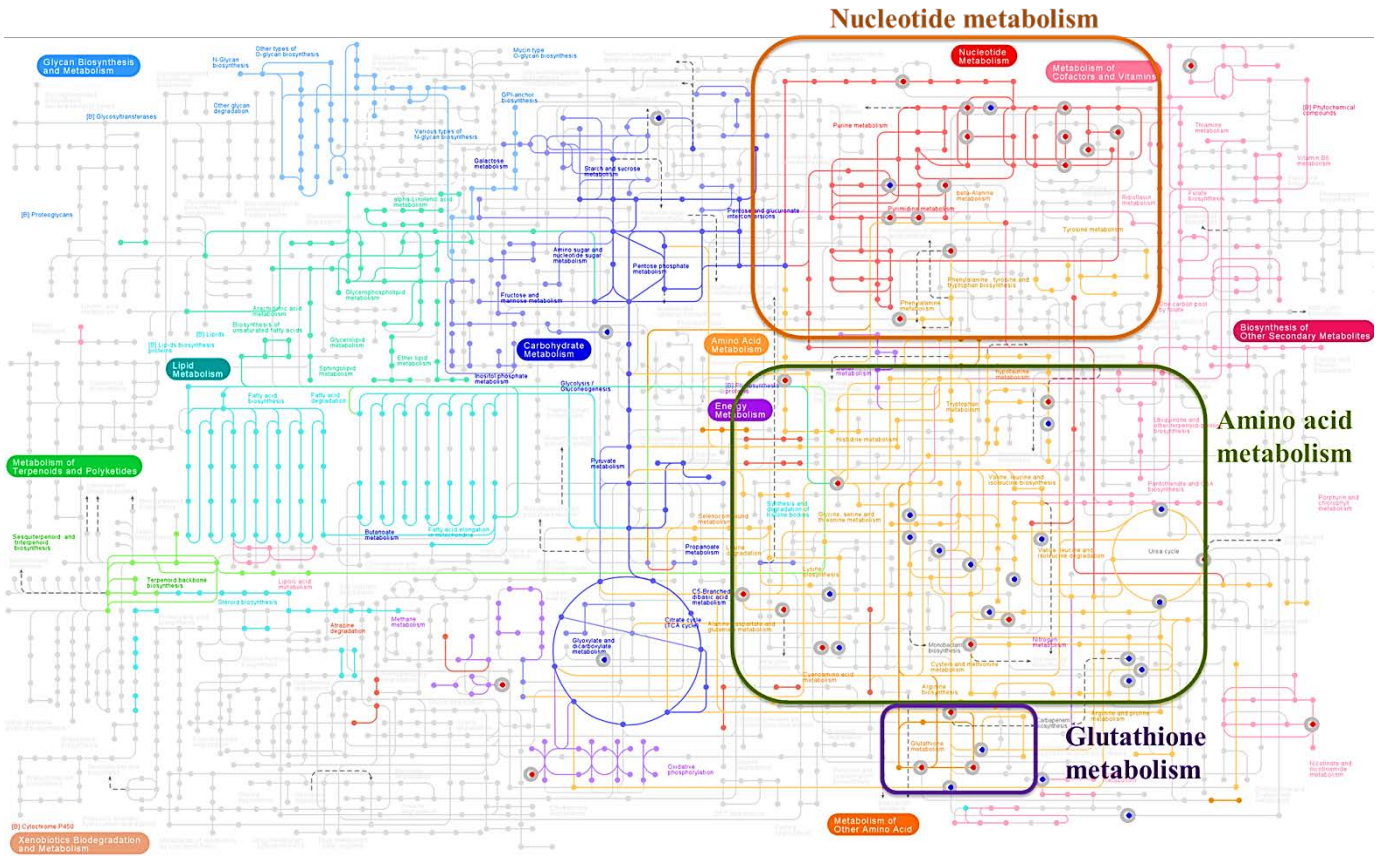


Figure S4. Metabolic map indicating the identified metabolites whose concentrations increased (in blue) or decreased (in red) in acetate-grown yeast samples compared to glucose-grown yeast samples. A grey circle is added to facilitate their location in the map. Names and KEGG symbols of the displayed metabolites are indicated in the Table 5 of the main text. Approximate location of nucleotide, amino acids, and glutathione metabolic pathways in the map are indicated by orange, green, and purple squares, respectively.

Map created using the KEGG Pathway Analysis tool (http://www.genome.jp/kegg/tool/map_pathway2.html).

PLS-DA evaluation of the features obtained by low-level data fusion and results integration approaches

In the main text, statistical assessment of the significant metabolites for sample differentiation was performed by using inferential tests (nonparametric Mann–Whitney U test with multiple hypothesis testing corrections). Alternatively, multivariate methods can also be used to evaluate both the sample differentiation (glucose- or acetate- grown samples) and the relevant features obtained when using the information provided by the MCR-ALS components. A common method to perform this type of analysis is Partial Least Squares Discriminant Analysis (PLS-DA).

Partial Least Squares Discriminant Analysis (PLS-DA) is the application of PLS method for discrimination purposes. In PLS-DA, the dependent variable (to be predicted), **Y**, is a vector or matrix that codifies the pertinence or not of a sample to a particular sample class or type. In this method, **X** contains the input information about the areas of the elution profiles resolved by MCR-ALS for each sample.

The PLS method constructs a set of loading weights (or weights) **W** which gives the relationships between **X** and **Y** during the regression process. From PLS weight vectors, Variable Importance on Projection (VIP) can be calculated to facilitate feature selection. VIP values provide a score value for each variable and rank them according to their importance in the projection used by the PLS model. In this way, the higher the VIP scores of a certain variable (generally a threshold value of one is used) are, the more importance of this variable for the discrimination model.

In this work, three PLS-DA were developed regarding:

1) PLS-DA of the MCR-ALS features for the independent analysis of LC-MS data
2) PLS-DA of the MCR-ALS features for the independent analysis of CE-MS data
3) PLS-DA of the MCR-ALS features for the low-level fused LC-MS and CE-MS data

In all cases, areas of the elution profiles resolved by MCR-ALS were autoscaled prior to the analysis. PLS-DA models were cross-validated using the leave-one-out method.

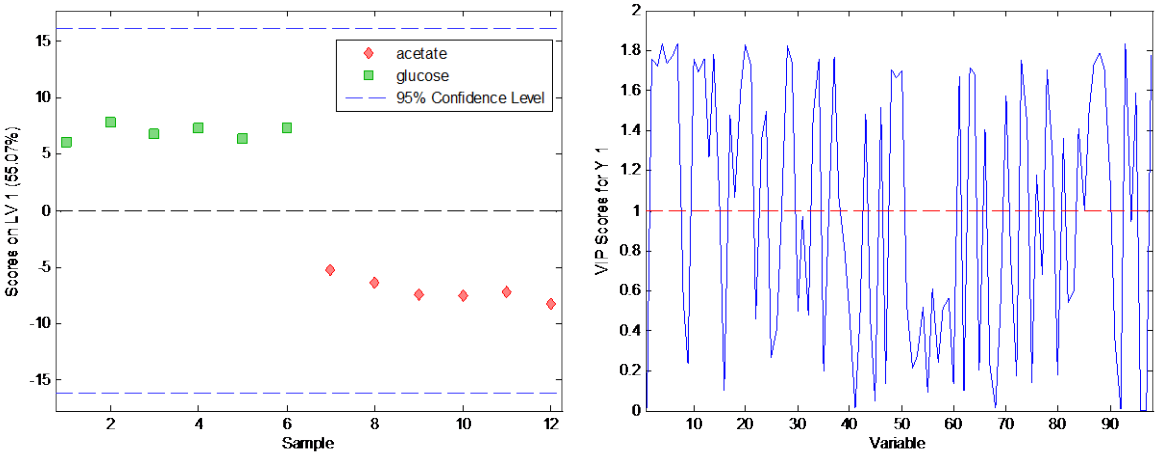
PLS-DA of the MCR-ALS features for the independent analysis of LC-MS data

PLS-DA model (glucose- and acetate- grown samples): 12 samples and 98 features

- 1 component
- Explained X-Variance: 55.07%
- Explained Y-Variance: 98.71%
- CV Sensitivity: 1.0
- CV Specificity: 1.0

Scores plot clearly distinguishes between the two groups of samples considering just the first latent variable (A). When considering VIP scores, the most relevant variables for the sample discrimination can be identified (B).

A) **B)**



From the VIP Scores, the 20 most relevant metabolites can be identified (two metabolites with VIP value higher than 1.70 could not be identified):

VIP Score	**Metabolite**
1.83	L-Carnitine
1.83	L-Proline
1.83	N-(L-Arginino)succinate
1.82	2-deoxy-D-ribose 1-phosphate
1.79	Uridine 5'-monophosphate (UMP)
1.78	Nicotinamide adenine dinucleotide (NAD+)
1.78	Adenosine monophosphate (AMP) or Deoxyguanosine monophosphate (dGMP)
1.77	L-Phenylalanine
1.77	Biotin
1.76	L-Glutamate
1.76	Pyroglutamic acid
1.75	Glutathione
1.75	N6-(L-1,3-Dicarboxylpropyl)-L-lysine
1.74	L-Tyrosine

1.73	L-Leucine
1.73	LysoPC(18:1(9Z))
1.72	Trehalose
1.72	N-acetyl-L-glutamate
1.72	Guanine
1.71	Glycerophosphocholine

All these metabolites were also found as statistically relevant by means of the inferential tests.

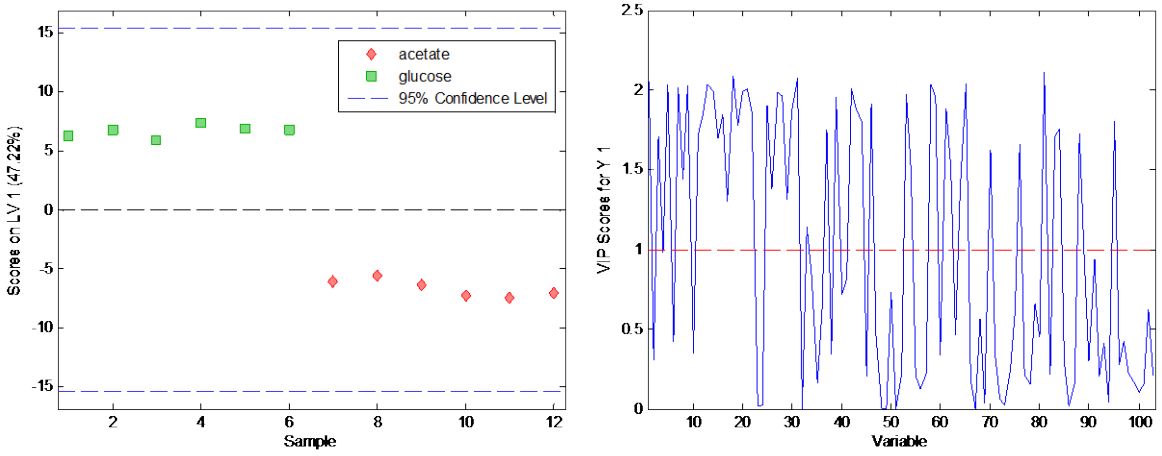
PLS-DA of the MCR-ALS features for the independent analysis of CE-MS data

PLS-DA model (glucose- and acetate- grown samples): 12 samples and 103 features

- 1 component
- Explained X-Variance: 47.22%
- Explained Y-Variance: 99.26%
- CV Sensitivity: 1.0
- CV Specificity: 1.0

Again, a clear sample discrimination can be observed in the scores plot (C) whereas VIP scores allowed selecting the most relevant features (D).

C) **D)**



From the VIP Scores, the 20 most relevant metabolites can be identified (four metabolites with VIP value higher than 1.87 could not be identified):

VIP Score	**Metabolite**
2.11	L-Valine
2.07	gamma-Amino-gamma-cyanobutanoate
2.05	L-Proline

2.04	Acetyllysine
2.03	L-Leucine
2.03	L-Serine
2.03	L-Lysine
2.01	L-Threonine
2.01	L-Cystathione
2.01	4-guanidinobutanoic acid
2.00	N-(L-Arginino)succinaate
1.99	4-Aminobutanoate (GABA)
1.97	Cytidine monophosphate (CMP)
1.96	5'-Methylthioadenosine
1.96	5-(2-Hydroxyethyl)4-methylthiazole
1.90	N6-(L-1,3-Dicarboxypropyl)-L-Lysine
1.88	N-acetyl-L-glutamate
1.88	2-deoxy-D-ribose
1.87	Glycerophosphocoline
1.86	L-Phenylalanine

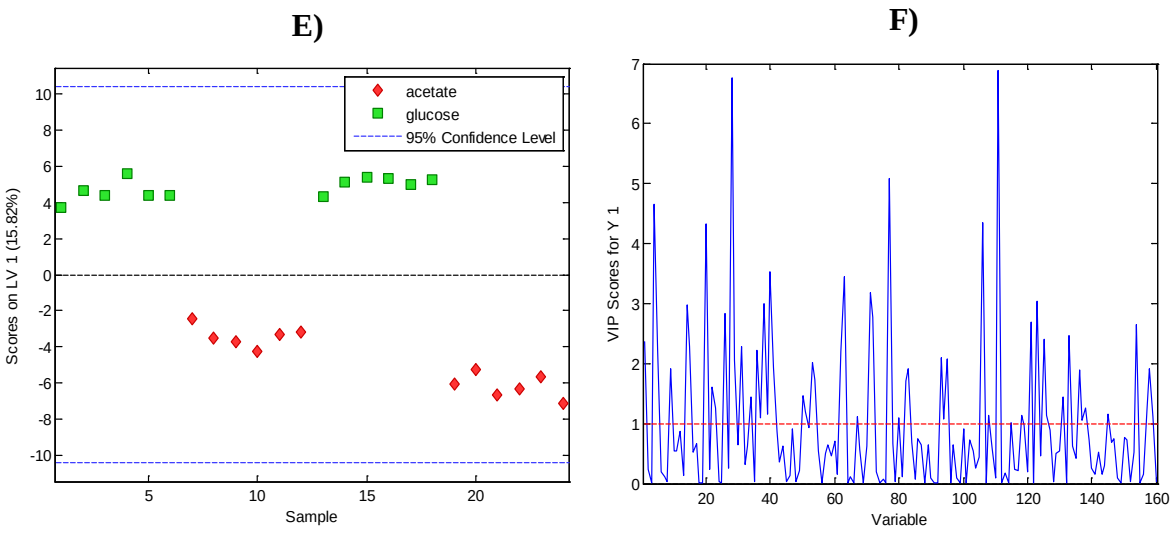
All these metabolites were also found as statistically relevant by means of the inferential tests.

PLS-DA of the MCR-ALS features for the low-level fused LC-MS and CE-MS data

PLS-DA model (glucose- and acetate- grown samples): 24 samples and 160 features

- 1 components
- Explained X-Variance: 15.82%
- Explained Y-Variance: 94.74%
- CV Sensitivity: 1.0
- CV Specificity: 1.0

Finally, the fused model allowed discriminating among samples for both LC-MS and CE-MS in the scores plot (E). Most relevant features can be selected from the VIP scores plot (F).



From the VIP Scores, the 20 most relevant metabolites can be identified (four metabolites with VIP value higher than 1.87 could not be identified):

VIP Score	**Metabolite**
6.88	gamma-Amino-gamma-cyanobutanoate
6.77	Phosphoglycolic acid
4.65	L-Proline
4.36	L-Glutamine
4.34	N6-(L-1,3-Dicarboxypropyl)-L-Lysine
3.54	L-Methionine
3.46	N-acetyl-L-glutamate
3.18	Adenylsuccinic acid
3.03	Guanine
2.99	N-(L-Arginino)succinate
2.99	L-Phenylalanine
2.84	4-Aminobutanoate (GABA)
2.65	Acetyllysine
2.48	L-Carnitine
2.40	Glycerophosphocholine
2.37	Trehalose
2.28	L-Valine
2.28	Adenosine monophosphate (AMP) or Deoxyguanosine monophosphate (dGMP)
2.23	4-guanidinobutanoic acid
2.23	L-Leucine

All these metabolites were also found as statistically relevant by means of the inferential tests.

3.3. DISCUSSIÓ DELS RESULTATS

En aquesta secció es discuteixen els resultats obtinguts en les publicacions incloses en aquest capítol. En concret, es presenta el desenvolupament i l'optimització de les metodologies posades a punt en aquesta Tesi, tant analítiques com quimiomètriques, per dur a terme estudis de metabolòmica no dirigida basats en MS.

A continuació, es mostra de forma resumida les diferents fases de treball que s'han emprat en l'establiment de les metodologies analítiques de HILIC-MS i CE-MS (**Figura 3.1**) i de tractament de dades presentades en els treballs anteriors, i que serveix també com a introducció de la discussió dels resultats d'aquest capítol.

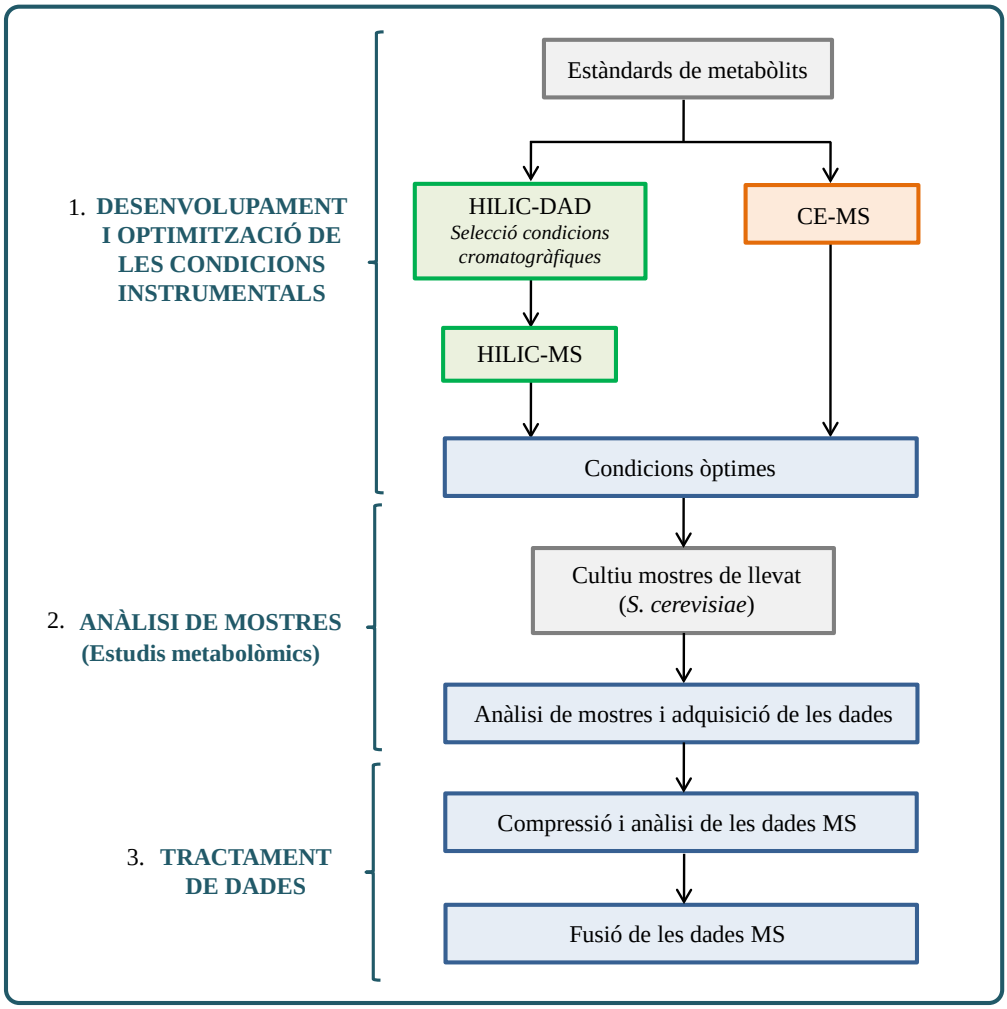


Figura 3.1. Fases de treball per a l'anàlisi no dirigida del metaboloma.

3.3.1. Anàlisi de metabòlits per LC i CE

La primera etapa en el desenvolupament de mètodes analítics per a obtenir els perfils metabòlics de sistemes biològics és l'anàlisi de patrons representatius de diverses famílies de metabòlits a diferents condicions experimentals. En les dues primeres publicacions d'aquest capítol es van avaluar diferents condicions experimentals de LC i CE per a l'anàlisi de metabòlits.

Les fases estacionàries HILIC amida i zwitteriònica són una gran alternativa en separacions de LC de metabolòmica no dirigida

En els estudis metabolòmics basats en LC primer cal escollir la fase estacionària HILIC més adequada en cada cas. Avui en dia, tal i com s'ha descrit en la introducció de la Tesi hi ha una gran varietat de columnes HILIC, incloent de sílice pura, químicament neutres, d'intercanvi iònic i fases estacionàries zwitteriòniques [10, 11]. A més, el mecanisme de retenció d'aquestes depèn de cada tipus de fase estacionària HILIC considerada i de les condicions experimentals [12, 13]. 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 d'aquesta. Així, en la primera publicació d'aquesta Tesi es va considerar un disseny experimental (DoE) factorial complet tenint en compte quatre factors experimentals per a l'anàlisi de metabòlits mitjançant LC-DAD: el tipus de fase estacionària HILIC, el modificador orgànic, el pH i la força iònica de la fase mòbil. Amb aquest objectiu, es va analitzar una mescla de 12 metabòlits de diverses famílies (aminoàcids, nucleòtids, nucleòsids, sucres i altres, veure secció 2.2 de l'article I) en 90 condicions cromatogràfiques diferents: cinc fases estacionàries HILIC (amida, BEH amida, amina, zwitteriònica i mode mixt diol), tres condicions de pH (àcid, moderadament àcid i neutre) i tres nivells de força iònica (baix, mig i alt). D'altra banda, es va escollir la millor configuració cromatogràfica per dur a terme estudis de metabolòmica no dirigida.

El dissenys experimental que es va emprar en aquesta Tesi es pot resumir de la següent manera (**Figura 3.2**):

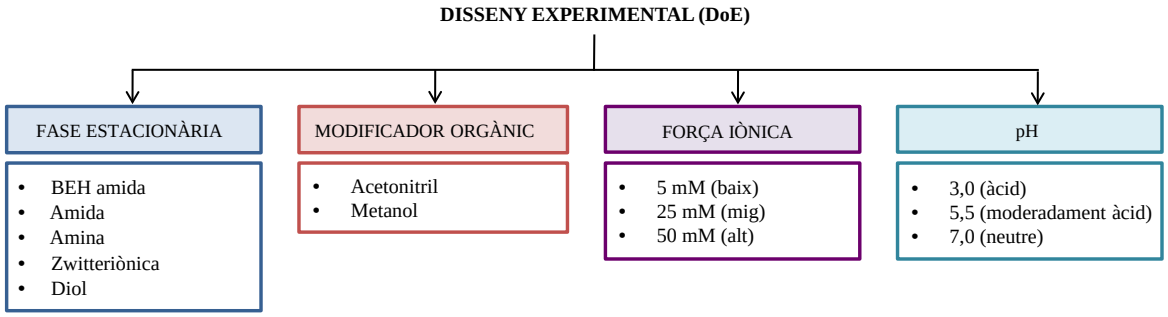


Figura 3.2. Representació gràfica del disseny experimental factorial complet emprat.

L'aplicació de diverses eines quimiomètriques va permetre determinar els factors experimentals més influents en l'anàlisi de metabòlits i trobar les millors condicions experimentals per a estudis metabolòmics no dirigits amb columnes HILIC. L'anàlisi simultània ANOVA de components (ASCA) i la funció resposta Berridge (CRF) [14] van servir per estimar la significació estadística dels factors estudiats i la interacció d'aquests en el disseny experimental. Els factors més rellevants (estadísticament significatius, $p<0.05$) són la fase estacionària, modificador orgànic i la interacció d'aquests dos factors (fase estacionària-modificador orgànic). En general, els efectes d'aquests dos factors són els únics que van resultar rellevants en el comportament dels metabòlits a les diferents condicions cromatogràfiques. A més, els valors CRF van permetre avaluar el comportament de les cinc fases estacionàries. Les dues fases estacionàries amida i lafase estacionària zwitteriònica van ser les que van proporcionar els valors CRF més alts, la qual cosa les apunta com les millors fases estacionàries per dur a terme estudis de metabolòmica [3]. Les diferències en el comportament de les dues fases estacionàries amida probablement són degudes a les diferents propietats químiques de la superfície d'aquestes fases estacionàries HILIC. En general, aquest resultats coincideixen amb els descrits recentment en altres estudis comparatius de HILIC [15-18]. En aquests treballs es va destacar el potencial de HILIC, recomanant l'ús de les fases estacionàries HILIC amida i zwitteriòniques per a la retenció de metabòlits enlloc de fases estacionàries amina i diol. Tot i així, la fase estacionària amina va mostrar en el nostre estudi resultats força acceptables. En canvi, la fase estacionària diol es va considerar la pitjor columna per a l'anàlisi de metabòlits polars, especialment quan s'utilitza metanol com a modificador orgànic (veure Figura 4 article I). Els tres mètodes emprats en aquest

treball (ASCA, anàlisi de components principals (PCA) i la funció CRF) van mostrar que la fase estacionària diol té un comportament molt diferents a la resta de fases estacionàries. Aquesta diferència en el comportament de la fase estacionària mode mixt diol s'explica per les propietats químiques duals que presenta (cadena hidrofòbica amb un grup diol). En canvi, la resta de fases estacionàries investigades presenten únicament propietats hidrofíliques (grups amida o sulfobetaïna). Per tant, les fases estacionàries més recomanables en termes de retenció i selectivitat dels metabòlits investigats són les fases estacionàries amida i zwitteriòniques. Tot i així, la fase estacionària diol podria se una bona opció per analitzar simultàniament compostos polar i no polars, com en el camp de la lipidòmica. Entre les columnes amida, la TSK Gel Amide-80 és avui en dia una de les columnes més emprades en el camp de la metabolòmica [19-22]. Per contra, entre les columnes zwitteriòniques es destaca l'ús freqüent de les columnes SeQuant® ZIC® HILIC: la ZIC-pHILIC [23-25] i, la més comuna i emprada en aquest treball, ZIC-HILIC [26-29].

A més, l'acetonitril es va considerar el millor modificador orgànic per dur a terme separacions cromatogràfiques HILIC segons els valors CRF. Això pot ser degut a que el metanol dificulta la formació de la capa d'aigua a la superfície de la fase estacionària [30]. Ara bé, l'ús de mètodes quimiomètrics com MCR-ALS poden ser una alternativa útil quan la qualitat de la separació cromatogràfica no és suficient, com en els experiments HILIC amb metanol, en que la resolució dels pics cromatogràfics per MCR-ALS millora els valors de CRF (similars als d'acetonitril). Finalment, la influència del pH i la força iònica va resultar ser menys important a la cromatografia HILIC ja que probablement siguin factors que depenen específicament de les propietats químiques dels metabòlits estudiats en cada situació en particular.

HILIC-MS és una tècnica molt potent en estudis de metabolòmica no dirigida

Seguint la línia de recerca, es van portar a terme estudis addicionals per a la separació de metabòlits amb columnes HILIC [31], els quals han permès confirmar la robustesa dels resultats obtinguts i les conclusions extretes en l'article I. En aquest treball, recentment publicat, es va analitzar un nombre major de metabòlits patró (54 metabòlits) mitjançant HILIC-MS. D'aquesta manera, es va acabar de

corroborar que el comportament de la fase estacionària amina i diol difereix considerablement de les fases estacionàries amida i zwitteriòniques, les quals presenten un mecanisme de retenció de metabòlits millor i similar.

A partir dels resultats obtinguts en els dos treballs mencionats, es va desenvolupar la metodologia HILIC-MS adequada per a l'anàlisi metabolòmica no dirigida. Es va escollir l'acetonitril com a modificador orgànic atès que millora el mecanisme de retenció dels metabòlits [30]. Entre les fases estacionàries estudiades, es va seleccionar la fase estacionària amida (TSK Gel Amide-80) ja que es va demostrar com una de les millor columnes HILIC per dur a terme estudis no dirigits de metabolòmica i, a més, va proporcionar majors valors de CRF amb acetonitril. Aquesta columna s'ha emprat en diversos estudis metabolòmics per a l'anàlisi d'una gran varietat de metabòlits en matrius biològiques, com aminoàcids, oligosacàrids, glucòsids, sucres, nucleòtids, pèptids, etc. [32-34]. També, es va decidir seleccionar com a fase aquosa l'acetat d'amoni 5 mM a pH 5,5. El contingut de sal o additiu de la fase mòbil no pot ser molt alt, ja que pot afectar a la ionització per electrosprai (ESI) i ocasionar supressió iònica [35, 36].

L'última etapa va consistir en l'optimització del gradient en aquestes condicions mitjançant la mateixa mescla de patrons (12 metabòlits) per a obtenir una bona separació cromatogràfica i un temps d'anàlisi raonable, el qual es va confirmar en l'anàlisi de mostres reals. El gradient d'elució òptim per a l'anàlisi no dirigidadel metaboloma dels organismes model posteriorment estudiats en aquesta Tesi amb solvent A (acetonitril) i solvent B (5 mM d'acetat d'amoni ajustat a pH 5,5 amb àcid acètic) és el següent: 0-8 min, gradient lineal del 25 al 30% B; 8-12 minuts, del 30 al 60% B; 12-17 minuts, 60% B; 17-20 min, es torna del 60% al 25% B; i de 20 a 27 minuts, 25% B. A la **Figura 3.3** es mostra la separació cromatogràfica de la mescla patró en les condicions experimentals optimitzades.

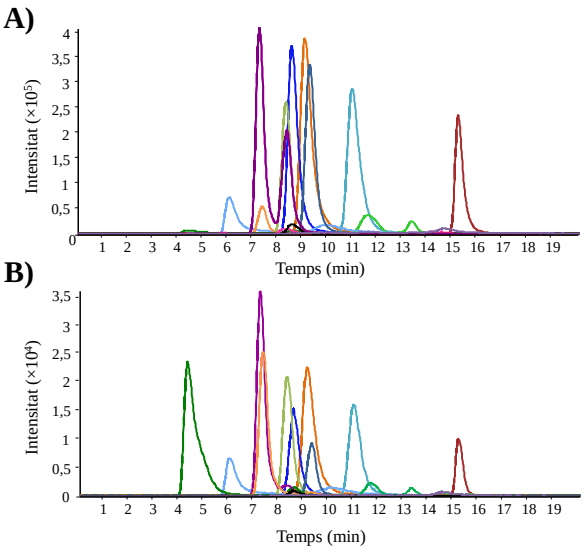


Figura 3.3. Cromatogrames d'ions extrets (EICs) dels metabòlits de la mescla patró a una concentració de 15 µg·mL^{-1} en (a) ESI positiu i (b) ESI negatiu.

Altres consideracions que cal tenir en compte per fer servir la cromatografia HILIC en estudis de metabolòmica és la importància dels temps d'estabilització de la columna per tal d'evitar problemes de poca reproductibilitat dels resultats [3]. A més, HILIC és un mode de separació que esta evolucionant molt ràpidament degut al seu potencial per a l'anàlisi de compostos molt i moderadament polars i iònics. En els últims cinc anys han aparegut noves columnes HILIC, com de gel sílice o polímers orgànics [37]. També, s'ha reduït la mida de partícula del rebliment de les columnes HILIC per tal d'augmentar l'eficàcia de la separació cromatogràfica. La majoria de les primeres columnes HILIC que van aparèixer i que s'han emprat en aquesta Tesi tenien una mida de partícula de 5 µm (cromatografia de líquids d'alta eficàcia, HPLC), mentre que actualment gairebé totes les cases comercials han desenvolupat fases estacionàries HILIC entre 2-3 µm de mida de partícula (cromatografia de líquids d'ultraalta eficàcia, UHPLC).

CE-MS és una alternativa poderosa per a l'anàlisi dels perfils metabòlics

En la segona publicació d'aquesta Tesi, es va investigar el potencial de la tècnica CE-MS en estudis de metabolòmica no dirigida. Es van provar diferents metodologies CE-MS per a una anàlisi adequada del metaboloma tenint en especial consideració l'optimització de les condicions de separació i detecció en ambdós modes d'ionització ESI (positiu i negatiu). Amb aquest objectiu, es van analitzar mescles complexes d'estàndards cobrint un ampli ventall de famílies de metabòlits amb propietats

fisicoquímiques (àcids orgànics, sucres, nucleòtids, nucleòsids i aminoàcids) i concentracions diferents (de 10 a 100 $\mu g \cdot mL^{-1}$) (veure apartat 2.1 de l'article II). Les condicions òptimes de CE-MS es van seleccionar especialment en termes de sensibilitat i reproductibilitat per assegurar una informació completa i fiable dels perfils metabòlics de les mostres d'interès.

Primer, es van investigar l'àcid acètic i l'àcid fòrmic a concentracions 0,5 i 1,0 M i una barreja 50 mM d'àcid acètic i 50 mM d'àcid fórmic com a electròlits de separació (BGE) àcids per dur a terme les separacions de CE en ESI positiu, proposats en treballs metabolòmics anteriors [38, 39]. En aquestes condicions, alguns àcids orgànics, nucleòtids i sucres no es van poder detectar en ESI positiu d'acord amb la seva càrrega elèctrica en aquests BGEs. En els cas dels metabòlits detectats en ESI positiu (per exemple, alguns àcids orgànics, nucleòsids i aminoàcids), es va demostrar que el BGE d'àcid acètic 1,0 M oferia els millors resultats en termes de sensibilitat i que generava una corrent elèctrica més baixa (<50 µA), evitant així, danyar el sistema electrònic de l'espectròmetre de masses (TOF). A més, com l'acoblament CE-MS es realitza amb una interfície de líquid auxiliar coaxial (*sheath flow*) es va optimitzar la composició d'aquest líquid. El líquid auxiliar ajuda a tancar el circuit elèctric a la sortida del capil·lar de separació i a augmentar el cabal proporcionat per la CE [40]. En conseqüència, es van comparar el 2-propanol i el metanol com a modificadors orgànics del líquid auxiliar a diferents percentatges entre el 50 i el 100% (v/v) amb un 0,05 o 0,5% (v/v) d'àcid fòrmic. La composició 60:40 (v/v) de 2-propanol: aigua amb un 0,05% (v/v) d'àcid fòrmic va oferir els millors resultats en termes de sensibilitat en ESI positiu.

D'altra banda, es van optimitzar les condicions en ESI negatiu ja que alguns metabòlits no es podien detectar en ESI positiu o la sensibilitat era baixa. Malgrat que molt estudis es limiten a estudiar el metaboloma en ESI positiu [41, 42]. Per contra, alguns treballs han proposat l'ús de capil·lars amb recobriments polimèrics carregats positivament i així invertir el flux electroosmòtic (EOF), polimèrics inerts o simplement capil·lars de sílice fosa. En el cas d'emprar capil·lars polimèrics iònics, com SMILE(+) [43, 44] o COSMO(+) [45], amb el temps s'ha vist que no són una bona alternativa degut a la possible adsorció dels anàlits de les mostres biològiques a les parets del capil·lar el qual ocasiona problemes en la migració dels metabòlits, eixamplament i asimetria de banda dels pics electroforètics

[46]. En canvi, els capil·lars recoberts de polímers inerts, com PEEK o PTFE [46], eliminen el fenomen d'adsorció a la paret i permeten obtenir una bona repetibilitat per als temps de migració en l'anàlisi metabolòmic. Malgrat això, aquest tipus de capil·lars no són una tecnologia econòmica. Per aquest motiu, en aquesta Tesi s'han utilitzat capil·lars de sílice fosa en ESI negatiu però tenint especial cura en l'etapa de filtrat de les mostres. Tanmateix, el desplaçament dels pics en els temps de migració no va ser molt gran ja que els experiments portats a terme no requerien d'un nombre elevat d'injeccions degut al nombre reduït de mostres i replicats tècnics. A més, altres treballs també han demostrat que es pot obtenir informació bona dels perfils metabòlics en ESI negatiu emprant capil·lars de sílice fosa [47-50].

En ESI negatiu, es recomana utilitzar un BGE dèbilment bàsic i volàtil de formiat o acetat d'amoni [46, 51-54]. Finalment, es va emprar un BGE d'acetat d'amoni 25 mM (pH 8,5) per analitzar les diferents mescles de metabòlits patró. A aquest valor de pH, els anions migren cap al càtode gràcies al flux electroosmòtic (EOF) i es pot obtenir una bona sensibilitat en ESI negatiu per a la majoria dels metabòlits que no es van poder detectar en ESI positiu. En aquest cas, es va investigar la composició del líquid auxiliar entre un 50 i un 100% (v/v) de 2-propanol i metanol amb l'addició d'amoníac entre el 0,05 i el 0,5% (v/v). La composició 60:40 (v/v) 2-propanol: aigua al 0,5% d'amoníac (v/v) va proporcionar les condicions òptimes de sensibilitat.

En resum, CE-MS sota les condicions mencionades va permetre obtenir una anàlisi reproduïble i sensible dels metabòlits tant catiònic com aniònics.

3.3.2. MCR-ALS: una eina útil per a l'anàlisi de dades metabolòmiques no dirigides

Els espectres d'alta resolució *full scan* en mode *profile* de metabolòmica no dirigida són molt voluminosos i generen grans matrius de dades. Per consegüent, és obligatòria la reducció de la mida d'aquestes dades metabolòmiques pel seu preprocessament i anàlisi. En la segona i tercera publicació d'aquest capítol s'han presentat diferents estratègies d'anàlisi de dades metabolòmiques no dirigides. En concret, es van analitzar les dades corresponents a dos estudis relacionats amb els efectes de les condicions de creixement de *S.cerevisiae*, com la temperatura i la font de carboni emprada. En aquests

estudis es va demostrar la utilitat de l'etapa de compressió de les dades i de l'anàlisi mitjançant el mètode MCR-ALS. D'una banda, la compressió i reducció de les dimensions de les matrius de dades té grans avantatges, com la disminució dels requeriments de memòria durant el processament de les dades i del seu temps d'anàlisi i, en el cas d'emprar el procediment de cerca de les regions d'interès (ROI), s'eviten els problemes derivats de la pèrdua de resolució i qualitat de la informació. D'altra banda, l'anàlisi mitjançant MCR-ALS permet obtenir els perfils de concentració i espectrals dels metabòlits, a partir del qual es pot realitzar les estimacions dels canvis en les concentracions relatives dels metabòlits estudiats i la seva identificació.

Estratègies de compressió de les dades metabolòmiques

Abans de l'anàlisi per MCR-ALS, és necessari reduir les dimensions de les matrius de dades de CE-MS i HILIC-MS. Hi ha diferents mètodes de compressió de les dades metabolòmiques. En la segona publicació d'aquest capítol (article II) la compressió de les dades es va realitzar mitjançant un procediment d'interpolació (veure secció 2.3.3 de la introducció de la Tesi) que es va aplicar als espectres MS obtinguts a cada temps de migració de les dades experimentals de CE-MS. D'aquesta manera, els espectres MS *raw* d'alta resolució obtinguts en els diferents temps de migració es converteixen en espectres de baixa resolució on els valors *m/z* resultants són iguals per tots els temps de migració. Com que el grau d'interpolació emprat era petit (una resolució de 0,01 Da/e), aquesta compressió inicial no va ser suficient per poder processar tot el conjunt de dades obtingudes en l'anàlisi simultània de múltiples experiments CE-MS, raó per la qual es va aplicar una divisió dels electroferogrames en diferents regions de temps (*time windowing*) i en diferents regions espectrals (*spectral windowing*). Tot aquest procés de compressió i de subdivisió de les dades obtingudes va permetre analitzar els electroferogrames de totes les mostres simultàniament per MCR-ALS (article II). No obstant això, aquesta estratègia va necessitar d'un nombre gran de models MCR-ALS (un per cada finestra espectral de cada regió de l'electroferograma). A més, per poder conèixer el valor de la massa exacta dels metabòlits d'interès per la seva correcta identificació es va requerir de la cerca de les masses del ions a les dades originals (*raw*), ja que degut al grau d'interpolació emprat (0,01 Da/e) sol es van conservar dos decimals en els valors de *m/z*. Aquest procediment de compressió és similar

al *binning*, que es va emprar en l'article V i que presenta avantatges i inconvenient similars a la interpolació. En canvi, en l'article III d'aquesta Tesi es presenta una aproximació diferent de compressió de les dades metabolòmiques no dirigides de CE-MS i HILIC-MS mitjançant el procediment de cerca i selecció de les regions d'interès (ROI) abans de la seva anàlisi per MCR-ALS (veure secció 2.3.3 de la introducció de la Tesi), emprant el mètode anomenat ROIMCR [55, 56]. L'aplicació d'aquest procediment ROI va comportar una enorme reducció de la dimensió espectral sense cap pèrdua de resolució de *m/z* (conserva la resolució espectral original), eliminant aquella informació que no era rellevant. L'únic inconvenient que suposa el procés de compressió ROI de les dades de l'article III és l'optimització dels paràmetres de l'algorisme per buscar els senyals MS significatius (valor llindar de relació senyal/soroll, error *m/z* i nombre mínim de punts per definir un pic cromatogràfic). Aquests paràmetres s'han d'optimitzar tenint en compte l'instrument MS emprat. A la **Figura 3.4** es mostra un exemple d'un cromatograma de LC-MS i un electroferograma de CE-MS després de dur a terme la compressió ROI en l'anàlisi d'una mostra de llevat.

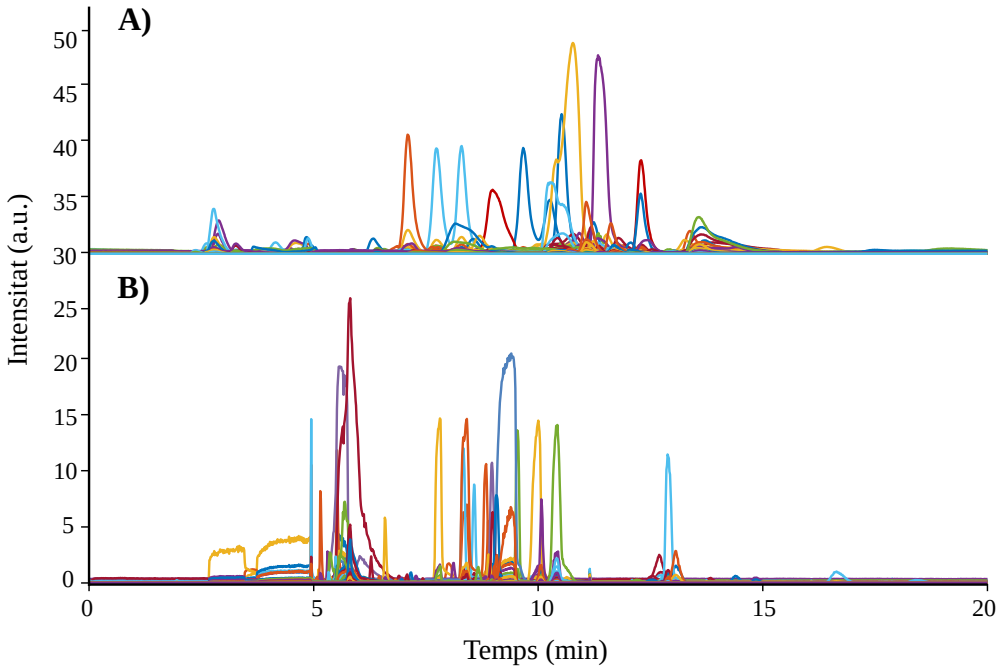


Figura 3.4. Compressió ROI d'un (a) cromatograma HILIC-MS (aproximadament 1600 valors *m/z*) (b) electroferograma CE-MS (aproximadament 1200 valors *m/z*) d'una mostra control de llevat cultivada a condicions òptimes de creixement.

Resolució dels metabòlits mitjançant MCR-ALS

Una vegada comprimides les dades es va realitzar la resolució dels metabòlits presents en les mostres analitzades cromatogràficament o electroforèticament mitjançat el mètode MCR-ALS. Aquest mètode és una eina eficaç per a la resolució dels perfil d'elució (LC-MS) o de migració (CE-MS) i dels espectres dels metabòlits presents en les mostres analitzades per mètodes cromatogràfics [57, 58] i electroforètics [59, 60]. En el cas de les dades metabolòmiques de HILIC-MS i CE-MS, aquest mètode va permetre resoldre els perfils purs dels metabòlits tot i la complexitat d'aquestes dades (múltiples coelucions i pics totalment solapats). A la **Figura 3.5** es representa un exemple de resolució de quatre perfil d'elució resolts pel procediment combinat ROIMCR quan s'analitzen 12 mostres simultàniament de HILIC-MS. A més, aquest procediment permet resoldre alguns problemes cromatogràfics i electroforètics addicionals, com ara l'eliminació de les contribucions de soroll de fons, la millora de les relacions senyal/soroll, la resolució de pics asimètrics i la no necessitat d'alineament dels pics entre cromatogrames o electroferogrames d'un mateix metabòlit. Aquestes dificultats són especialment crítiques en les separacions de CE.

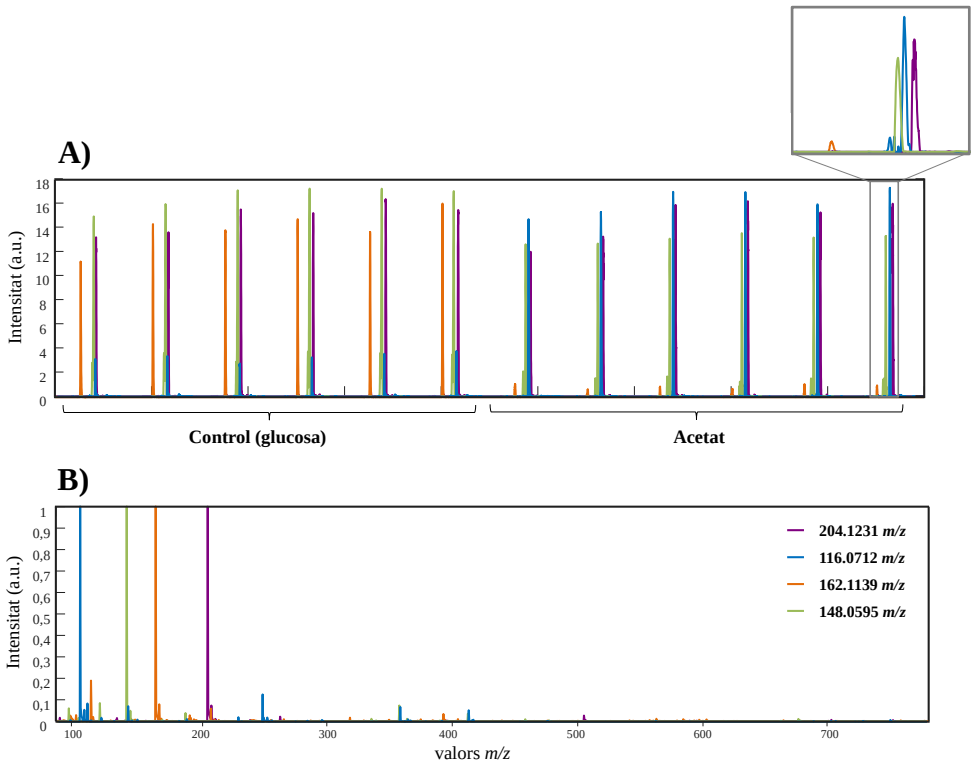


Figura 3.5. Exemple de quatre (a) perfils d'elusió i (b) espectrals resolts per MCR-ALS de les mostres de llevat cultivades amb una font de carboni fermentable (glucosa) i no fermentable (acetat).

En els articles II i III s'ha destacat el potencial de la combinació d'eines de compressió (com interpolació, *binning* o ROI) amb el mètode MCR-ALS per a la detecció de biomarcadors en estudis òmics. Un cop resolts els perfils d'elució dels diferents metabòlits es poden calcular les seves àrees. Diverses publicacions recents han aprofitat la metodologia interpolació-MCR o *binning*-MCR per a la investigació de biomarcadors biològics per LC-MS [61-63], CE-MS [41], per espectroscòpia de ressonància magnètica nuclear (RMN) [64], i a partir d'imatges d'espectrometria de masses (MSI) [65] i hiperespectrals [66, 67]. Tot i així, el procediment ROIMCR resulta més eficaç que les eines més tradicionals ja que permet minimitzar els requisits d'emmagatzematge i el temps d'anàlisi i evita les pèrdues d'exactitud de massa dels valors de *m/z*. Conseqüentment, la metodologia ROIMCR s'ha consolidat per a l'anàlisi de dades de LC-MS [68-70] i de dades més complexes i voluminoses com són les dades de cromatografia de líquids bidimensional acoblada a l'espectrometria de masses (LC×LC-MS) [71] o imatges d'espectrometria de masses (MSI) [72].

3.3.3. Fusió de dades amb el procediment ROIMCR: anàlisi de conjunts massius de dades per a una millor comprensió dels processos biològics

En aquesta Tesi també s'han proposat dues estratègies d'integració de dades metabolòmiques no dirigides basades en el procediment ROIMCR per a la resolució dels perfils d'elució o migració dels metabòlits. En concret, s'ha demostrat la utilitat d'aquestes metodologies per a la fusió de dades de dues plataformes diferents d'espectrometria de masses (HILIC-MS i CE-MS). Un dels gran avantatges dels procediments de fusió de dades és que permeten obtenir una interpretació biològica més completa dels canvis donats en els processos biològics. A partir d'aquests procediments les debilitats d'una plataforma analítica es poden compensar amb les fortaleses de l'altra aprofitant els diferents mecanismes de separació d'aquestes.

D'una banda, es possible fer la integració de la informació a nivell conceptual (fusió de dades de nivell alt), on es realitza l'anàlisi independent de cadascun dels conjunt de dades i, finalment, es combina la informació extreta obtinguda a partir de cada conjunt o bloc de dades i s'obté una interpretació biològica conjunta millorada (**Figura 3.6a**). D'altra banda, es proposa fer una fusió

directa de les dades originals de HILIC-MS i CE-MS (fusió de dades de nivell baix). En aquest cas, s'aprofita que el mode espectral entre els dos blocs de dades és comú i s'analitzen conjuntament els dos conjunts de dades mitjançant la metodologia ROIMCR per aconseguir una interpretació biològica més precisa (**Figura 3.6b**).

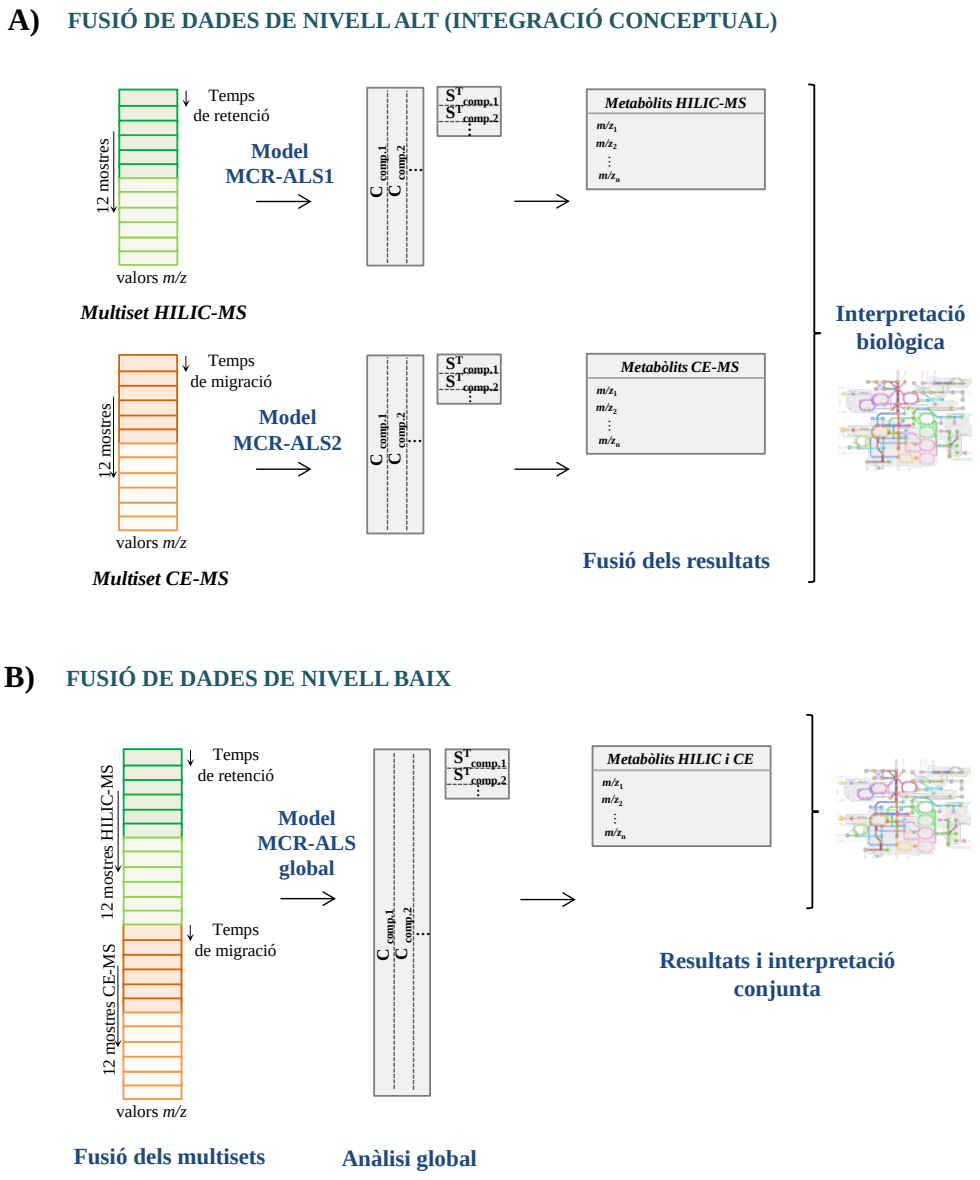


Figura 3.6. Representació esquemàtica de les estratègies d'integració basades en ROIMCR.

A partir dels resultats obtinguts amb les dues estratègies presentades (**Figura 3.6**), es va veure que les dues metodologies permetien confirmar gairebé els mateixos metabòlits rellevants (biomarcadors), la qual cosa reforçava l'explicació i interpretació de la resposta biològica de *S. Cerevisiae* quan s'estudia

en condicions de creixement adverses. Tanmateix, la fusió de dades de nivell baix va ampliar la cobertura de metabòlits estadísticament significatius respecte a la fusió de dades de nivell alt (60 vs. 52) (**Figura 3.7a** i **b**). La integració conceptual (de nivell alt) va proporcionar únicament 2 metabòlits nous respecte els que es van resoldre mitjançant la fusió de nivell baix, mentre que aquest enfocament va permetre descobrir 10 metabòlits rellevants que no s'havien detectat amb l'anàlisi independent dels blocs de dades (**Figura 3.7c**). A més, al dur a terme la integració de les dades originals amb un únic model MCR-ALS global es van poder detectar alguns metabòlits que semblaven ser específics de l'anàlisi HILIC-MS o de l'anàlisi CE-MS (veure Figura 3 article III). En concret, la fusió de dades de nivell baix va resultar ser una estratègia més poderosa per a la resolució d'aquells metabòlits d'interès que es troben a concentracions més baixes (o pels quals l'anàlisi és de sensibilitat menor). Per exemple, alguns senyals MS de baixa intensitat de metabòlits que es podien superposar amb els senyals d'altres contribucions, com els del soroll de fons, només es van resoldre apropiadament quan es va realitzar l'anàlisi simultània de les dades de HILIC-MS i CE-MS. Els metabòlits que no es detecten bé per una de les tècniques degut a que les condicions experimentals no són favorables (per exemple, per supressió iònica o perquè la seva concentració en la mostra és molt baixa) es podran resoldre millor si la seva detecció en l'altra tècnica és més bona. Així doncs, amb l'estratègia d'integració de baix nivell es combinen els avantatges i es minimitzen els inconvenients de cadascuna de les tècniques individuals per millorar la qualitat dels resultats. Per tant, aquesta estratègia ofereix una millor cobertura del metabolisme en relació a la combinació directa de la informació extreta de cadascuna de les tècniques per separat o integració conceptual.

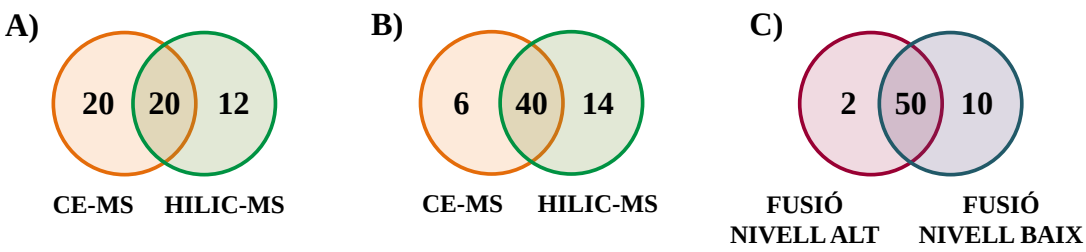


Figura 3.7. Diagrama de Venn dels biomarcadors identificats quan es porta a terme la (a) fusió de nivell alt, (b) fusió de nivell alt i (c) comparativa de les dues estratègies de fusió de dades.

En el cas concret d'utilitzar les tècniques de CE-MS i HILIC-MS, la majoria de metabòlits identificats es van poder detectar amb ambdues plataformes (veure Taula 4 article III). En aquesta tercera publicació es va confirma la idoneïtat de les estratègies de fusió de dades presentades, i en especial, de la fusió de nivell baix per a l'obtenció d'informació més precisa i amb menor incertesa i així, poder possibilitar una interpretació biològica més acurada. A més, aquesta estratègia presentada dóna lloc a noves possibilitats de fusió de dades de MS, com per exemple la plataforma de CE-MS i LC-MS de fase invertida, que sens dubte permetria augmentar la cobertura de metabòlits detectats. A mode de resum, a la **Taula 3.1** es mostren els principals avantatges i inconvenients de les dues metodologies de fusió de dades presentades en aquesta Tesi.

Taula 3.1. Avantatges i inconvenients de les dues estratègies de fusió de dades basades en ROIMCR.

Fusió	Avantatges	Inconvenients
Nivell alt (conceptual)	Anàlisi de cada bloc de dades simple.	Major nombre d'anàlisis independents. Temps de computació més llarg.
Nivell baix	Un únic anàlisi. Temps de computació curt. Major cobertura de metabòlits. Millor interpretació de les dades.	Major complexitat en l'anàlisi de les dades. Requereix d'escalat de les dades (si l'escala dels blocs és diferent).

En resum, els resultats obtinguts en la tercera publicació d'aquesta Tesi van demostrar que ambdues estratègies basades en ROIMCR per a la fusió de les dades són útils per a realitzar la integració dels resultats obtinguts mitjançant estudis de metabolòmica no dirigida i, en especial, la integració de nivell baix. Aquests procediments permeten una millor detecció i caracterització dels biomarcadors i interpretació dels canvis bioquímics dels sistemes biològics. A més, són una possible alternativa als mètodes de fusió de dades que generalment s'utilitzen en el camp de l'òmica per a la integració de dades a partir de mètodes en els quals es considera un nombre reduït de variables o components.

3.4. REFERÈNCIES

1. T'Kindt, R., Storme, M., Deforce, D., Van Bocxlaer, J., Evaluation of hydrophilic interaction chromatography versus reversed-phase chromatography in a plant metabolomics perspective, *Journal of Separation Science*. 2008, *31*, 1609-1614.
2. 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.
3. Wernisch, S., Pennathur, S., Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics, *Analytical and Bioanalytical Chemistry*. 2016, *408*, 6079-6091.
4. Hirayama, A., Wakayama, M., Soga, T., Metabolome analysis based on capillary electrophoresis-mass spectrometry, *TrAC Trends in Analytical Chemistry*. 2014, *61*, 215-222.
5. Bouhifd, M., Hartung, T., Hogberg, H. T., Kleensang, A., Zhao, L., Review: toxicometabolomics, *Journal of Applied Toxicology:JAT*. 2013, *33*, 1365-1383.
6. Madsen, R., Lundstedt, T., Trygg, J., Chemometrics in metabolomics-A review in human disease diagnosis, *Analytica Chimica Acta*. 2010, *659*, 23-33.
7. Richards, S. E., Dumas, M. E., Fonville, J. M., Ebbels, T. M. D., Holmes, E., Nicholson, J. K., Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework, *Chemometrics and Intelligent Laboratory Systems*. 2010, *104*, 121-131.
8. Gligorijević, V., Pržulj, N., Methods for biological data integration: perspectives and challenges, *Journal of the Royal Society Interface*. 2015, *12*, 20150571.
9. Jewett, M. C., Hansen, M., Nielsen, J., in: Jewett, M. C., Nielsen, J. (Eds.), Data acquisition, analysis, and mining: Integrative tools for discerning metabolic function in Saccharomyces cerevisiae. *In Metabolomics: A Powerful Tool in Systems Biology. Topics in Current Genetics.* Springer, Berlin, Heidelberg. 2007, pp. 159-187.
10. Jandera, P., Stationary and mobile phases in hydrophilic interaction chromatography: a review, *Analytica Chimica Acta*. 2011, *692*, 1-25.
11. Buszewski, B., Noga, S., Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique, *Analytical and Bioanalytical Chemistry*. 2012, *402*, 231-247.
12. Greco, G., Letzel, T., Main interactions and influences of the chromatographic parameters in HILIC separations, *Journal of Chromatographic Science*. 2013, *51*, 684-693.
13. Rakic, T., Jancic Stojanovic, B., Malenovic, A., Ivanovic, D., Medenica, M., Improved chromatographic response function in HILIC analysis: application to mixture of antidepressants, *Talanta*. 2012, *98*, 54-61.
14. Berridge, J. C., Unattended optimisation of reversed-phase high-performance liquid chromatographic separations using the modified simplex algorithm, *Journal of Chromatography A*. 1982, *244*, 1-14.
15. 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.
16. 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.
17. 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.
18. 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.
19. Van der Werf, M. J., Overkamp, K. M., Muilwijk, B., Coulier, L., Hankemeier, T., Microbial metabolomics: toward a platform with full metabolome coverage, *Analytical Biochemistry*. 2007, *370*, 17-25.
20. Xu, Y., Yang, L., Yang, F., Xiong, Y., Wang, Z., Hu, Z., Metabolic profiling of fifteen amino acids in serum of chemical-induced liver injured rats by hydrophilic interaction liquid chromatography coupled with tandem mass spectrometry, *Metabolomics*. 2012, *8*, 475-483.

21. Ducruix, C., Junot, C., Fievet, J. B., Villiers, F., Ezan, E., Bourguignon, J., New insights into the regulation of phytochelatin biosynthesis in A. thaliana cells from metabolite profiling analyses, *Biochimie*. 2006, *88*, 1733-1742.
22. Navarro-Reig, M., Jaumot, J., Pina, B., Moyano, E., Galceran, M. T., Tauler, R., 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, *Metallomics : Integrated Biometal Science*. 2017, *9*, 660-675.
23. Zhang, T., Watson, D. G., Evaluation of the technical variations and the suitability of a hydrophilic interaction liquid chromatography-high resolution mass spectrometry (ZIC-pHILIC-Exactive orbitrap) for clinical urinary metabolomics study, *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences*. 2016, *1022*, 199-205.
24. Zhang, R., Watson, D. G., Wang, L., Westrop, G. D., Coombs, G. H., Zhang, T., Evaluation of mobile phase characteristics on three zwitterionic columns in hydrophilic interaction liquid chromatography mode for liquid chromatography-high resolution mass spectrometry based untargeted metabolite profiling of Leishmania parasites, *Journal of Chromatography A*. 2014, *1362*, 168-179.
25. Qin, F., Zhao, Y. Y., Sawyer, M. B., Li, X. F., Column-switching reversed phase-hydrophilic interaction liquid chromatography/tandem mass spectrometry method for determination of free estrogens and their conjugates in river water, *Analytica Chimica Acta*. 2008, *627*, 91-98.
26. Kovaříková, P., Stariat, J., Klimeš, J., Hrušková, K., Vávrová, K., Hydrophilic interaction liquid chromatography in the separation of a moderately lipophilic drug from its highly polar metabolites-the cardioprotectant dexrazoxane as a model case, *Journal of Chromatography A*. 2011, *1218*, 416-426.
27. Rodriguez-Gonzalo, E., Garcia-Gomez, D., Carabias-Martinez, R., Study of retention behaviour and mass spectrometry compatibility in zwitterionic hydrophilic interaction chromatography for the separation of modified nucleosides and nucleobases, *Journal of Chromatography A*. 2011, *1218*, 3994-4001.
28. Zhang, T., Creek, D. J., Barrett, M. P., Blackburn, G., Watson, D. G., Evaluation of coupling reversed phase, aqueous normal phase, and hydrophilic interaction liquid chromatography with Orbitrap mass spectrometry for metabolomic studies of human urine, *Analytical Chemistry*. 2012, *84*, 1994-2001.
29. Cubbon, S., Bradbury, T., Wilson, J., Thomas-Oates, J., Hydrophilic Interaction Chromatography for Mass Spectrometric Metabonomic Studies of Urine, *Analytical Chemistry*. 2007, *79*, 8911-8918.
30. 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.
31. Navarro-Reig, M., Ortiz-Villanueva, E., Tauler, R., Jaumot, J., Modelling of Hydrophilic Interaction Liquid Chromatography Stationary Phases Using Chemometric Approaches, *Metabolites*. 2017, *7*, 54.
32. Tolstikov, 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.
33. 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.
34. Karlsson, G., Winge, S., Sandberg, H., Separation of monosaccharides by hydrophilic interaction chromatography with evaporative light scattering detection, *Journal of Chromatography A*. 2005, *1092*, 246-249.
35. Chen, Y., Mori, M., Pastusek, A. C., Schug, K. A., Dasgupta, P. K., On-Line Electrodialytic Salt Removal in Electrospray Ionization Mass Spectrometry of Proteins, *Analytical Chemistry*. 2011, *83*, 1015-1021.
36. Constantopoulos, T. L., Jackson, G. S., Enke, C. G., Effects of salt concentration on analyte response using electrospray ionization mass spectrometry, *Journal of the American Society for Mass Spectrometry*. 1999, *10*, 625-634.
37. Jandera, P., Janas, P., Recent advances in stationary phases and understanding of retention in hydrophilic interaction chromatography. A review, *Analytica Chimica Acta*. 2017, *967*, 12-32.
38. Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., Nishioka, T., Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry, *Journal of Proteome Research*. 2003, *2*, 488-494.

39. Benavente, F., van der Heijden, R., Tjaden Ubbo, R., van der Greef, J., Hankemeier, T., Metabolite profiling of human urine by CE-ESI-MS using separation electrolytes at low pH, *Electrophoresis*. 2006, *27*, 4570-4584.
40. Smith, R. D., Olivares, J. A., Nguyen, N. T., Udseth, H. R., Capillary zone electrophoresis-mass spectrometry using an electrospray ionization interface, *Analytical chemistry*. 1988, *60*, 436-441.
41. Pont, L., Benavente, F., Jaumot, J., Tauler, R., Alberch, J., Ginés, S., Barbosa, J., Sanz-Nebot, V., Metabolic profiling for the identification of Huntington biomarkers by on-line solid-phase extraction capillary electrophoresis mass spectrometry combined with advanced data analysis tools, *Electrophoresis*. 2016, *37*, 795-808.
42. González-Domínguez, R., García, A., García-Barrera, T., Barbas, C., Gómez-Ariza José, L., Metabolomic profiling of serum in the progression of Alzheimer's disease by capillary electrophoresis-mass spectrometry, *Electrophoresis*. 2014, *35*, 3321-3330.
43. Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M., Nishioka, T., Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry, *Analytical Chemistry*. 2002, *74*, 2233-2239.
44. Mishima, E., Fukuda, S., Mukawa, C., Yuri, A., Kanemitsu, Y., Matsumoto, Y., Akiyama, Y., Fukuda, N. N., Tsukamoto, H., Asaji, K., Shima, H., Kikuchi, K., Suzuki, C., Suzuki, T., Tomioka, Y., Soga, T., Ito, S., Abe, T., Evaluation of the impact of gut microbiota on uremic solute accumulation by a CE-TOFMS-based metabolomics approach, *Kidney International*. 2017, *92*, 634-645.
45. Sasidharan, K., Soga, T., Tomita, M., Murray, D. B., A Yeast metabolite extraction protocol optimised for time-series analyses, *PLoS One*. 2012, *7*, e44283.
46. Tanaka, Y., Higashi, T., Rakwal, R., Wakida, S. i., Iwahashi, H., Development of a capillary electrophoresis-mass spectrometry method using polymer capillaries for metabolomic analysis of yeast, *Electrophoresis*. 2008, *29*, 2016-2023.
47. Liu, S., Wang, L., Hu, C., Huang, X., Liu, H., Xuan, Q., Lin, X., Peng, X., Lu, X., Chang, M., Xu, G., Plasma metabolomics profiling of maintenance hemodialysis based on capillary electrophoresis - time of flight mass spectrometry, *Scientific Reports*. 2017, *7*, 8150.
48. Das, G., Patra, J. K., Lee, S.-Y., Kim, C., Park, J. G., Baek, K.-H., Analysis of metabolomic profile of fermented Orostachys japonicus A. Berger by capillary electrophoresis time of flight mass spectrometry, *PLoS One*. 2017, *12*, e0181280.
49. Bennuru, S., Lustigman, S., Abraham, D., Nutman, T. B., Metabolite profiling of infection-associated metabolic markers of onchocerciasis, *Molecular and Biochemical Parasitology*. 2017, *215*, 58-69.
50. Soga, T., Ishikawa, T., Igarashi, S., Sugawara, K., Kakazu, Y., Tomita, M., Analysis of nucleotides by pressure-assisted capillary electrophoresis-mass spectrometry using silanol mask technique, *Journal of Chromatography A*. 2007, *1159*, 125-133.
51. Stutz, H., Advances in the analysis of proteins and peptides by capillary electrophoresis with matrix-assisted laser desorption/ionization and electrospray-mass spectrometry detection, *Electrophoresis*. 2005, *26*, 1254-1290.
52. Shamsi Shahab, A., Miller Blair, E., Capillary electrophoresis-mass spectrometry: Recent advances to the analysis of small achiral and chiral solutes, *Electrophoresis*. 2004, *25*, 3927-3961.
53. Benavente, F., Sanz-Nebot, V., Barbosa, J., van der Heijden, R., van der Greef, J., Hankemeier, T., CE-ESI-MS of biological anions in plastic capillaries at high pH, *Electrophoresis*. 2007, *28*, 944-949.
54. Liu, J. X., Aerts, J. T., Rubakhin, S. S., Zhang, X. X., Sweedler, J. V., Analysis of endogenous nucleotides by single cell capillary electrophoresis-mass spectrometry, *The Analyst*. 2014, *139*, 5835-5842.
55. 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.
56. Dalmau, N., Bedia, C., Tauler, R., Validation of the regions of interest multivariate curve resolution (roimcr) procedure for untargeted lc-ms lipidomic analysis, *Analytica Chimica Acta*. 2018, *1025*, 80-91.
57. Parastar, H., Jalali-Heravi, M., Sereshti, H., Mani-Varnosfaderani, A., Chromatographic fingerprint analysis of secondary metabolites in citrus fruits peels using gas chromatography-mass spectrometry combined with advanced chemometric methods, *Journal of Chromatography A*. 2012, *1251*, 176-187.

58. Siano, G. G., Pérez, I. S., García, M. D. G., Galera, M. M., Goicoechea, H. C., Multivariate curve resolution modeling of liquid chromatography-mass spectrometry data in a comparative study of the different endogenous metabolites behavior in two tomato cultivars treated with carbofuran pesticide, *Talanta*. 2011, *85*, 264-275.
59. Szymańska, E., Markuszewski Michał, J., Vander Heyden, Y., Kaliszan, R., Efficient recovery of electrophoretic profiles of nucleoside metabolites from urine samples by multivariate curve resolution, *Electrophoresis*. 2009, *30*, 3573-3581.
60. Saurina, J., in: Hanrahan, G., Gomez, F. A. (Eds.), *Chemometric Methods in Capillary Electrophoresis*, John Wiley & Sons, New Jersey 2009, pp. 199-226.
61. 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.
62. Farres, M., Pina, B., Tauler, R., Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS, *Metabolomics*. 2015, *11*, 210-224.
63. Navarro-Reig, M., Jaumot, J., Garcia-Reiriz, A., Tauler, R., Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, *Analytical and Bioanalytical Chemistry*. 2015, *407*, 8835-8847.
64. Motegi, H., Tsuboi, Y., Saga, A., Kagami, T., Inoue, M., Toki, H., Minowa, O., Noda, T., Kikuchi, J., Identification of reliable components in multivariate curve resolution-alternating least squares (MCR-ALS): a Data-Driven Approach across Metabolic Processes, *Scientific Reports*. 2015, *5*, 15710.
65. Jaumot, J., Tauler, R., Potential use of multivariate curve resolution for the analysis of mass spectrometry images, *The Analyst*. 2015, *140*, 837-846.
66. Piqueras, S., Krafft, C., Beleites, C., Egodage, K., von Eggeling, F., Guntinas-Lichius, O., Popp, J., Tauler, R., de Juan, A., Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues, *Analytica Chimica Acta*. 2015, *881*, 24-36.
67. Scurr, D. J., Hook, A. L., Burley, J. A., Williams, P. M., Anderson, D. G., Langer, R. C., Davies, M. C., Alexander, M. R., Strategies for MCR image analysis of large hyperspectral data-sets, *Surface and Interface Analysis:SIA*. 2013, *45*, 466-470.
68. 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.
69. Farres, M., Pina, B., Tauler, R., LC-MS based metabolomics and chemometrics study of the toxic effects of copper on Saccharomyces cerevisiae, *Metallomics*. 2016, *8*, 790-798.
70. 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.
71. Navarro-Reig, M., Jaumot, J., Baglai, A., Vivo-Truyols, G., Schoenmakers, P. J., Tauler, R., Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution, *Analytical Chemistry*. 2017, *89*, 7675-7683.
72. 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.

CAPÍTOL 4.

Estudi dels efectes de compostos disruptors endocrins en embrions de peix zebra

4.1. INTRODUCCIÓ

En els darrers anys ha crescut la preocupació per l'impacte de l'alliberament constant al medi ambient de compostos sintètics procedents de l'activitat industrial. En aquest context, aquest capítol es basa en l'aplicació de metodologies metabolòmiques i transcriptòmiques no dirigides per tal d'elucidar potencials biomarcadors relacionats amb l'exposició d'embrions de peix zebra (*Danio rerio*) a compostos disruptors endocrins (*endocrine disruptors compounds*, EDCs). A més, de les metodologies analítiques i de tractament de dades presentades en el capítol anterior, en aquest capítol, s'han investigat els efectes de diferents EDCs àmpliament trobats en el medi ambient sobre les rutes metabòliques dels embrions de peix zebra, i així poder avaluar també les possibles conseqüències sobre la població humana [1]. En particular, els EDCs que s'han considerat en aquesta Tesi són el bisfenol A (BPA), el sulfonat de perfluorooctà (PFOS) i el tributilestany (TBT). Tot i que es coneix que els EDCs són capaços de modular i interrompre el sistema endocrí dels éssers vius i ocasionar canvis en la seva reproducció i desenvolupament [2, 3], encara no estan prou clars quins són els seus efectes sobre les rutes metabòliques [4].

En aquesta Tesi, s'ha utilitzat una estratègia metabolòmica no dirigida per a obtenir una visió global dels canvis en els perfils metabòlics dels embrions de peix zebra mitjançant el mètode HILIC-MS presentat en el capítol anterior. D'aquesta manera i aprofitant els avantatges del mètode quimiomètric MCR-ALS, s'han comparat els efectes dels diferents disruptors endocrins en embrions de peix zebra. El mecanisme d'acció dels EDCs sobre el metabolisme dels organismes és molt complex i requereix d'anàlisis òmiques complementàries que permetin extreure una explicació biològica global. Per aquest motiu, en la segona part d'aquest capítol (article V) s'ha emprat també la tècnica de seqüenciació de nova generació RNA-Seq que ha permès l'anàlisi no dirigida del transcriptoma dels embrions exposats a BPA. A partir de la combinació dels canvis en el transcriptoma i el metaboloma, s'ha investigat el potencial de la integració conceptual dels resultats obtinguts en els dos nivells òmics (fusió de nivell alt) [5] dels gens alterats a nivell de RNA i dels metabòlits afectats, per així poder obtenir una millor caracterització global de la disrupció causada pel BPA.

4.2. PUBLICACIONS

- **Article científic IV.** Assessment of endocrine disruptors effects on zebrafish (*Danio rerio*) embryos by untargeted LC-HRMS metabolomic analysis. **E. Ortiz-Villanueva**, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler. *Science of the Total Environment* 635 (2018) 156-166.

En aquest article s'avaluen els efectes individuals de tres EDCs diferents (BPA, PFOS i TBT) sobre el metaboloma d'embrions de peix zebra exposats des de les 48 a les 120 hores després de la seva fertilització (hpf). Els canvis en els perfils metabòlics dels embrions de peix zebra s'han estudiat mitjançant un procediment analític no dirigit HILIC-MS. L'anàlisi funcional dels biomarcadors metabòlics detectats va suggerir una resposta similar dels embrions al tractament amb els tres EDCs. També, es van detectar algunes rutes metabòliques específiques que posen de manifest modes d'acció diferenciadors dels contaminants estudiats.

- **Article científic V.** Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach. **E. Ortiz-Villanueva**, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler. *Environmental Pollution* 231 (2017) 22-36.

En aquest article s'estudia en més detall els efectes de l'exposició d'embrions de peix zebra a diferents dosis de BPA durant les primeres 120 hpf. Amb aquest objectiu, es va emprar una anàlisi metabolòmica no dirigida HILIC-MS en combinació amb una anàlisi transcriptòmica basada en la seqüenciació RNA-Seq. La integració dels efectes observats en els dos nivells òmics va permetre obtenir una millor caracterització dels biomarcadors moleculars rellevants i de les rutes metabòliques subjacents afectades per la toxicitat del BPA. A més, va permetre realitzar una comprensió més completa de la disrupció del BPA a quan es realitza l'anàlisi d'un únic nivell òmic.

4.2.1. Article científic IV.

Assessment of endocrine disruptors effects on zebrafish (*Danio rerio*) embryos by untargeted LC-HRMS metabolomic analysis.

E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler.

Science of the Total Environment 635 (2018) 156-166.

Science of the Total Environment 635 (2018) 156–166

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Assessment of endocrine disruptors effects on zebrafish (*Danio rerio*) embryos by untargeted LC-HRMS metabolomic analysis

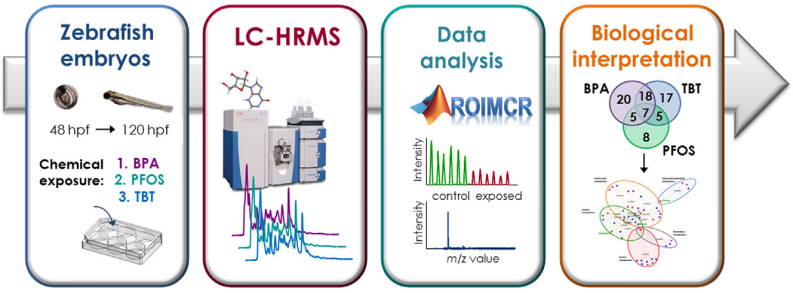
Elena Ortiz-Villanueva, Joaquim Jaumot, Rubén Martínez, Laia Navarro-Martín, Benjamin Piña, Romà Tauler *

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

HIGHLIGHTS

- Metabolic disruption in zebrafish embryos by environmentally relevant EDCs.
- BPA, PFOS and TBT effects were revealed by LC-HRMS metabolomics.
- Chemometric analysis allowed the assessment of EDCs effects.
- EDC treatments showed a considerable overlap of the most altered metabolic pathways.
- Specific metabolic disruption of these EDCs was also identified.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 February 2018
Received in revised form 29 March 2018
Accepted 30 March 2018
Available online 13 April 2018

Editor: D. Barcelo

Keywords:
Endocrine disrupting chemicals
Bisphenol A
Perfluorooctane sulfonate
Tributyltin
Untargeted metabolomics
Zebrafish embryos

ABSTRACT

Bisphenol A (BPA), perfluorooctane sulfonate (PFOS), and tributyltin (TBT) are emerging endocrine disruptors (EDCs) with still poorly defined mechanisms of toxicity and metabolic effects in aquatic organisms. We used an untargeted liquid chromatography-high resolution mass spectrometry (LC-HRMS) metabolomic approach to study the effects of sub-lethal doses of these three EDCs on the metabolic profiles of zebrafish embryos exposed from 48 to 120 hpf (hours post fertilization). Advanced chemometric data analysis methods were used to reveal effects on the subjacent regulatory pathways. EDC treatments induced changes in concentrations of about 50 metabolites for TBT and BPA, and of 25 metabolites for PFOS. The analysis of the corresponding metabolic changes suggested the presence of similar underlying zebrafish responses to BPA, TBT and PFOS affecting the metabolism of glycerophospholipids, amino acids, purines and 2-oxocarboxylic acids. We related the changes in glycerophospholipid metabolism to alterations in absorption of the yolk sack, the main source of nutrients (including lipids) for the developing embryo, linking the molecular markers with adverse phenotypic effects. We propose a general mode of action for all three chemical compounds, probably related to their already described interaction with the PPAR/RXR complex, combined with specific effects on different signaling pathways resulting in particular alterations in the zebrafish embryos metabolism.

© 2018 Elsevier B.V. All rights reserved.

Abbreviations: AIF, All-ion fragmentation; BPA, Bisphenol A; CE-MS, Capillary electrophoresis-mass spectrometry; EDC, Endocrine disrupting chemical; GC-MS, Gas chromatography-mass spectrometry; HCA, Hierarchical cluster analysis; FTMS, Hybrid Fourier transform mass spectrometry; IS, Internal standard; LC-MS, Liquid chromatography-mass spectrometry; LC-HRMS, Liquid chromatography-high resolution mass spectrometry; LOEC, Lowest observed effect concentration value; MCR-ALS, Multivariate curve resolution by alternating least squares; PFOS, Perfluorooctane sulfonate; Q-TOF, Quadrupole time-of-flight; rMANOVA, Regularized multivariate analysis of variance; ROI, regions of interest; TBT, Tributyltin; WWTP, Wastewater treatment plant; YSA, Yolk sack area.

* Corresponding author.
E-mail address: roma.tauler@idaea.csic.es. (R. Tauler).

https://doi.org/10.1016/j.scitotenv.2018.03.369
0048-9697/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

The concern about endocrine disrupting chemicals (EDCs) effects on human and wildlife health is currently increasing (Diamanti-Kandarakis et al., 2009). EDCs are exogenous compounds that initiate abnormal endocrine system processes by activating or inactivating endocrine target receptors or disturbing hormonal balances and metabolism, eventually disrupting homeostatic mechanisms, and finally affecting reproduction and development (Mallozzi et al., 2017). There is a broad range of potential EDCs, including organochlorines, dioxins, organotins (*e.g.,* tributyltin, TBT), polyfluoroalkyl compounds (*e.g.,* perfluorooctane sulfonate, PFOS), brominated flame retardants, alkylphenols, bisphenols (*e.g.,* Bisphenol A, BPA) and phthalates (Casals-Casas and Desvergne, 2011; Oliveira et al., 2016), in addition to natural or synthetic hormones (*e.g.,* estradiol, estrone or ethinyl estradiol), which are being released into the environment. These compounds could reach organisms through oral ingestion, respiratory inhalation or dermal absorption, depending on their physicochemical properties (Woodruff, 2011). Aquatic organisms can uptake them directly from water by gills or skin, *via* uptake of suspended particles or through the consumption of contaminated food (van der Oost et al., 2003). EDCs are considered an important threat to the human health, wildlife and environment at environmentally relevant concentrations (Xu et al., 2013). For instance, PFOS is generally found in the environment at low concentrations. However, concentrations up to 600 $ng \cdot L^{-1}$ have been reported in the Tennessee River downstream (Huang et al., 2010; Hansen et al., 2002).

EDCs are intensely investigated in ecotoxicological and environmental fields in an attempt to discover their possible adverse effects on aquatic organisms, and also to evaluate the consequences of the environmental exposures to these pollutants for the human population. These different families of chemicals are extensively used for industrial applications (including food and clothing industry), ending up in wastewater treatment plants (WWTPs). As most of WWTPs are unable to eliminate these compounds totally, EDCs finally arrive to the environment causing diverse physiological effects in many aquatic species (Chang et al., 2016; Mihaich et al., 2009). This harmful impact on marine and freshwater ecosystems leads to different restrictions and banning legislations through the world (Martínez et al., 2017; Grün et al., 2006). Despite these limitations, remnant levels in aquatic systems could still promote some endocrine disruption effects, such as adipogenic activity upon TBT exposure (Grün et al., 2006; Inadera and Shimomura, 2005).

Owing to the extensive list of described metabolic disorders, including endocrine, reproductive, hormone balance, and immune systems and the toxicity effects caused by EDCs in wildlife (Gore et al., 2015; Giulivo et al., 2016), a more comprehensive view of their complex toxic effects using several methodologies is needed in biological and environmental research. Omic approaches play a crucial role in the understanding of the toxic effects and mechanisms of action of EDCs. Accordingly, they are becoming a key field in the assessment of the occurring processes to characterize specific metabolic disruption (Messerlian et al., 2017). Among omic sciences, the untargeted analysis of changes on the concentration of metabolites (metabolomics) has emerged as a powerful tool to study signaling pathways (Bundy et al., 2008; Baxter et al., 2007), leading to complex metabolic responses in model organisms, such as the zebrafish (*Danio rerio*) and its embryos.

Metabolites are considered as the downstream product of gene expression revealing relative changes at transcriptomic and proteomic levels (Urbanczyk-Wochniak et al., 2003). Therefore, metabolomics reflects the underlying biochemical activity and it gives insight into the understanding of the physiology and molecular phenotype of the investigated organisms (Viant, 2007). However, metabolomic research is a complex field due to the broad range of low molecular weight compounds (metabolites, mass range of 50–1500 dalton (Da)) with large structural diversity due to the different involved cell biological processes. Hence, development and application of different high-throughput analytical platforms are essential for metabolomic studies. Among these analytical platforms, hyphenated mass spectrometry-based techniques present advantageous properties for metabolomics, since they provide higher sensitivity and selectivity than other used platforms. Liquid chromatography-mass spectrometry (LC-MS), gas chromatography–mass spectrometry (GC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) offer easy and reliable metabolic profiling of biological systems (Zhang et al., 2012). Moreover, recent advances in MS-based technologies have allowed covering a larger variety of chemical compounds, which is especially useful in untargeted omics. In fact, in untargeted metabolomics, high-resolution mass spectrometry techniques such as, time-of-flight (TOF) (Wilson et al., 2005), quadrupole time-of-flight (Q-TOF) (Weaver et al., 2007) or hybrid Fourier transform mass spectrometry (FTMS) have gained importance in relation to the conventional low-resolution mass spectrometry platforms. For instance, Orbitrap instruments provide the very advantageous all-ion fragmentation (AIF) option for metabolite identification without the need of selecting the precursor ion and of reanalyzing samples (Geiger et al., 2010).

However, data analysis in untargeted omic studies is still a challenging issue. Untargeted metabolomics generates massive amounts of full scan MS data, requiring the use of advanced data analysis methods. Thereby, several data analysis tools have emerged to handle these large datasets. The "regions of interest" (ROI) type of approaches have been proposed with the goal of reducing the size of metabolomics datasets without any loss of mass accuracy (Tautenhahn et al., 2008). In this way, the ROI compression pretreatment in combination with the multivariate curve resolution alternating least squares (MCR-ALS) procedure (in the so-called ROIMCR procedure) has demonstrated to be a powerful strategy to get very complete metabolic profiling of the investigated systems (Gorrochategui et al., 2016; Gorrochategui et al., 2015a).

The main aim of this work is to assess metabolomic responses of zebrafish (*D. rerio*) embryos exposed to sub-lethal doses of EDCs. Untargeted LC-HRMS ROIMCR metabolomic approach allows identifying potential biomarkers related to toxicity mechanisms of investigated pollutants in these aquatic organisms.

2. Materials and methods

2.1. Chemicals and reagents

All chemicals used in the preparation of buffers and solutions were analytical reagent grade. Acetic acid (glacial), methanol (HPLC grade) and acetonitrile (HPLC and MS grade) were purchased from Merck (Darmstadt, Germany). Chloroform was supplied by Carlo Erba (Peypin, France). Ammonium acetate (MS grade), dimethyl sulfoxide (DMSO), water (HPLC and MS grade), calcium sulfate dihydrate ($CaSO_4 \cdot 2H_2O$), bisphenol A (BPA), perfluorooctane sulfonate (PFOS), tributyltin (TBT) and methionine sulfone and piperazine-1,4-bis(2-ethanesulfonic acid) (PIPES), used as the internal standards (IS), were provided by Sigma-Aldrich (St. Louis, MO, USA).

2.2. Animal maintenance and rearing conditions

Adult zebrafish were maintained under standard conditions in fish water, composed of reverse-osmosis purified water containing 90 $\mu g \cdot mL^{-1}$ of Instant Ocean (Aquarium Systems, Sarrebourg, France) and 0.58 mM $CaSO_4 \cdot 2H_2O$ at a temperature of 28 (±1°C). Fish were fed twice a day with dry flakes (TetraMin, Tetra, Germany). Embryos from wild-type zebrafish were obtained by natural mating placing six males and three females on 4-L breeding tanks with a mesh bottom. At 2 hpf (hours post fertilization), eggs were collected and rinsed. Fertilized viable eggs were then randomly distributed in 6-well multiplates (10 embryos/well). Embryos were raised at 28.5 °C with a 12 Light:12 Dark photoperiods in fish water (3 mL/well). All experiments were

carried out in accordance with the institutional guidelines under a license from the local government (DAMM 7669, 7964), and were approved by the Institutional Animal Care and Use Committees at the Research and Development Centre of the Spanish Research Council, CID-CSIC.

2.3. Toxicological and morphological analyses

2.3.1. Exposure protocols

BPA, PFOS and TBT stock solutions were prepared in DMSO on the day of the experiment. Experimental solutions with the same final concentration of DMSO (0.2%) were obtained by dissolving the stocks with fish water. Exposure concentrations of BPA, PFOS and TBT were chosen by a preliminary range-finding test (see Section 2.3.2). Embryos were kept in clean fish water during the first 48 hpf, to avoid early embryonic processes, and then exposed to the different chosen concentrations until 120 hpf (see Table 1).

Water solutions were prepared and changed every day to assure continuous exposure to the contaminants until embryo collection. Chemical analysis of media water solutions was not performed as these chemicals have been shown to be stable for up to 48 h in water solutions (Jordão et al., 2016).

2.3.2. LOEC determination

Lowest observed effect concentration values (LOECs) were chosen based in the effects of each EDC in larval morphology using a preliminary range-finding test (120 hpf) including a wide range of concentrations for each compound (see Supplementary Table S1). Five subsets of zebrafish embryos were exposed to fish water (1.125 $g \cdot L^{-1}$ Instant Ocean® + 250 $\mu g \cdot L^{-1}$ $CaSO_4$) containing the individual chemical standards (using DMSO as a vehicle) or only DMSO (0.2%). Anatomical development of embryos was followed daily during the exposure as described by Kimmel et al. (1995) under a stereomicroscope Nikon SMZ1500 equipped with a Nikon digital sight DS-Ri1 digital camera. Effects in mortality (24, 48, 72, 96 and 120 hpf) and hatching (72, 96 and 120 hpf) during the exposure protocol was negligible and sub-lethal endpoints such as coagulated embryos, lack of somite formation, non-detachment of the tail or lack of heartbeat were not observed. Only a decrease in the swim bladder inflation was observed in the highest concentration (LOEC) of BPA and TBT. In the case of BPA highest concentration, exposed embryos showed a slightly darker pigmentation than controls. Survival, hatching, swim bladder inflation and tail malformations (lateral and dorsal deformities) at 120 hpf were considered for LOEC value estimation. The lowest concentration showing a statistically significant effect in at least one of these four parameters (Fisher Exact Probability Test; $p < 0.05$) was taken as LOEC; the calculated values were 17.5 μM for BPA, 2.0 μM for PFOS, and 0.1 μM for TBT (see Supplementary Table S1). These concentrations were used as maximal exposure values for metabolomic studies, to avoid interferences in the molecular analyses from alterations in development or in larvae viability.

2.3.3. Determination of yolk sack area (YSA)

Zebrafish embryos were exposed to a large range of concentrations (up to lethality) in groups of 50 individuals per condition (chosen concentrations for each compound in Fig. 6). Exposed embryos were fixed in 4% paraformaldehyde (PFA) overnight at 4 °C, washed several times with PBS and gradually transferred to glycerol 90% for conservation (Raldua et al., 2008). Fixed specimens were recorded using a stereomicroscope Nikon SMZ1500 equipped with a Nikon digital Sight DS-Ri1 camera, and remains of yolk sack were measured using the free graphical analysis software ImageJ (National Institutes of Health, Bethesda, MD, USA). Results were expressed as mm^2 of yolk sack area, or YSA.

Table 1
BPA, PFOS and TBT water concentrations used in this metabolomic study.

	Control		LOEC/10		LOEC
BPA	0 μM	0.44 μM	1.75 μM	4.4 μM	17.5 μM
PFOS	0 μM	0.06 μM	0.2 μM	0.6 μM	2.0 μM
TBT	0 μM	0.003 μM	0.01 μM	0.03μM	0.10 μM

2.4. Metabolomic analyses

2.4.1. Exposure protocol

Zebrafish embryos were exposed in pools of 20 individuals as described to either vehicle (0.2% DMSO) or to a gradient of dilutions with LOECs as highest concentrations (Table 1). Five biological replicates per treatment were used for LC-HRMS analyses, totaling 25 samples per chemical compound (five replicates per condition, control (circles), high (stars), medium-high (triangles), medium-low (squares) and low concentrations (diamonds)). Exposed embryos were collected and washed twice with 0.5 mL HPLC grade water, snap-frozen in dry ice, and stored at −80 °C until further analysis.

2.4.2. Metabolite extraction

For metabolite extraction, individual pools of zebrafish embryos were thawed in a water bath at room temperature. Embryo metabolites were first extracted with 900 μL of methanol containing methionine sulfone (surrogate) at a concentration of 5 $\mu g \cdot mL^{-1}$. After vortexing 15 s, the mixture was sonicated for 15 min and vortexed again 15 s. Samples were then centrifuged at 23,500*g* for 10 min at 4 °C to isolate the supernatant, and 500 μL of water and 300 μL of chloroform were added. Samples were vortexed 15 s, placed on ice for 10 min and centrifuged again at 23,500*g* for 10 min at 4 °C. Finally, the aqueous fractions were evaporated to dryness under nitrogen gas and reconstituted with 100 μL of 1:1 v/v acetonitrile:water containing PIPES (IS) at a concentration of 5 $\mu g \cdot mL^{-1}$. Prior to injection, zebrafish embryo extracts were filtrated through a 0.22 μm filters (Ultrafree®-MC, Millipore Bedford, MA, US) at 11,000*g* for 4 min and stored at −80 °C until the analysis. All centrifugations were performed at 4 °C in a Serie Digicen 21 centrifuge (Ortoalresa, Madrid, Spain).

For each internal standard, an aqueous solution (1000 $\mu g \cdot mL^{-1}$) was prepared and stored in the freezer at −20 °C until its use. Working standard solutions were obtained by diluting the stock solutions with water. Diluted standard solutions were used to spike embryo extract samples. Quality control (QC) samples were generated by pooling 10 μL of all the studied samples (extracts).

2.5. LC-HRMS analysis

Chromatographic separations were carried on an Accela UHPLC system (Thermo Scientific, Hemel Hempstead, UK) using a hydrophilic interaction liquid chromatography (HILIC) column (TSK Gel Amide-80 column: 250 mm length, 2.1 mm inner diameter and 5 μm particle size) from Tosoh Bioscience (Tokyo, Japan) at room temperature.

Elution gradient was performed using solvent **A** (acetonitrile) and solvent **B** (5 mM of ammonium acetate adjusted to pH 5.5 with acetic acid) as follows: 0–8 min, linear gradient from 25 to 30% **B**; 8–10 min, from 30 to 60% **B**; 10–12 min, 60% **B**; 12–14 min, back from 60% to 25% **B**; and from 14 to 20 min, 25% **B**. The mobile phase flow rate was 0.15 $mL \cdot min^{-1}$ and the injection volume was 5 μL. Sample injection was performed with an autosampler at 10 °C. Solvents were degassed for 15 min by sonication before use.

An Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Hemel Hempstead, UK) equipped with an HTC PAL autosampler and a Surveyor MS Plus pump was used. The ionization source employed was a heated electrospray (HESI) source operated in positive and negative mode separately alternating MS scans of the precursor ions and all ion fragmentation (AIF) scans where metabolites were fragmented in

HCD collision cell. Mass spectra were acquired in profile mode at a resolution of 50,000 FWHM (full width half maximum) at *m/z* 200. Working parameters were as follows: electrospray voltage, 3.0 kV; sheath gas flow rate, 45 arbitrary units (a.u.); auxiliary gas flow rate, 10 a.u.; heated capillary temperature, 300 °C; automatic gain control (AGC), $1 \cdot 10^6$; and maximum injection time was set at 250 ms with two microscans/scan. Full scan mass range was from *m/z* 80 to 1000. AIF was also performed with normalized collision energy (NCE) of 25 eV.

2.6. Data analysis

LC-HRMS data were analyzed by a combination of multivariate analysis tools with the goal of evaluating the most significant metabolic changes produced by the studied EDCs (BPA, PFOS and TBT).

2.6.1. Data conversion, compression and data matrices arrangements

Each LC-HRMS chromatographic run was converted to a NetCDF data file through Thermo Xcalibur 2.1 software (Thermo Scientific, San Jose, CA) and imported into MATLAB R2016a environment (The Mathworks Inc. Natick, MA, USA) using the corresponding Bioinformatic Toolbox functions. Due to the vast size of the full scan MS files in profile mode and to the computer storage limitations, input full resolution data were processed using the "regions of interest" procedure and in-house written routines (MSROI approach) (Gorrochategui et al., 2016; Ortiz-Villanueva et al., 2017a). ROI data matrices ($\mathbf{D_k}$, k = 1, …,K), are of much reduced size compared to raw original mass spectral data without any loss of mass spectral accuracy (see Supplementary Material, Regions of interest (MSROI) approach section). After this compression, ROI reduced data matrices were then arranged in six different column-wise augmented data matrices ($\mathbf{D_{augBPA}}$, $\mathbf{D_{augPFOS}}$ and $\mathbf{D_{augTBT}}$, one for each EDC in both ESI modes), containing the information about 25 zebrafish embryo samples (5 replicates per each chemical dose). Column-wise augmented data matrices ($\mathbf{D_{aug}}$) covered a particular *m/z* range and were built up by arranging the individual ROI data matrices ($\mathbf{D_k}$) corresponding to all embryo samples, one on top of each other. One example of the augmented data matrix arrangement built with the information of all embryo samples in the analysis of the effects of PFOS for negative ESI mode is depicted in Fig. S1.

2.6.2. ROIMCR resolution procedure, metabolite detection and statistical assessment

Once augmented data matrices were ROI compressed, $\mathbf{D_{augBPA}}$, $\mathbf{D_{augPFOS}}$ and $\mathbf{D_{augTBT}}$ matrices were independently analyzed by the adaptation of the MCR-ALS method (Jaumot et al., 2005; Jaumot et al., 2015), which is called ROIMCR procedure (Ortiz-Villanueva et al., 2017a). Application of ROIMCR allowed the resolution of the elution profiles of the resolved components (possible zebrafish metabolites) as well as their corresponding mass spectra profiles (Ortiz-Villanueva et al., 2017a; Farres et al., 2016). Peak areas of the resolved elution profiles for each of the ROIMCR components in the analyzed samples (control and exposed samples for each EDC) were obtained. More details about MCR-ALS method are described in the Supplementary Material. After applying MCR-ALS, resolved component peak areas were normalized taking into account the mean of the internal standards peak areas to correct the instrumental intensity drifts among injections. For a detailed examination of concentration changes, normalized peak areas of resolved components in control and exposed samples were individually evaluated applying two complementary methods. First, regularized multivariate analysis of variance (rMANOVA) was used to analyze ROIMCR resolved component peak areas for each EDC treatment. This approach determines whether the groups differ statistically considering exposure level as a factor. rMANOVA method combines the advantages of ASCA and MANOVA models (Engel et al., 2015) allowing the estimation of the statistical significance of the considered factor (chemical exposure) by using a permutation test (here, number of permutations was set to 10,000). More details about the rMANOVA method can be found in the work from Engel (Engel et al., 2015). rMANOVA also showed the most important components (metabolites) whose peak areas (relative concentrations) changed and differentiated among groups of samples according to the studied factor level. These metabolites were considered for further analyses as specific biomarkers for each particular EDC treatment. In addition, one-way univariate ANOVA with Benjamini-Hochberg multiple comparison testing procedure (Benjamini and Hochberg, 1995) was performed to compare and confirm one by one those previously discovered potential biomarkers. Those with an adjusted *p*-value lower than 0.05 were considered statistically significant. Only those components showing significant concentration changes by both mentioned approaches were selected as possible metabolite biomarkers. Benjamini-Hochberg multiple comparison tests were performed using MATLAB Statistics and Machine Learning Toolbox™. Changes in component peak areas among different groups of samples were also examined by hierarchical cluster analysis (HCA) (Schonlau, 2004) combined with heatmap display of autoscaled peak areas in order to show clusters of metabolites with similar behavior after EDCs exposure.

2.6.3. Metabolite identification

High-resolution spectra profiles (four decimal digits) of recovered MCR-ALS components whose peak areas changed significantly (see above) allowed performing initial metabolite identification. This estimation was performed based on the comparison of the accurate molecular mass measured by LC-HRMS with the corresponding values found in online database resources, such as MassBank (Horai et al., 2010), METLIN Metabolite Database (Smith et al., 2005), Human Metabolome Database (HMDB) (Wishart et al., 2007) and MetaCyc database (Caspi et al., 2014). The relative mass error was required to be lower than 5 ppm in agreement to the Directive 2002/657/CE for mass spectrometric detection performance criteria and requirements. Since HESI is a soft ionization source, protonated and deprotonated molecules were chosen as the most frequent option.

Besides the hypothesized *m/z* values identities, a metabolite confirmation by comparison of the experimental AIF mass spectrum with the corresponding theoretical MS/MS spectrum of the suspected metabolites from the library of National Institute of Standards and Technology (NIST) (www.nist.gov/srd/nist1a.html) and MassBank databases (Horai et al., 2010) was performed. The AIF scan mode of Exactive Orbitrap was applied for unequivocal metabolite identification. This strategy enabled obtaining the product ion spectra of all detected ions without the preselection of their precursor ions in the quadrupole. Therefore, a complete identification was achieved following the requirements of the previously mentioned Directive 2002/657/CE. More than four identification points were accomplished for the majority of metabolites to ensure a reliable identification. Since mass spectra were acquired with a resolution higher than 20,000 FWHM, the HRMS precursor ion and two products ions reach 2 and 2.5 identification points, respectively. The accurate mass of both the precursor and the product ions were required to have a relative mass error lower than 5 ppm.

Finally, the list of all previously identified metabolites was searched in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2012) mapping them on to the whole *D. rerio* pathway dataset to investigate the possible metabolic processes affected by every EDC used in this work.

3. Results and discussion

Due to the large size of untargeted metabolomics data files in profile mode and to the limited amounts of computer storage available, data were firstly processed and compressed by the MSROI approach. After data compression, column-wise augmented data matrices for each one EDC treatments had 26,250 rows (elution times) and approximately 600 columns (*m/z* values) for ESI+ and; 27,500 rows and approximately 500 columns for ESI−. As an example of the results of the application of

the ROIMCR procedure, Fig. 1 shows the full scan LC-HRMS data of the 25 embryo extracts from PFOS experiment (column-wise augmented data matrix $\mathbf{D}_{\mathbf{augPFOS}}$) in ESI− after ROI compression (see also Supplementary Material Fig. S1). In the case of PFOS, five groups of samples can be distinguished: controls and PFOS exposed samples at the levels of 0.06, 0.2, 0.6 and 2.0 μM. LC-HRMS chromatograms presented complex profiles with multiple-coeluted compounds at a broad range of concentrations (see the zoomed sample, elution profiles are drawn in different colors). For this reason, we proposed the use of ROIMCR procedure to facilitate the complete resolution of the elution and mass spectra profiles of all the resolved components. These profiles can provide the information about the identity and concentration changes of the metabolites from the different investigated samples.

3.1. Detection and identification of potential biomarkers

MCR-ALS was applied independently to the six ROI compressed augmented data matrices ($\mathbf{D}_{\mathbf{augBPA}}$, $\mathbf{D}_{\mathbf{augPFOS}}$ and $\mathbf{D}_{\mathbf{augTBT}}$ for the positive and negative ESI modes) with the information of the samples simultaneously analyzed. Despite the complexity of untargeted LC-HRMS data, the MCR-ALS procedure could properly resolve a large number of components from embryo extracts. The quality of MCR-ALS models was evaluated by explained variances (R^2) and lack of fit (*LOF*) values (defined in Supplementary Material MCR-ALS section). Final models showed R^2 values higher than 99.5% and *LOF* values lower than 7% considering a number of components between 86 to110 depending on the case. MCR-ALS results (R^2, *LOF* and the number of resolved components) for each augmented data matrix ($\mathbf{D}_{\mathbf{augBPA}}$, $\mathbf{D}_{\mathbf{augPFOS}}$ and $\mathbf{D}_{\mathbf{augTBT}}$ for positive and negative ESI mode) are given in Supplementary Material Table S2.

An example of the resolution of two components by the ROIMCR procedure is shown in Supplementary Material Fig. S2. Fig. S2A depicts the elution profiles of the two components resolved by MCR-ALS. As it can be seen, the peak areas of these two components changed compared to control samples upon PFOS exposure. For instance, purple elution profile was up-regulated against PFOS exposure while blue elution profile was down-regulated. These two components could be preliminarily identified as metabolites by their corresponding mass spectra ($\mathbf{S}^{\mathbf{T}}$) (Fig. S2B) using the *m/z* value of their precursor ions. For instance, ions at *m/z* 111.0197 and *m/z* 611.1429 were tentatively identified as uracil and oxidized glutathione, respectively.

Statistical analysis of the changes in the peak areas of MCR-ALS resolved elution profiles was performed to investigate the possible metabolite biomarkers related to the different chemical exposures. rMANOVA was used to assess the multivariate (all peak changes simultaneously) statistical significance of the observed differences by every EDC treatment. Results showed that three EDCs in both ESI modes affected significantly ($p < 0.0001$) the experimental variance. Fig. 2A–F shows the rMANOVA scores plots of the first two components in the two ionization modes (ESI+ and ESI−), where PC1 separates control samples from the highest EDC treated samples and PC2 differentiates among the samples at lower treatment levels. Larger PC1 loading values could be related to components (metabolites) allowing the differentiation between control and EDCs treated samples. In addition, largest PC2 loading values could be used to distinguish among lower exposed groups. One-way ANOVA with Benjamini-Hochberg multiple comparison testing was also applied to corroborate encountered metabolite biomarkers by rMANOVA analysis. Only those variables (metabolites) showing significant differences between the mean of control samples and the mean of any of the four exposed levels (*p*-value < 0.05) were finally selected as potential metabolite biomarkers. After that, most of the relevant biomarkers were tentatively identified according to their molecular mass values retrieved from the MCR-ALS resolved pure mass spectra ($\mathbf{S}^{\mathbf{T}}$), by searching them in online databases. Finally, identities of these biomarkers were confirmed using the acquired AIF scan mode data. One example of AIF confirmation is given in Supplementary Material Fig. S3.

The total number of possible metabolite biomarkers identified was 50 for BPA, 25 for PFOS, and 47 for TBT. Metabolite identity, ion assignment, measured mass, *m/z* error (≤5 ppm), elution time, fold-changes, folding trends and the product ions used for the metabolite confirmation are given in Tables S3, S4 and S5 for BPA, PFOS and TBT, respectively. In some cases, metabolites were only tentatively identified because fragment ions could not be detected using the mass scanning range conditions used in this study or presented only one predicted product ion.

Hierarchical clustering analysis (HCA) of auto-scaled metabolite peak areas was applied to evaluate similarities among resolved

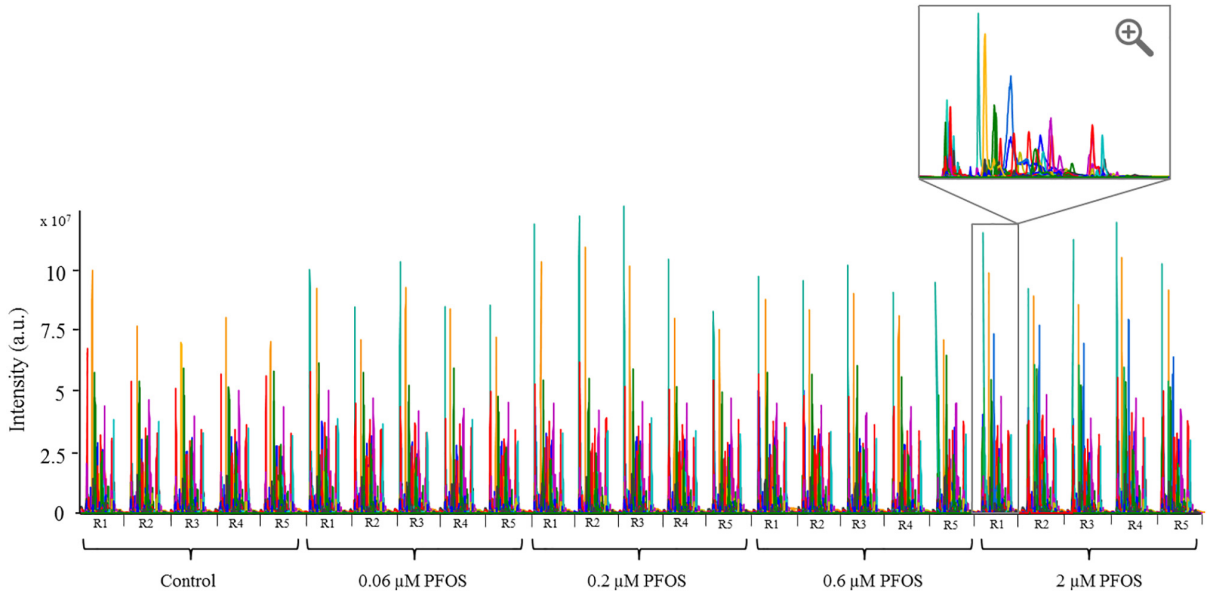


Fig. 1. LC-MS chromatograms of all the investigated samples in the PFOS experiment after ROI compression: controls and the four concentration levels of 0.06, 0.2, 0.6 and 2 μM (5 biological replicates at each condition). An example of the 2 μM PFOS sample is given in a zoomed view where elution profiles are drawn in different colors.

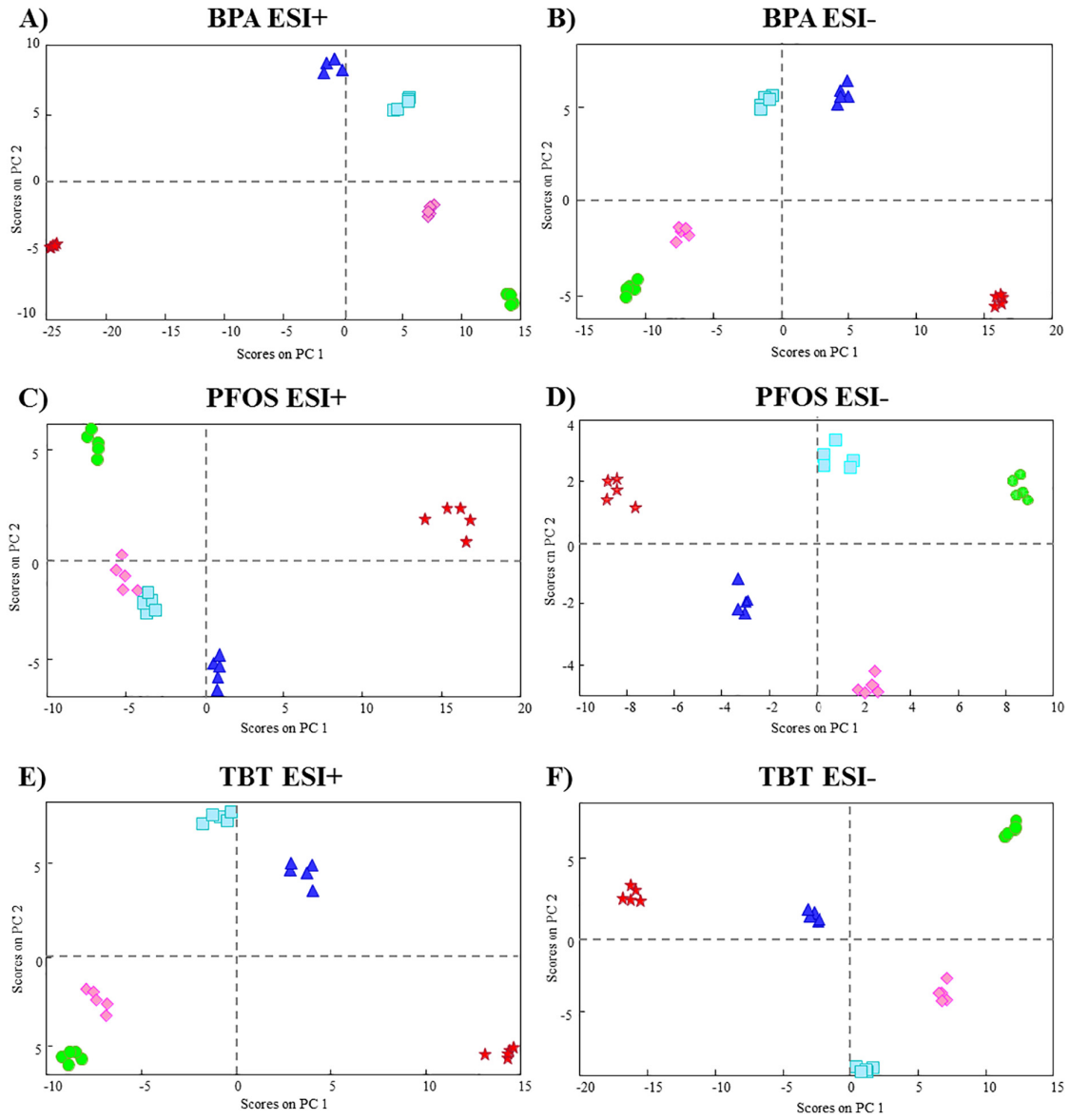


Fig. 2. rMANOVA scores plots of MCR-ALS resolved peak areas in positive and negative ESI mode for: (A–B) BPA, (C–D) PFOS and (E–F) TBT treatment. Green circles are control samples, pink diamonds are embryo samples exposed to lowest doses, cyan squares are embryo samples exposed to the second lowest doses, blue triangles are embryo samples exposed to intermediate doses and red stars are embryo samples exposed to highest doses.

metabolites after each specific EDC treatment. Heatmap representations of metabolite peak area changes related to BPA (Fig. 3), PFOS (Fig. S4A) and TBT (Fig. S4B) showed two main clusters. Clusters A included metabolites increasing their concentrations upon raising chemical exposure level. In contrast, metabolites in clusters B depleted their concentrations upon increasing chemical exposure. BPA pattern appear to be more similar to TBT than to PFOS pattern (see Figs. 3 and S4). BPA and TBT show larger effects on metabolite concentrations (peak areas) at intermediate exposures.

3.2. Functional analysis

Identified biomarkers related to BPA, PFOS and TBT exposures allowed the determination of affected metabolic pathways using the KEGG database to postulate adverse outcomes attributed to each EDC. A total of 80 different metabolites were identified as potential EDCs exposure biomarkers (50 metabolites for BPA exposure, 47 for TBT and 25 for PFOS). As it can be seen, both BPA and TBT retrieved a larger number of biomarkers, whereas PFOS appeared to detect a narrower range of adverse effects at metabolome level. Fig. 4A summarizes in a Venn diagram the relationships among identified biomarkers for each EDC, highlighting those common and uncommon (specific) biomarkers among studied EDCs.

Seven biomarkers (creatine, adenosine monophosphate, L-proline L-tyrosine, inosine 5β-cyprinolsulfate, 2-oxo-3-hydroxy-4-phosphobutanoic acid) were common for all three EDCs, suggesting a general effect of EDCs in amino acid, purine and 2-oxocarboxylic acid metabolism. The comparison of common metabolites for each possible pair combination showed that BPA and TBT presented 18 metabolites affected by both exposures. In contrast, PFOS and BPA,

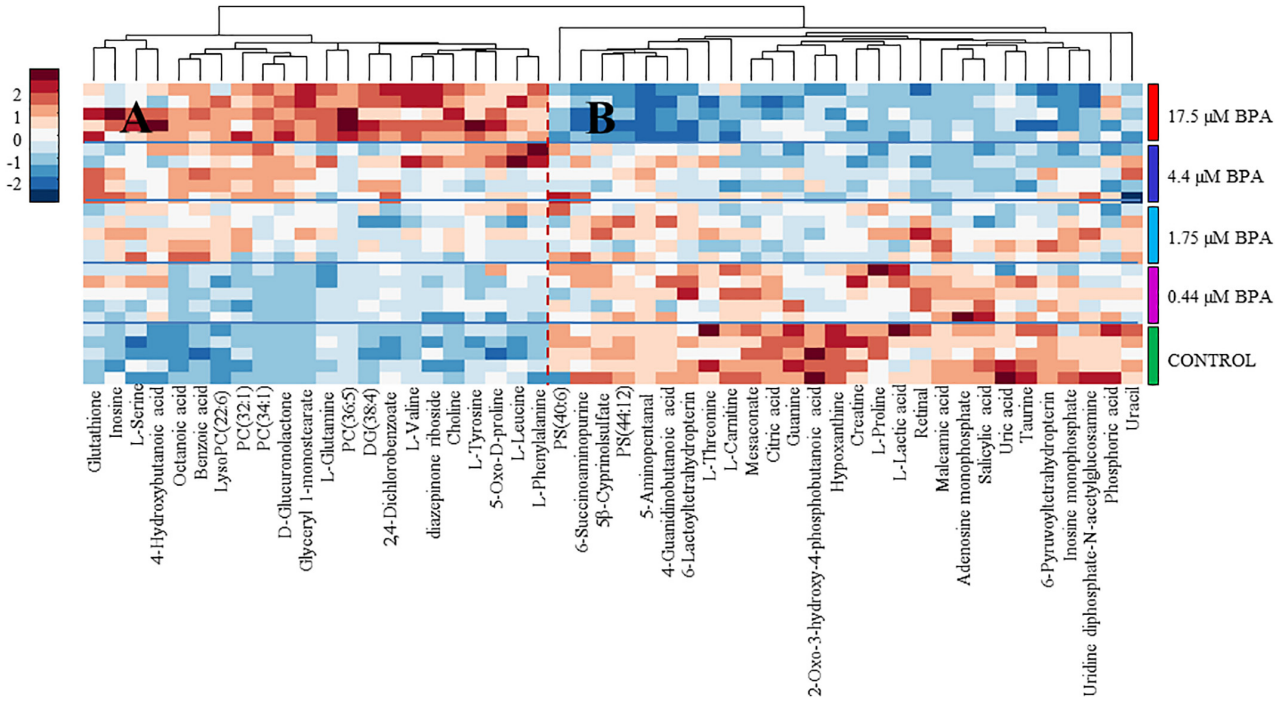


Fig. 3. Hierarchical clustering heatmap of the peak areas (autoscaled data) of the statistically significant identified biomarkers whose concentration changed between control and BPA exposed samples. Cell colors represent auto-scaled relative abundances of each metabolite, indicating metabolites up-regulated (red) and down-regulated going (blue). Color intensity codes are given in the color bar at the left side of the figures. Different clusters containing the metabolites with the same behavior are outlined.

and PFOS and TBT only had five metabolites in common for each pair. From these results, BPA and TBT exposure exhibit a more similar mode of action whereas PFOS treatment produced lower effects and did not allow establishing a definite relation to the other studied EDCs. Nevertheless, a more comprehensive differentiation among exposures was achieved by taking into account specific metabolites in each case, 20 metabolites for BPA, eight for PFOS, and 17 for TBT.

KEGG pathway analysis of common and uncommon potential metabolite biomarkers of BPA, PFOS and TBT treatments reflected different adverse effects. BPA results showed 55 possible functional modules affected; nine of them with more than four hits and 25 with more than two hits. The most significant cluster of biomarkers corresponded to the amino acid metabolism and the second most populated cluster included metabolites related to purine metabolism (adenosine monophosphate, L-glutamine, guanine, inosine monophosphate, hypoxanthine, inosine and uric acid). Furthermore, KEGG analysis exposed clusters linked to vitamin/cofactor metabolism (folate), signaling pathways (retinol) and lipid metabolism (glycerophospholipids and fatty acids). Similarly, TBT showed 51 possibly functional modules affected: six modules with more than four hits and 26 with more than two hits. The most affected metabolic pathways corresponded to amino acid metabolism as well as in the case of BPA exposure. Purine metabolism was again the second more densely populated cluster. In addition, KEGG analysis revealed other clusters possibly affected related to carbon and 2-oxocarboxylic acid metabolism, signaling pathways (taurine, butanoate) and lipid metabolism (glycerophospholipids and lipids associated to the biosynthesis of unsaturated fatty acids). Finally, PFOS analysis reflected a lower metabolic disruption than BPA and TBT, showing

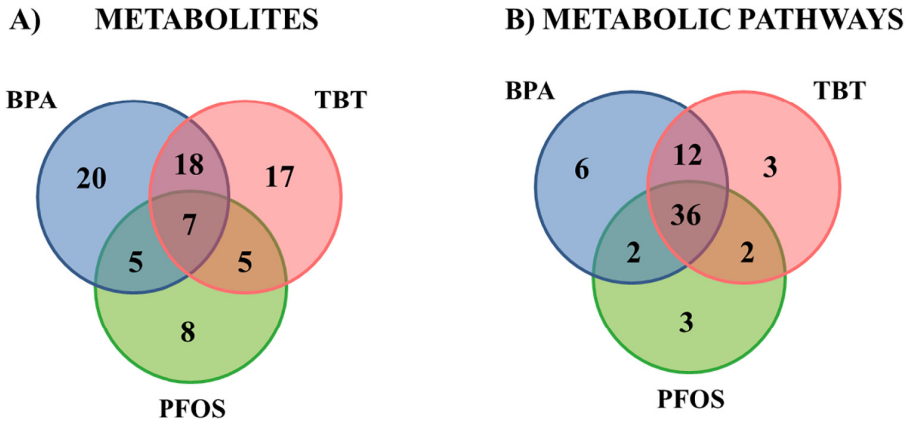


Fig. 4. Venn diagram of the identified (a) biomarkers and (b) metabolic pathways upon BPA, PFOS and TBT exposures. Number of specific and common metabolites or metabolic pathways between/among treatments are indicated.

43 identified functional modules. Within these modules, only four presented more than four hits and 16 more than two hits. The most affected pathways were those related to amino acid, purine, carbon and 2-oxocarboxylic acid metabolism, as in the case of BPA and TBT exposures. However, it is important to note that, in the PFOS case, lipid disruption was not significantly observed in contrast with BPA and TBT exposures that showed changes in the lipid metabolism which may play a crucial role in membrane structure and functions.

Comparison of the metabolic pathways affected by the three treatments showed a much higher degree of overlap between them than when considering individual metabolites (compare Fig. 4A and B). About 80% of affected pathways are common to at least two of the treatments, and about 60% of them are common to the three (Fig. 4B). These results suggest an underlying common mode of action, independently of differential effects on particular metabolites. These mutual relationships can be further explored in the diagram of Fig. 5, which show the number of affected metabolites for each treatment included in the 20 KEGG modules with at least five hits. Note that some of the metabolites are repeated, both vertically (the same metabolite could be detected in more than one treatment) and horizontally (a metabolite can be included in several KEGG modules). Color codes reflect the relative importance of a given module in the total number of metabolites affected by a given treatment. Except for modules dre00654 and dre02010, which appear underrepresented in the PFOS dataset, the relative importance of the rest of modules looks similar for the three treatments. This similarity holds not only for the number of metabolites in each module but also in the sense of the observed changes.

All but two of the metabolites (uracil and L-valine) included in the modules shown in Fig. 5 showed the same trend (up or down) in at least two out of the three treatments (Table S6). The representation of these changes over the actual metabolic pathways, the functional similarity (or equivalence) of the effects brought about by the three treatments becomes now more evident. For example, we observed the increase in the concentration of nucleosides (guanosine, inosine) from the purine metabolism pathway and a decrease in the concentrations of nucleotides (GMP, IMP, AMP) and of nitrogen bases or related metabolites (guanine, hypoxanthine and urate, Fig. S5A). While this trend was also observed individually for all three treatments, the combination of all of them allowed for a more complete understanding of global effects. A similar situation occurred when analyzing the taurine and hypotaurine pathway: the general trend of concentration increase of hypotaurine while concentration of taurine and related metabolites decrease was more clearly observed when the effects of all three types of treatments were considered jointly (Fig. S5B). Finally, changes observed in the glycerophospholipid pathway upon treatment with either BPA or TBT (but not PFOS) is interpreted as an enhancement of the production of phosphocholine lipids and a decrease in phosphoserine lipids. Again, the combination of the results from two of the treatments provided a better understanding of the observed changes (Fig. S5C).

3.3. Biological relevance of affected metabolic pathways

Owing to the observed metabolic disruption trends, it is concluded that the studied EDCs produced toxic effects and alterations in cell proliferation. These significant changes in the concentrations of amino acids confirmed the presence of oxidative stress produced by EDCs in zebrafish embryos, in accordance with previous studies (Soboń et al., 2016; Huang et al., 2017). For instance, relative concentration changes of L-proline are considered to be an essential response of organisms to oxidative stress. Also, levels of alanine, L-glutamine and 4-aminobutanoate are linked with oxidative stress (Soboń et al., 2016). A similar conclusion can be drawn from uric acid levels associated with hypoxanthine concentration changes in BPA- and TBT-exposed samples (Yoon et al., 2017). Besides the importance of L-glutamine levels suggesting oxidative stress in zebrafish embryos, its concentration changes could also be related to disturbances in purine metabolism, as it has been already reported on the fish metabolome upon BPA exposure (Yoon et al., 2017).

BPA and TBT treatments resulted in an increase of essentially all lipid families present in zebrafish embryos (phosphatidylcholines (PCs), lysophosphatidylcholines (LysoPCs), diacylglycerols (DGs)), except for phosphatidylserines (PSs), which decreased in concentration. These results are in agreement with a previous study of their effects in cells (Gorrochategui et al., 2015b), where TBT effects in lipid metabolism were large in comparison to perfluorinated compounds, such as PFOS. In the case of zebrafish embryos, lipid metabolism is strongly influenced by the absorption of the yolk sack, the main reservoir of lipids and other nutrients that are required for embryo development, and which becomes practically exhausted a 120 hpf, just before the onset of the self-feeding larval stage. Microscopic observation revealed that exposure to LOEC levels of BPA significantly increases the area of the remaining yolk sack at 120 hpf, an effect only partially observed for TBT, and completely absent in PFOS-treated embryos (Fig. 6). Therefore, these results show a correlation between the absorption of the yolk sack and the observed effects in lipid metabolism related pathways (see Table S6). It is especially remarkable the increase of the yolk sac remnants in BPA and TBT along the increase in phosphatidylcholines, the predominant lipid class in the yolk sac at 120 hpf (Fraher et al., 2016). In contrast,

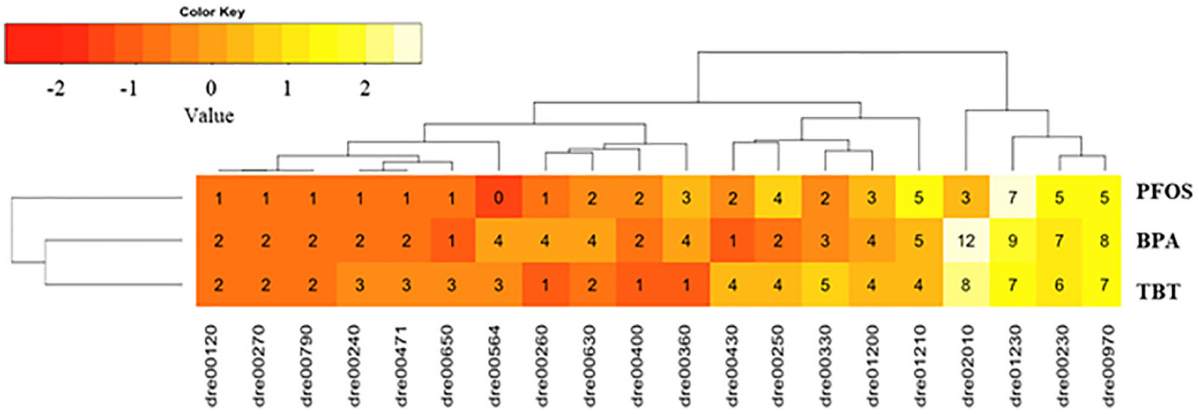


Fig. 5. Effects of the three EDC treatments on the concentrations of metabolites associated to different KEGG modules. Numbers indicated the number of metabolites affected in each module by every treatment. The same metabolite can be present in more than one pathway, and being affected by more than one treatment (see Tables S3, S4 and S5 for details). Colors represent the relative importance of metabolites associated to each pathway for each treatment; two squares of the same shade of color correspond to identical fraction of metabolites affected by the treatment. The analysis only includes metabolic modules with at least five hits (all treatments considered, including common metabolites).

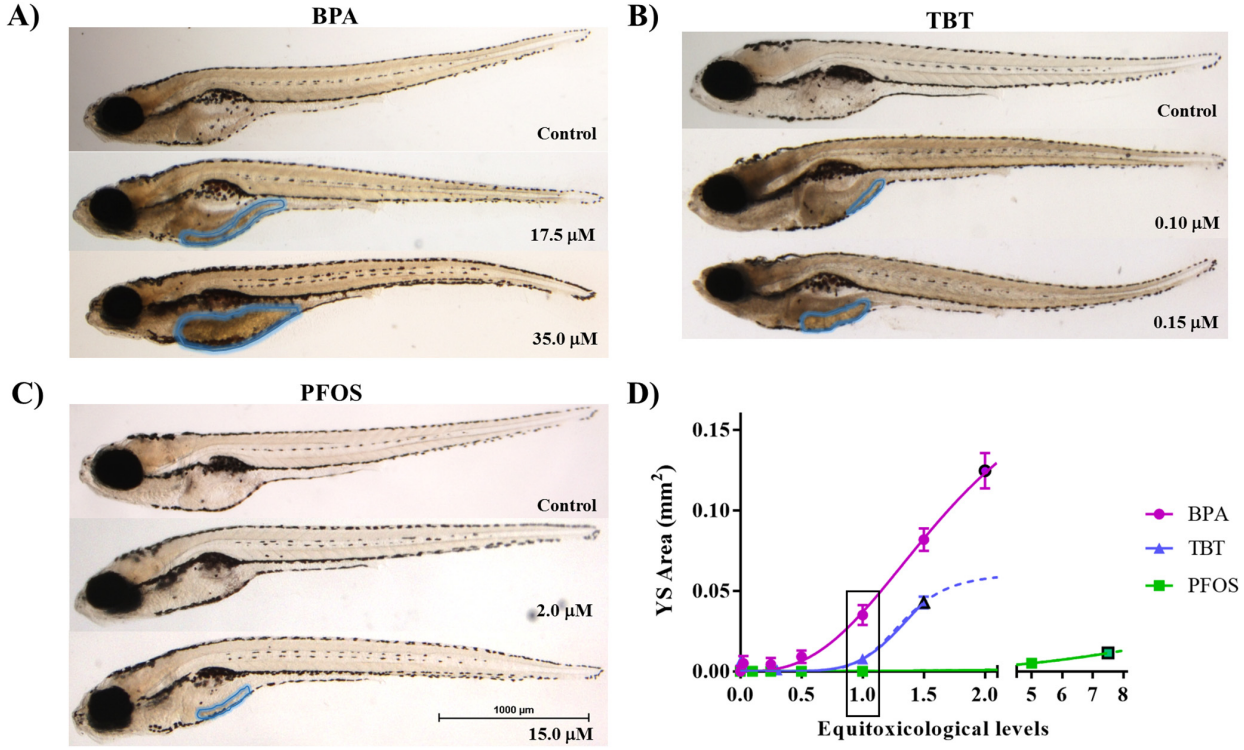


Fig. 6. Effects of the three EDC treatments on the yolk sac area of the larvae at 120 hpf. Representative larvae of the yolk sac area (YSA) means for the control, LOEC (lowest observed effect concentration) and highest concentration groups are shown for the exposure to (A) BPA, (B) TBT and (C) PFOS. YSA are blue encircled. D) YSA mean ± SEM ($n = 50$) at each condition, setting control values to zero for each experimental set. Concentrations represented referred to the LOEC values for each EDC compound (set as 1), at equitoxic levels. The highest concentrations used for BPA, TBT and PFOS in the range finding tests are marked with black-border symbols.

the observation that increasing concentrations of TBT and PFOS also result in increased yolk sack area (Fig. 6) suggests that the underlying mechanism applies to the three compounds. BPA seems to be the most efficient compound of these three regarding the yolk sac bad-absorption, at least in terms of equitoxic levels, since TBT and PFOS effects on the YSA are observed at concentrations very close already to lethality doses. Even at these doses, BPA effects on YSA were stronger than TBT or PFOS, which only induced a slight increase of yolk sack remnants (Fig. 6, black-bordered points).

TBT produced a substantial imbalance in taurine metabolism responsible for the protection of cells against membrane disintegration. These results suggest that taurine metabolism could be also altered to compensate possible cell membrane damage linked to lipid disruption. These effects are also only weakly observed in BPA exposed samples. Taurine has been described to be an essential amino acid for normal development of the nervous system (Ye et al., 2016).

Changes in metabolites related to energy metabolism were also observed upon TBT exposure. Increase in α-D-glucose levels together with a decrease in L-glutamate and alanine indicated carbohydrate metabolic disturbances (Zhou et al., 2010; Andreassen et al., 2005). In this study, down-regulation of L-glutamate and alanine suggested that glycogen-derived glucose could be necessary to balance energy needs of TBT-exposed zebrafish embryos to maintain homeostasis (Zhou et al., 2010; Zhang et al., 2016).

Probably, the most striking observation was the metabolic disruption in the phototransduction module on BPA-exposed embryos directly linked to the retinoid mechanism, a pathway not affected by TBT or PFOS exposures. The disruption of retinoid homeostasis by chemicals is a process that can take place at multiple metabolic and functional levels (Shmarakov, 2015). The mechanism of action for the proposed retinoid-disrupting effects of BPA is still unclear, but it has already reported in previous works (Shmarakov, 2015; Ortiz-Villanueva et al., 2017b). Our data indicate that this specific feature of BPA toxicity is not shared by TBT and PFOS.

The high similarity among the metabolic effects produced by BPA, TBT, and PFOS treatments in contrast to their apparent molecular structural dissimilarity would suggest the presence of a common cellular target. Although our data does not allow the possible identification of this common putative target, recent reports suggest the PPAR/RXR signaling system as a strong candidate, especially considering the implication of this complex in the regulation of multiple metabolic pathways (Lempradl et al., 2015; Huang and Chen, 2017). Since PPAR/RXR-regulated pathways are implicated in many human metabolic diseases (Huang and Chen, 2017), the identification of these putative effectors can have profound consequences in environmental risk assessment studies of the effects of EDCs pollution.

4. Conclusions

Untargeted LC-HRMS metabolomic approach allowed the proposal of potential biomarkers related to the effects caused by BPA, PFOS and TBT at sub-lethal doses on zebrafish embryos metabolism.

BPA and TBT exposures showed more severe effects than PFOS exposure, as they disrupted a larger number of pathways in the zebrafish embryos metabolism. The comparison of the effects of all three EDC treatments defined a considerable overlap of the most altered biochemical pathways, although the affected metabolites did not necessarily coincide. When they did, the observed changes followed the same trend (up or down) in most cases, suggesting a similar mode of action for all three EDC. Nevertheless, some evidence for specific effects on particular signaling pathways was also detected, being the one on lipids or retinoid, the most conspicuous one.

Despite the potential of untargeted LC-HRMS metabolomics analyses to assess and distinguish among different metabolic disruptions, the mechanistic interpretation of the observed changes (and therefore, of the potential toxic effects on aquatic fauna and, ultimately, on human populations) would still require its confirmation using combined studies with other omic technologies, such as transcriptomics or lipidomics. Therefore, additional work is pursued in this direction at present.

Acknowledgements

This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) / ERC Grant Agreement n. 320737.

Conflict of interest statement

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2018.03.369.

References

Andreassen, T.K., Skjoedt, K., Korsgaard, B., 2005. Upregulation of estrogen receptor α and vitellogenin in eelpout (*Zoarces viviparus*) by waterborne exposure to 4-tert-octylphenol and 17β-estradiol. Comp. Biochem. Physiol., Part C: Toxicol. Pharmacol. 140, 340–346.

Baxter, C.J., Redestig, H., Schauer, N., Repsilber, D., Patil, K.R., Nielsen, J., Selbig, J., Liu, J., Fernie, A.R., Sweetlove, L.J., 2007. The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. Plant Physiol. 143, 312–325.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.

Bundy, J.G., Davey, M.P., Viant, M.R., 2008. Environmental metabolomics: a critical review and future perspectives. Metabolomics 5, 3.

Casals-Casas, C., Desvergne, B., 2011. Endocrine disruptors: from endocrine to metabolic disruption. Annu. Rev. Physiol. 73, 135–162.

Caspi, R., Altman, T., et al., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 42, D459–D471.

Chang, E.T., Adami, H.-O., Boffetta, P., Wedner, H.J., Mandel, J.S., 2016. A critical review of perfluorooctanoate and perfluorooctanesulfonate exposure and immunological health conditions in humans. Crit. Rev. Toxicol. 46, 279–331.

Diamanti-Kandarakis, E., Bourguignon, J.-P., Giudice, L.C., Hauser, R., Prins, G.S., Soto, A.M., Zoeller, R.T., Gore, A.C., 2009. Endocrine-disrupting chemicals: an Endocrine Society scientific statement. Endocr. Rev. 30, 293–342.

Engel, J., Blanchet, L., Bloemen, B., van den Heuvel, L.P., Engelke, U.H.F., Wevers, R.A., Buydens, L.M.C., 2015. Regularized MANOVA (rMANOVA) in untargeted metabolomics. Anal. Chim. Acta 899, 1–12.

Farres, M., Pina, B., Tauler, R., 2016. LC-MS based metabolomics and chemometrics study of the toxic effects of copper on *Saccharomyces cerevisiae*. Metallomics 8, 790–798.

Fraher, D., Sanigorski, A., Mellett, N.A., Meikle, P.J., Sinclair, A.J., Gibert, Y., 2016. Zebrafish embryonic lipidomic analysis reveals that the yolk cell is metabolically active in processing lipid. Cell Rep. 14, 1317–1329.

Geiger, T., Cox, J., Mann, M., 2010. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol. Cell. Proteomics 9, 2252–2261.

Giulivo, M., Lopez de Alda, M., Capri, E., Barceló, D., 2016. Human exposure to endocrine disrupting compounds: their role in reproductive systems, metabolic syndrome and breast cancer. A review. Environ. Res. 151, 251–264.

Gore, A.C., Chappell, V.A., Fenton, S.E., Flaws, J.A., Nadal, A., Prins, G.S., Toppari, J., Zoeller, R.T., 2015. Executive summary to EDC-2: the Endocrine Society's second scientific statement on endocrine-disrupting chemicals. Endocr. Rev. 36, 593–602.

Gorrochategui, E., Jaumot, J., Tauler, R., A protocol for LC-MS metabolomic data processing using chemometric tools. 2015a.

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

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. TrAC Trends Anal. Chem. 82, 425–442.

Grün, F., Watanabe, H., Zamanian, Z., Maeda, L., Arima, K., Cubacha, R., Gardiner, D.M., Kanno, J., Iguchi, T., Blumberg, B., 2006. Endocrine-disrupting organotin compounds are potent inducers of Adipogenesis in vertebrates. Mol. Endocrinol. 20, 2141–2155.

Hansen, K.J., Johnson, H.O., Eldridge, J.S., Butenhoff, J.L., Dick, L.A., 2002. Quantitative characterization of trace levels of PFOS and PFOA in the Tennessee River. Environ. Sci. Technol. 36, 1681–1685.

Horai, H., Arita, M., et al., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714.

Huang, Q., Chen, Q., 2017. Mediating roles of PPARs in the effects of environmental chemicals on sex steroids. PPAR Res. 2017, 3203161.

Huang, H., Huang, C., Wang, L., Ye, X., Bai, C., Simonich, M.T., Tanguay, R.L., Dong, Q., 2010. Toxicity, uptake kinetics and behavior assessment in zebrafish embryos following exposure to perfluorooctanesulphonicacid (PFOS). Aquat. Toxicol. 98, 139–147.

Huang, S.S.Y., Benskin, J.P., Veldhoen, N., Chandramouli, B., Butler, H., Helbing, C.C., Cosgrove, J.R., 2017. A multi-omic approach to elucidate low-dose effects of xenobiotics in zebrafish (*Danio rerio*) larvae. Aquat. Toxicol. 182, 102–112.

Inadera, H., Shimomura, A., 2005. Environmental chemical tributyltin augments adipocyte differentiation. Toxicol. Lett. 159, 226–234.

Jaumot, J., Gargallo, R., de Juan, A., Tauler, R., 2005. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. Chemom. Intell. Lab. Syst. 76, 101–110.

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

Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016. Compounds altering fat storage in Daphnia magna. Sci. Total Environ. 545, 127–136.

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, 109–114.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F., 1995. Stages of embryonic development of the zebrafish. Dev. Dyn. 203, 253–310.

Lempradl, A., Pospisilik, J.A., Penninger, J.M., 2015. Exploring the emerging complexity in transcriptional regulation of energy homeostasis. Nat. Rev. Genet. 16, 665.

Mallozzi, M., Leone, C., Manurita, F., Bellati, F., Caserta, D., 2017. Endocrine disrupting chemicals and endometrial cancer: an overview of recent laboratory evidence and epidemiological studies. Int. J. Environ. Res. Public Health 14, 334.

Martínez, M.L., Piol, M.N., Sbarbati Nudelman, N., Verrengia Guerrero, N.R., 2017. Tributyltin bioaccumulation and toxic effects in freshwater gastropods Pomacea canaliculata after a chronic exposure: field and laboratory studies. Ecotoxicology 26, 691–701.

Messerlian, C., Martinez, R. M., Hauser, R., Baccarelli, A. A., 'Omics' and endocrine-disrupting chemicals [mdash] new paths forward, *Nat Rev Endocrinol*. 2017, *advance online publication*.

Mihaich, E.M., Friederich, U., Caspers, N., Hall, A.T., Klecka, G.M., Dimond, S.S., Staples, C.A., Ortego, L.S., Hentges, S.G., 2009. Acute and chronic toxicity testing of bisphenol A with aquatic invertebrates and plants. Ecotoxicol. Environ. Saf. 72, 1392–1399.

Oliveira, E., Barata, C., Piña, B., 2016. Endocrine disruption in the omics era: new views, new hazards, new approaches. Open Biotechnol. J. 10.

Ortiz-Villanueva, E., Benavente, F., Piña, B., Sanz-Nebot, V., Tauler, R., Jaumot, J., 2017a. Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data. Anal. Chim. Acta 978, 10–23.

Ortiz-Villanueva, E., Navarro-Martin, L., Jaumot, J., Benavente, F., Sanz-Nebot, V., Pina, B., Tauler, R., 2017b. Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach. Environ. Pollut. 231, 22–36.

Raldua, D., Andre, M., Babin, P.J., 2008. Clofibrate and gemfibrozil induce an embryonic malabsorption syndrome in zebrafish. Toxicol. Appl. Pharmacol. 228, 301–314.

Schonlau, M., 2004. Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Comput. Stat. 19, 95–111.

Shmarakov, I.O., 2015. Retinoid-xenobiotic interactions: the ying and the yang. Hepatobiliary Surgery and Nutrition. 4, 243–267.

Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751.

Soboń, A., Szewczyk, R., Długoński, J., 2016. Tributyltin (TBT) biodegradation induces oxidative stress of Cunninghamella echinulata. Int. Biodeterior. Biodegrad. 107, 92–101.

Tautenhahn, R., Böttcher, C., Neumann, S., 2008. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinf. 9, 504.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., Fernie, A.R., 2003. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO Rep. 4, 989–993.

van der Oost, R., Beyer, J., Vermeulen, N.P., 2003. Fish bioaccumulation and biomarkers in environmental risk assessment: a review. Environ. Toxicol. Pharmacol. 13, 57–149.

Viant, M.R., 2007. Metabolomics of aquatic organisms: the new 'omics' on the block. Mar. Ecol. Prog. Ser. 332, 301–306.

Weaver, P.J., Laures, A.M.F., Wolff, J.-C., 2007. Investigation of the advanced functionalities of a hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer. Rapid Commun. Mass Spectrom. 21, 2415–2421.

Wilson, I.D., Nicholson, J.K., Castro-Perez, J., Granger, J.H., Johnson, K.A., Smith, B.W., Plumb, R.S., 2005. High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. J. Proteome Res. 4, 591–598.

Wishart, D.S., Tzur, D., et al., 2007. HMDB: the human metabolome database. Nucleic Acids Res. 35, D521–D526.

Woodruff, T.J., 2011. Bridging epidemiology and model organisms to increase understanding of endocrine disrupting chemicals and human health effects. J. Steroid Biochem. Mol. Biol. 127, 108–117.

Xu, H., Yang, M., Qiu, W., Pan, C., Wu, M., 2013. The impact of endocrine-disrupting chemicals on oxidative stress and innate immune response in zebrafish embryos. Environ. Toxicol. Chem. 32, 1793–1799.

Ye, G., Chen, Y., Wang, H.o., Ye, T., Lin, Y., Huang, Q., Chi, Y., Dong, S., 2016. Metabolomics approach reveals metabolic disorders and potential biomarkers associated with the developmental toxicity of tetrabromobisphenol A and tetrachlorobisphenol A. 6, 35257.

Yoon, C., Yoon, D., Cho, J., Kim, S., Lee, H., Choi, H., Kim, S., 2017. 1H-NMR-based metabolomic studies of bisphenol A in zebrafish (*Danio rerio*). J. Environ. Sci. Health B 52, 282–289.
Zhang, A., Sun, H., Wang, P., Han, Y., Wang, X., 2012. Modern analytical techniques in metabolomics analysis. Analyst 137, 293–300.
Zhang, J., Sun, P., Yang, F., Kong, T., Zhang, R., 2016. Tributyltin disrupts feeding and energy metabolism in the goldfish (Carassius auratus). Chemosphere 152, 221–228.
Zhou, J., Zhu, X.-s., Cai, Z.-h., 2010. Tributyltin toxicity in abalone (Haliotis diversicolor supertexta) assessed by antioxidant enzyme activity, metabolic response, and histopathology. J. Hazard. Mater. 183, 428–433.

Informació suplementària de l'article científic IV.

Assessment of endocrine disruptors effects on zebrafish (*Danio rerio*) embryos by untargeted LC-HRMS metabolomic analysis.

E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler.

Science of the Total Environment 635 (2018) 156-166.

Untargeted metabolomic data analysis

Regions of interest (ROI) approach

The ROI compression method used in this work searches for significant mass traces and high mass densities among the different analyzed chromatograms, arranging them in data matrices. See [1; 2] for more details about the proposed procedure. ROI approach requires the input of a signal-to-noise threshold value, mass accuracy (*i.e. m/z* error) and the minimum number of elution times to be considered as a peak for each ROI for every sample. In this case, mentioned parameters were respectively set to 5.10^4 (threshold), 0.025 Da/e (mass accuracy) and 250 (minimum number of elution times in a peak for all samples). When this compression approach was applied, a relatively low number of *m/z* values (approximately 500-600) was considered to be further investigated. The resulted column-wise MSROI data matrices were obtained by a pairwise search of the common and uncommon ROI values among the individual data matrices (corresponding to each chromatographic run). In this way, ROI values present in all considered samples were identified. MSROI data matrices further analyzed by MCR-ALS (see below) had an equal number of *m/z* values (column of the final augmented data matrix), but they can have different number of chromatographic elution times (rows of the final augmented data matrix). The dimensions of the ROI augmented data matrices were the total number of elution times considered in the whole set of samples (rows) and the total number of considered *m/z* ROI values (columns).

Multivariate curve resolution alternating least squares (MCR-ALS)

MCR-ALS is a powerful chemometric method for the analysis of unresolved multicomponent mixtures with strongly overlapped signals of the chemical constituents, as frequently are present in untargeted metabolomic LC-HRMS analysis [3]. In the case of the analysis of a single sample, a single LC-MS metabolomic data matrix **D** is obtained. The rows of this matrix have the experimental mass spectra at all elution times and the columns have the chromatograms at every *m/z* value. MCR-ALS decomposed an individual MSROI compressed data matrix **D** using

a bilinear model that produced the concentration profiles and MS spectra of the resolved contributions. MCR-ALS analysis of the data matrix **D** gave two factor matrices, **C** and $\mathbf{S^T}$, as in Eq. 1:

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

where **C** is the matrix of elution profiles of the resolved contributions (components), $\mathbf{S^T}$ contains their mass spectra, and **E** is the matrix containing the residuals unexplained by the model. Peaks resolved in matrix **C** are allowed to vary in position (shifts) and shape among samples.

In the case of the analysis of multiple samples (multiple data matrices), a column-wise augmented data matrix $\mathbf{D_{aug}}$ containing information about the different analyzed samples (Eq. 2), a common matrix of the mass spectra of the resolved components ($\mathbf{S^T}$) for all samples, and a set of matrices of describing the resolved elution profiles ($\mathbf{C_{aug}}$) in every sample are obtained. For instance, in the case of the LC-HRMS analysis of the samples in the PFOS experiments using ESI- the extended bilinear model is given in Eq.2:

$$\mathbf{D_{augPFOS\_ESI-}} = \begin{bmatrix} \mathbf{D_{control1}} \\ \mathbf{D_{control2}} \\ \vdots \\ \mathbf{D_{2\ \mu M\ PFOS\ 5}} \end{bmatrix} = \begin{bmatrix} \mathbf{C_{control1}} \\ \mathbf{C_{control2}} \\ \vdots \\ \mathbf{C_{2\ \mu M\ PFOS\ 5}} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_{control1}} \\ \mathbf{E_{control2}} \\ \vdots \\ \mathbf{E_{2\ \mu M\ PFOS\ 5}} \end{bmatrix} = \mathbf{C_{aug}}\mathbf{S^T} + \mathbf{E_{aug}} \quad (2)$$

An initial guess of the number of components needed to explain the observed data variance can be obtained using the singular value decomposition method [4]. Initial estimations of either **C** or $\mathbf{S^T}$ can be then obtained using pure variable detection method [5]). Equations 1 or 2 can be solved using an alternating least squares (ALS) optimization under non-negativity constraints for elution ($\mathbf{C_{aug}}$) and spectra ($\mathbf{S^T}$) profiles, as well as spectral normalization (equal height) to provide chemical meaning to the resolved elution and mass spectra profiles and minimize possible intensity or rotational ambiguities. MCR-ALS procedure has been explained in more detail in previous works and it is not given here for brevity. See references [6; 7] for more details. The fit quality was evaluated from the explained data variance (R^2) and the percentage of lack of fit (*LOF*), defined in Eq. 3 and 4 [8]:

$$R^2\,(\%) = 100\,\frac{\sum_{i,j} d_{ij}^2 - \sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2} \quad (3)$$

$$\text{Lof}\,(\%) = 100\,\sqrt{\frac{\sum_{i,j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i,j} d_{ij}^2}} \quad (4)$$

where i=1,...,I and j=1,...,J, d_{ij} is an element of the original data matrix **D**, e_{ij} is the residuals of d_{ij}; and $\hat{d}_{ij}$ is the adjusted element by ALS. The final selection of the number of components was performed considering the change in R^2 and *LOF* values when more components are added, and the quality of the resolved chromatographic and spectra profiles.

MCR-ALS analysis was carried out using the MCR-ALS toolbox freely available at www.mcrals.info.

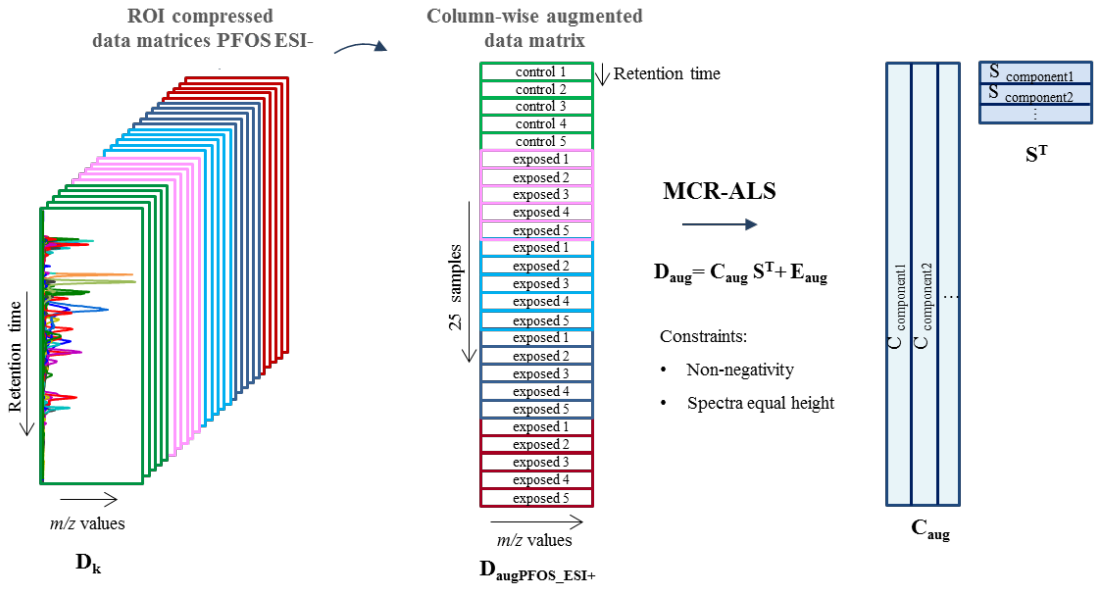


Figure S1. Schematic representation of data arrangement and MCR-ALS decomposition for PFOS data in negative ESI mode.

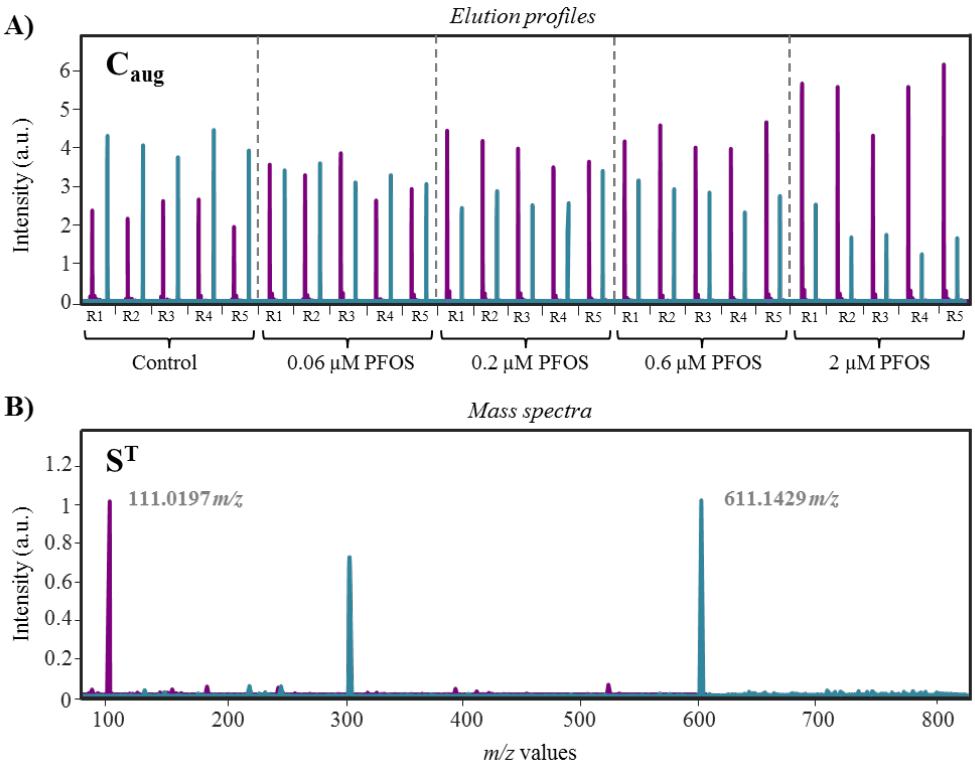


Figure S2. An example of two resolved components by MCR-ALS of the column-wise augmented data matrix $\mathbf{D}_{\mathbf{augPFOS}}$ in ESI-; (A) Elution profiles for the different samples and (B) Mass spectra profiles for the two resolved components.

Metabolite identity confirmation

An example of AIF confirmation for the case of a chromatographic peak at the retention time 9.44 minutes in the full scan data in ESI-, which showed the ion at *m/z* 347.0393. This ion was tentatively identified as the deprotonated molecule ($[M-H]^-$) of inosine monophosphate (IMP) with 1.2 ppm of relative mass error, which has an exact mass of 347.0398 Da. Figure S2A displays the experimental AIF high-resolution mass spectrum corresponding to the chromatographic peak at 9.44 minutes. Figure S2B corresponds to the MS/MS spectrum of IMP obtained from NIST database, which shows the precursor ion at 347.0394 *m/z* and several product ions. As it can be seen, the most intense ions were at 96.9699 and 135.0317 *m/z*. The experimental AIF spectrum (Figure S2A) also showed these two major product ions at 96.9696 and 135.0311 *m/z* with less than 3.1 ppm of relative error. Accordingly, the metabolite was unequivocally identified as IMP with 4.5 identification points (2 reached for the HRMS precursor ion and 2.5 for the two product ions) and fully accomplishment of the identification criteria recommended by Directive 2002/657/CE.

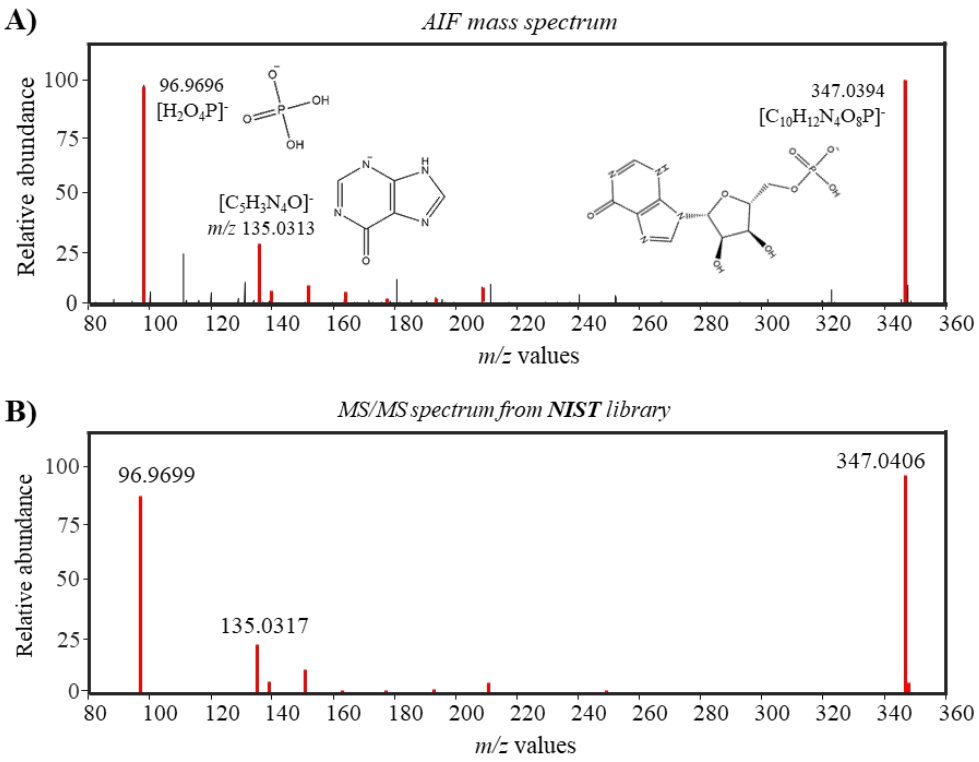


Figure S3. Example of identification of metabolite *m/z* 347.0393: (A) Experimental AIF spectrum at retention time 9.44 min and (B) MS/MS NIST spectrum of inosine monophosphate (IMP) after fragmentation of precursor ion *m/z* 347 in HCD collision cell.

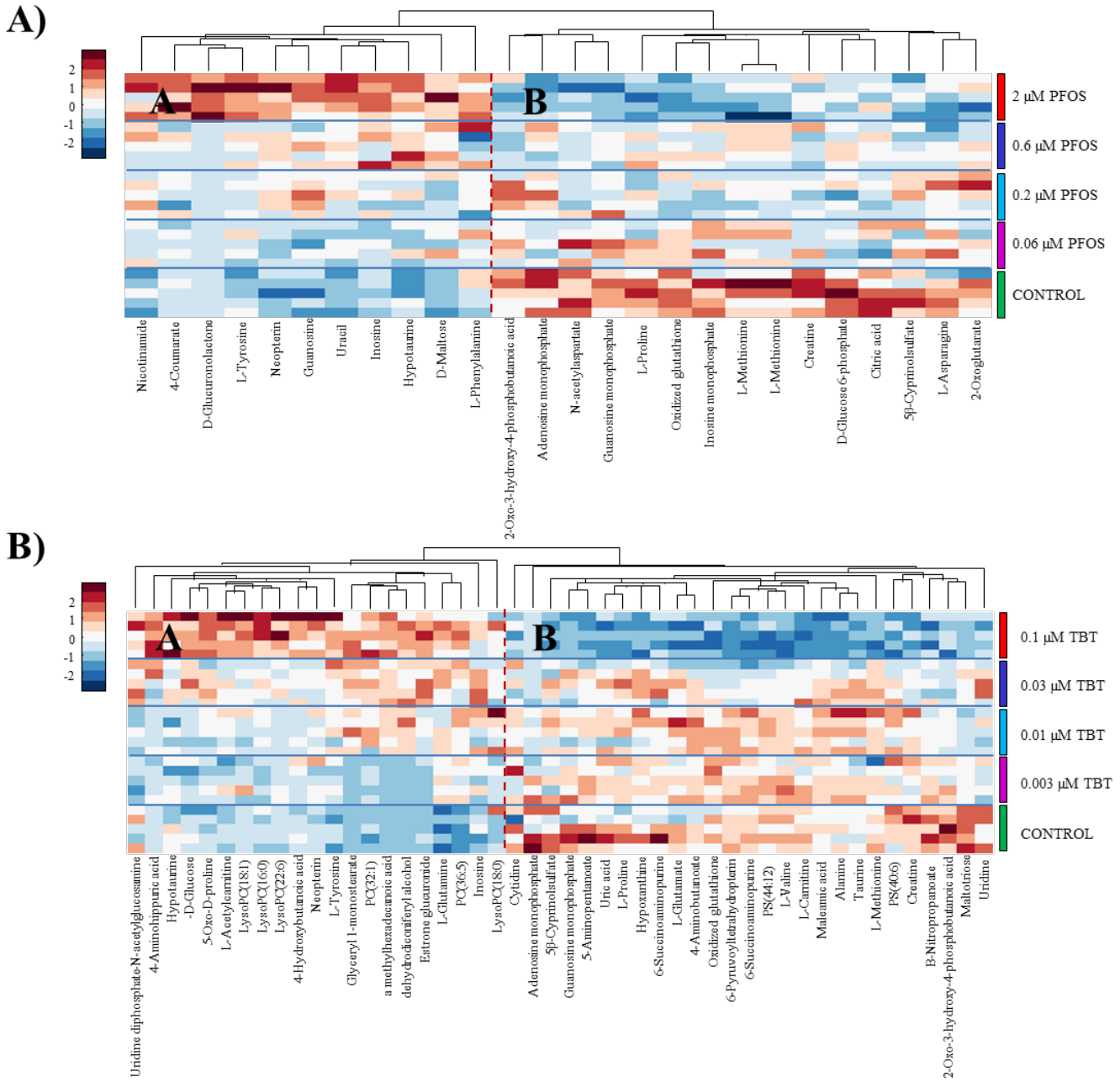
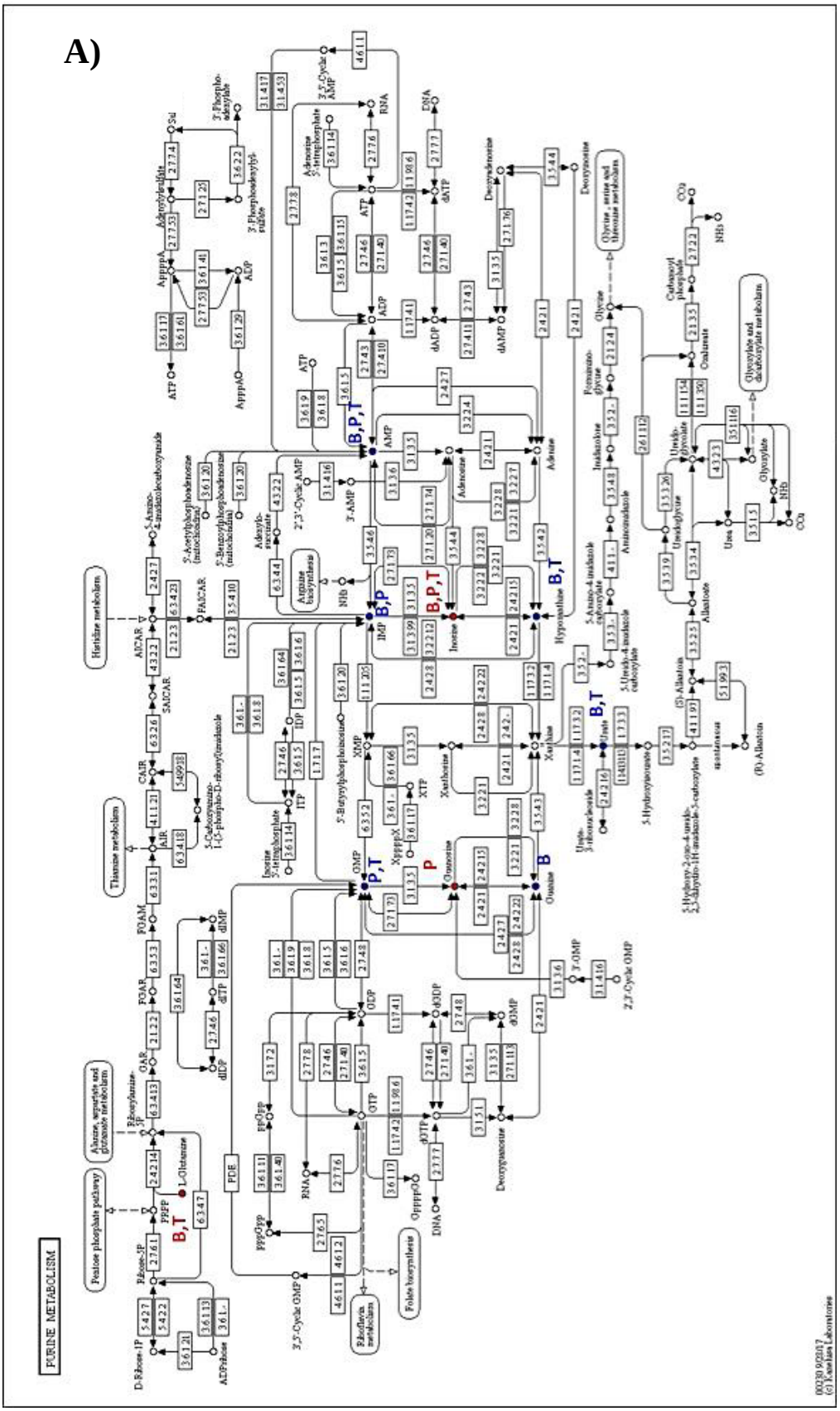
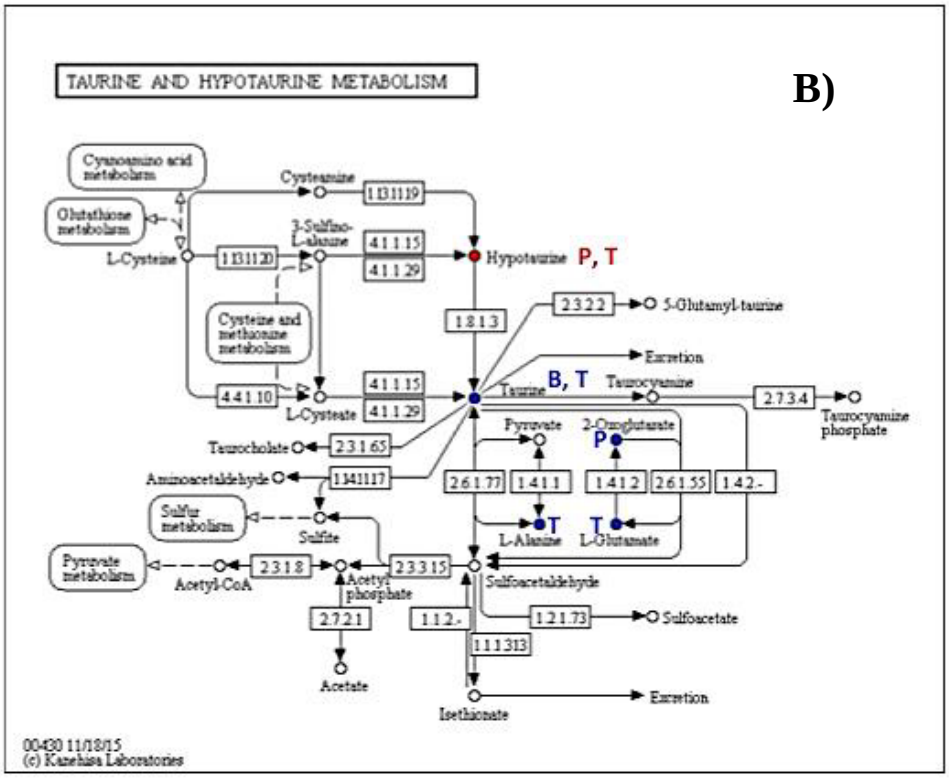


Figure S4. Hierarchical clustering heatmap of the peak areas (autoscaled data) of the statistically significant identified biomarkers whose concentration changed between control and BPA exposed samples by: (A) PFOS and (B) TBT. Cells colors represent auto-scaled relative abundances of each metabolite, indicating metabolites up-regulated (red) and down-regulated going (blue). Color intensity codes are given in the color bar at the left side of the figures. Different clusters containing the metabolites with the same behavior are outlined.


A)
PURINE METABOLISM


TAURINE AND HYPOTAURINE METABOLISM
B)
Cyanoamino acid metabolism
Glutathione metabolism
L-Cysteine
Cysteamine
1.13.11.19
3-Sulfino-L-alanine
4.1.1.15
4.1.1.29
1.13.11.20
Hypotaurine P, T
Cysteine and methionine metabolism
1.8.1.3
2.3.2.2
5-Glutamyl-taurine
Excretion
4.4.1.10
L-Cysteate
Taurine B, T
Taurocyamine
2.7.3.4
Taurocyamine phosphate
Taurocholate
2.3.1.65
Pyruvate
2-Oxoglutarate P
Aminoacetaldehyde
1.3.4.11.17
2.6.1.77
1.4.1.1
1.4.1.2
2.6.1.55
1.4.2.-
Sulfur metabolism
Sulfite
L-Alanine T
L-Glutamate T
Pyruvate metabolism
Acetyl-CoA
2.3.1.8
2.3.3.15
Acetyl phosphate
Sulfoacetaldehyde
2.7.2.1
1.1.2.-
1.2.1.73
Sulfoacetate
1.1.1.313
Acetate
Isethionate
Excretion
00430 11/18/15
(c) Kanehisa Laboratories

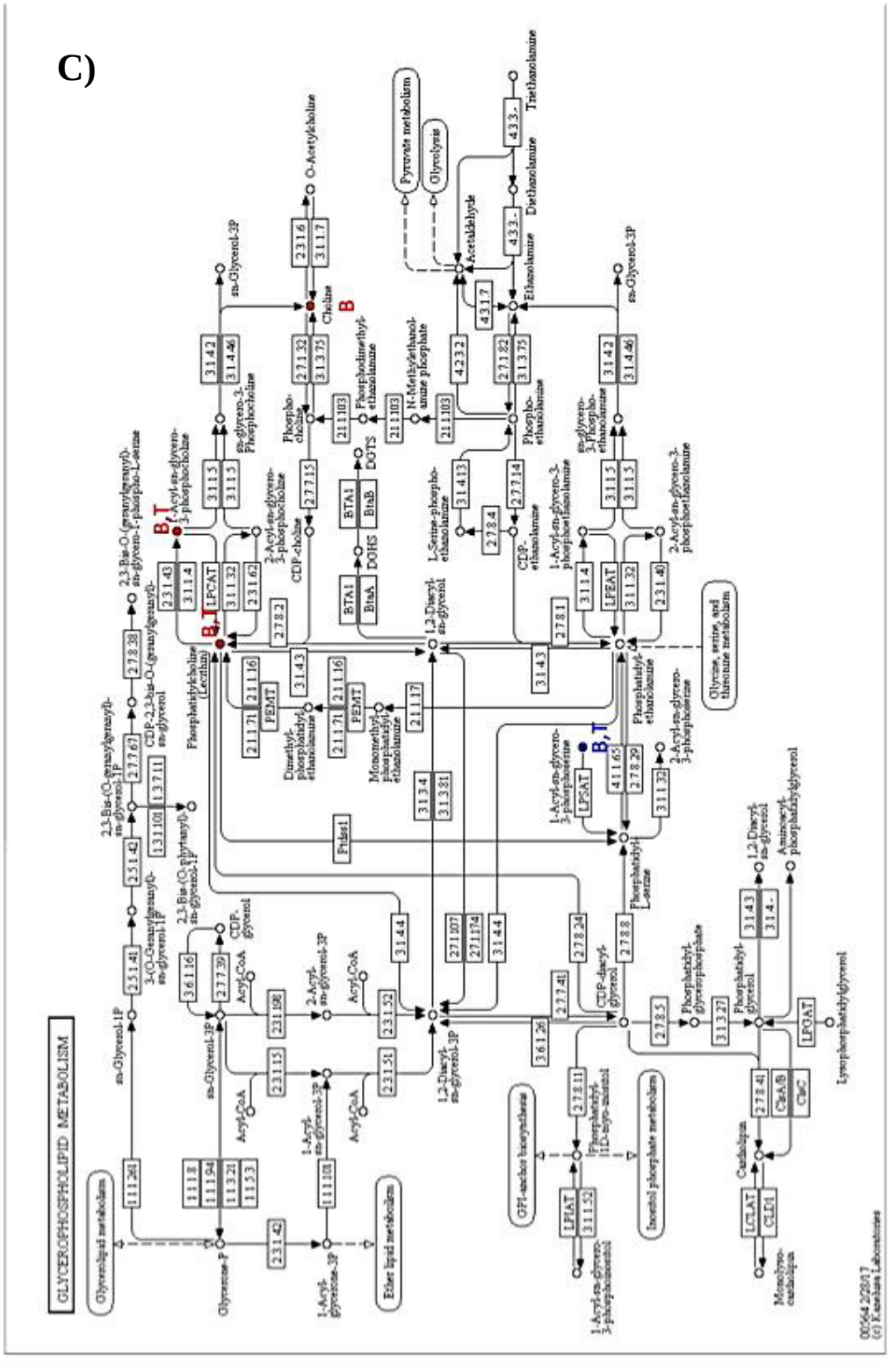


Figure S5. Scheme of three biochemical pathways affected by any of the three treatments: (A) purine metabolism, (B) taurine and hypotaurine metabolism and (C) glyderophospholipid metabolism. Red and blue dots indicate an increase or a decrease of metabolite concentration upon investigated treatments; specific EDC for which these changes were observed are

indicated by letters (B, BPA; P, PFOS; T, TBT). For easier understanding of the observed effects, letters follow the same color code as the dots. Diagrams are from KEGG database (http://www.genome.jp/kegg/tool/map_pathway2.html).

Table S1. Developmental effects of BPA, PFOS and TBT on zebrafish embryos: survival, hatching, inflated swim bladder and tail malformations (lateral and dorsoventral deformities) at 120 hpf. Fisher Exact Probability Test: $p < 0.05$ (∗), $p < 0.01$ (∗∗), $p < 0.001$ (∗∗∗).

		Survival		**Hatching**		**Inflated swim bladder**		**Tail malformations**	
	μM	**n/total**	**Statistical significance**	**n/total**	**Statistical significance**	**n/total**	**Statistical significance**	**n/total**	**Statistical significance**
BPA	**Control**	49/49	n/a	49/49	n/a	48/49	n/a	16/98	**n/a**
	0.44	50/50		50/50		48/50		14/100	
	4.4	50/50		50/50		49/50		24/100	
	8.8	49/49		49/49		46/49		24/98	
	17.5	49/50		50/50		44/50		29/100	∗
	26.3	49/49		49/49		18/49	∗∗∗	42/98	∗∗∗
	35.0	51/51		47/51		1/51	∗∗∗	54/102	∗∗∗
	43.8	50/50		41/50	∗∗	0/50	∗∗∗	66/100	∗∗∗
	219.0	0/50	∗∗∗	0/50	∗∗∗	0/50	∗∗∗	ND	n/a
	438.0	0/50	∗∗∗	0/50	∗∗∗	0/50	∗∗∗	ND	n/a
PFOS	**Control**	50/50	n/a	50/50	n/a	50/50	n/a	4/100	n/a
	0.2	50/50		50/50		50/50		4/100	
	0.5	49/49		48/49		48/49		5/98	
	1.0	50/50		50/50		50/50		6/100	
	2.0	50/50		50/50		49/50		12/100	∗
	5.0	50/50		50/50		33/50	∗∗∗	17/100	∗∗
	10.0	50/50		50/50		1/50	∗∗∗	42/100	∗∗∗
	15.0	46/48		48/48		0/48	∗∗∗	45/96	∗∗∗
	20.0	9/49	∗∗∗	49/49		0/49	∗∗∗	ND	n/a
	200.0	0/47	∗∗∗	47/47		0/47	∗∗∗	ND	n/a
TBT	**Control**	50/50	n/a	50/50	n/a	50/50	n/a	7/100	n/a
	$3.0 \cdot 10^{-4}$	50/50		50/50		50/50		10/100	
	$3.0 \cdot 10^{-3}$	49/49		49/49		49/49		11/98	
	$3.0 \cdot 10^{-2}$	50/50		50/50		49/50		11/100	
	0.10	49/49		49/49		43/49	∗	37/98	∗∗∗
	0.15	42/49	∗∗	49/49		0/49	∗∗∗	59/98	∗∗∗
	0.20	2/49	∗∗∗	49/49		0/49	∗∗∗	ND	n/a
	0.25	0/50	∗∗∗	50/50		0/50	∗∗∗	ND	n/a
	0.30	0/39	∗∗∗	39/39		0/39	∗∗∗	ND	n/a

n/a = Not applicable
ND= Not determined

Table S2. MCR-ALS data fitting results for the analysis of ROI augmented data matrices.

MCR-ALS results				
EDC	**Data**	R^2 **(%)**	***LOF*** **(%)**	**No. of components**
BPA	$\mathbf{D}_{\mathbf{augBPA\_ESI+}}$	99.5	6.8	90
	$\mathbf{D}_{\mathbf{augBPA\_ESI-}}$	99.7	5.6	110
PFOS	$\mathbf{D}_{\mathbf{augPFOS\_ESI+}}$	99.7	5.8	86
	$\mathbf{D}_{\mathbf{augPFOS\_ESI-}}$	99.8	4.7	102
TBT	$\mathbf{D}_{\mathbf{augTBT\_ESI+}}$	99.5	6.9	98
	$\mathbf{D}_{\mathbf{augTBT\_ESI-}}$	99.8	3.8	101

Table S3. List of identified biomarkers upon BPA treatment (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] = 5 ppm).

Compound	Ion Assigment	Measured mass (Da)	Error (ppm)	AIF (*m/z*)	Fold-change ctrl vs. 0.44 µM	Fold-change ctrl vs. 1.75 µM	Fold-change ctrl vs. 4.4 µM	Fold-change ctrl vs. 17.5 µM	Trend	t_r (min)	*p*-adj	KEGG
ESI+ mode												
5-Aminopentanal	$[M+H]^+$	102.0914	-1.0	84.0808/85.0648/102.0914	1.0	1.0	1.0	0.1	DOWN	11.84	1.1E-13	C12455
Choline	$[M+H]^+$	104.1069	1.0	104.1070	1.0	1.1	1.1	1.2	UP	14.72	3.7E-05	C00114
[a]L-Proline	$[M+H]^+$	116.0707	0.0	116.0707	0.9	0.9	0.8	0.8	DOWN	9.35	1.1E-03	C00148
L-Threonine	$[M+H]^+$	120.0656	-0.8	84.0444/102.0550/120.0656	0.9	0.8	0.8	0.6	DOWN	10.97	7.2E-03	C00188
[b]Taurine	$[M+H]^+$	126.0220	-0.8	108.0112/126.0220	0.9	0.9	0.9	0.8	DOWN	8.38	5.9E-05	C00245
[b]5-Oxo-D-proline	$[M+H]^+$	130.0500	-0.8	84.0443/102.0549/130.0499	1.0	1.1	1.2	1.2	UP	11.41	7.9E-03	C02237
[a]Creatine	$[M+H]^+$	132.0768	0.0	90.0549/132.0768	0.9	0.8	0.7	0.7	DOWN	10.65	1.6E-04	C00300
[b]Hypoxanthine	$[M+H]^+$	137.0459	-0.7	94.0400/110.349/119.0352/137.0459	0.8	0.7	0.7	0.6	DOWN	6.58	2.2E-09	C00262
4-Guanidinobutanoic acid	$[M+H]^+$	146.0926	-1.4	86.0604/87.0443111.0553/146.0925	1.0	1.0	0.9	0.5	DOWN	11.12	3.9E-05	C01035
[b]L-Glutamine	$[M+H]^+$	147.0764	0.0	84.0444/86.0600/101.0710/102.0550/129.0659/130.0500/147.0762	0.9	1.0	1.0	1.3	UP	11.44	1.6E-04	C00064
Guanine	$[M+H]^+$	152.0569	-1.3	82.0399/110.0349/135.0302/152.0568	0.8	0.7	0.7	0.7	DOWN	6.86	5.9E-05	C00242
[b]L-Carnitine	$[M+H]^+$	162.1127	-1.2	85.0289/102.0915/103.0389/162.1127	1.0	0.9	0.9	0.9	DOWN	12.78	5.6E-03	C00318
[c]L-Phenylalanine	$[M+H]^+$	166.0864	-0.6	93.0700/103.0542/120.0808	1.6	1.6	1.9	1.9	UP	6.76	1.0E-03	C00079
[a]L-Tyrosine	$[M+H]^+$	182.0813	-0.5	91.0542/95.0491/107.0491/119.0492/123.0441/136.0759/147.0442	1.0	1.5	1.5	2.0	UP	8.81	4.5E-04	C00082
[c]D-Glucuronolactone	$[M+NH_4]^+$	194.0656	1.5	159.0294/177.0395	23.7	110.5	239.1	338.6	UP	5.62	1.8E-10	C02670
[b]6-Pyruvoyltetrahydropterin	$[M+H]^+$	238.0937	-0.8	112.0506/140.0568/166.0724/168.0769/194.0674/220.0831/221.0672/238.0936	0.9	0.9	0.9	0.7	DOWN	5.88	5.6E-03	C03684
6-Lactoyltetrahydropterin	$[M+H]^+$	240.1093	-1.0	166.0725/222.0989/240.1093	1.0	0.9	0.8	0.5	DOWN	7.79	2.1E-04	C04244
diazepinone riboside	$[M+H]^+$	245.1134	-0.8	245.1132/305.1031	1.0	1.2	1.4	1.7	UP	10.88	9.2E-05	-
[a]Inosine	$[M+H]^+$	269.0883	-1.1	94.0400/110.0349/119.0353/137.0459	1.5	1.1	1.6	2.3	UP	6.58	6.3E-03	C00294
Glutathione	$[M+H]^+$	308.0911	0.6	162.0230/179.0487/116.0165/130.0509	1.3	1.3	1.6	1.8	UP	8.89	9.8E-03	C00051
[b]LysoPC(22:6)	$[M+H]^+$	568.3403	-0.9	86.0964/184.0733/240.0999/269.2264/283.2418/311.2373/385.2737/550.3290	1.7	1.7	2.0	2.2	UP	3.29	1.8E-05	C04230
DG(38:4)	$[M+H-H_2O]^+$	627.5358	-1.0	247.2423/265.2527/289.2528/363.2894/627.5352	1.6	1.7	1.9	2.7	UP	2.68	1.3E-05	-
[b]PC(32:1)	$[M+H]^+$	732.5554	-2.2	184.0736/732.5563	1.5	2.9	3.4	3.7	UP	2.93	5.9E-05	C00157
PC(34:1)	$[M+H]^+$	760.5852	-0.1	184.0736/760.5876	1.4	2.0	3.4	3.6	UP	2.94	2.3E-13	C00157
[b]PC(36:5)	$[M+H]^+$	780.5550	-1.5	84.0806/184.0732/245.2264	2.7	3.0	2.7	3.1	UP	2.92	1.2E-09	C00157
[b]PS(44:12)	$[M+H]^+$	880.5126	-0.3	88.0395/269.2265/283.2424/311.2371	1.0	0.9	0.9	0.7	DOWN	2.64	1.9E-04	-
ESI- mode												
L-Lactic acid	$[M-H]^-$	89.0242	2.2	89.0242	1.0	0.9	0.8	0.7	DOWN	5.05	4.5E-02	C00186
[b]4-Hydroxybutanoic acid	$[M-H]^-$	103.0398	2.9	85.0292/103.0397	1.4	1.4	1.4	1.6	UP	4.54	4.0E-03	C00989
L-Serine	$[M-H]^-$	104.0351	1.9	104.035	1.1	1.2	1.2	1.3	UP	14.25	3.4E-02	C00065
[c]Uracil	$[M-H]^-$	111.0197	2.7	111.0196	0.9	0.9	0.8	0.7	DOWN	5.61	4.7E-02	C00106
[b]Maleamic acid	$[M-H]^-$	114.0194	2.6	98.0245/99.0088	1.0	0.9	0.5	0.4	DOWN	6.19	1.3E-05	C01596
[a]L-Proline	$[M-H]^-$	114.0558	2.8	114.0557	0.9	0.9	0.8	0.7	DOWN	11.83	3.5E-03	C00148
[b]L-Valine	$[M-H]^-$	116.0714	2.6	116.0713	1.2	1.2	1.5	2.0	UP	10.98	2.7E-04	C00183
L-Threonine	$[M-H]^-$	118.0507	2.5	118.0506	0.9	0.8	0.8	6.0	DOWN	13.91	4.6E-03	C00188
Benzoic acid	$[M-H]^-$	121.0291	3.3	121.0289	1.3	1.9	1.8	2.0	UP	3.47	5.5E-06	C00180
[b]Taurine	$[M-H]^-$	124.0071	2.4	94.9808/106.9808/124.0070	0.9	0.8	0.7	0.7	DOWN	10.35	2.4E-04	C00245
Mesaconate	$[M-H]^-$	129.0191	1.6	85.0295/129.0187	0.9	0.7	0.7	0.7	DOWN	4.37	4.0E-05	C01732
[a]Creatine	$[M-H]^-$	130.0619	2.3	88.0402/97.0270	1.0	1.0	0.8	0.8	DOWN	13.42	4.7E-02	C00300
L-Leucine	$[M-H]^-$	130.0871	2.3	130.0870	1.2	1.2	1.4	1.4	UP	8.71	2.7E-02	C00123
[b]Hypoxanthine	$[M-H]^-$	135.0309	2.2	92.0251/106.0280/133.0137/135.0308	0.9	0.9	0.8	0.7	DOWN	6.67	4.3E-02	C00262
Salicylic acid	$[M-H]^-$	137.0243	0.7	93.0346/137.0246	1.0	0.8	0.7	0.7	DOWN	3.02	9.6E-05	C00805
Octanoic acid	$[M-H]^-$	143.1073	3.5	143.1071	1.2	2.1	2.2	2.2	UP	3.22	8.9E-07	C06423
[b]Uric acid	$[M-H]^-$	167.0207	2.4	96.0201/97.0040/124.0149/167.0206	0.9	0.9	0.8	0.9	DOWN	6.29	2.5E-02	C00366
[a]L-Tyrosine	$[M-H]^-$	180.0662	2.2	93.0343/119.0498/163.0395/180.0660	1.0	1.5	1.5	2.1	UP	9.8	4.5E-04	C00082

2,4-Dichlorobenzoate	$[M-H]^-$	188.9512	2.1	129.0552/144.9612	1.1	1.0	1.3	1.4	UP	3.13	3.7E-04	C06670
[c]Citric acid	$[M-H]^-$	191.0194	1.6	85.0292/87.0085/111.0082	0.9	0.7	0.7	0.6	DOWN	4.39	3.9E-05	C00158
Phosphoric acid	$[2M-H]^-$	194.9461	0.5	96.9693/158.9248	0.7	0.4	0.5	0.6	DOWN	9.79	1.8E-02	C00009
[a]Inosine	$[M-H]^-$	267.0731	1.5	92.0252/108.0200/135.0308/267.0729	1.2	1.4	1.4	1.5	UP	7.64	4.9E-02	C00294
[b]6-Succinoaminopurine	$[M-H+HAc]^-$	294.0833	3.7	101.0240/134.0468	0.9	1.0	0.8	0.7	DOWN	5.51	1.6E-03	-
Glutathione	$[M-H]^-$	306.0759	2.0	99.0561/128.0349/143.0457/210.0877	1.2	1.5	1.7	1.8	UP	10.06	2.2E-05	C00051
Retinal	$[M-H+HAc]^-$	343.2271	2.3	121.1019/161.0960	1.0	1.0	0.6	0.6	DOWN	2.84	3.5E-03	C00376/ C02110/ C16681
[a]Adenosine monophosphate	$[M-H]^-$	346.0558	0.0	96.9694/107.0358/134.0468	0.9	0.6	0.4	0.4	DOWN	8.45	3.9E-05	C00020
[c]Inosine monophosphate	$[M-H]^-$	347.0393	1.4	96.9693/135.0308/150.9796/211.0006/347.391	1.0	0.9	0.8	0.5	DOWN	9.16	3.1E-02	C00130
[b]Glyceryl 1-monostearate	$[M-H+HAc]^-$	417.3215	1.7	283.264	1.1	30.5	36.2	67.4	UP	2.93	4.8E-08	-
[a]2-Oxo-3-hydroxy-4-phosphobutanoic acid	$[2M-H]^-$	426.9670	3.3	96.9694/117.0190	0.8	0.8	0.8	0.8	DOWN	2.66	1.2E-05	C06054
[a]5β-Cyprinolsulfate	$[M-H]^-$	531.2995	0.3	501.2890/515.3044	0.9	1.0	0.8	0.6	DOWN	2.9	2.4E-04	C05468
[b]Uridine diphosphate-N-acetylglucosamine	$[M-H]^-$	606.0743	-1.3	282.0380/362.0034/384.9829	0.9	0.9	0.9	0.7	DOWN	8.65	1.7E-03	C00043
[b]PS(40:6)	$[M-H]^-$	834.5288	0.4	419.2573/747.4979/834.5305	1.0	1.0	1.0	0.8	DOWN	2.63	1.3E-02	-
[b]PS(44:12)	$[M-H]^-$	878.4973	0.6	463.2246/791.1603	1.0	0.8	0.8	0.8	DOWN	2.59	5.1E-07	-

[a] Common biomarkers between three EDCs.

[b] Common biomarkers between BPA and TBT.

[c] Common biomarkers between BPA and PFOS.

Table S4. List of identified biomarkers upon PFOS treatment (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] = 5 ppm).

Compound	Ion assignation	Measured mass (Da)	Error (ppm)	AIF (*m/z*)	Fold-change ctrl vs. 0.06 µM	Fold-change ctrl vs. 0.2 µM	Fold-change ctrl vs.0.6 µM	Fold-change ctrl vs. 2 µM	Trend	t_r (min)	*p*-adj	KEGG
ESI+ mode												
[b]D-Glucuronolactone	$[M+NH_4]^+$	194.0665	3.1	159.0294/177.0395	0.8	1.9	1.3	44.1	UP	6.56	2.8E-10	C02670
[c]Guanosine monophosphate	$[M+H]^+$	364.0657	1.1	97.0284/135.0303/152.0568	0.8	0.8	0.8	0.6	DOWN	12.62	5.0E-04	C00144
[c]Neopterin	$[M+H]^+$	254.0888	1.6	106.0402/109.0653	1.2	1.3	1.7	1.4	UP	9.76	1.6E-04	C05926
[a]L-Proline	$[M+H]^+$	116.0707	0.9	116.0707	0.9	0.9	0.9	0.8	DOWN	13.03	7.5E-03	C00148
[c]Oxidized glutathione	$[M+H]^+$	613.1595	0.5	231.0435/355.0742/409.0848/484.1168/538.1276/595.1491/613.1595	0.8	0.7	0.7	0.6	DOWN	14.48	7.7E-03	C00127
[b]Inosine monophosphate	$[2M+H]^+$	697.1019	0.6	97.0283/137.0458/233.0669/349.0543/485.0927/561.0629/679.1017	0.9	0.8	0.7	0.6	DOWN	11.05	1.3E-02	C00130
[a]Creatine	$[M+H]^+$	132.0768	0.0	87.0788/90.0549/114.0662	0.4	0.4	0.4	0.4	DOWN	14.15	1.8E-02	C00300
[a]L-Tyrosine	$[M+H]^+$	182.0812	0.0	91.0542/95.0492/107.0491/119.0492/123.0441/136.0759/147.0442	1.0	1.0	1.1	1.4	UP	10.71	7.5E-03	C00082
[b]L-Phenylalanine	$[M+H]^+$	166.0865	1.2	91.0545/93.0700/103.0542/107.0492/120.0808	1.0	1.0	1.0	1.2	UP	8.73	1.8E-02	C00079
Nicotinamide	$[M+H]^+$	123.0554	0.8	80.495/96.0450/123.0555	1.1	1.2	1.2	1.4	UP	5.27	6.4E-03	C00153
[c]L-Methionine	$[M+H]^+$	150.0585	1.3	87.0262/102.0549/104.0528/133.0315	0.9	0.8	0.9	0.7	DOWN	10.69	3.2E-02	C00073
D-Maltose	$[M+H-H_2O]^+$	325.1132	0.9	85.0284/97.0284/109.0285/127.0391/145.0497	1.3	1.2	1.4	1.5	UP	15.61	2.6E-03	C00208
4-Coumarate	$[M+H]^+$	165.0548	1.2	91.0542/95.0491/103.0542/119.0492/123.0441/147.0442	1.0	1.0	1.0	1.3	UP	10.73	1.4E-02	C00811
[a]5β-Cyprinolsulfate	$[M+NH_4]^+$	550.3415	1.3	415.3209/417.3364/435.3470/497.2933	0.9	0.9	0.8	0.7	DOWN	2.95	3.2E-02	C05468
[a]Inosine	$[M+H]^+$	269.0883	1.1	94.0400/110.0349/119.0353/137.0459	1.1	1.4	1.8	2.3	UP	8.21	5.0E-04	C00294
[a]Adenosine monophosphate	$[M+H]^+$	348.0707	0.9	97.0284/119.0354/136.0620/250.0938/348.0706	0.5	0.7	0.6	0.5	DOWN	10.03	1.4E-02	C00020
ESI- mode												
[c]Hypotaurine	$[M-H]^-$	108.0121	0.9	108.0121	1.1	1.1	1.5	1.5	UP	12.67	1.4E-02	C00519
[b]Uracil	$[M-H]^-$	111.0197	2.7	111.0197	1.6	1.8	2.0	2.7	UP	5.59	2.0E-05	C00106
L-Asparagine	$[M-H]^-$	131.0459	2.3	95.0248/113.0352/114.0192	1.0	1.0	0.8	0.8	DOWN	14.72	2.5E-02	C00152
2-Oxoglutarate	$[M-H]^-$	145.0139	2.1	101.023	1.1	0.8	0.7	0.8	DOWN	5.33	4.8E-02	C00026
[c]L-Methionine	$[M-H]^-$	148.0434	2.7	100.0402/148.0432	0.9	0.8	0.9	0.7	DOWN	9.68	2.9E-02	C00073
[b]L-Phenylalanine	$[M-H]^-$	164.0715	1.2	91.0553/103.0553/147.0449/164.0715	1.2	1.2	1.2	1.3	UP	8.05	1.8E-02	C00079
[a]L-Tyrosine	$[M-H]^-$	180.0661	0.0	93.0343/106.0419/107.0498/119.0498/134.0603/163.0396/180.0661	1.0	1.0	1.1	1.4	UP	9.67	7.5E-03	C00082
[b]Citric acid	$[M-H]^-$	191.0194	0.5	85.0292/87.0085/111.0083/129.0188	0.7	0.5	0.7	0.7	DOWN	4.34	2.2E-02	C00158
N-acetylaspartate	$[M-H+HAc]^-$	234.0626	3.0	88.0402/115.0034	0.9	0.8	0.8	0.7	DOWN	5.62	4.2E-03	C01042
[c]Neopterin	$[M-H]^-$	252.0734	1.6	91.0405/150.0419	1.1	1.2	1.2	1.3	UP	8.92	7.5E-03	C05926
D-Glucose 6-phosphate	$[M-H]^-$	259.022	1.5	96.9693/138.9796/259.0215	0.4	0.5	0.6	0.5	DOWN	12.39	1.4E-02	C00668
[a]Inosine	$[M-H]^-$	267.0729	2.2	92.0251/108.0199/135.0308/267.0729	1.2	1.1	1.3	1.5	UP	7.61	1.1E-02	C00294
Guanosine	$[M-H]^-$	282.0838	2.1	108.0195/133.0151/150.0413	1.1	1.6	1.3	1.6	UP	8.67	1.6E-03	C00387
D-Maltose	$[M-H]^-$	341.108	2.6	85.0293/97.0281/109.0405/127.0507	1.1	1.5	3.1	2.7	UP	14.36	9.3E-03	C00208
[a]Adenosine monophosphate	$[M-H]^-$	346.0551	2.0	96.9693/134.0466/150.9795/346.0548	0.6	0.8	0.7	0.5	DOWN	7.42	2.2E-03	C00020
[b]Inosine monophosphate	$[M-H]^-$	347.0395	0.9	96.9696/135.0313/150.9801/347.0394	1.0	0.8	0.9	0.8	DOWN	9.27	1.6E-02	C00130
[a]2-Oxo-3-hydroxy-4-phosphobutanoic acid	$[2M-H]^-$	426.967	0.2	96.9694/117.0190	0.9	1.0	0.5	0.5	DOWN	2.65	1.9E-02	C06054
[a]5β-Cyprinolsulfate	$[M-H]^-$	531.2991	1.1	501.2888/515.3045	0.9	0.7	0.7	0.8	DOWN	2.90	4.7E-02	C05468
[c]Oxidized glutathione	$[M-H]^-$	611.1429	1.5	272.0880/306.0757/338.0476/611.1429	0.8	0.6	0.7	0.4	DOWN	14.11	3.5E-06	C00127

[a] Common biomarkers between three EDCs.
[b] Common biomarkers between PFOS and BPA.
[c] Common biomarkers between PFOS and TBT.

Table S5. List of identified biomarkers upon TBT treatment (mass accuracy = [((exact mass-measured mass)/exact mass) × 10^6] = 5 ppm).

Compound	Ion assignation	Measured mass (Da)	Error (ppm)	AIF (m/z)	Fold-change ctrl vs. 0.003 µM	Fold-change ctrl vs. 0.01 µM	Fold-change ctrl vs. 0.03 µM	Fold-change ctrl vs. 0.1 µM	Trend	t_r (min)	*p*-adj	KEGG
ESI+ mode												
Alanine	$[M+H]^+$	90.0548	2.2	90.0547	1.0	1.0	0.9	0.8	DOWN	10.94	5.0E-05	C01401
4-Aminobutanoate	$[M+H]^+$	104.0705	1.0	86.0600/87.0440/104.0705	1.0	1.1	1.0	0.8	DOWN	12.34	6.1E-03	C00334
[b]L-Valine	$[M+H]^+$	118.0860	2.5	118.0861	1.0	1.0	0.9	0.6	DOWN	9.47	6.5E-04	C00183
[b]Taurine	$[M+H]^+$	126.0219	0.1	108.0113/126.0219	0.9	1.1	1.0	0.7	DOWN	8.63	5.8E-03	C00245
[b]5-Oxo-D-proline	$[M+H]^+$	130.0499	0.0	84.0442/102.0549/130.0499	1.1	1.2	1.4	1.8	UP	11.70	1.2E-05	C02237
[a]Creatine	$[M+H]^+$	132.0766	1.5	87.0786/90.0549/114.0662	0.9	0.9	0.8	0.7	DOWN	10.94	2.8E-04	C00300
[b]Hypoxanthine	$[M+H]^+$	137.0457	0.7	94.0400/110.347/119.0351/137.0457	1.1	1.1	1.0	0.9	DOWN	6.08	2.8E-04	C00262
[b]L-Glutamine	$[M+H]^+$	147.0764	0.0	84.0443/86.0600/101.0709/102.0549/129.0658/130.0499/147.0764	1.1	1.2	1.6	1.6	UP	11.83	3.6E-05	C00064
[c]L-Methionine	$[M+H]^+$	150.0583	0.0	87.0262/102.0549/104.0528/133.0315	0.9	1.0	1.0	0.7	DOWN	8.06	1.1E-03	C00073
[b]L-Carnitine	$[M+H]^+$	162.1124	0.6	85.0288/102.0913/103.0389/162.1125	1.0	1.0	0.9	0.7	DOWN	13.18	4.4E-05	C00318
[a]L-Tyrosine	$[M+H]^+$	182.0810	1.1	91.0542/95.0491/107.0491/119.0492/123.0441/136.0759/147.0442	1.3	0.9	0.9	1.5	UP	10.45	1.6E-02	C00082
L-Acetylcarnitine	$[M+H]^+$	204.1231	0.5	85.0283/129.0785/144.1019/145.0496/204.1231	1.0	1.2	1.2	1.6	UP	10.67	3.5E-04	C02571
4-Aminohippuric acid	$[M+NH4]^+$	212.1029	0.5	94.0651/151.0868/195.0766	1.0	1.0	1.2	1.7	UP	11.34	9.9E-05	-
[b]6-Succinoaminopurine	$[M+H]^+$	236.0775	1.3	103.0390/136.0618	1.0	1.0	0.8	0.7	DOWN	5.14	1.2E-05	-
[b]6-Pyruvoyltetrahydropterin	$[M+H]^+$	238.0934	0.4	166.0724/169.0673/220.0831/239.0934	1.1	1.1	0.7	0.4	DOWN	5.94	3.2E-06	C03684
[c]Neopterin	$[M+H]^+$	254.0883	0.4	164.0566/178.0722/192.0515	1.2	1.1	1.1	1.5	UP	7.87	6.0E-03	C05926
[a]Inosine	$[M+H]^+$	269.0886	2.1	94.0400/110.0349/119.0353/137.0459	1.3	1.5	1.9	1.4	UP	6.71	1.3E-02	C00294
a methylhexadecanoic acid	$[M+NH4]^+$	288.2896	0.3	83.0855/97.1012/103.0754/117.0911	0.9	15.0	15.3	21.8	UP	4.35	3.7E-10	-
[a]Adenosine monophosphate	$[M+H]^+$	348.0704	0.0	97.0283/98.9841/136.0617/348.0702	0.5	0.4	0.3	0.5	DOWN	7.61	1.5E-03	C00020
dehydrodiconiferyl alcohol	$[M+H]^+$	359.1487	0.6	137.0597/151.0754	0.8	12.7	10.9	12.1	UP	3.18	7.3E-08	-
[c]Guanosine monophosphate	$[M+H]^+$	364.0650	0.8	97.0283/135.0301/152.0566	1.0	0.9	0.8	0.7	DOWN	9.44	1.8E-03	C00144
LysoPC(18:1)	$[M+H]^+$	522.3555	0.1	265.2524/283.2632/339.2892/504.3452	1.0	1.1	1.1	2.0	UP	3.30	6.7E-05	C04230
LysoPC(18:0)	$[M+H]^+$	524.3710	0.2	86.0964/104.1070/184.0733/506.3606/524.3712	1.4	1.4	11.7	2.1	UP	3.26	3.2E-03	C04230
[a]5β-Cyprinolsulfate	$[M+NH4]^+$	550.3406	0.4	415.3206/417.3360/431.3164/433.3314/435.3466/497.2925	0.9	1.0	0.6	0.3	DOWN	3.02	3.8E-05	C05468
[c]Oxidized glutathione	$[M+H]^+$	613.1595	0.5	231.0435/355.0742/409.0848/484.1168/538.1276/595.1491/613.1595	1.2	1.1	0.9	0.4	DOWN	13.98	6.7E-05	C00127
[b]PC(32:1)	$[M+H]^+$	732.5541	0.4	184.0731/732.5534	1.5	2.9	3.4	3.7	UP	2.94	6.7E-05	C00157
[b]PC(36:5)	$[M+H]^+$	780.5538	1.5	84.0806/184.0732/245.2264	2.6	2.8	2.7	3.2	UP	2.90	4.9E-05	C00157
[b]PS(44:12)	$[M+H]^+$	880.5122	0.1	88.0395/269.2265/283.2424/311.2371	1.1	1.1	0.9	0.7	DOWN	2.64	2.0E-02	-
ESI- mode												
4-Aminobutanoate	$[M-H]^-$	102.0559	2.0	102.0558	1.1	1.1	1.0	0.6	DOWN	14.17	6.1E-03	C00334
[b]4-Hydroxybutanoic acid	$[M-H]^-$	103.0399	1.9	85.0292/103.0398	1.2	1.1	1.3	2.6	UP	4.51	6.1E-04	C00989
[c]Hypotaurine	$[M-H]^-$	108.0122	2.8	108.012	0.9	0.8	1.1	1.7	UP	12.73	2.3E-03	C00519
[a]L-Proline	$[M-H]^-$	114.0558	2.6	114.0557	1.0	1.0	1.0	0.4	DOWN	11.86	2.8E-03	C00148
[b]L-Valine	$[M-H]^-$	116.0714	2.6	116.0713	0.9	0.9	1.0	0.7	DOWN	10.96	1.6E-03	C00183
5-Oxo-D-proline	$[M-H]^-$	128.0349	3.1	82.0296/128.0349	1..0	1.1	1.5	1.7	UP	5.83	5.3E-06	C02237
[a]Creatine	$[M-H]^-$	130.0619	2.3	88.0402/97.0270	1.2	1.3	1.2	0.9	DOWN	13.44	4.3E-03	C00300
[b]Hypoxanthine	$[M-H]^-$	135.0308	3.0	92.0252/133.0138/135.0308	1.0	1.1	1.1	0.6	DOWN	6.65	8.8E-03	C00262
[b]L-Glutamine	$[M-H]^-$	145.0615	2.8	84.0452/101.0720/109.0408/127.0512/128.0354/145.0619	1.1	1.1	1.4	1.8	UP	14.10	4.3E-03	C00064
L-Glutamate	$[M-H]^-$	146.0456	2.1	102.0558/128.0349	1.0	1.1	1.0	0.7	DOWN	10.28	2.4E-02	C00025
[c]L-Methionine	$[M-H]^-$	148.0434	2.7	100.0402/148.0432	0.9	0.8	0.9	0.4	DOWN	9.81	1.2E-02	C00073
[b]Uric acid	$[M-H]^-$	167.0207	2.4	96.0201/97.0040/124.0149/167.0207	1.0	0.9	0.9	0.5	DOWN	6.32	6.9E-04	C00366
[b]Maleamic acid	$[M-H+HAc]^-$	174.0403	2.9	98.0246/99.0085	1.0	1.1	1.0	0.7	DOWN	6.22	3.1E-06	C01596
5-Aminopentanoate	$[M-H+HAc]^-$	176.0924	2.3	99.9261/115.9203/116.0714	0.9	0.7	0.6	0.2	DOWN	11.40	7.7E-04	C00431
β-Nitropropanoate	$[M-H+HAc]^-$	178.0365	0.0	86.0245/104.0346	0.9	0.8	0.8	0.6	DOWN	7.79	9.2E-04	C05669
α-D-Glucose	$[M-H]^-$	179.0557	2.2	85.0292/89.0241/96.9598	1.0	1.0	1.6	2.0	UP	11.55	6.7E-05	C00267

[a]L-Tyrosine	$[M-H]^-$	180.0662	2.2	93.0343/119.0501/163.0400/180.0665	1.1	0.9	1.0	2.0	UP	9.76	1.6E-03	C00082
Cytidine	$[M-H]^-$	242.0776	2.5	81.0455/109.0400	1.0	1.1	1.0	0.8	DOWN	8.52	3.7E-02	C00475
Uridine	$[M-H]^-$	243.0618	2.1	82.0298/110.0246/122.0242/152.0346/153.0300	0.7	0.7	0.9	0.5	DOWN	6.81	6.0E-03	C00299
[b]6-Succinoaminopurine	$[M-H+HAc]^-$	294.0839	1.7	101.0240/134.0468	0.9	0.9	0.9	0.6	DOWN	5.50	4.8E-02	-
[a]Adenosine monophosphate	$[M-H]^-$	346.0554	1.2	96.9694/107.0358/134.0467	0.7	0.6	0.5	0.5	DOWN	8.36	2.8E-02	C00020
[c]Guanosine monophosphate	$[M-H]^-$	362.0497	2.8	96.9693/150.0416/211.0001/362.0496	0.9	0.8	0.7	0.5	DOWN	10.60	2.0E-02	C00144
[b]Glyceryl 1-monostearate	$[M-H+HAc]^-$	417.3215	1.7	283.2637	0.9	54.5	57.0	63.7	UP	2.91	2.7E-05	-
[a]2-Oxo-3-hydroxy-4-phosphobutanoic acid	$[2M-H]^-$	426.9675	1.2	96.9693/117.0189	0.8	0.8	0.7	0.6	DOWN	2.63	1.9E-04	C06054
Estrone glucuronide	$[M-H]^-$	445.1859	2.0	85.0292/99.0085/103.0240	0.9	2.4	4.8	5.4	UP	3.13	7.2E-06	C11133
[a]5β-Cyprinolsulfate	$[M-H]^-$	531.2990	1.3	501.2889/515.3045	0.7	0.8	0.7	0.4	DOWN	2.89	1.2E-03	C05468
LysoPC(16:0)	$[M-H+HAc]^-$	554.3458	0.9	255.2324/480.3089	0.9	1.1	1.2	2.0	UP	3.58	3.8E-05	C04230
Maltotriose	$[M-H+HAc]^-$	563.1826	0.5	97.0292/161.0451/179.0550	0.5	0.4	0.6	0.3	DOWN	15.28	3.7E-10	C01835
[b]Uridine diphosphate-N-acetylglucosamine	$[M-H]^-$	606.0737	1.0	282.0378/362.0037/384.9830	1.0	1.1	1.1	1.2	UP	8.30	4.8E-02	C00043
[c]Oxidized glutathione	$[M-H]^-$	611.1438	1.5	272.0881/306.0758/338.0479/611.1433	1.1	1.0	0.8	0.4	DOWN	14.16	3.2E-04	C00127
[b]LysoPC(22:6)	$[M-H+HAc]^-$	626.3453	1.6	283.2424/552.3086	1.5	1.3	1.8	4.5	UP	3.48	1.0E-03	C04230
[b]PS(40:6)	$[M-H]^-$	834.5281	1.2	419.2557/747.4964/834.5302	1.0	1.0	0.9	0.7	DOWN	2.61	1.9E-03	-
[b]PS(44:12)	$[M-H]^-$	878.4984	0.7	791.1604/463.2243	0.9	0.7	0.8	0.7	DOWN	2.60	4.8E-02	-

[a] Common biomarkers between three EDCs.

[b] Common biomarkers between TBT and BPA.

[c] Common biomarkers between TBT and PFOS.

Table S6. Summary of observed trends in the concentrations of the metabolites associated to the metabolic pathways included in Figure 5 for each EDC treatment.

Pathway	KEGG code	Compound name	BPA	PFOS	TBT
dre00120 Primary bile acid biosynthesis	C00245	Taurine	DOWN		DOWN
	C05468	5beta-Cyprinolsulfate	DOWN	DOWN	DOWN
dre00230 Purine metabolism	C00020	AMP	DOWN	DOWN	DOWN
	C00064	L-Glutamine	UP		UP
	C00130	IMP	DOWN	DOWN	
	C00144	GMP		DOWN	DOWN
	C00242	Guanine	DOWN		
	C00262	Hypoxanthine	DOWN		DOWN
	C00294	Inosine	UP	UP	UP
	C00366	Urate	DOWN		DOWN
	C00387	Guanosine		UP	
dre00240 Pyrimidine metabolism	C00064	L-Glutamine	UP		UP
	C00106	Uracil	DOWN	UP	
	C00299	Uridine			DOWN
	C00475	Cytidine			DOWN
dre00250 Alanine, aspartate and glutamate metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00041	L-Alanine			DOWN
	C00064	L-Glutamine	UP		UP
	C00152	L-Asparagine		DOWN	
	C00158	Citrate	DOWN	DOWN	
	C00334	4-Aminobutanoate			DOWN
	C01042	N-Acetyl-L-aspartate		DOWN	
dre00260 Glycine, serine and threonine metabolism	C00065	L-Serine	UP		
	C00114	Choline	UP		
	C00188	L-Threonine	DOWN		
	C00300	Creatine	DOWN	DOWN	DOWN
dre00270 Cysteine and methionine metabolism	C00041	L-Alanine			DOWN
	C00051	Glutathione	UP		
	C00065	L-Serine	UP		
	C00073	L-Methionine		DOWN	DOWN
dre00330 Arginine and proline metabolism	C00025	L-Glutamate			DOWN
	C00148	L-Proline	DOWN	DOWN	DOWN
	C00300	Creatine	DOWN	DOWN	DOWN
	C00334	4-Aminobutanoate			DOWN
	C00431	5-Aminopentanoate			DOWN
	C01035	4-Guanidinobutanoate	DOWN		
dre00360 Phenylalanine metabolism	C00079	L-Phenylalanine	UP	UP	
	C00082	L-Tyrosine	UP	UP	UP
	C00180	Benzoate	UP		
	C00805	Salicylate	DOWN		
	C00811	4-Coumarate		UP	
dre00430 Taurine and hypotaurine metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00041	L-Alanine			DOWN
	C00245	Taurine	DOWN		DOWN
	C00519	Hypotaurine		UP	UP
dre00471	C00025	L-Glutamate			DOWN

D-Glutamine and D-glutamate metabolism	C00026	2-Oxoglutarate		DOWN	
	C00064	L-Glutamine	UP		UP
	C02237	5-Oxo-D-proline	UP		UP
dre00564 Glycerophospholipid metabolism	C00114	Choline	UP		
	C00157	Phosphatidylcholine	UP		UP
	C04230	1-Acyl-sn-glycero-3-phosphocholine	UP		UP
	C18125	1-Acyl-sn-glycero-3-phosphoserine	DOWN		DOWN
dre00630 Glyoxylate and dicarboxylate metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00064	L-Glutamine	UP		UP
	C00065	L-Serine	UP		
	C00158	Citrate	DOWN	DOWN	
	C01732	Mesaconate	DOWN		
dre00650 Butanoate metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00334	4-Aminobutanoate			DOWN
	C00989	4-Hydroxybutanoic acid	UP		UP
dre00790 Folate biosynthesis	C03684	6-Pyruvoyltetrahydropterin	DOWN		DOWN
	C04244	6-Lactoyl-5,6,7,8-tetrahydropterin	DOWN		
	C05926	Neopterin		UP	UP
dre00970 Aminoacyl-tRNA biosynthesis	C00025	L-Glutamate			DOWN
	C00041	L-Alanine			DOWN
	C00064	L-Glutamine	UP		UP
	C00065	L-Serine	UP		
	C00073	L-Methionine		DOWN	DOWN
	C00079	L-Phenylalanine	UP	UP	
	C00082	L-Tyrosine	UP	UP	UP
	C00123	L-Leucine	UP		
	C00148	L-Proline	DOWN	DOWN	DOWN
	C00152	L-Asparagine		DOWN	
	C00183	L-Valine	UP		DOWN
	C00188	L-Threonine	DOWN		
dre01200 Carbon metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00041	L-Alanine			DOWN
	C00065	L-Serine	UP		
	C00158	Citrate	DOWN	DOWN	
	C00267	alpha-D-Glucose			UP
	C00668	alpha-D-Glucose 6-phosphate		DOWN	
	C00989	4-Hydroxybutanoic acid	UP		UP
	C01732	Mesaconate	DOWN		
dre01210 2-Oxocarboxylic acid metabolism	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00073	L-Methionine		DOWN	DOWN
	C00079	L-Phenylalanine	UP	UP	
	C00082	L-Tyrosine	UP	UP	UP
	C00123	L-Leucine	UP		
	C00158	Citrate	DOWN	DOWN	
	C00183	L-Valine	UP		DOWN
dre01230 Biosynthesis of amino acids	C00025	L-Glutamate			DOWN
	C00026	2-Oxoglutarate		DOWN	
	C00041	L-Alanine			DOWN
	C00064	L-Glutamine	UP		UP
	C00065	L-Serine	UP		
	C00073	L-Methionine		DOWN	DOWN
	C00079	L-Phenylalanine	UP	UP	

	C00082	L-Tyrosine	UP	UP	UP
	C00123	L-Leucine	UP		
	C00148	L-Proline	DOWN	DOWN	DOWN
	C00152	L-Asparagine		DOWN	
	C00158	Citrate	DOWN	DOWN	
	C00183	L-Valine	UP		DOWN
	C00188	L-Threonine	DOWN		
dre02010 ABC transporters	C00009	Orthophosphate	DOWN		
	C00025	L-Glutamate			DOWN
	C00041	L-Alanine			DOWN
	C00051	Glutathione	UP		
	C00064	L-Glutamine	UP		UP
	C00065	L-Serine	UP		
	C00079	L-Phenylalanine	UP	UP	
	C00114	Choline	UP		
	C00123	L-Leucine	UP		
	C00148	L-Proline	DOWN	DOWN	DOWN
	C00183	L-Valine	UP		DOWN
	C00188	L-Threonine	DOWN		
	C00208	Maltose		UP	
	C00245	Taurine	DOWN		DOWN
	C00487	Carnitine	DOWN		DOWN
	C01835	Maltotriose			DOWN

References

[1] Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. TrAC Trends in Analytical Chemistry 82, 425-442.
[2] Ortiz-Villanueva, E., Benavente, F., Piña, B., Sanz-Nebot, V., Tauler, R., Jaumot, J., 2017. Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data. Analytica Chimica Acta 978, 10-23.
[3] 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.
[4] Golub, G., Solna, K., Dooren, P.V., 2000. Computing the SVD of a General Matrix Product/Quotient. SIAM J. Matrix Anal. Appl. 22, 1-19.
[5] Windig, W., Guilment, J., 1991. Interactive self-modeling mixture analysis. Analytical Chemistry 63, 1425-1432.
[6] Farrés, M., Piña, B., Tauler, R., 2015. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS. Metabolomics 11, 210-224.
[7] Gorrochategui, E., Casas, J., Porte, C., Lacorte, S., Tauler, R., 2015. Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells. Analytica Chimica Acta 854, 20-33.
[8] de Juan, A., Jaumot, J., Tauler, R., 2014. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Analytical Methods 6, 4964-4976.

4.2.2. Article científic V.

Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach.

E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler. *Environmental Pollution* 231 (2017) 22-36.

Environmental Pollution 231 (2017) 22–36

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach☆

Elena Ortiz-Villanueva [a], Laia Navarro-Martín [a], Joaquim Jaumot [a], Fernando Benavente [b], Victoria Sanz-Nebot [b], Benjamín Piña [a], Romà Tauler [a, *]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 645, 08028 Barcelona, Spain

ARTICLE INFO

Article history:
Received 26 January 2017
Received in revised form
27 July 2017
Accepted 28 July 2017

Keywords:
Bisphenol A
Metabolic disruption
Non-targeted metabolomics
Transcriptomics
Zebrafish

ABSTRACT

Although bisphenol A (BPA) is commonly recognized as an endocrine disruptor, the metabolic consequences of its exposure are still poorly understood. In this study, we present a non-targeted LC-MS based metabolomic analysis in combination with a full-genome, high-throughput RNA sequencing (RNA-Seq) to reveal the metabolic effects and the subjacent regulatory pathways of exposing zebrafish embryos to BPA during the first 120 hours post-fertilization. We applied multivariate data analysis methods to extract biochemical information from the LC-MS and RNA-Seq complex datasets and to perform testable predictions of the phenotypic adverse effects. Metabolomic and transcriptomic data revealed a similar subset of altered pathways, despite the large difference in the number of identified biomarkers (around 50 metabolites and more than 1000 genes). These results suggest that even a moderate coverage of zebrafish metabolome may be representative of the global metabolic changes. These multi-*omic* responses indicate a specific metabolic disruption by BPA affecting different signaling pathways, such as retinoid and prostaglandin metabolism. The combination of transcriptomic and metabolomic data allowed a dynamic interpretation of the results that could not be drawn from either single dataset. These results illustrate the utility of -*omic* integrative analyses for characterizing the physiological effects of toxicants beyond the mere indication of the affected pathways.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Bisphenol A (BPA) is a common industrial chemical amply used in the manufacture of polycarbonate plastics and epoxy resins (Wang et al., 2015). Due to their unbeatable resistance and temperature tolerance, BPA-derived plastics are usually employed in the production of a broad range of goods, including containers, electronic products, and medical devices (Huang et al., 2012). BPA leaks from the original plastics over time by the hydrolysis of the ester bonds linking it to plastic monomers. As a consequence, large amounts of BPA are continuously released into the environment, especially to surface and groundwater, ending up in wastewater treatment plants and being finally deposited in sediments (Flint et al., 2012; Kang et al., 2006). According to the United States Environmental Protection Agency (EPA), the annual release of BPA into the environment exceeded 500 tons in 2012. The highest BPA levels ($mg \cdot L^{-1}$) are usually found in landfill leachate and mill effluents, while usual concentrations in surface waters are at the $\mu g \cdot L^{-1}$ levels (Canesi and Fabbri, 2015; Flint et al., 2012; Kolpin et al., 2002). Nevertheless, BPA is considered an important threat to the human health, wildlife and environment, owing to its estrogenic activity at environmentally relevant concentrations (Xu et al., 2013). In addition to its assumed endocrine and reproductive toxicity (Crain et al., 2007; Kang et al., 2007; Villeneuve et al., 2012), several authors have already described that BPA exposure disrupts hormone balances (Arase et al., 2011; Careghini et al., 2014), alters gonadal functions (Chen et al., 2017; Choi et al., 2016; Ekman et al., 2012; Lee et al., 2003), and affects hepatic vitellogenin production in fish and other organisms (Lindholst et al., 2000; Rankouhi et al., 2002). However, the mode of action of some of these effects is not entirely understood. The European Food Safety Authority fixed a temporary tolerable daily intake (t-(TDI)) of 4 $\mu g \cdot kg^{-1}$ body weight (BW) for humans, which could be reached in some cases even from exclusively environmental, non-

☆ This paper has been recommended for acceptance by Maria Cristina Fossi.
* Corresponding author.
E-mail address: roma.tauler@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.envpol.2017.07.095
0269-7491/© 2017 Elsevier Ltd. All rights reserved.

dietary exposures (European Food Safety and A, 2016; Lejonklou et al., 2017).

Diverse organisms and biological systems have been used to evaluate the toxicological effects of BPA exposure, including rats (Chen et al., 2014; Geetharathan and Josthna, 2016; Li et al., 2016a; Tremblay-Franco et al., 2015), mice (Javurek et al., 2016) and mammal cells (Stiefel et al., 2016). However, aquatic model organisms confer the most valuable option for toxicological tests in environmental and biomedical research, especially crustaceans (*Daphnia magna*) (Jordão et al., 2016; Nagato et al., 2016) and zebrafish (*Danio rerio*) (Chen et al., 2012; Lam et al., 2011). Zebrafish has been recognized as a privileged vertebrate model due to its relative genetic proximity to humans and the extensive characterization of its genome and transcriptome. Zebrafish embryos up to 120 hours post-fertilization (hpf) are considered as replacement method in toxicology according to the stringent European legislation on the protection of animals used for scientific purposes (Directive 2010/63/EU, 2010). Other advantages, such as its small size, wide distribution, and easy growth conditions offer the possibility to perform small-scale and high-throughput analyses for *-omic* studies, including metabolic profiling (Olivares et al., 2013; Xu et al., 2013). Zebrafish has been deeply studied at least at four *-omic* levels: genomics, transcriptomics, proteomics and metabolomics (Mushtaq et al., 2013). In addition, several environmental studies have assessed BPA toxic effects in zebrafish (Chen et al., 2015, 2017; Laing et al., 2016; Qiu et al., 2015; Santangeli et al., 2016).

At present, there is an increasing trend to use integration approaches in an attempt to obtain complementary information for better interpretation. For instance, information coming from the analysis of same samples at different *-omic* levels could be integrated (Cavill et al., 2016; Huang et al., 2017; Katsiadaki et al., 2010; Santos et al., 2010), such as transcriptomic (mRNA) and metabolomic (metabolites) data. Alternatively, integration of the information from the analysis of same samples at the same *-omic* level using different analytical platforms is also common (Soanes et al., 2011), such as the joint analysis of one-dimensional ^{1}H nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography-mass spectrometry (LC-MS) data. In the environmental field, several studies dealing with the combination of multi-*omic* levels to evaluate the effects of toxicants by using quantitative real-time polymerase chain reaction (qPCR) or DNA microarrays (genomic or transcriptomic information) and NMR or targeted LC-MS (metabolomic information) have been reported (Benskin et al., 2014; Katsiadaki et al., 2010; Santos et al., 2010; Williams et al., 2009). To the best of our knowledge, only few studies focus on the pure integration of transcriptomic (qPCR) and metabolomic (targeted LC-MS) data (Huang et al., 2017). Recently, the combination of information from other platforms has also been proposed (Song et al., 2016; Yang et al., 2017). These integration efforts aim to improve the abilities to identify and to connect the different molecular pathways affected by a particular stimulus or stressing condition (Gligorijević et al., 2016; Huang et al., 2017).

The main aim of this paper is to study the effects of sub-lethal BPA exposures of early zebrafish embryos up to 120 hpf, particularly in metabolic pathways, by using a combination of metabolomic and transcriptomic data. On the one hand, a non-targeted metabolomic approach based on LC-MS was chosen for a more comprehensive evaluation of metabolic alterations at the metabolome level. This method is appropriated due to its great sensitivity and selectivity. On the other hand, mRNA abundances were investigated using high-performance RNA sequencing (RNA-Seq). Results from both assays were further analyzed using multivariate chemometric models and standard data analysis tools to assess the metabolomic and transcriptomic effects of BPA separately. Finally, integration of the results from both *-omic* levels was used both to characterize relevant molecular biomarkers with high reliability and to identify the subjacent regulatory pathways underlying the observed BPA toxicity-induced changes in zebrafish embryos.

2. Materials and methods

2.1. Chemicals and reagents

All chemicals used in the preparation of buffers and solutions were analytical reagent grade. Acetic acid (glacial), methanol (HPLC grade) and acetonitrile (HPLC and MS grade) were purchased from Merck (Darmstadt, Germany). Chloroform was supplied by Carlo Erba (Peypin, France). Ammonium acetate (MS grade), dimethyl sulfoxide (DMSO), water (HPLC and MS grade), calcium sulfate dihydrate ($CaSO_4 \cdot 2H_2O$) and bisphenol A (BPA) were provided by Sigma-Aldrich (St. Louis, MO, USA). L-methionine sulfone and Piperazine-1,4-bis(2-ethanesulfonic acid) (PIPES), used as internal standards (IS), were also supplied by Sigma-Aldrich (St. Louis, MO, USA).

2.2. Animals and rearing conditions

Adult wild-type zebrafish were maintained under standard conditions in fish water, composed of reverse-osmosis purified water containing 90 $\mu g \cdot mL^{-1}$ of Instant Ocean (Aquarium Systems, Sarrebourg, France) and 0.58 mM $CaSO_4 \cdot 2H_2O$ at a temperature of 28 (±1 °C). Adults were fed twice a day with dry flakes (TetraMin, Tetra, Germany). Zebrafish embryos were obtained by natural mating by placing 6 males and 3 females on 4-L mesh-bottom breeding tanks. At 2 hpf, eggs were collected and rinsed. Fertilized viable eggs were randomly distributed in 6-well multiplates (10 embryos/well). Embryos were raised at 28.5 °C with a 12 Light:12 Dark photoperiods in fish water (3 mL/well). All procedures were conducted in accordance with the institutional guidelines under a license from the local government (DAMM 7669, 7964) and were approved by the Institutional Animal Care and Use Committees at the Research and Development Center of the Spanish Research Council, CID-CSIC.

2.3. Zebrafish embryo BPA treatments and sample collection

BPA stock solutions were prepared in DMSO (carrier) on the day of the experiment. Dosing solutions with the same final concentration of DMSO (0.2% v/v) were obtained by dissolving the stocks with fish water. Zebrafish embryos were exposed to contaminated water samples at 2 hpf. A preliminary range-finding test was performed to select BPA concentrations. The highest selected BPA concentration (17.5 μM) corresponds to the lowest observed effect concentration (LOEC) of gross morphological effects calculated experimentally. The concentrations used in the present study were chosen to be below the LOEC, hardly even any phenotypic effect at the morphological level was observed at either life stage. This selection was done to avoid detecting molecular events directly related to alterations in development, malformations or larvae viability is coincident with previous data (Berghmans et al., 2008; Pelayo et al., 2012).

Zebrafish embryos were exposed to fish water (1.125 $g \cdot L^{-1}$ Instant Ocean® + 250 $\mu g \cdot L^{-1}$ $CaSO_4$) containing DMSO (0.2% v/v) as the control group, 4.4, 8.8 and 17.5 μM of BPA (using DMSO as a carrier). Anatomical development of embryos was followed daily during the exposure as described by Kimmel et al. (1995) under a stereomicroscope Nikon SMZ1500 equipped with a Nikon digital sight DS-Ri1 digital camera. Mortality/survival (24, 48, 72, 96 and 120 hpf), hatching (72, 96 and 120 hpf) and inflated swim bladder

(96 and 120 hpf) rates were recorded for each concentration (see Supplementary Material Fig. S1). Mortality during the exposure protocol was negligible and sub-lethal endpoints such as coagulated embryos, lack of somite formation, non-detachment of the tail, lack of heartbeat were not observed. Non-hatched eggs were insignificant for either control and exposed groups. However, embryos exposed to 17.5 μM of BPA presented lower rates of inflated swim bladder at 120 hpf (56%) and slightly darker pigmentation. Continuous exposure to BPA was assured by preparing and renewing water solutions every day until embryo collection at 120 hpf. Chemical analysis of media water solutions was not performed in accordance to the high stability of BPA in water previously reported by Jordão et al. (2016), where the measured concentration of BPA in old test solutions (after 48 h) resulted only 5% lower than freshly prepared ones.

Pools of 20 and 8 zebrafish embryos were gathered for each biological replicate to ensure sufficient amount of metabolites and RNA, respectively. A total number of six biological replicates per treatment were used for LC-MS analyses and three for RNA-Seq analyses. A rinsing step was included for the samples collected for metabolite extraction, which consisted in washing the embryos twice with 0.5 mL HPLC grade water. In both cases, embryo pools were snap-frozen in dry ice and stored at -80 °C until further analysis.

2.4. Sample preparation and analysis

2.4.1. Metabolite extraction and LC-MS analysis

Each pool of 20 zebrafish embryos was thawed in a water bath at room temperature. Metabolites were first extracted with 900 μL of methanol containing the internal standard (IS1) (L-methionine sulfone) at a final concentration of 5 $\mu g \cdot mL^{-1}$. After vortexing 15 s, the mixture was sonicated for 15 min and vortexed again 15 s. Samples were then centrifuged at 23500 g for 10 min at 4 °C to isolate the supernatant and subsequently 500 μL of water and 300 μL of chloroform were added. Samples were vortexed 15 s, placed on ice for 10 min and centrifuged again at 23500 g for 10 min at 4 °C. Finally, the aqueous fractions were evaporated to dryness under nitrogen gas and reconstituted with 100 μL of 1:1 v/v acetonitrile:water containing PIPES (IS2) at a concentration of 5 $\mu g \cdot mL^{-1}$. Prior to injection, zebrafish embryo extracts were filtered through 0.22 μm filters (Ultrafree®-MC, Millipore Bedford, MA, USA) at 11000 g for 4 min and stored at -80 °C until their analysis. All centrifugations were performed at 4 °C in a Serie Digicen 21 centrifuge (Ortoalresa, Madrid, Spain).

LC-MS analyses were carried out using an Agilent Infinity 1200 series LC system coupled with an orthogonal G1385-44300 interface (Agilent Technologies, Waldbronn, Germany) to a 6220 oa-TOF LC/MS mass spectrometer (Agilent Technologies). LC control and separation data acquisition were performed using ChemStation software (Agilent Technologies) that was running in combination with the MassHunter workstation software (Agilent Technologies) for control and data acquisition of the TOF mass spectrometer. For the chromatographic separations, an HILIC TSK Gel Amide-80 column (250 mm length, 2.1 mm inner diameter and 5 μm particle size, Tosoh Bioscience, Tokyo, Japan) was used at 25 °C with gradient elution at a flow rate of 0.15 $mL \cdot min^{-1}$. Elution gradient was performed using solvent A (acetonitrile) and solvent B (5 mM of ammonium acetate adjusted to pH 5.5 with acetic acid) as follows: 0–8 min, linear gradient from 25 to 30% B; 8–12 min, from 30 to 60% B; 12–17 min, 60% B; 17–20 min, back linearly from 60% to 25% B; and from 20 to 27 min, 25% B. Solvents were degassed for 15 min by sonication before use. Sample injection was performed with an autosampler at 4 °C, and the injection volume was 5 μL. All samples (six replicates per treatment: control, 4.4 μM BPA, 8.8 μM BPA and 17.5 μM BPA) were randomly injected. Several blank samples and calibration standards were also randomly injected to further assess the stability of the instrument among runs.

The TOF mass spectrometer operated both in positive and negative mode using the following parameters: capillary voltage 4000 V, drying gas temperature 350 °C, drying gas flow rate 8 $L \cdot min^{-1}$, nebulizer gas 32 psi, fragmentor voltage 150 V, skimmer voltage 65 V and OCT 1 RF Vpp voltage 300 V. Data were collected in profile mode at 1 spectrum/s (approximately 10 000 transients/spectrum) with an *m/z* range of 85–1000 working in the extended dynamic range mode (2 GHz) with the mass range set to standard.

2.4.2. RNA extraction, library construction and sequencing

Total RNA was isolated from three pools of 8 zebrafish embryos using AllPrep DNA/RNA Mini Kit (Qiagen, CA, USA) as described by the manufacturer. Extracted RNA was reconstituted in RNase-free water. Then, total RNA was assayed for quantity and quality using Qubit® RNA BR Assay kit (Thermo Fisher Scientific) and RNA 6000 Nano Assay on a Bioanalyzer 2100 (Agilent Technologies). Given the monotonic response of the metabolome to BPA, we opted to focus the transcriptomic analysis comparing control and high dose (17.5 μM BPA) groups. High-quality RNA biological replicates were sent for sequencing to the National Center for Genomic Analysis (CNAG, Barcelona, Spain). In all cases, RNA concentrations ranged between 50 and 200 $ng \cdot \mu L^{-1}$, free of DNA and with RNA integrity (RIN) number >8.

The RNA-Seq libraries were prepared from total RNA using KAPA Stranded mRNA-Seq Kit Illumina® Platforms (Kapa Biosystems) with minor modifications. Briefly, after poly-A based mRNA enrichment with oligo-dT magnetic beads and 500 ng of total RNA as the input material, the mRNA was fragmented (resulting RNA fragment size was 80–250 nt, with the major peak at 130 nt). The second strand cDNA synthesis was performed in the presence of dUTP instead of dTTP, to achieve the strand specificity. The blunt-ended double-stranded cDNA was 3′adenylated and Illumina indexed adapters (Illumina) were ligated. The ligation product was enriched with 15 PCR cycles and the final library was validated on an Agilent 2100 Bioanalyzer with the DNA 7500 assay.

Each library was sequenced using TruSeq SBS Kit v3-HS, in the paired-end mode with a read length of 2 × 76bp. We generated on average 40 million paired-end reads for each sample in a fraction of a sequencing lane on HiSeq2000 (Illumina) following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files by CASAVA.

2.5. Data analysis

2.5.1. LC-MS metabolomic data analysis

LC-MS data were analyzed by a combination of multivariate analysis tools to evaluate the most significant metabolic changes caused by BPA treatment. The data analysis workflow used in this study was carried out as described in previous studies (Navarro-Reig et al., 2015; Ortiz-Villanueva et al., 2015) with some modifications.

Total ion chromatograms (TIC) of the 24 samples were analyzed using data exploration chemometric tools implemented in the PLS Toolbox version 7.3.1 (Eigenvector Research Inc., Wenatchee, WA, USA). First, the intensity scale of every TIC was normalized by the area of PIPES (IS2). Chromatographic peaks were aligned in the time domain to correct the small variations in retention times of metabolites among different chromatographic runs using the correlation optimized warping (COW) method (Tomasi et al., 2004). Chromatograms were also baseline corrected by automatic

weighted least squares and mean-centered. Finally, unsupervised principal component analysis (PCA) was applied to differentiate between embryo samples according to BPA exposure and to detect possible outlier samples.

Non-targeted analysis of LC-MS raw data was performed using the multivariate curve resolution alternating least squares (MCR-ALS) method (Jaumot et al., 2015) following the same procedure as described in previous studies (Farrés et al., 2015; Ortiz-Villanueva et al., 2015) (see details in Supplementary material). The application of this approach allowed the resolution of the chromatographic elution profiles of the metabolites present in every zebrafish embryo sample as well as of their corresponding mass spectra (Farrés et al., 2015; Ortiz-Villanueva et al., 2015). Prior to MCR-ALS analysis, blank samples were employed to estimate the noise threshold used as a low filter in the MinMax algorithm (Ortiz-Villanueva et al., 2015), used for peak normalization (see Supplementary material, MCR-ALS section). The threshold was set at 1e2 for negative ionization and 2e3 for positive ionization. The intensity scale of every chromatogram was also normalized dividing by the area of PIPES (IS2).

After applying MCR-ALS, the significance of the concentration changes upon BPA treatment of resolved elution profiles (detected metabolites) in the different zebrafish embryo samples was assessed using statistical tests. With this aim, the chromatographic peak areas of the resolved metabolites by MCR-ALS were calculated. The obtained peak area values from MassHunter software were normalized by the peak area of the IS2 (PIPES), for each sample. Then, changes in the normalized peak areas of control and exposed samples for each metabolite were individually evaluated applying one-way ANOVA, considering exposure level as a factor, and correcting *p*-values according to the Benjamini-Hochberg multiple comparison testing procedure (Benjamini and Hochberg, 1995). Metabolites with an adjusted *p*-value lower than 0.05 were selected as possible biomarkers of the investigated BPA exposure effects and considered for further analysis. These multiple comparison tests were performed using MATLAB Statistics and Machine Learning Toolbox™ and Benjamini-Hochberg FDR function described in (Groppe et al., 2011). Changes in metabolite areas among different groups of samples were also examined by hierarchical cluster analysis (HCA) (Schonlau, 2004) combined with the heatmap display of the auto-scaled metabolite concentrations in order to show groups (clusters) of metabolites with similar behavior after BPA exposure.

Only the metabolites that showed a statistically significant difference between means of groups were selected as potential biomarkers of BPA exposure and were tentatively identified. This identification was based on the comparison of the accurate molecular mass measured by LC-MS with the corresponding values found in online database resources, such as METLIN Metabolite Database (Smith et al., 2005), Human Metabolome Database (Wishart et al., 2007) and MetaCyc database (Caspi et al., 2014).

2.5.2. RNA sequencing data analysis

RNA-Seq reads were aligned to the *D. rerio* reference genome (GRCz10) using STAR version 2.5.1b (Dobin et al., 2012). Genes annotated in GRCz10.84 were quantified using RSEM version 1.2.28 (Li and Dewey, 2011) with default parameters. Differential expression analysis between control and 17.5 μM BPA treatment was performed with the DESeq2 (v.1.10.1) R package with the default Walt test option that uses a variant of scaling factor normalization based on the assumption that most genes are not differentially expressed (Anders and Huber, 2010; Love et al., 2014). This normalization is performed dividing each sample by the median of each gene scaling ratio. Variability between replica batches was taken into account in the design matrix. Genes with an adjusted *p*-value lower than 0.05 (after Walt test with Benjamini-Hochberg multiple comparison testing) were selected for further study.

2.5.3. Functional analysis of integrated metabolomic and transcriptomic data

Pathway-based integration of metabolomic and transcriptomic datasets was performed through metabolic pathways' knowledge defined by KEGG database. The joint list of identified markers from both metabolomic and transcriptomic data was introduced as input data into the KEGG database (*D. rerio* pathway dataset) to investigate common metabolic pathways and mechanisms affected by the BPA treatment. The analysis was limited to pathways related to metabolism (modules dre00001-dre01999). Identified pathways with at least two hits were included in the network analysis, using the reshape2 and igraph packages in R (RCoreTeam, 2014). Bipartite graphs were drawn from an incidence table of metabolites (represented as KEGG C-codes) versus genes (represented by their official gene names, ZFIN.org). Any given metabolite was considered linked to a given gene if they share at least one common KEGG pathway. The general pathway dre01100 (Metabolic pathways) was excluded from the analysis as it essentially included all tested biomarkers (see Table S1).

3. Results and discussion

3.1. Metabolomics of BPA exposure

Preliminary exploration of the TIC data of the 24 embryos extracts (six replicates per condition: control, 4.4 μM BPA, 8.8 μM BPA and 17.5 μM BPA) by PCA allowed detecting general patterns and differentiating the samples according to BPA treatment. Examples of LC-MS chromatograms belonging to the embryos extract in negative and positive ESI (ESI- and ESI+) modes are given in Fig. 1A and B, respectively. Chromatograms showed a complex profile with several coeluting compounds in the first 18 min. In ESI- data, the second component (PC2) separated the embryo samples in relation to the BPA exposure in the PCA scores plot (Fig. 1C), showing noticeable BPA-dose effects. In ESI+ data, PCA also highlights differences in samples in accordance with the BPA exposure when considering the first two components, with control, 4.4 μM and 8.8 μM BPA samples partially overlapped (Fig. 1D).

Non-targeted analysis of full scan metabolomic LC-MS raw data is challenging due to the large number of multiple coeluted peaks (Ortiz-Villanueva et al., 2015). To solve this problem and obtain more reliable metabolite profiling, the strategy based on the application of the MCR-ALS method was proposed (Ortiz-Villanueva et al., 2015). The simultaneous MCR-ALS analysis of the 24 embryo extracts, allowed a fast systematic and reliable resolution of the chromatographic profiles (elution) and mass spectra of most of the sample metabolites. MCR-ALS resolved profiles explained most of the experimental data variance (R^2 higher than 96%). In MCR-ALS, the number of components is used to allow modeling possible solvent and background contributions, which are resolved (subtracted) from the metabolite contributions and not evaluated in the next steps of the data analysis workflow. For instance, components giving elution profiles with unreasonable chromatographic peak shapes or noisy mass spectra were not considered in the following steps of the data analysis workflow. Every MCR-ALS resolved component was defined by a dyad of elution and mass spectrum profiles, which could be associated with one single metabolite. Several ion features derived from the same metabolite were grouped in the same single MCR-ALS component. These ion features usually correspond to the different isotopic signals of an ion with a particular nominal *m/z* value, due to the

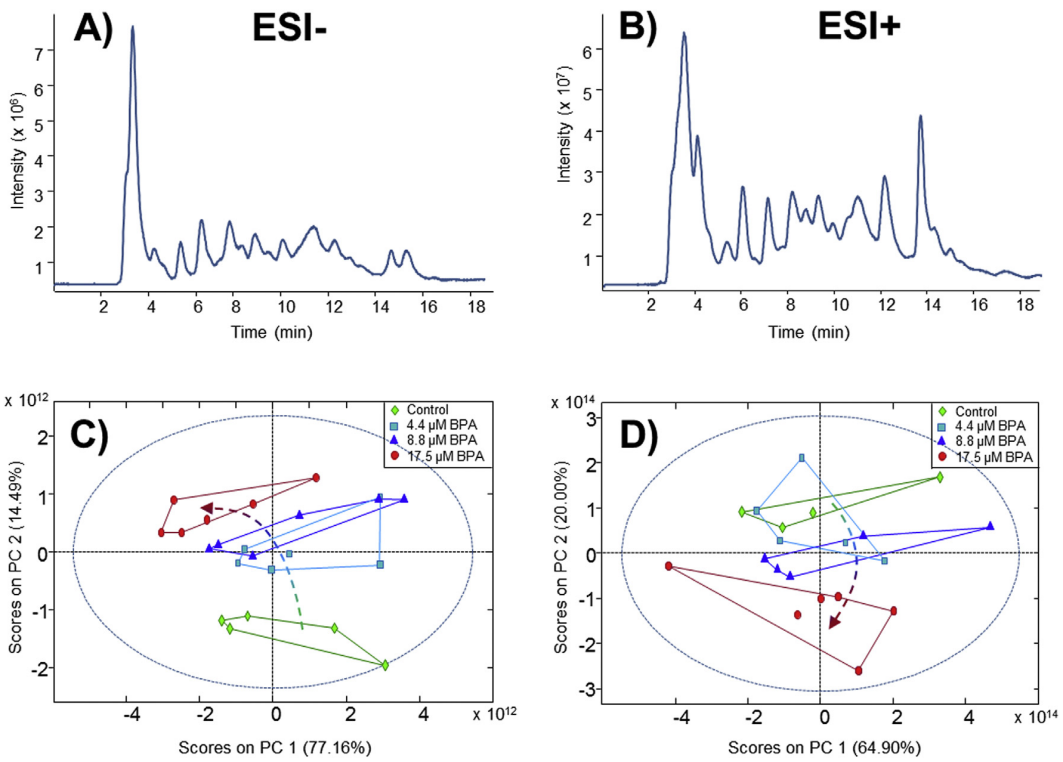


Fig. 1. Total ion chromatogram (TIC) of a control embryo extract analyzed by LC-MS using: (A) ESI- and (B) ESI+ mode. PCA scores plots of the TICs of the 24 embryo extracts: (C) in ESI- and (D) ESI+; green diamonds are control samples, blue squares are embryo samples exposed to BPA at 4.4 µM, purple triangles are samples exposed to BPA at 8.8 µM and red circles are samples exposed to BPA at 17.5 µM. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hydrogen cations/anions and other adducts. More details about MCR-ALS method are described in the Supplementary Material.

Statistical assessment of the metabolite changes based on a comparison of peak areas identified 29 metabolites from ESI- and 30 metabolites from ESI+ ionization modes as significant contributions. These metabolites were tentatively identified based on the comparison of their accurate experimental molecular mass values with the corresponding molecular mass values reported in different online databases (mass error of the possible molecular formulas $\leq$ 10 ppm). Tables 1 and 2 show the lists of tentatively identified metabolites, which differentiate between control and BPA-exposed samples in ESI- and ESI+ mode, respectively. Metabolite identity, detected ion, measured molecular mass, relative mass error, retention time, adjusted *p*-value (only when *p*-value $\leq$ 0.05), fold-change, folding trend and KEGG C-code are included in the tables. Nine metabolites were detected in both ESI- and ESI+ mode. Fold changes were calculated dividing the mean of the peak areas of the metabolites in each dose-exposed group by the mean of the peak areas of the same metabolites in the control group, and the trend was determined taking as a reference the response of control samples.

Hierarchical clustering analysis of the peak areas of these metabolites (biomarkers) grouped the samples that presented similar behavior according to the observed metabolite changes (heatmap in Fig. 2). Four groups of samples were outlined, correctly reflecting BPA-dose effects (controls, 4.4 µM BPA, 8.8 µM BPA and 17.5 µM BPA) and revealing slight variations between biological replicates possibly due to natural biological variation. In Fig. 2, the relative change in the metabolite concentrations is also depicted with colors in the heatmap. Up-regulation and down-regulation of metabolites are represented in red and blue, respectively. Hierarchical clustering of metabolites revealed two main clusters, one including metabolites whose concentrations increased upon BPA treatment (Cluster A in Fig. 2), and the second one including metabolites that become depleted by the treatment. The heatmap also suggests a monotonic response, as samples treated with the lowest BPA concentration (4.4 µM) grouped with the untreated ones (similar patterns), whereas samples corresponding to the two highest concentrations tested clustered together (see the dendrogram on the top of Fig. 2).

Functional analysis revealed that most metabolites affected by BPA were related to amino acids or nucleotide/nucleosides metabolic pathways (Fig. 2, Tables 1 and 2). Relevant changes in amino acid metabolism as a response to BPA treatment have been already reported in zebrafish embryos (Huang et al., 2016, 2017) and Daphnia (Nagato et al., 2016), suggesting that this is a rather general effect. The observed decrease on nucleotide-related metabolites by BPA treatment, particularly hypoxanthine and uric acid, may be interpreted as changes in signaling pathways. For example, an imbalance in uric acid production associated with hypoxanthine concentration changes has been related to oxidative stress in zebrafish embryos (Yoon et al., 2017). Purine metabolism, which was mainly down-regulated upon BPA treatment, may be linked to the observed increase in glutamate levels (see Fig. 2, Tables 1 and 2), an effect already observed in the fish metabolome upon BPA exposure (Yoon et al., 2017). In addition to these two major groups, we observed changes in biochemical pathways related to lipids, sugars and their derivatives, and to steroid metabolism, in agreement with previous transcriptomic (Boucher et al., 2016; Villeneuve et al., 2012) and metabolomic studies (Yoon et al.,

Table 1
List of tentatively identified metabolites statistically significant to differentiate between control and BPA exposed zebrafish embryos samples in ESI-.

	Compound	Molecular formula	Ion assignment	Measured molecular mass (Da)	[a]Mass error (ppm)	Fold-change ctrl vs. 4.4 μM	Fold-change ctrl vs. 8.8 μM	Fold-change ctrl vs. 17.5 μM	Trend	t_r (min)	[b]*p*-adj	[c]KEGG
Cluster A	L-Valine	C5H11NO2	$[M-H]^-$	116.0707	6.9	1.1	1.2	1.3	UP	12.5	1.34E-03	C00183
	Benzoic acid	C7H6O2	$[M-H]^-$	121.0288	5.8	1.0	1.6	1.5	UP	3.8	1.74E-08	C00180
	[d] L-Leucine	C6H13NO2	$[M-H]^-$	130.0868	4.6	1.1	1.2	1.1	UP	9.5	3.42E-02	C00123
	[d] L-Glutamate	C5H9NO4	$[M-H]^-$	146.0450	6.2	1.2	1.1	1.3	UP	11	3.82E-03	C00025
	L-Lyxonate	C5H10O6	$[M-H]^-$	165.0404	0.6	1.2	2.5	2.3	UP	3.9	2.64E-06	C05412
	2,4-Dichlorobenzoate	C7H4Cl2O2	$[M-H]^-$	188.9504	6.4	1.1	1.4	1.7	UP	3.5	1.28E-07	C06670
	[e] 10,20-Dihydroxyeicosanoic acid	C20H40O4	$[M-H+HAc]^-$	403.3052	3.2	299.8	307.8	414.4	UP	3.7	2.61E-10	-
	[e] Glyceryl 1-monostearate	C21H42O4	$[M-H+HAc]^-$	417.3217	1.7	139.0	128.6	141.7	UP	3.4	4.35E-13	-
Cluster B	L-Lactic acid	C3H6O3	$[M-H]^-$	89.0238	6.7	0.9	0.7	0.8	DOWN	5.3	1.47E-04	C00186
	Uracil	C4H4N2O2	$[M-H]^-$	111.0195	4.5	1.0	0.7	0.7	DOWN	6.1	4.99E-04	C00106
	[d] L-Proline	C5H9NO2	$[M-H]^-$	114.0556	4.4	0.9	0.8	0.9	DOWN	12.8	3.70E-02	C00148
	[d] L-Threonine	C4H9NO3	$[M-H]^-$	118.0501	7.6	0.9	0.7	0.8	DOWN	14.5	2.45E-02	C00188
	[d] Taurine	C2H7NO3S	$[M-H]^-$	124.0073	0.8	1.0	0.8	0.8	DOWN	11.3	1.34E-03	C00245
	[d] Creatine	C4H9N3O2	$[M-H]^-$	130.0618	3.1	1.0	0.8	0.8	DOWN	14.1	5.92E-03	C00300
	[d] Hypoxanthine	C5H4N4O	$[M-H]^-$	135.0305	5.2	0.9	0.7	0.7	DOWN	7.3	1.74E-05	C00262
	Quinolinate	C7H5NO4	$[M-H]^-$	166.0141	3.0	0.7	0.5	0.4	DOWN	4.4	7.61E-07	C03722
	Uric acid	C5H4N4O3	$[M-H]^-$	167.0205	3.6	0.9	0.9	0.8	DOWN	6.5	1.21E-03	C00366
	N-Acetyl-L-aspartic acid	C6H9NO5	$[M-H]^-$	174.0405	1.7	1.0	0.9	0.8	DOWN	6.2	3.09E-05	C01042
	Homomethionine	C6H13NO2S	$[M-H+HAc]^-$	222.0799	3.2	0.5	0.4	0.5	DOWN	3.5	7.98E-10	C17213
	p-Hydroxyphenylacetothiohydroximate	C8H9NO2S	$[M-H+HAc]^-$	242.0489	1.7	0.4	0.2	0.2	DOWN	3.5	4.83E-10	C17239
	S-(Hydroxyphenylacetothiohydroximoyl)-L-cysteine	C11H14N2O4S	$[M-H]^-$	269.0599	1.1	0.9	0.6	0.5	DOWN	3.6	4.48E-06	C17238
	[d] Linoleate	C18H32O2	$[M-H]^-$	279.2329	0.4	0.9	0.7	0.6	DOWN	3.4	5.92E-03	C01595
	Stearate	C18H36O2	$[M-H]^-$	283.2645	0.7	0.6	0.7	0.7	DOWN	3.2	3.09E-05	C01530
	[e] 6-Succinoaminopurine	C9H9N5O3	$[M-H+HAc]^-$	294.0833	3.7	0.9	0.8	0.8	DOWN	6.1	2.58E-02	-
	all-trans-Retinal	C20H28O	$[M-H+HAc]^-$	343.2288	2.6	1.1	0.9	0.6	DOWN	3.3	5.04E-03	C00376
	11-cis-Retinaldehyde	C20H28O	$[M-H+HAc]^-$	343.2288	2.6	1.1	0.9	0.6	DOWN	3.3	5.04E-03	C02110
	9-cis-Retinal	C20H28O	$[M-H+HAc]^-$	343.2288	2.6	1.1	0.9	0.6	DOWN	3.3	5.04E-03	C16681
	Prostaglandin G2	C20H32O6	$[M-H]^-$	367.2113	3.5	1.0	0.8	0.6	DOWN	3.8	4.95E-04	C05956
	6-Ketoprostaglandin E1	C20H32O6	$[M-H]^-$	367.2113	3.5	1.0	0.8	0.6	DOWN	3.8	4.95E-04	C05962
	2-Oxo-3-hydroxy-4-phosphobutanoic acid	C4H7O8P	$[2M-H]^-$	426.9688	0.9	0.9	0.7	0.6	DOWN	3	3.09E-05	C06054
	[d] 5β-Cyprinolsulfate	C27H48O8S	$[M-H]^-$	531.2977	3.8	1.2	0.7	0.6	DOWN	3.3	1.28E-07	C05468
	Uridine diphosphate glucuronic acid	C15H22N2O18P2	$[M-H]^-$	579.0273	0.5	0.8	0.7	0.4	DOWN	16.7	1.77E-06	C00167

[a] Mass accuracy = $[((\text{exact mass-measured mass})/\text{exact mass}) \times 10^6] \leq 10$ ppm.
[b] Obtained *p*-value using one-way ANOVA with Benjamini-Hochberg multiple testing correction.
[c] Metabolite KEGG C-code.
[d] Metabolites which were also detected in ESI+.
[e] Metabolites without KEGG C-code.

Table 2
List of tentatively identified metabolites statistically significant to differentiate between control and BPA exposed zebrafish embryos samples in ESI+.

	Compound	Molecular formula	Ion assignment	Measured molecular mass (Da)	[a]Mass error (ppm)	Fold-change ctrl vs. 4.4 μM	Fold-change ctrl vs. 8.8 μM	Fold-change ctrl vs. 17.5 μM	Trend	t_r (min)	[b]*p*-adj	[c]KEGG
Cluster A	[d] L-Leucine	C6H13NO2	$[M+H]^+$	132.1012	5.3	0.9	1.1	1.3	UP	9.5	3.02E-06	C00123
	[d] L-Glutamate	C5H9NO4	$[M+H]^+$	148.0606	1.4	1.2	1.1	1.2	UP	11.0	3.53E-02	C00025
	L-Phenylalanine	C9H11NO2	$[M+H]^+$	166.0870	4.2	0.9	1.4	1.3	UP	9.0	5.58E-07	C00079
	L-Tyrosine	C9H11NO3	$[M+H]^+$	182.0821	4.9	1.3	1.0	1.3	UP	10.8	4.96E-05	C00082
	D-Glucuronolactone	C6H8O6	$[M+NH_4]^+$	194.0656	1.5	106.6	122.7	137.6	UP	3.7	3.10E-10	C02670
	6-phospho D-glucono-1,5-lactone	C6H11O9P	$[M+H]^+$	259.0212	0.4	1.5	2.5	2.4	UP	3.6	2.21E-04	C01236
	Inosine	C10H12N4O5	$[M+H]^+$	269.0884	1.5	1.1	1.2	1.3	UP	8.4	2.37E-03	C00294
	LysoPC(15:0)	C23H48NO7P	$[M+H]^+$	482.3255	2.9	1.0	1.6	1.4	UP	4.1	1.36E-02	C04230
	LysoPC(22:6)	C30H50NO7P	$[M+H]^+$	568.3373	4.4	0.9	1.0	1.3	UP	4.2	2.18E-02	C04230
Cluster B	[d] L-Proline	C5H9NO2	$[M+H]^+$	116.0713	6.0	0.9	0.7	0.7	DOWN	12.8	3.70E-02	C00148
	[d] L-Threonine	C4H9NO3	$[M+H]^+$	120.0654	0.8	0.9	0.7	0.8	DOWN	14.5	1.55E-02	C00188
	[d] Taurine	C2H7NO3S	$[M+H]^+$	126.0224	4.0	0.9	0.8	0.7	DOWN	11.3	3.10E-02	C00245
	[d] Creatine	C4H9N3O2	$[M+H]^+$	132.0775	5.3	1.0	0.8	0.8	DOWN	14.1	6.59E-04	C00300
	[d] Hypoxanthine	C5H4N4O	$[M+H]^+$	137.0457	0.7	0.9	0.8	0.7	DOWN	7.3	3.62E-02	C00262
	1-Naphthylamine	C10H9N	$[M+H]^+$	144.0809	0.7	1.2	0.7	0.7	DOWN	8.6	2.21E-04	C14790
	4-Guanidinobutanoate	C5H11N3O2	$[M+H]^+$	146.0926	1.4	0.7	0.5	0.6	DOWN	14.6	6.20E-06	C01035
	L-Methionine	C5H11NO2S	$[M+H]^+$	150.0584	0.7	1.1	0.8	0.8	DOWN	10.7	1.20E-06	C00073
	Guanine	C5H5N5O	$[M+H]^+$	152.0563	2.6	1.1	0.6	0.7	DOWN	8.7	1.80E-04	C00242
	L-alpha-Aminoadipate	C6H11NO4	$[M+H]^+$	162.0763	1.2	0.9	0.8	0.6	DOWN	12.0	5.46E-03	C00956
	2-Amino-4-hydroxy-6-(hydroxymethyl)-7,8-dihydropteridine	C7H9N5O2	$[M+H]^+$	196.0832	1.5	1.3	0.8	0.4	DOWN	10.1	1.43E-05	C01300
	L-Histidinol phosphate	C6H12N3O4P	$[M+H]^+$	222.0635	1.4	0.6	0.7	0.7	DOWN	5.5	5.43E-03	C01100
	6-pyruvoyl tetrahydropterin	C9H11N5O3	$[M+H]^+$	238.0937	0.8	0.7	0.8	0.8	DOWN	7.2	2.76E-03	C03684
	6-Lactoyl-5,6,7,8-tetrahydropterin	C9H13N5O3	$[M+H]^+$	240.1093	0.8	1.2	0.7	0.4	DOWN	10.2	3.10E-05	C04244
	[d] Linoleate	C18H32O2	$[M+NH_4]^+$	298.2740	0.3	0.9	0.8	0.7	DOWN	3.4	1.29E-02	C01595
	[e] Glycerol 1-palmitate	C19H38O4	$[M+H]^+$	331.2820	6.9	0.7	0.7	0.4	DOWN	3.5	2.21E-04	-
	Inosine monophosphate (IMP)	C10H13N4O8P	$[M+H]^+$	349.0541	1.4	1.1	0.9	0.7	DOWN	10.2	8.60E-03	C00130
	Trehalose	C12H22O11	$[M+NH4]^+$	360.1487	3.6	1.1	0.9	0.7	DOWN	14.9	9.43E-03	C01083
	Dihydrofolate	C19H21N7O6	$[M+H\text{-}H_2O]^+$	426.1490	5.9	0.9	0.7	0.6	DOWN	3.4	6.20E-06	C00415
	Glycerylphosphorylethanolamine	C5H14NO6P	$[2M+H]^+$	431.1192	0.5	1.1	0.5	0.4	DOWN	7.9	6.10E-07	C01233
	[d] 5β-Cyprinolsulfate	C27H48O8S	$[M+NH_4]^+$	550.3409	0.2	1.3	0.9	0.7	DOWN	3.3	3.72E-04	C05468

[a] Mass accuracy = $[((\text{exact mass-measured mass})/\text{exact mass}) \times 10^6] \leq 10$ ppm.
[b] Obtained *p*-value using one-way ANOVA with Benjamini-Hochberg multiple testing correction.
[c] Metabolite KEGG C-code.
[d] Metabolites which were also detected in ESI-.
[e] Metabolites without KEGG C-code.

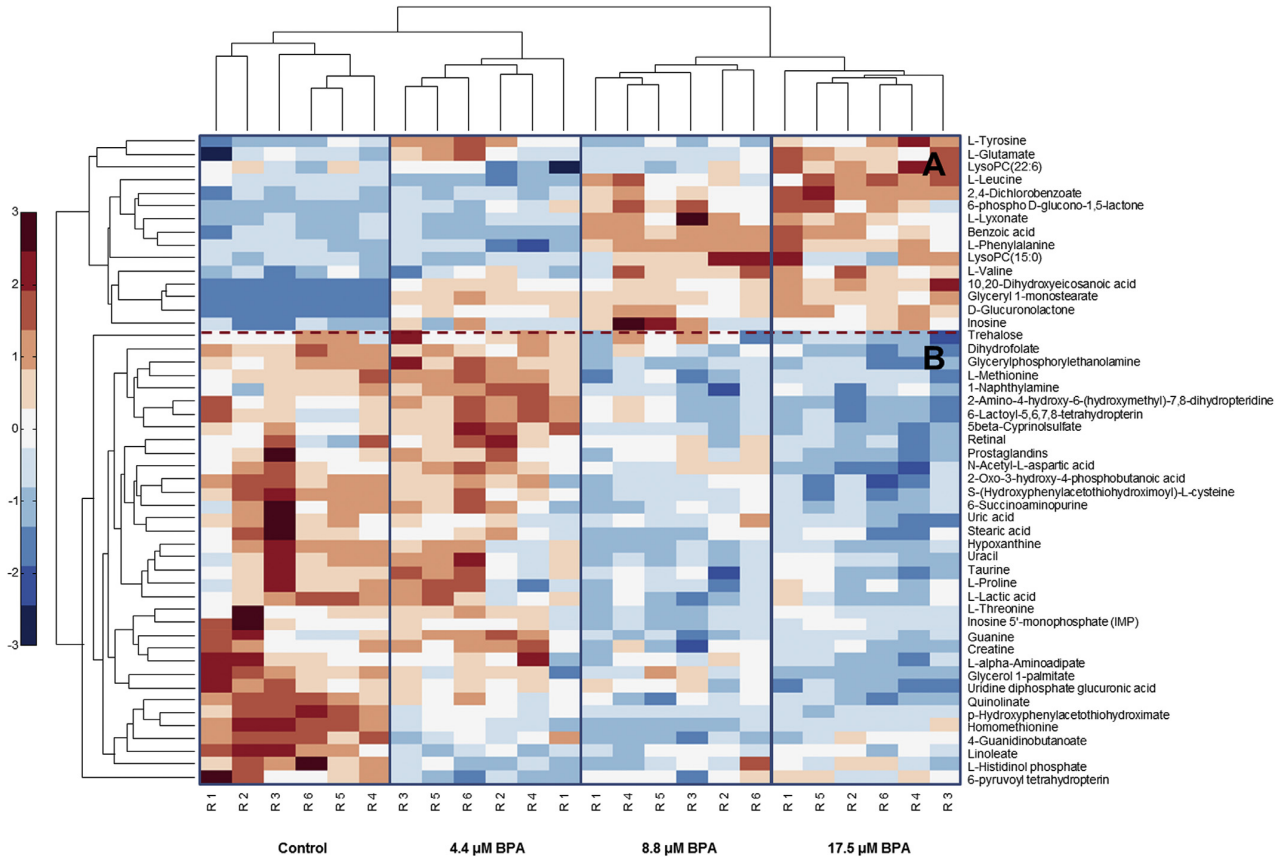


Fig. 2. Heatmap of the statistically significant tentatively identified metabolites whose concentrations changed between control and BPA-exposed samples (4.4, 8.8 and 17.5 µM): (A) up-regulated metabolites; (B) down-regulated metabolites. Cells colours represent autoscaled relative abundances of each metabolite, going from less (blue) to more (red). Colour codes are indicated in the color bar at the left side of the figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2017). Particularly, changes in glycerophospholipid metabolism suggest cell membrane damage by BPA (Huang et al., 2016; Yoon et al., 2017). Finally, and related to these primary cellular components, we observed changes in secondary metabolism, in pathways related to nicotinamide, linoleic and arachidonic acids, retinol metabolism, and cofactors and vitamins.

3.2. Transcriptomics of BPA exposure

Statistical assessment of changes in the mRNA levels between control and 17.5 µM BPA groups revealed 1381 differentially expressed genes (adjusted *p*-value <0.05), with 765 up-regulated and 616 down-regulated genes.

Hierarchical clustering of differentially expressed mRNA levels identified two main clusters according to the expression changes upon BPA exposure (A and B, Fig. 3). Cluster A comprises the expressed genes that were up-regulated by BPA exposure, whereas cluster B includes down-regulated genes. Genes with the highest differential expression (Top-twenty genes) are listed in Table 3. As it is shown, these top-twenty genes were related to response to different biological processes including estrogen stimulus and xenobiotics, oxidation-reduction process, metabolic process, fin regeneration, tissue development and cell proliferation.

Most of these altered biological processes by BPA were reported in previous literature. For instance, brain aromatase (*cyp19a1b*) is a well-known marker for estrogen exposure in fish embryos, and its increased expression in BPA-treated samples is consistent with the estrogenic activity of BPA in vertebrates (Chen et al., 2017). BPA alters cytochrome P450 enzymes activities, which play a crucial role in oxidative metabolism (Huang et al., 2016). BPA disruption on different cellular functions involved in cell proliferation and tissue development were also described (Lam et al., 2011, 2016); we consider that the observed alterations in nucleotide and amino acid synthesis pathways may be at least partially related to these effects. Furthermore, up-regulation of *gstp2* mRNA levels (see Table 3) may correlate to an increase in GST activity as a response to an increased oxidative stress and apoptosis (Hassan et al., 2012). In fact, changes in GST activity has been considered as a sign of contaminated aquatic ecosystems with BPA (Li et al., 2008).

3.3. Identification of altered pathways with metabolomic and transcriptomic pathway-based integration

Integration of metabolomic and transcriptomic datasets improves understanding of underlying biological processes to gain insight into a system and its mechanism of action. However, this pathway-level integration and corresponding biological interpretation of data is not trivial (Cavill et al., 2016). In this work, datasets of statistically significant affected genes and metabolites were evaluated using the existing biological knowledge through the metabolic pathways defined by KEGG database. KEGG pathway analysis of the aggregated metabolite and gene biomarkers (1431

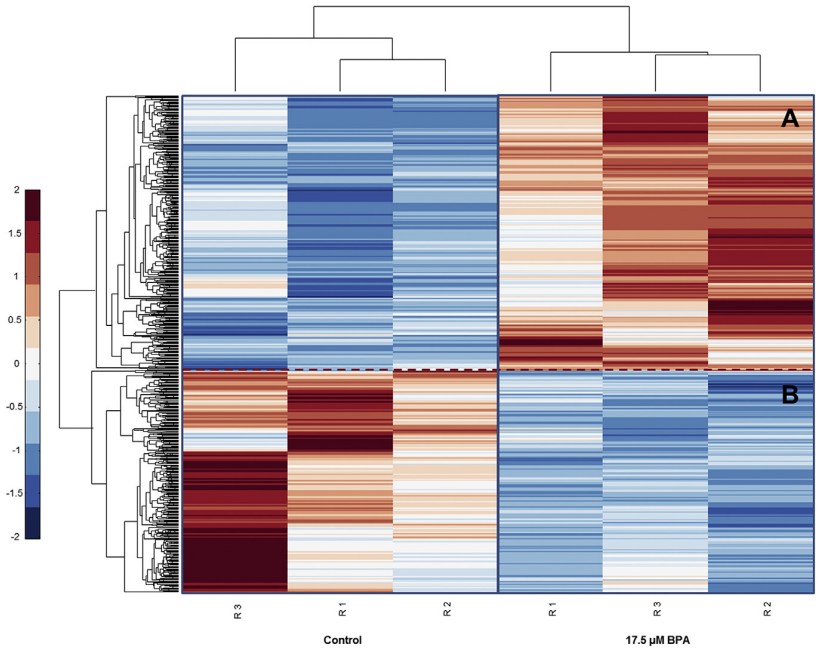


Fig. 3. Heatmap of the differentially expressed genes after BPA exposure (17.5 µM). Cells colours are coded as in Fig. 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different entries in total) identified 70 functional modules related to metabolism that included at least two biomarkers, 38 of them with more than 4 hits (Table S1). All of these 38 functional modules included at least two genes and only two of them did not contain at least one metabolite, one related to drug metabolism (*i.e.* exogenous compounds, dre00982) and a second one related to N-glycan biosynthesis (dre00510), a pathway for which few detectable metabolites should be expected (Table S1). This result could be considered as indicative that results derived from both metabolomic and transcriptomic data revealed additional information (Cavill et al., 2016), showing similar physiological effects that essentially affected the same set of metabolic pathways (dre01100 module). Among the less-represented pathways (four hits or less), those related to hydrophobic metabolites (lipids, steroids) or exogenous compounds were also mainly represented by genes. At this point, it is important to note that the experimental procedure for metabolite extraction did not favor the recovery of hydrophobic compounds; hence lipids and related metabolites were likely underrepresented in the LC-MS analysis. Further screening efforts would be required to cover this limitation and assess possible effects of BPA treatment on zebrafish embryos lipidome. Therefore, ascertainment of possible changes in lipid metabolism could be achieved maximizing the interpretation obtained at the transcriptome level. Nonetheless, non-targeted approaches used herein presented clear advantages regarding metabolite and gene coverage over targeted methods usually employed (Huang et al., 2017).

3.4. Biological relevance of observed metabolic disruption

After screening out those biomarkers only present in the dre01100 module, 48 metabolites and 119 genes were used to construct the bipartite graph shown in Fig. 4. The largest cluster of biomarkers corresponded to genes and metabolites related to amino acid metabolism (red oval in Fig. 4), whereas a second, densely populated cluster included biomarkers related to purine metabolism (brown oval in Fig. 4). Therefore, most metabolic changes were related to protein and amino acid synthesis, major components of the cells and likely related to proliferation/toxic effects of the BPA treatment (Lam et al., 2011, 2016). In addition, pathway-based integration analysis revealed different clusters or biomarkers related to vitamin/cofactor biosynthesis (folate, ascorbate), to signaling pathways (retinol, glutamate, taurine, prostaglandins), and to lipid metabolism (glycerophospholipids, stearate, linoleate). At least some of these clusters may be reflecting the disruption of signaling/regulatory functions of the fish embryos by BPA, an activity linked to its potential as endocrine disruptor (Chen and Reese, 2013; LaKind et al., 2014; Pelayo et al., 2012; Porreca et al., 2017; Rochester, 2013).

The physiological meaning of the observed changes in the levels of the different biomarkers can be inferred by detecting which genes or metabolites increased (red symbols) or decreased (blue symbols) upon treatment with BPA (Fig. 4). Most metabolites related to purine, retinol, folate and lipid metabolism decreased their concentrations upon BPA treatment, whereas several amino acid concentrations increased (Figs. 2 and 4). These variations can be correlated to the changes in mRNA levels of the genes included in the same metabolic pathways, although this correlation was not necessarily a direct one. Perturbation of aminoacyl-tRNA synthesis related to the increase in amino acid concentrations in BPA-treated embryos could be linked to the reduction of energy demand for maintaining homeostasis (Huang et al., 2017). Amino acid concentration changes affect several downstream dependent metabolic and biosynthesis pathways. Potentially affected pathway of folate (folate deficiency) on embryo development could be related to the blocking process of BPA effects on the organism, probably associated with depressive symptoms (Geng et al., 2015). Furthermore, dysregulation of linoleate has also been previously described that could be commonly related to several neurological disorders including depression (Huang et al., 2017).

Table 3
Biological processes and molecular functions of the most differentially expressed genes (top-twenty) after BPA exposure.

ENSEMBL ID	[a]ZFIN ID	Gene name	[b]baseMean	[c]lfc	[d]lfcSE	[e]*p*-adj	Biological Process	Molecular Function
Up-regulated								
ENSDARG00000098360	*cyp19a1b*	Cytochrome P450, family 19, subfamily A, polypeptide 1b	339.3	2.86	0.21	1.62E-36	Response to estrogen stimulus	Heme binding
ENSDARG00000045453	*f13a1a.1*	Coagulation factor XIII, A1 polypeptide a, tandem duplicate 1	64.9	1.84	0.22	1.01E-13	Peptide cross-linking	Protein-glutamine gamma-glutamyltransferase activity
ENSDARG00000104593	*cyp2k18*	Cytochrome P450, family 2, subfamily K, polypeptide 18	522.8	1.81	0.21	4.09E-15	Oxidation-reduction process	Heme binding
ENSDARG00000011573	*abcb11a*	ATP-binding cassette, sub-family B (MDR/TAP), member 11a	397	1.72	0.19	2.07E-15	Transmembrane transport	ATP binding
ENSDARG00000103019	*gstp2*	Glutathione S-transferase pi 2	555.3	1.68	0.21	5.90E-12	Response to toxic substance	Glutathione transferase activity
ENSDARG00000019838	*ugdh*	UDP-glucose 6-dehydrogenase	1294.3	1.68	0.19	6.25E-16	Atrioventricular valve formation	NAD binding
ENSDARG00000068493	*si:zfos-411a11.2*	Cytochrome family	552.6	1.68	0.18	9.51E-18	Oxidation-reduction process	Heme binding
ENSDARG00000087017	*ptgr1*	Prostaglandin reductase 1	258.2	1.6	0.2	7.41E-13	Oxidation-reduction process	Oxidoreductase activity
ENSDARG00000099525	*si:ch1073-13h15.3*	Unknown	1415.9	1.51	0.18	1.32E-13	Oxidation-reduction process	GDP-dissociation inhibitor activity
ENSDARG00000002394	*ugt5d1*	UDP glucuronosyltransferase 5 family, polypeptide D1	216.1	1.46	0.21	4.76E-09	Metabolic process	Glucuronosyltransferase activity
Down-regulated								
ENSDARG00000041433	*si:dkey-7c18.24*	Unknown	3990.1	−2.06	0.18	1.29E-26	Unknown	Unknown
ENSDARG00000094104	*si:ch211-213i16.3*	Linc RNA	683.2	−1.58	0.21	1.62E-11	Unknown	Unknown
ENSDARG00000094041	*krt17*	Keratin 17	80 473.8	−1.45	0.14	5.59E-22	Fin regeneration	Structural molecule activity
ENSDARG00000104359	*anxa1c*	Annexin a1c	3474.7	−1.43	0.21	3.57E-09	Unknown	Calcium ion binding
ENSDARG00000071216	*si:ch211-133n4.9*	Unknown	44.3	−1.25	0.22	2.54E-06	Hydrolase activity	Unknown
ENSDARG00000102057	*si:ch73-329n5.2*	Unknown	80.9	−1.23	0.22	3.74E-06	Cell-matrix adhesion	Unknown
ENSDARG00000093628	*s100a11*	S100 calcium binding protein A11	7171.6	−1.21	0.17	1.71E-09	Regulation of cell proliferation	Calcium ion binding
ENSDARG00000069293	*ahsg2*	Alpha-2-HS-glycoprotein 2	499.7	−1.19	0.19	1.07E-07	Negative regulation of biomineral tissue development	Cysteine-type endopeptidase inhibitor activity
ENSDARG00000076269	*zgc:172 131*	Unknown	61.9	−1.16	0.21	1.15E-05	Unknown	GTP binding
ENSDARG00000026726	*anxa1a*	Annexin a1a	12 680.2	−1.1	0.16	5.89E-09	Fin regeneration	Calcium ion binding

[a] ZFIN Gene official name.
[b] Mean of normalized counts of all samples.
[c] Log2 fold change.
[d] Log2 fold change standard error.
[e] Obtained *p*-value using Wald test with Benjamini-Hochberg multiple testing correction.

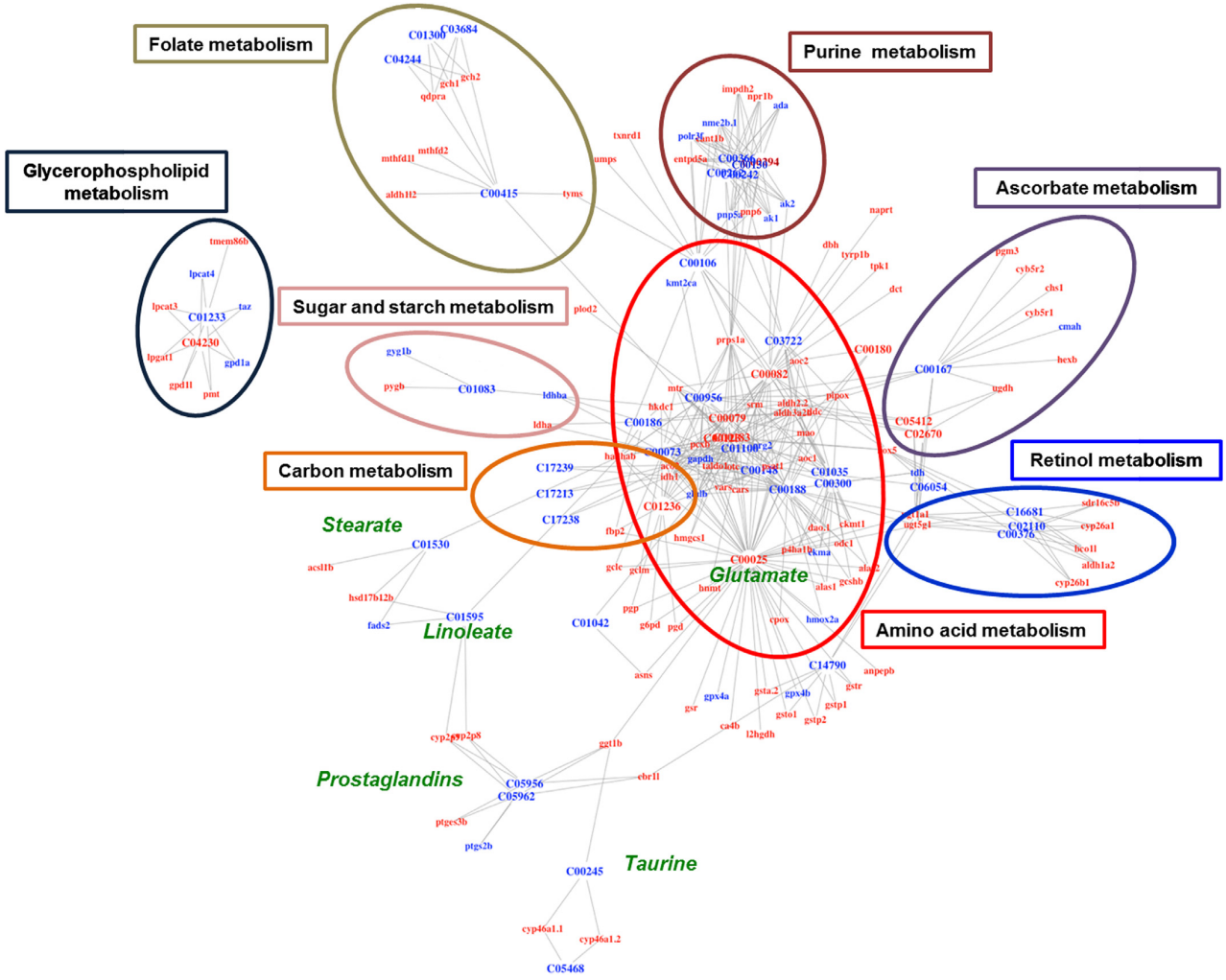


Fig. 4. Integrated network of functional interactions between metabolites (represented by their KEGG C-codes) and genes (represented by their ZFIN official names) whose levels were affected in zebrafish embryos by the BPA treatment. Metabolites and genes are connected if they share at least one common KEGG pathway, excluding the general dre01100 (Table S1). Increased or decreased metabolite or mRNA abundances (genes) in BPA-exposed samples are indicated by red and blue symbols, respectively. Standard names of some relevant metabolites are shown in green just beside their corresponding C-code. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

On the other hand, total retinoid levels (9-cis, 11-cis, and all-trans retinal) decreased upon BPA treatment, whereas up to 7 genes related to their metabolic pathway (*aldh1a2, bco1l, cyp26b1, si:ch1073-13h15.3, rdh12, ugt1a1, ugt5g1*) appear overexpressed (Fig. 4 and, in detail, in Fig. 5A). Previous studies have shown that a general increase of the retinoid degradation pathway can be observed upon zebrafish embryo exposures to physiological retinoids (Oliveira et al., 2013). In addition, a recent study carried out in mice indicates that retinoid acid enables transcriptional and post-transcriptional responses that are necessary for BPA biodegradation (Shmarakov et al., 2017). These results are consistent with a retinoid-mimicking effect of BPA, which in turn would trigger the feedback mechanisms that physiologically control the retinoid concentrations in the cells. The precise mode of action for this proposed retinoid-disrupting effect of BPA is still to be investigated but is well-known that retinoid levels exert relevant effects on embryonic development, cell growth, differentiation and apoptosis (Bushue and Wan, 2010).

Similarly, the decrease in the concentrations of the prostaglandins (either PGG_2 or 6-keto-PGE_1) in BPA-treated embryo samples can be explained by the observed reduction of mRNA abundance of their producing enzyme *ptgs2b*. This effect is combined with the increase in expression of up to 7 enzymes catalyzing alternative pathways either from the common precursor, arachidonic acid, or the intermediate metabolite PGH_2 (Fig. 4 and, in detail, in Fig. 5B). The disruption of the prostaglandin pathway in BPA-treated organisms could be related to the inflammatory response, but the meaning of this observation is still unclear. Only few studies on the effects of BPA on prostaglandins synthesis have been up to now, and all of them are referred to adult tissues (Rossitto et al., 2015). Although multiple roles of prostaglandins signaling pathway in reproduction have been described (Rossitto et al., 2015), the role of prostaglandins during early development is still unknown. BPA was shown to affect ovarian function via prostaglandin synthesis in mammalian granulosa cells (OVS et al., 2016), but it is difficult to find a relationship with our study using fish embryos. Our results suggest that further studies could be carried out to elucidate the mode of action of BPA on prostaglandin synthesis in early development.

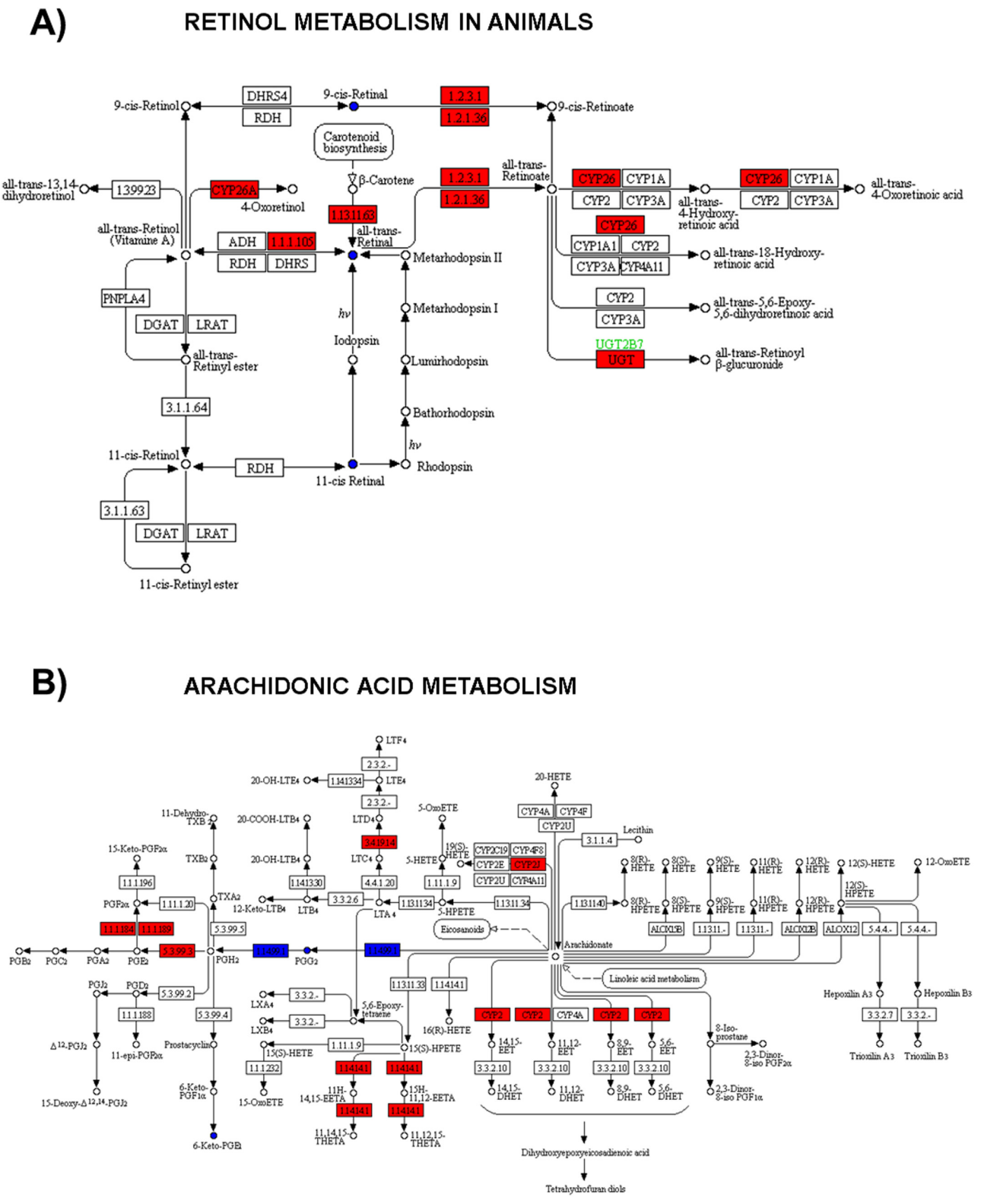


Fig. 5. Metabolic KEGG maps corresponding to (A) retinol and (B) arachidonic metabolism. Increased (red) or decreased (blue) metabolite concentrations are represented by dots, and enzymes/genes by boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Conclusions

Non-targeted LC-MS metabolomic and RNA-Seq transcriptomic analyses reflected a defined metabolic disruption of exposing zebrafish embryos to BPA. These non-targeted approaches allowed identifying a complete disruption of BPA compared to the usually used targeted techniques with limited profiling capability. The application of multivariate chemometric models and other data analysis tools allowed extracting relevant biochemical information from these datasets and obtain the phenotypic adverse effects of BPA in zebrafish embryos. Integration of both transcriptomic and metabolomic results at the pathway-level revealed the presence of affected metabolic routes in zebrafish embryos after BPA exposure. Our results suggested a significant effect on several signaling pathways, with possible indicators of metabolic disruption by BPA. This conclusion is in agreement with different reports indicating a rather pleiotropic toxic effect of BPA and BPA-related compounds, affecting estrogenic, androgenic, thyroid, retinoid and prostaglandin signaling pathways and cholesterol and lipid homeostasis (Boucher et al., 2016; Li et al., 2016b; Thompson et al., 2015; Wetherill et al., 2007). Several of these already recognized as BPA-affected pathways have also been directly (retinoid, estrogenic, lipids) or indirectly (arachidonic/prostaglandin metabolism) detected by our metabolomic or transcriptomic analyses, or by both. In addition, we observed effects in previously unexpected pathways, like the folate or ascorbate metabolism pathways, information that can help in further studies to understand the complexities of BPA toxicity. While clearly illustrating the utility of high-performance analytical procedures and the application of multivariate data analysis methods to interpret the complex experimental data, the conclusions of our work have also shown how metabolomic and transcriptomic results can be coupled to fully characterize the effects of toxicants beyond the simple indication of the affected pathways. For example, the interpretation of the changes in the retinoid metabolic pathway could only be done by combining results from both *-omic* analyses, as neither method allows predicting the activation of retinoids' degradation, rather than inhibition of their synthesis, by itself. Only a solid understanding of the organisms' metabolism will eventually allow the full interpretation of the observed changes, which is an absolute requirement to translate experimental data from laboratory assays to useful information for the protection of the environment and human health. Therefore, pathway-based integration of metabolomic and transcriptomic data offers evidence of BPA effects on the system highlighting the correlation between pathways behavior, despite the lack of observable phenotypic effects. Further analytical studies should be performed to correlate gene changes related to lipid metabolism upon BPA exposure, such as using LC-MS lipidomic analyses. Although it has been demonstrated the robustness of a pathway-based strategy for metabolomic and transcriptomic datasets integration, more efforts are pursued in multivariate-based integration (Cavill et al., 2016).

Acknowledgements

This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 320737. Some part of this study was also supported by a grant from the Spanish Ministry of Economy and Competitiveness (CTQ2014-56777-R). LNM was supported by a Beatriu de Pinos Postdoctoral Fellow (2013BP-B-00088) awarded by the Secretary for Universities and Research of the Ministry of Economy and Knowledge of the Government of Catalonia and the Cofund programme of the Marie Curie Actions of the 7th R&D Framework Programme of the European Union.

Conflict of interest statement

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envpol.2017.07.095.

References

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.
Arase, S., Ishii, K., Igarashi, K., Aisaki, K., Yoshio, Y., Matsushima, A., Shimohigashi, Y., Arima, K., Kanno, J., Sugimura, Y., 2011. Endocrine disrupter bisphenol a increases in situ estrogen production in the mouse urogenital sinus. Biol. Reproduction 84, 734–742.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.
Benskin, J.P., Ikonomou, M.G., Liu, J., Veldhoen, N., Dubetz, C., Helbing, C.C., Cosgrove, J.R., 2014. Distinctive metabolite profiles in in-migrating sockeye salmon suggest sex-linked endocrine perturbation. Environ. Sci. Technol. 48, 11670–11678.
Berghmans, S., Butler, P., Goldsmith, P., Waldron, G., Gardner, I., Golder, Z., Richards, F.M., Kimber, G., Roach, A., Alderton, W., Fleming, A., 2008. Zebrafish based assays for the assessment of cardiac, visual and gut function - potential safety screens for early drug discovery. J. Pharmacol. Toxicol. Methods 58, 59–68.
Boucher, J.G., Gagné, R., Rowan-Carroll, A., Boudreau, A., Yauk, C.L., Atlas, E., 2016. Bisphenol a and bisphenol S induce distinct transcriptional profiles in differentiating human primary preadipocytes. PLoS ONE 11, e0163318.
Bushue, N., Wan, Y.-J.Y., 2010. Retinoid pathway and cancer therapeutics. Adv. Drug Deliv. Rev. 62, 1285–1298.
Canesi, L., Fabbri, E., 2015. Environmental effects of BPA: focus on aquatic species. Dose-Response 13, 1559325815598304.
Careghini, A., Mastorgio, A.F., Saponaro, S., Sezenna, E., 2014. Bisphenol A, nonylphenols, benzophenones, and benzotriazoles in soils, groundwater, surface water, sediments, and food: a review. Environ. Sci. Pollut. Res. 22, 5711–5741.
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., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42, D459–D471.
Cavill, R., Jennen, D., Kleinjans, J., Briedé, J.J., 2016. Transcriptomic and metabolomic data integration. Briefings Bioinforma. 17, 891–901.
Chen, Y., Reese, D.H., 2013. A screen for disruptors of the retinol (vitamin a) signaling pathway. Birth Defects Res. Part B Dev. Reproductive Toxicol. 98, 276–282.
Chen, X., Hang, X., Ke, W., Ma, Z., Sun, Y., 2012. Acute and subacute toxicity of Bisphenol A on zebrafish (Danio rerio). Adv. Mater. Res. 356–360, 138–141.
Chen, M., Zhou, K., Chen, X., Qiao, S., Hu, Y., Xu, B., Xu, B., Han, X., Tang, R., Mao, Z., Dong, C., Wu, D., Wang, Y., Wang, S., Zhou, Z., Xia, Y., Wang, X., 2014. Metabolomic analysis reveals metabolic changes caused by bisphenol a in rats. Toxicol. Sci. 138, 256–267.
Chen, J., Xiao, Y., Gai, Z., Li, R., Zhu, Z., Bai, C., Tanguay, R.L., Xu, X., Huang, C., Dong, Q., 2015. Reproductive toxicity of low level bisphenol A exposures in a two-generation zebrafish assay: evidence of male-specific effects. Aquat. Toxicol. 169, 204–214.
Chen, J., Saili, K.S., Liu, Y., Li, L., Zhao, Y., Jia, Y., Bai, C., Tanguay, R.L., Dong, Q., Huang, C., 2017. Developmental bisphenol A exposure impairs sperm function and reproduction in zebrafish. Chemosphere 169, 262–270.
Choi, B.-I., Harvey, A.J., Green, M.P., 2016. Bisphenol A affects early bovine embryo development and metabolism that is negated by an oestrogen receptor inhibitor. Sci. Rep. 6, 29318.
Crain, D.A., Eriksen, M., Iguchi, T., Jobling, S., Laufer, H., LeBlanc, G.A., Guillette Jr., L.J., 2007. An ecological assessment of bisphenol-A: evidence from comparative biology. Reprod. Toxicol. 24, 225–239.
Directive 2010/63/EU, 2010. Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official J. Eur. Union L 276, 33–79, 20 October 2010.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.
Ekman, D.R., Hartig, P.C., Cardon, M., Skelton, D.M., Teng, Q., Durhan, E.J., Jensen, K.M., Kahl, M.D., Villeneuve, D.L., Gray, L.E., Collette, T.W., Ankley, G.T., 2012. Metabolite profiling and a transcriptional activation assay provide direct evidence of androgen receptor antagonism by bisphenol a in fish. Environ. Sci. Technol. 46, 9673–9680.
European Food Safety, A, 2016. Overview of existing methodologies for the

estimation of non-dietary exposure to chemicals from the use of consumer products and via the environment. EFSA J. 14, e04525 n/a.
Farrés, M., Piña, B., Tauler, R., 2015. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS. Metabolomics 11, 210–224.
Flint, S., Markle, T., Thompson, S., Wallace, E., 2012. Bisphenol A exposure, effects, and policy: a wildlife perspective. J. Environ. Manag. 104, 19–34.
Geetharathan, T., Josthna, P., 2016. Effect of BPA on protein, lipid profile and immuno-histo chemical changes in placenta and uterine tissues of albino rat. Int. J. Pharm. Clin. Res. 8, 260–268.
Geng, Y., Gao, R., Chen, X., Liu, X., Liao, X., Li, Y., Liu, S., Ding, Y., Wang, Y., He, J., 2015. Folate deficiency impairs decidualization and alters methylation patterns of the genome in mice. MHR Basic Sci. reproductive Med. 21, 844–856.
Gligorijević, V., Malod-Dognin, N., Pržulj, N., 2016. Integrative methods for analyzing big data in precision medicine. Proteomics 16, 741–758.
Groppe, D.M., Urbach, T.P., Kutas, M., 2011. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48, 1711–1725.
Hassan, Z.K., Elobeid, M.A., Virk, P., Omer, S.A., ElAmin, M., Daghestani, M.H., AlOlayan, E.M., 2012. Bisphenol a induces hepatotoxicity through oxidative stress in rat model. Oxidative Med. Cell. Longev. 2012, 194829.
Huang, Y.Q., Wong, C.K.C., Zheng, J.S., Bouwman, H., Barra, R., Wahlström, B., Neretin, L., Wong, M.H., 2012. Bisphenol A (BPA) in China: a review of sources, environmental levels, and potential human health impacts. Environ. Int. 42, 91–99.
Huang, S.S.Y., Benskin, J.P., Chandramouli, B., Butler, H., Helbing, C.C., Cosgrove, J.R., 2016. Xenobiotics produce distinct metabolomic responses in zebrafish larvae (Danio rerio). Environ. Sci. Technol. 50, 6526–6535.
Huang, S.S.Y., Benskin, J.P., Veldhoen, N., Chandramouli, B., Butler, H., Helbing, C.C., Cosgrove, J.R., 2017. A multi-omic approach to elucidate low-dose effects of xenobiotics in zebrafish (Danio rerio) larvae. Aquat. Toxicol. 182, 102–112.
Jaumot, J., de Juan, A., Tauler, R., 2015. MCR-ALS GUI 2.0: new features and applications. Chemom. Intelligent Laboratory Syst. 140, 1–12.
Javurek, A.B., Spollen, W.G., Johnson, S.A., Bivens, N.J., Bromert, K.H., Givan, S.A., Rosenfeld, C.S., 2016. Effects of exposure to bisphenol A and ethinyl estradiol on the gut microbiota of parents and their offspring in a rodent model. Gut Microbes 7, 471–485.
Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016. Compounds altering fat storage in Daphnia magna. Sci. Total Environ. 545, 127–136.
Kang, J.-H., Kondo, F., Katayama, Y., 2006. Human exposure to bisphenol A. Toxicology 226, 79–89.
Kang, J.-H., Aasi, D., Katayama, Y., 2007. Bisphenol a in the aquatic environment and its endocrine-disruptive effects on aquatic organisms. Crit. Rev. Toxicol. 37, 607–625.
Katsiadaki, I., Williams, T.D., Ball, J.S., Bean, T.P., Sanders, M.B., Wu, H., Santos, E.M., Brown, M.M., Baker, P., Ortega, F., Falciani, F., Craft, J.A., Tyler, C.R., Viant, M.R., Chipman, J.K., 2010. Hepatic transcriptomic and metabolomic responses in the Stickleback (Gasterosteus aculeatus) exposed to ethinyl-estradiol. Aquat. Toxicol. 97, 174–187.
Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F., 1995. Stages of embryonic development of the zebrafish. Dev. Dyn. 203, 253–310.
Kolpin, D.W., Furlong, E.T., Meyer, M.T., Thurman, E.M., Zaugg, S.D., Barber, L.B., Buxton, H.T., 2002. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. Streams, 1999–2000: a national reconnaissance. Environ. Sci. Technol. 36, 1202–1211.
Laing, L.V., Viana, J., Dempster, E.L., Trznadel, M., Trunkfield, L.A., Uren Webster, T.M., van Aerle, R., Paull, G.C., Wilson, R.J., Mill, J., Santos, E.M., 2016. Bisphenol A causes reproductive toxicity, decreases dnmt1 transcription, and reduces global DNA methylation in breeding zebrafish (Danio rerio). Epigenetics 11, 526–538.
LaKind, J.S., Goodman, M., Mattison, D.R., 2014. Bisphenol A and indicators of obesity, glucose metabolism/type 2 diabetes and cardiovascular disease: a systematic review of epidemiologic research. Crit. Rev. Toxicol. 44, 121–150.
Lam, S.H., Hlaing, M.M., Zhang, X., Yan, C., Duan, Z., Zhu, L., Ung, C.Y., Mathavan, S., Ong, C.N., Gong, Z., 2011. Toxicogenomic and phenotypic analyses of bisphenol-a early-life exposure toxicity in zebrafish. PLoS ONE 6, e28273.
Lam, H.-M., Ho, S.-M., Chen, J., Medvedovic, M., Tam, N.N.C., 2016. Bisphenol a disrupts HNF4α-regulated gene networks linking to prostate preneoplasia and immune disruption in noble rats. Endocrinology 157, 207–219.
Lee, H.J., Chattopadhyay, S., Gong, E.-Y., Ahn, R.S., Lee, K., 2003. Antiandrogenic effects of bisphenol a and nonylphenol on the function of androgen receptor. Toxicol. Sci. 75, 40–46.
Lejonklou, M.H., Dunder, L., Bladin, E., Pettersson, V., Ronn, M., Lind, L., Walden, T.B., Lind, P.M., 2017. Effects of low-dose developmental bisphenol a exposure on metabolic parameters and gene expression in male and female fischer 344 rat offspring. Environ. Health Perspect. 125, 067018.
Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma. 12, 323.
Li, X., Lin, L., Luan, T., Yang, L., Lan, C., 2008. Effects of landfill leachate effluent and bisphenol A on glutathione and glutathione-related enzymes in the gills and digestive glands of the freshwater snail Bellamya purificata. Chemosphere 70, 1903–1909.
Li, S., Jin, Y., Wang, J., Tang, Z., Xu, S., Wang, T., Cai, Z., 2016. Urinary profiling of cis-diol-containing metabolites in rats with bisphenol A exposure by liquid chromatography-mass spectrometry and isotope labeling. Analyst 141, 1144–1153.
Li, N., Jiang, W., Ma, M., Wang, D., Wang, Z., 2016. Chlorination by-products of bisphenol A enhanced retinoid X receptor disrupting effects. J. Hazard. Mater. 320, 289–295.
Lindholst, C., Pedersen, K.L., Pedersen, S.N., 2000. Estrogenic response of bisphenol A in rainbow trout (Oncorhynchus mykiss). Aquat. Toxicol. 48, 87–94.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.
Mushtaq, M.Y., Verpoorte, R., Kim, H.K., 2013. Zebrafish as a model for systems biology. Biotechnol. Genet. Eng. Rev. 29, 187–205.
Nagato, E.G., Simpson, A.J., Simpson, M.J., 2016. Metabolomics reveals energetic impairments in Daphnia magna exposed to diazinon, malathion and bisphenol-A. Aquat. Toxicol. 170, 175–186.
Navarro-Reig, M., Jaumot, J., García-Reiriz, A., Tauler, R., 2015. 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. 407, 8835–8847.
Olivares, A., van Drooge, B.L., Casado, M., Prats, E., Serra, M., van der Ven, L.T., Kamstra, J.H., Hamers, T., Hermsen, S., Grimalt, J.O., Piña, B., 2013. Developmental effects of aerosols and coal burning particles in zebrafish embryos. Environ. Pollut. 178, 72–79.
Oliveira, E., Casado, M., Raldúa, D., Soares, A., Barata, C., Piña, B., 2013. Retinoic acid receptors' expression and function during zebrafish early development. J. steroid Biochem. Mol. Biol. 138, 143–151.
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.
OVS, M., KHN, A., MAM, I., Kodithuwakku, L.S., 2016. Endocrine Disruptor Bisphenol-A (BPA) alters the Prostaglandins Synthesis Cascade Enzyme gene expression in Porcine Granulosa Cells in vitro, Wayamba University Sri Lanka International Conference 2016.
Pelayo, S., Oliveira, E., Thienpont, B., Babin, P.J., Raldúa, D., André, M., Piña, B., 2012. Triiodothyronine-induced changes in the zebrafish transcriptome during the eleutheroembryonic stage: implications for bisphenol A developmental toxicity. Aquat. Toxicol. 110–111, 114–122.
Porreca, I., Ulloa-Severino, L., Almeida, P., Cuomo, D., Nardone, A., Falco, G., Mallardo, M., Ambrosino, C., 2017. Molecular targets of developmental exposure to bisphenol A in diabesity: a focus on endoderm-derived organs. Obes. Rev. 18, 99–108.
Qiu, W., Zhao, Y., Yang, M., Farajzadeh, M., Pan, C., Wayne, N.L., 2015. Actions of bisphenol a and bisphenol S on the reproductive neuroendocrine system during early development in zebrafish. Endocrinology 157, 636–647.
Rankouhi, T.R., van Holsteijn, I., Letcher, R., Giesy, J.P., van den Berg, M., 2002. Effects of primary exposure to environmental and natural estrogens on vitellogenin production in carp (Cyprinus carpio) hepatocytes. Toxicol. Sci. 67, 75–80.
RCoreTeam, 2014. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rochester, J.R., 2013. Bisphenol A and human health: a review of the literature. Reprod. Toxicol. 42, 132–155.
Rossitto, M., Ujjan, S., Poulat, F., Boizet-Bonhoure, B., 2015. Multiple roles of the prostaglandin D2 signaling pathway in reproduction. Reproduction 149, R49–R58.
Santangeli, S., Maradonna, F., Gioacchini, G., Cobellis, G., Piccinetti, C.C., Dalla Valle, L., Carnevali, O., 2016. BPA-induced deregulation of epigenetic patterns: effects on female zebrafish reproduction. Sci. Rep. 6, 21982.
Santos, E.M., Ball, J.S., Williams, T.D., Wu, H., Ortega, F., van Aerle, R., Katsiadaki, I., Falciani, F., Viant, M.R., Chipman, J.K., Tyler, C.R., 2010. Identifying health impacts of exposure to copper using transcriptomics and metabolomics in a fish model. Environ. Sci. Technol. 44, 820–826.
Schonlau, M., 2004. Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Comput. Stat. 19, 95–111.
Shmarakov, I.O., Borschovetska, V.L., Blaner, W.S., 2017. Hepatic detoxification of bisphenol a is retinoid-dependent. Toxicol. Sci. 157, 141–155.
Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751.
Soanes, K.H., Achenbach, J.C., Burton, I.W., Hui, J.P.M., Penny, S.L., Karakach, T.K., 2011. Molecular characterization of zebrafish embryogenesis via DNA microarrays and multiplatform time course metabolomics studies. J. Proteome Res. 10, 5102–5117.
Song, Q., Chen, H., Li, Y., Zhou, H., Han, Q., Diao, X., 2016. Toxicological effects of benzo(a)pyrene, DDT and their mixture on the green mussel Perna viridis revealed by proteomic and metabolomic approaches. Chemosphere 144, 214–224.
Stiefel, F., Paul, A.J., Jacopo, T., Sgueglia, A., Stützle, M., Herold, E.M., Hesse, F., 2016. The influence of bisphenol A on mammalian cell cultivation. Appl. Microbiol. Biotechnol. 100, 113–124.
Thompson, P.A., Khatami, M., Baglole, C.J., Sun, J., Harris, S., Moon, E.-Y., Al-Mulla, F., Al-Temaimi, R., Brown, D., Colacci, A., Mondello, C., Raju, J., Ryan, E., Woodrick, J., Scovassi, I., Singh, N., Vaccari, M., Roy, R., Forte, S., Memeo, L., Salem, H.K., Amedei, A., Hamid, R.A., Lowe, L., Guarnieri, T., Bisson, W.H., 2015. Environmental immune disruptors, inflammation and cancer risk. Carcinogenesis 36, S232–S253.
Tomasi, G., van den Berg, F., Andersson, C., 2004. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. J. Chemom. 18, 231–241.

Tremblay-Franco, M., Cabaton, N.J., Canlet, C., Gautier, R., Schaeberle, C.M., Jourdan, F., Sonnenschein, C., Vinson, F., Soto, A.M., Zalko, D., 2015. Dynamic metabolic disruption in rats perinatally exposed to low doses of bisphenol-a. PLoS ONE 10, e0141698.

Villeneuve, D.L., Garcia-Reyero, N., Escalon, B.L., Jensen, K.M., Cavallin, J.E., Makynen, E.A., Durhan, E.J., Kahl, M.D., Thomas, L.M., Perkins, E.J., Ankley, G.T., 2012. Ecotoxicogenomics to support ecological risk assessment: a case study with bisphenol a in fish. Environ. Sci. Technol. 46, 51–59.

Wang, S., Wang, L., Hua, W., Zhou, M., Wang, Q., Zhou, Q., Huang, X., 2015. Effects of bisphenol A, an environmental endocrine disruptor, on the endogenous hormones of plants. Environ. Sci. Pollut. Res. 22, 17653–17662.

Wetherill, Y.B., Akingbemi, B.T., Kanno, J., McLachlan, J.A., Nadal, A., Sonnenschein, C., Watson, C.S., Zoeller, R.T., Belcher, S.M., 2007. In vitro molecular mechanisms of bisphenol A action. Reprod. Toxicol. 24, 178–198.

Williams, T.D., Wu, H., Santos, E.M., Ball, J., Katsiadaki, I., Brown, M.M., Baker, P., Ortega, F., Falciani, F., Craft, J.A., Tyler, C.R., Chipman, J.K., Viant, M.R., 2009. Hepatic transcriptomic and metabolomic responses in the stickleback (Gasterosteus aculeatus) exposed to environmentally relevant concentrations of dibenzanthracene. Environ. Sci. Technol. 43, 6341–6348.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., MacInnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., Querengesser, L., 2007. HMDB: the human metabolome database. Nucleic Acids Res. 35, D521–D526.

Xu, H., Yang, M., Qiu, W., Pan, C., Wu, M., 2013. The impact of endocrine-disrupting chemicals on oxidative stress and innate immune response in zebrafish embryos. Environ. Toxicol. Chem. 32, 1793–1799.

Yang, B., Guan, Q., Tian, J., Komatsu, S., 2017. Transcriptomic and proteomic analyses of leaves from Clematis terniflora DC. under high level of ultraviolet-B irradiation followed by dark treatment. J. Proteomics 150, 323–340.

Yoon, C., Yoon, D., Cho, J., Kim, S., Lee, H., Choi, H., Kim, S., 2017. 1H-NMR-based metabolomic studies of bisphenol A in zebrafish (Danio rerio). J. Environ. Sci. Health, Part B 52, 282–289.

Informació suplementària a l'article científic V.

Metabolic disruption of zebrafish (*Danio rerio*) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach.

E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler. *Environmental Pollution* 231 (2017) 22-36.

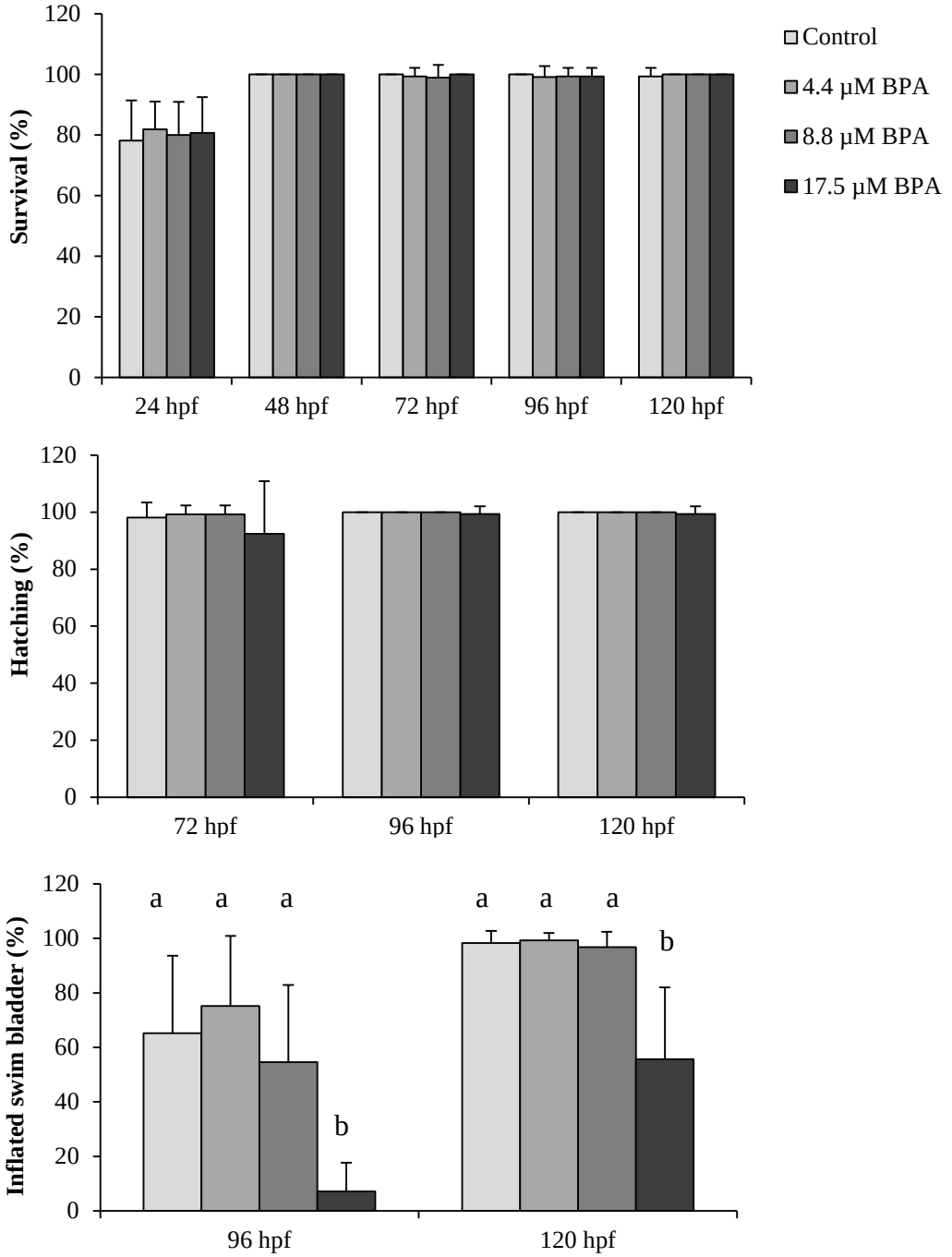


Figure S1. Development effects of BPA on zebrafish embryos: survival (%), hatching (%) and inflated swim bladder (%). Means (bars) + standard deviation (whiskers). Data from each stage have been analyzed by the non-parametric Kruskal–Wallis test with pairwise multiple comparisons and significances at $p < 0.05$. Letters at the bottom figure indicate similar (a) and different (b) behavior between treatments at each stage.

Table S1. Pathway analysis (KEGG) for metabolites and genes significantly affected by BPA treatment [a].

Pathway	Description	[b]Metabolites/[c]Genes
dre01100	Metabolic pathways	C00025, C00073, C00079, C00082, C00106, C00123, C00130, C00148, C00167, C00180, C00183, C00186, C00188, C00242, C00245, C00262, C00294, C00300, C00366, C00376, C00415, C00956, C01035, C01042, C01083, C01100, C01233, C01236, C01300, C01595, C02110, C02670, C03684, C03722, C04244, C05956, C05962, C06054, C06670, *aco2, acsl1b, ada, ak1, ak2, alas1, alas2, aldh1a2, aldh2.2, aldh3a2b, alg2, anpepb, aoc1, aoc2, aox5, arg2, asns, atp5c1, atp5g3b, b3gat3, b4galnt3b, bco1l, cbr1l, ckma, ckmt1, cox6b1, cox6b2, cpox, cyp19a1b, cyp26a1, cyp26b1, cyp2p8, cyp2p9, dad1, dao.1, dbh, dct, ddc, ddost, degs2, ebp, fbp2, g6pd, galnt8a.1, gapdh, gba2, gch1, gch2, gclc, gclm, gcshb, ggt1b, glulb, gyg1b, hadhab, hexb, hkdc1, hmgcs1, hsd17b12b, idh1, impdh2, inpp5jb, ipmka, ldha, ldhba, lpcat4, mao, mogs, mthfd1l, mthfd2, mtr, naprt, nme2b.1, odc1, otc, p4ha1b, pcxb, pgd, pgm3, pgp, pigf, pigs, pipox, pnp5a, pnp6, polr3f,*
dre01230	Biosynthesis of amino acids	C00025, C00073, C00079, C00082, C00123, C00148, C00183, C00188, C00956, C01100, *aco2, arg2, gapdh, glulb, idh1, mtr, otc, pcxb, prps1a, psat1, taldo1*
dre00480	Glutathione metabolism	C00025, *anpepb, g6pd, gclc, gclm, ggt1b, gpx4a, gpx4b, gsr, gsta.2, gsto1, gstp1, gstp2, gstr, idh1, odc1, pgd, srm*
dre00230	Purine metabolism	C00130, C00242, C00262, C00294, C00366, *ada, ak1, ak2, cant1b, entpd5a, impdh2, nme2b.1, npr1b, pnp5a, pnp6, polr3f,*
dre01200	Carbon metabolism	C00025, C01236, *aco2, fbp2, g6pd, gapdh, hadhab, hkdc1, idh1, pcxb, pgd, pgp, prps1a, psat1, taldo1*
dre00330	Arginine and proline metabolism	C00025, C00148, C00300, C01035, *aldh2.2, aldh3a2b, aoc1, arg2, ckma, ckmt1, dao.1, mao, odc1, p4ha1b, srm*
dre01210	2-Oxocarboxylic acid metabolism	C00025, C00073, C00079, C00082, C00123, C00183, C00956, C17213, C17238, C17239, *aco2, idh1*
dre00830	Retinol metabolism	C00376, C02110, C16681, *aldh1a2, aox5, bco1l, cyp26a1, cyp26b1, sdr16c5b, ugt1a1, ugt5g1*
dre00260	Glycine, serine and threonine metabolism	C00188, C00300, *alas1, alas2, aoc2, dao.1, gcshb, mao, pipox, psat1, tdh*
dre00240	Pyrimidine metabolism	C00106, *cant1b, entpd5a, nme2b.1, pnp5a, pnp6, polr3f, txnrd1, tyms, umps*
dre00970	Aminoacyl-tRNA biosynthesis	C00025, C00073, C00079, C00082, C00123, C00148, C00183, C00188, *cars, vars*
dre00980	Metabolism of xenobiotics by cytochrome P450	C14790, *cbr1l, gsta.2, gsto1, gstp1, gstp2, gstr, ugt1a1, ugt5g1*
dre00982	Drug metabolism	*aox5, gsta.2, gsto1, gstp1, gstp2, gstr, mao, ugt1a1, ugt5g1*
dre00564	Glycerophospholipid metabolism	C01233, C04230, *gpd1a, gpd1l, lpcat3, lpcat4, lpgat1, pmt, taz*
dre00520	Amino sugar and nucleotide sugar metabolism	C00167, *chs1, cmah, cyb5r1, cyb5r2, hexb, hkdc1, pgm3, ugdh*
dre00010	Glycolysis / Gluconeogenesis	C00186, *aldh2.2, aldh3a2b, fbp2, gapdh, hkdc1, ldha, ldhba*
dre00340	Histidine metabolism	C00025, *C01100, aldh2.2, aldh3a2b, aoc1, ddc, hnmt, mao*
dre00380	Tryptophan metabolism	C03722, *aldh2.2, aldh3a2b, aoc1, aox5, ddc, hadhab, mao*
dre00350	Tyrosine metabolism	C00082, *aoc2, aox5, dbh, dct, ddc, mao, tyrp1b*
dre00590	Arachidonic acid metabolism	C05956, C05962, *cbr1l, cyp2p8, cyp2p9, ggt1b, ptges3b, ptgs2b*
dre00053	Ascorbate and aldarate metabolism	C00167, C02670, C05412, *aldh2.2, aldh3a2b, ugdh, ugt1a1, ugt5g1*
dre00860	Porphyrin and chlorophyll metabolism	C00025, C00188, *alas1, alas2, cpox, hmox2a, ugt1a1, ugt5g1*
dre00790	Folate biosynthesis	C00415, C01300, C03684, C04244, *gch1, gch2, qdpra*
dre00310	Lysine degradation	C00956, *aldh2.2, aldh3a2b, hadhab, kmt2ca, pipox, plod2*
dre00410	beta-Alanine metabolism	C00106, C03722, *aldh2.2, aldh3a2b, aoc2, hadhab, srm*
dre00270	Cysteine and methionine metabolism	C00073, *gclc, gclm, ldha, ldhba, mtr, srm*
dre00280	Valine, leucine and isoleucine degradation	C00123, C00183, *aldh2.2, aldh3a2b, aox5, hadhab, hmgcs1*
dre00510	N-Glycan biosynthesis	*alg2, dad1, ddost, mogs, rpn1, stt3a*
dre00360	Phenylalanine metabolism	C00079, C00082, C00180, *aoc2, ddc, mao*
dre00030	Pentose phosphate pathway	C01236, *fbp2, g6pd, pgd, prps1a, taldo1*
dre00620	Pyruvate metabolism	C00186, *aldh2.2, aldh3a2b, ldha, ldhba, pcxb*
dre00670	One carbon pool by folate	C00415, *aldh1l2, mthfd1l, mthfd2, mtr, tyms*

dre00500	Starch and sucrose metabolism	C01083, *gyg1b, hkdc1, pygb*
dre01212	Fatty acid metabolism	*acsl1b, cpt2, fads2, hadhab, hsd17b12b*
dre00760	Nicotinate and nicotinamide metabolism	C03722, *aox5, naprt, pnp5a, pnp6*
dre00071	Fatty acid degradation	*acsl1b, aldh2.2, aldh3a2b, cpt2, hadhab*
dre00630	Glyoxylate and dicarboxylate metabolism	C00025, *aco2, gcshb, glulb, pgp*
dre00040	Pentose and glucuronate interconversions	C00167, C05412, *ugdh, ugt1a1, ugt5g1*
dre01040	Biosynthesis of unsaturated fatty acids	C01530, C01595, *fads2, hadhab, hsd17b12b*
dre00640	Propanoate metabolism	C00186, *hadhab, ldha, ldhba*
dre00600	Sphingolipid metabolism	*degs2, gba2, neu3.2, sptlc3*
dre00140	Steroid hormone biosynthesis	*cyp19a1b, hsd17b12b, ugt1a1, ugt5g1*
dre00120	Primary bile acid biosynthesis	C00245, C05468, *cyp46a1.1, cyp46a1.2*
dre00983	Drug metabolism	*impdh2, ugt1a1, ugt5g1, umps*
dre00650	Butanoate metabolism	C00025, *hadhab, hmgcs1, l2hgdh*
dre00730	Thiamine metabolism	C00082, *ak1, ak2, tpk1*
dre00250	Alanine, aspartate and glutamate metabolism	C00025, C01042, *asns, glulb*
dre00220	Arginine biosynthesis	C00025, *arg2, glulb, otc*
dre00190	Oxidative phosphorylation	*atp5c1, atp5g3b, cox6b1, cox6b2*
dre00020	Citrate cycle (TCA cycle)	*aco2, idh1, pcxb*
dre00750	Vitamin B6 metabolism	C06054, *aox5, psat1*
dre00900	Terpenoid backbone biosynthesis	*fntb, hmgcs1, pdss2*
dre00450	Selenocompound metabolism	*mtr, sephs1, txnrd1*
dre00051	Fructose and mannose metabolism	C00186, *fbp2, hkdc1*
dre00511	Other glycan degradation	*gba2, hexb, neu3.2*
dre00910	Nitrogen metabolism	C00025, *ca4b, glulb*
dre00290	Valine, leucine and isoleucine biosynthesis	C00123, C00183, C00188
dre00591	Linoleic acid metabolism	C01595, *cyp2p8, cyp2p9*
dre00430	Taurine and hypotaurine metabolism	C00025, C00245, *ggt1b*
dre00562	Inositol phosphate metabolism	*inpp5jb, ipmka*
dre00400	Phenylalanine, tyrosine and tryptophan biosynthesis	C00079, C00082
dre00563	Glycosylphosphatidylinositol (GPI)-anchor	*pigf, pigs*
dre00561	Glycerolipid metabolism	*aldh2.2, aldh3a2b*
dre00061	Fatty acid biosynthesis	C01530, *acsl1b*
dre00603	Glycosphingolipid biosynthesis	*hexb, naga*
dre00100	Steroid biosynthesis	*ebp, soat1*
dre00524	Neomycin, kanamycin and gentamicin biosynthesis	C00025, *hkdc1*
dre00062	Fatty acid elongation	*hadhab, hsd17b12b*
dre00770	Pantothenate and CoA biosynthesis	C00106, C00183

[a] All identified pathways with more than one hit are included.
[b] Metabolites are represented by KEGG C-code.
[c] Genes are represented by ZFIN Gene official name.

Multivariate curve resolution alternating least squares (MCR-ALS) method

MCR-ALS is a chemometric method especially suitable for the analysis of multicomponent systems with strongly coeluted and overlapped contributions, such as in LC-MS, GC-MS and CE-MS metabolomics data files. In the case of LC-MS, full-scan MS data matrix **D** contains the experimental mass spectra at all retention times in their rows and the chromatograms at all *m/z* values in their columns. MCR-ALS decomposes every individual experimental data matrix **D** into the elution profiles and mass spectra of the resolved contributions according to the following bilinear model (Eq.1):

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

where matrix **C** has the chromatographic profiles of the resolved contributions, $\mathbf{S}^{\mathbf{T}}$ has their corresponding mass spectra and **E** contains the residuals not explained by the model.

In this work, MCR-ALS was simultaneously applied to the analysis of several data matrices altogether, arranging the data matrices ($\mathbf{D}_{\mathbf{k}}$, corresponding to the 24 embryo samples) one at the top of each other in a column-wise augmented data matrix configuration ($\mathbf{D}_{\mathbf{aug}}$). In the case of ESI+ data, MCR-ALS decomposes the augmented data matrix $\mathbf{D}_{\mathbf{aug\_ESI+}}$ by using the same bilinear model, as is shown in Eq. 2:

$$\mathbf{D}_{\mathbf{aug\_ESI+}} = \begin{bmatrix} \boldsymbol{D_1} \\ \boldsymbol{D_2} \\ \vdots \\ \boldsymbol{D_{24}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C_1} \\ \boldsymbol{C_2} \\ \vdots \\ \boldsymbol{C_{24}} \end{bmatrix} \mathbf{S}^{\mathbf{T}} + \begin{bmatrix} \boldsymbol{E_1} \\ \boldsymbol{E_2} \\ \vdots \\ \boldsymbol{E_{24}} \end{bmatrix} = \mathbf{C}_{\mathbf{aug}}\mathbf{S}^{\mathbf{T}} + \mathbf{E}_{\mathbf{aug}} \quad \textbf{(2)}$$

In this case, MCR-ALS provided the common mass spectra of the resolved components ($\mathbf{S}^{\mathbf{T}}$) for all samples and the corresponding resolved chromatographic profiles ($\mathbf{C}_{\mathbf{aug}}$) in every sample. In this way, metabolite profiles between classes of samples can be resolved and compared. Prior to MCR-ALS analysis, the intensity scale of every chromatogram was normalized by the area of the internal standard (PIPES) and the data matrix $\mathbf{D}_{\mathbf{aug\_ESI+}}$ was then subdivided into different time windows (chromatographic regions) due to its huge size and computer processing limitations. A total number of

10 column-wise augmented data matrices were independently analyzed by MCR-ALS. Analogously, the ESI- augmented data matrix ($\mathbf{D}_{aug\_ESI-}$) was also subdivided into 10 column-wise augmented data matrices.

These 20 augmented data matrices, resulting from the previously explained time window reduction, were also normalized using an adaptation of the MinMax transformation described elsewhere [1]. This normalization allows the resolution of low concentrated metabolites as well as the major metabolites during MCR-ALS analysis.

Singular value decomposition (SVD) [2] method was then used to establish the initial guess of the number of components (independent contributions) of each augmented data matrix ($\mathbf{D}_{aug}$). After this initial approximation, the final number of components is decided by taking into account data fitting results and the reliability of the resolved profiles [1]. Subsequently, initial estimations of either elution profiles (**C** matrix) or mass spectra ($\mathbf{S}^T$ matrix), for the initially proposed number of components, were obtained using the purest variables of the experimental data [3]. Finally, the ALS optimization was performed under non-negativity constraints for chromatographic (**C**) and spectra ($\mathbf{S}^T$) profiles, and spectral normalization (equal height). In this way, resolved elution and mass spectra profiles had chemical meaning and intensity/rotational ambiguities are minimized [4].

References

[1] 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.
[2] Golub, G., Solna, K., Dooren, P.V., 2000. Computing the SVD of a General Matrix Product/Quotient. SIAM J. Matrix Anal. Appl. 22, 1-19.
[3] Windig, W., Guilment, J., 1991. Interactive self-modeling mixture analysis. Anal. Chem. 63, 1425-1432.
[4] de Juan, A., Jaumot, J., Tauler, R., 2014. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Analytical Methods 6, 4964-4976.

4.3. DISCUSSIÓ DELS RESULTATS

En aquesta secció es discuteixen i comparen els resultats obtinguts en els treballs inclosos en aquest capítol. En primer lloc, es mostren les pertorbacions a nivell metabòlic dels embrions de peix zebra (*D. rerio*) degudes a l'exposició als tres disruptors endocrins esmentats: el BPA, el PFOS i el TBT. En particular, es presenten els efectes d'aquests tres compostos químics mitjançant una anàlisi metabolòmica no dirigida. A continuació, es completa l'estudi dels efectes causats pel BPA mitjançant la integració dels resultats obtinguts en estudis no dirigits de transcriptòmica i metabolòmica. Això, permet obtenir una comprensió més global de la disrupció metabòlica d'aquest compost en els embrions de peix zebra.

4.3.1. Disrupció metabòlica de compostos disruptors endocrins

Els EDCs provoquen efectes rellevants en el metaboloma dels embrions de peix zebra

Per tal de poder descobrir la disrupció metabòlica dels EDCs investigats (BPA, PFOS i TBT) es van determinar cinc nivells d'exposició no tòxics per cadascun, és a dir, aquells nivells que poden causar afectació sense ocasionar la mortalitat de l'organisme (veure **Taula 4.1**). En concret, en els tres casos es van seleccionar concentracions per sota o iguals a la concentració més baixa en la que s'observaven efectes (*lowest observed effect concentration*, LOEC), similars a les trobades en el medi ambient. Aquestes concentracions es van avaluar inicialment mitjançant els corresponents assajos toxicològics. A la Taula S1 de l'article IV s'il·lustren els efectes en les primeres fases del desenvolupament embrionari (a les 120 hpf) deguts als tractaments amb BPA, PFOS i TBT a totes les concentracions investigades. Es van registrar els percentatges de supervivència, eclosió, inflació de la bufeta i les corresponents malformacions de la cua (lateral i dorsoventral) dels embrions.

Taula 4.1. Concentracions emprades per a l'exposició dels embrions de peix zebra per cada EDC estudiat.

BPA	PFOS	TBT
0 µM	0 µM	0 µM
0,44 µM	0,06 µM	0,003 µM
1,75 µM	0,2 µM	0,01 µM
4,4 µM	0,6 µM	0,03 µM
17,5 µM*	2,0 µM*	0,10 µM*

*Concentració LOEC.

En els assajos finals a les concentracions escollides, els embrions exposats a 17,5 µM de BPA i de 0,1 µM de TBT presentaven taxes més baixes de bufeta inflada (veure **Figura 4.1a** i **b**). En canvi en el cas del PFOS, alguns dels embrions exposats a 2,0 µM van mostrar malformacions a la cua, tal i com es mostra a la **Figura 4.1c**.

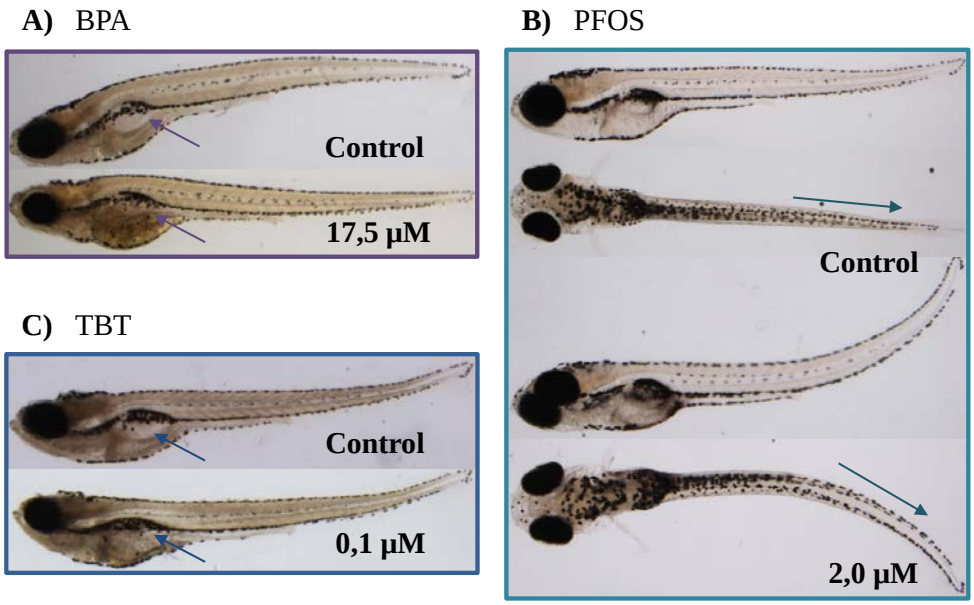


Figura 4.1. Efectes dels tres disruptors endocrins en el desenvolupament embrionari a les concentracions LOEC. Les fletxes indiquen les malformació dels embrions degudes a cada tractament.

Una vegada recol·lectades les mostres, els embrions es van analitzar mitjançant la metodologia no dirigida HILIC-MS. L'anàlisi multivariant basada en ROIMCR conjuntament amb l'anàlisi estadística mitjançant l'anàlisi de la variància multivariant regularitzada (rMANOVA) i ANOVA van permetre detectar un total de 50 biomarcadors metabòlics pel BPA, 47 pel TBT i 25 pel PFOS (veure Taules

S2, S3 i S4 de l'article IV) presents en el mòdul dre01100 (mapa de les rutes metabòliques de *D. rerio*) de la base de dades KEGG. Cal destacar que rMANOVA és una eina estadística multivariant relativament nova, per tant aquest estudi és un dels primers en els que s'ha aplicat aquesta estratègia per a la detecció de biomarcadors metabòlics rellevants. A la **Figura 4.2a** es mostra el nombre de metabòlits comuns i específics de cadascun dels tractaments investigats.

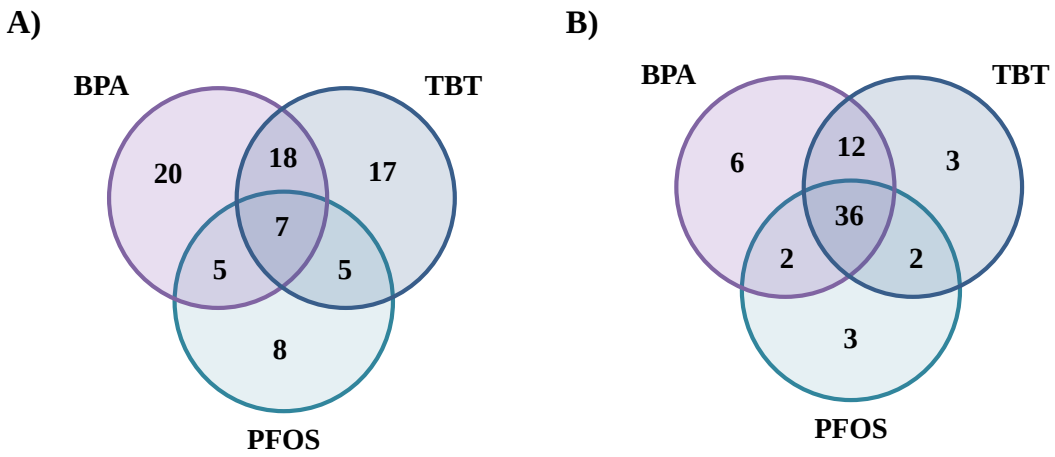


Figura 4.2. Diagrama de Venn (a) dels biomarcadors identificats i (b) de les rutes metabòliques afectades per BPA, PFOS i TBT. S'indica el nombre de metabòlits o rutes metabòliques específiques i comuns entre els tres tractaments.

Per obtenir una diferenciació integral dels tres tractaments cal tenir en compte els biomarcadors de cadascun i el seu anàlisi funcional. A la **Figura 4.2b** es resumeix el nombre de rutes metabòliques afectades comuns i específiques per cada tractament, les quals es descriuen en detall a continuació. La identificació de les rutes metabòliques afectades en els embrions exposats a BPA mitjançant la base de dades KEGG va reflectir que el grup més important de biomarcadors corresponia al metabolisme dels aminoàcids i el segon grup al metabolisme de les purines (adenosina monofosfat, L-glutamina, guanina, inosina monofosfat, hipoxantina, inosina i àcid úric). A més, l'anàlisi KEGG també va indicar un efecte rellevant en el metabolisme de les vitamines-cofactors (folat), en vies de senyalització (retinol) i en el metabolisme dels lípids (glicerofosfolípids i àcids grassos) (veure **Figura 4.3**). Tant el TBT com el BPA van mostrar que les dues rutes metabòliques més afectades corresponien al metabolisme dels aminoàcids i de les purines. A més, l'anàlisi mitjançant KEGG va

revelar altres rutes metabòliques afectades en el cas del TBT, com el metabolisme del carboni i l'àcid 2-oxocarboxílic, vies de senyalització (taurina, àcid butanoic) i el metabolisme dels lípids (glicerofosfolípids i lípids associats a la biosíntesi dels àcids grassos insaturats). Finalment, l'anàlisi funcional del biomarcadors deguts al tractament dels embrions amb PFOS va reflectir, en general, una disrupció metabòlica menor que el BPA i el TBT. En aquest cas, es va determinar un efecte predominant en el metabolisme dels aminoàcids, de les purines, del carboni i l'àcid 2-oxocarboxílic de manera similar a les anteriors exposicions. Tanmateix, és important tenir en compte que la disrupció dels lípids en aquest cas no va resultar significativa. Tot i així, cal incidir en que les condicions experimentals d'extracció dels metabòlits i d'anàlisi emprades en els estudis metabolòmics estan optimitzades per a la detecció de compostos de baix pes molecular, per tant, és normal que en general es detectin pocs efectes a nivell de lípids.

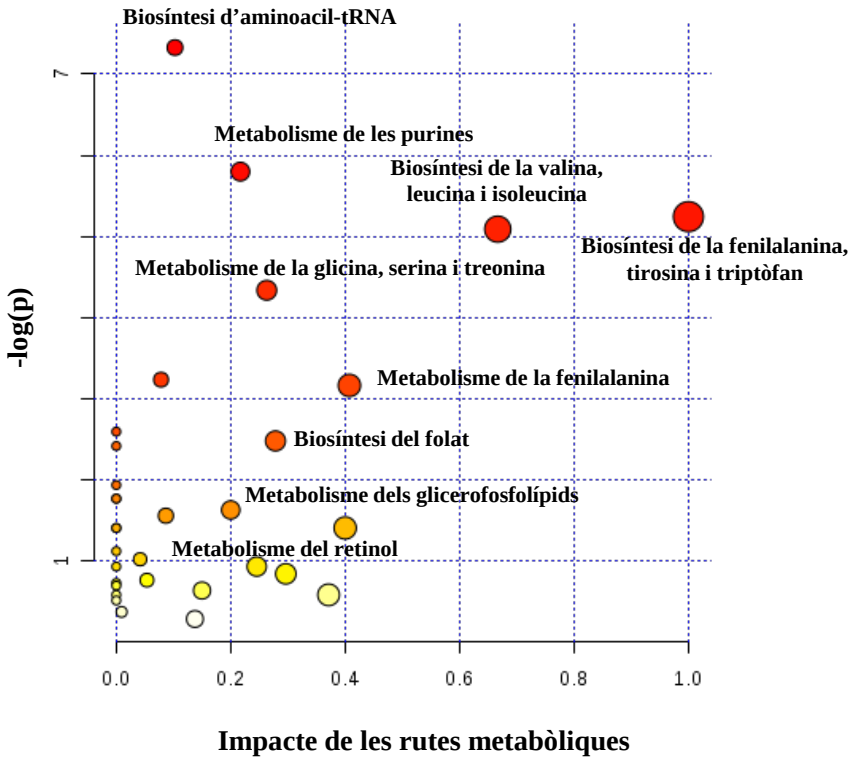


Figura 4.3. Representació gràfica de les principals rutes metabòliques afectades pel BPA en embrions de peix zebra. Els valors *p* de les corresponents rutes metabòliques s'han obtingut mitjançant el test exacte de Fisher [6]. Imatge generada amb el model d'anàlisi de rutes del programa Metaboanalyst [7].

La simple comparació de les rutes metabòliques afectades pels tres tractaments va mostrar un grau de superposició molt més gran dels efectes dels tres EDCs que quan s'observen els biomarcadors biològics en particular (**Figura 4.2**). En la majoria de casos, la importància relativa dels mòduls KEGG notablement afectats va resultar similar pels tres EDCs, tant pel que fa al nombre de metabòlits de cada mòdul com per la tendència dels canvis de concentració observats en els metabòlits. Per exemple, es va determinar un augment general de la concentració de nucleòsids (guanosina, inosina) en el metabolisme de les purines en contrast a la disminució dels nucleòtids (guanosina monofosfat, inosina monofosfat, adenosina monofosfat), de les bases nitrogenades i metabòlits relacionats (guanina, hipoxantina i àcid úric). La pertorbació en la biosíntesi de nucleòtids és una resposta cel·lular comuna contra l'estrès causat per l'exposició a xenobiòtics, tal i com s'ha vist recentment en el fong *Cunninghamella echinulata* exposat a TBT [8]. Aquest fet es va observar de forma individual pels tres tractaments, però per comprendre millor els efectes dels EDCs en el metaboloma dels embrions del peix zebra es va requerir de la combinació dels biomarcadors dels tres tractaments. De manera similar, l'anàlisi conjunta de les dades corresponents als tres contaminants va permetre determinar efectes en el metabolisme de la taurina i hipotaurina, generant un augment de la concentració d'hipotaurina i una disminució de la taurina i d'altres metabòlits relacionats. A més, a partir de l'alteració en el metabolisme dels glicerofosfolípids per l'exposició a BPA i TBT es va observar una millora en la producció de lípids de fosfocolina i una disminució de les fosfoserines.

Dels canvis de concentració detectats es pot extreure que els EDCs produeixen efectes tòxics i alteracions en la proliferació cel·lular, deduïts principalment pels canvis en la concentració dels metabòlits relacionats amb el metabolisme dels aminoàcids i la biosíntesi de proteïnes.

D'una banda, els canvis significatius observats en la concentració d'aminoàcids confirma la presència d'estrès oxidatiu produït pels EDCs en els embrions de peix zebra. Aquest fenomen també s'ha donat en estudis previs [9, 10]. Per exemple, els canvis en els nivells d'alanina, L-glutamina, 4-aminobutanoat i L-prolina es consideren una resposta essencial dels organismes en estrès oxidatiu [10, 11]. Yoon *et al*. van determinar que es pot extreure una conclusió similar dels canvis de concentració de l'àcid úric i del desajust de la concentració d'hipoxantina en mostres exposades a BPA i TBT [12].

També, la L-glutamina està relacionada amb la pertorbació en el metabolisme de les purines ocasionada per BPA, que es manifesta en la disminució en la producció de L-glutamat [12]. Lu *et al.* han definit el L-glutamat com un potencial biomarcador de l'exposició del mol·lusc *Haliotis diversicolor* a TBT [13]. En aquest treball, aquestes fluctuacions en la concentració de L-glutamat també s'han detectat en les exposicions a TBT.

D'altra banda, les concentracions de lípids en embrions de peix zebra (fosfatidilcolines (PC), lisofosfatidilcolines (LysoPCs), diacilglicerols (DG)) augmenten per l'exposició a BPA i TBT, excepte les fosfatidilserines (PS) que disminueixen. El metabolisme dels lípids juga un paper crucial en la integritat estructural i les funcions de la membrana cel·lular. Per tant, aquestes alteracions en els nivells de fosfolípids produïdes per BPA i TBT indiquen possibles danys a la membrana cel·lular i una possible amenaça de la viabilitat cel·lular [13, 14]. Bernat *et al.* ja han descrit que el TBT mostra una gran afinitat per la membrana cel·lular arribant a interferir la seva integritat [15]. En concret, Bernat *et al.* van determinar que el TBT no tan sols afecta els nivells de fosfolípids, els quals és considera que són els més sensibles a les alteracions del medi [16], sinó també dels d'àcids grassos del fong *Cunninghamella elegants.* En aquest cas, el BPA va mostrar un efecte sobre la composició d'àcids grassos lliures de *D. rerio*, mentre que el TBT va produir alteracions en les vies relacionades amb la biosíntesi d'àcids grassos insaturats. En general, aquests resultats coincideixen amb els presentats anteriorment per Gorrochategui *et al.* [17] sobre cultius cel·lulars, on els efectes del TBT en el metabolisme dels lípids van resultar ser superiors en comparació amb els produïts pels compostos perfluorats (PFCs), com el PFOS. A més, s'ha confirmat que el TBT dóna lloc a estrès oxidatiu en animals [13] i microorganismes [18], i que pot danyar proteïnes i lípids.

En aquest estudi, també es va detectar que el TBT produeix un desequilibri en el metabolisme de la taurina, el qual és un dels principals responsables de la protecció cel·lular, ja que evita la degradació de la membrana. Probablement, aquesta alteració podria donar-se per compensar els possibles danys de la membrana cel·lular reflectida per la disrupció dels lípids. Aquesta pertorbació també s'ha observat lleugerament en els embrions tractats amb BPA. A més, la taurina ha estat descrita com un aminoàcid essencial per al desenvolupament normal del sistema nerviós [19]. Paral·lelament, el TBT

ocasiona també canvis relacionats amb el metabolisme energètic, tal i com ha estat descrit per Soboń *et al.*[8]. En concret, es va detectar un augment dels nivells de α-D-glucosa en *D. rerio* juntament amb una disminució de L-glutamat i alanina, la qual cosa indica la presència de certes pertorbacions metabòliques dels carbohidrats [20, 21]. Aquesta reducció de L-glutamat i alanina suggereix que la glucosa derivada del glicogen podria ser necessària per equilibrar les necessitats energètiques i així mantenir l'homeòstasi dels embrions exposats a TBT [21, 22].

Finalment, un dels efectes més rellevants del BPA en embrions de peix zebra ha estat la disrupció del metabolisme del retinol vinculat amb el procés de fototransducció. La pertorbació de l'homeòstasi de retinoïdes per contaminants és un procés que pot tenir lloc a múltiples nivells metabòlics i funcionals [23]. El mecanisme d'acció del BPA en el metabolisme del retinol i el seu significat biològic és avui en dia controvertit, però ja s'ha detectat en diversos treballs recents [23].

En resum, l'anàlisi metabolòmica no dirigida HILIC-MS és una estratègia prometedora que ens ha permès entendre millor l'actuació tòxica dels contaminants ambientals. En concret, s'ha pogut avaluar i distingir les diferents alteracions metabòliques i els possibles efectes tòxics del BPA, TBT i PFOS en organismes aquàtics. No obstant, es requereixen més estudis òmics complementaris, com ara estudis transcriptòmics i/o lipidòmics, per a poder confirmar els efectes dels EDCs en *D. rerio* i així, interpretar millor els canvis observats en els perfils metabòlics.

Una millor comprensió global de la disrupció metabòlica del bisfenol A s'obté mitjançant la combinació d'una aproximació metabolòmica i transcriptòmica conjunta

Tot i que s'ha demostrat que la metabolòmica és una eina poderosa per estudiar la resposta dels organismes aquàtics enfront exposicions químiques, la integració de diferents nivells òmics permet entendre millor els efectes d'aquests compostos en les rutes metabòliques [5]. Així, es pot avaluar el risc ambiental de manera mes acurada i completa. La integració del coneixement de diferents nivells òmics s'aplica freqüentment en la investigació biomèdica i farmacèutica per a propòsits de diagnòstic i identificació dirigida o *target* [24], però avui en dia aquest enfocament continua sent poc explorat en l'àmbit de l'ecotoxicologia. En el camp de la toxicologia ambiental, existeixen encara pocs treballs on

es faci la integració de dades metabolòmiques i transcriptòmiques [25]. En la segona publicació d'aquest capítol (article V), s'ha dut a terme la integració de les respostes a nivell metabolòmic i transcriptòmic per a l'avaluació de la toxicitat del BPA, i així intentar aconseguir una comprensió més àmplia dels processos moleculars implicats. A partir de la integració, es pot obtenir simultàniament informació dels biomarcadors metabòlics i dels canvis en l'expressió gènica a nivell de mRNA, la qual cosa pot ajudar a extreure una explicació biològica més completa del comportament dels organismes enfront a l'exposició de BPA. En concret, en la última publicació d'aquesta Tesi s'han emprat dues metodologies no dirigides basades en HILIC-MS i en la seqüenciació massiva de RNA-Seq. L'anàlisi de les mostres i el processament de les dades de RNA-Seq es va dur a terme en el Centre Nacional d'Anàlisi Genòmica (CNAG) en col·laboració amb el grup d'investigació del Departament de Toxicologia Ambiental de l'Institut de Diagnosi Ambiental i Estudis de l'Aigua (IDAEA).

En aquest treball es van exposar embrions de peix zebra durant les primeres 120 hpf atès que es va considerar que és en aquesta finestra d'exposició quan es poden produir més efectes sobre el desenvolupament embrionari dels organismes aquàtics que es troben en aigües ambientals contaminades. En aquest cas, també es van realitzar els assajos toxicològics pertinents per escollir les concentracions d'exposició d'interès. En concret, es van seleccionar la concentració LOEC com la concentració d'exposició més alta (17,5 μM) i dues concentracions addicionals per sota d'aquesta (4,4 i 8,8 μM). La concentració LOEC es va determinar a partir dels efectes morfològics observats. D'aquesta manera, es van evitar alteracions importants en el desenvolupament o en la viabilitat dels embrions. Com es pot veure a la **Figura 4.4a**, els embrions exposats a una concentració de 17,5 μM de BPA presentaven un percentatge d'inflació de la bufeta natatòria inferior a la produïda per la resta de concentracions respecte les mostres control, tant a les 96 com 120 hpf. A la **Figura 4.4b** es mostra un exemple d'embrió control i un altre d'exposat a 17,5 μM de BPA, on es detecta la bufeta inflada sol en el primer cas.

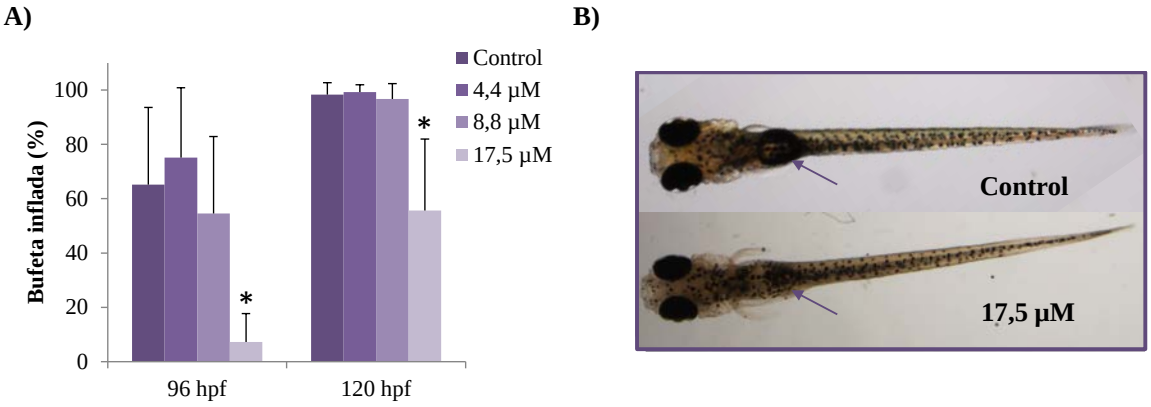


Figura 4.4. Efectes del BPA en el desenvolupament embrionari. (a) Taxa de bufetes inflades (%). Mitjana ± desviació estàndard (n=16). Les dades de cada etapa es van analitzar mitjançant el test no paramètric Kruskal-Wallis amb una significació de valor p<0,05. (b) Fotografia d'un embrió control i d'un d'exposat a 17,5 µM de BPA a les 120 hpf. Les fletxes indiquen la presència i l'absència de la bufeta inflada en els embrions control i exposats, respectivament.

Seguidament, es va avaluar la disrupció metabòlica dels embrions de *D. rerio* tractats amb BPA. La metodologia metabòlomica no dirigida de HILIC-MS i l'anàlisi multivariant basada principalment en la combinació *binning*-MCR (descrita en el capítol 3) van permetre determinar el conjunt de metabòlits dels embrions. L'anàlisi estadística d'ANOVA va permetre detectar aquells metabòlits les concentracions dels quals variaven significativament degut a l'exposició a BPA. Es van poder identificar un total de 50 metabòlits com a biomarcadors rellevants dels efectes d'aquest compost químic. A més, a partir dels canvis observats en les àrees dels perfils cromatogràfics d'aquests metabòlits es van poder agrupar les mostres segons els diferents nivells d'exposició (controls, 4,4 µM, 8,8 µM i 17,5 µM). Les mostres tractades amb la menor concentració de BPA (4,4 µM) es van agrupar amb els controls, mentre que les mostres corresponents a les dues dosis més altes també es van agrupar entre si (veure dendrograma de la Figura 2 de l'article V).

L'anàlisi funcional dels 50 biomarcadors metabòlics va revelar un efecte general del BPA sobre el metabolisme dels aminoàcids i dels nucleòtids i nucleòsids, tal i com s'ha descrit recentment [26]. Aquests resultats es poden contrastar amb altres estudis relacionats amb els efectes del BPA en el metaboloma d'embrions de peix zebra [9] i del crustaci *Daphnia Magna* [27]. Concretament, es va

observar una disminució en la concentració dels metabòlits relacionats amb el metabolisme dels nucleòtids, com la hipoxantina i l'àcid úric, que indiquen la presència d'efectes en vies de senyalització, com també s'ha descrit per l'exposició de rates a BPA [28]. Yoon *et al*. han assenyalat recentment que aquest desequilibri en la producció d'àcid úric associat amb una alteració en la concentració d'hipoxantina està relacionat amb l'estrès oxidatiu produït sobre els embrions [12]. A més, ja s'ha comentat que la disminució general del metabolisme de les purines per BPA pot estar vinculat a l'increment de L-glutamat [12, 28]. Així mateix, també es van observar canvis en les rutes metabòliques dels lípids, de sucres i d'esteroides com s'ha exposat en estudis anteriors de transcriptòmica [29, 30] i metabolòmica [12]. En particular, els canvis en el metabolisme del glicerofosfolípids suggereix que el BPA danya la membrana cel·lular [9, 12]. Finalment, es van observar canvis en el metabolisme secundari, en les vies relacionades amb la nicotinamida, l'àcid linoleic i araquidònic, el metabolisme del retinol i els cofactors i les vitamines.

D'altra banda, simultàniament es va dur a terme l'anàlisi de RNA-Seq de les mostres control i de les exposades a la concentració més alta de BPA (17,5 μM), per tal d'investigar quins eren els efectes del BPA en l'expressió gènica a nivell de mRNA. L'anàlisi de les lectures de cada gen va permetre detectar un total de 1381 gens estadísticament significatius. L'anàlisi preliminar dels gens associats als transcrits afectats amb una expressió més diferenciada entre les mostres control i les exposades a BPA es va relacionar amb els diferents possibles processos biològics afectats, incloent l'activitat estrogènica, els processos d'oxidació-reducció, la disrupció dels processos metabòlics, la regeneració de les aletes de la cua, el desenvolupament dels teixits i la proliferació cel·lular. Un dels gens més significatius i indicatius de l'activitat estrogènica del BPA en els vertebrats és l'aromatasa cerebral (*cyp19a1b*) [31]. La major expressió d'aquest gen esta d'acord amb els resultats descrits prèviament per Kishida *et al.*[32] i Chen *et al.*[33]. En concret, el BPA modifica l'activitat de l'enzim citocrom P450 el qual té un paper molt important en el metabolisme oxidatiu [9]. A més, Lam *et al*. han descrit recentment que el BPA altera diferents funcions cel·lulars involucrades en la proliferació cel·lular i en el desenvolupament del teixits [34, 35]. Finalment, l'augment de *gstp2* es correlaciona amb un

increment de l'activitat del Glutatió-S-Transferasa (GST) relacionada amb un increment de l'estrès oxidatiu i l'apoptosi [36].

Per tal d'extreure una explicació biològica més completa de la disrupció metabòlica observada a través de les anàlisis dels resultats de la metabolòmica i la transcriptòmica és essencial la integració de la informació dels dos conjunts de dades. Tanmateix, la integració dels canvis en els perfils metabòlics i transcriptòmics a nivell conceptual no és una tasca senzilla [5]. En l'article V es va dur a terme la integració dels metabòlits i dels gens afectats per l'exposició de BPA utilitzant el coneixement biològic previ de les rutes metabòliques del peix zebra definides en la base de dades KEGG (*pathway-level integration*). D'aquesta manera, es pot millorar la comprensió dels processos biològics afectats pel BPA i així obtenir informació sobre el seu mecanisme d'acció. En total, es van tenir en compte 48 metabòlits i 119 gens estadísticament significatius que es trobaven presents en el mòdul del peix zebra (dre01100) de la base de dades KEGG. L'anàlisi funcional conjunta dels gens i dels metabòlits va revelar que les dades metabolòmiques i transcriptòmiques ofereixen informació complementària, afectant bàsicament el mateix conjunt de rutes metabòliques i produint efectes fisiològics similars. La majoria de biomarcadors del tractament amb BPA corresponien a gens i metabòlits relacionats amb el metabolisme d'aminoàcids i de les purines, com s'ha vist també en l'article IV. És a dir, els canvis metabòlics i els efectes tòxics del tractament amb BPA estan principalment relacionats amb la síntesi de proteïnes i aminoàcids, components essencials de les cèl·lules [34, 35]. L'anàlisi conjunta dels gens i dels metabòlits afectats va permetre determinar altres rutes afectades significativament, com per exemple la biosíntesis de vitamines-cofactors (folat, àcid ascòrbic), les vies de senyalització (retinol, glutamat, taurina, prostaglandines) i el metabolisme dels lípids (glicerofosfolípids, àcid esteàric, àcid linoleic). Algunes d'aquestes rutes reflecteixen la disrupció de les funcions de senyalització-regulació dels embrions de peix zebra pel BPA, activitat que ja s'ha vinculat anteriorment amb el potencial del BPA com a disruptor endocrí [37-41].

El següent pas va consistir en l'anàlisi exhaustiva i conjunta de les tendències observades en les abundàncies relatives dels metabòlits i gens causades per l'exposició a BPA, la qual cosa va permetre extreure una explicació millor a nivell fisiològic d'aquestes pertorbacions.

Com ja s'ha indicat, les concentracions de la majoria de metabòlits relacionats amb el metabolisme de les purines, retinol, folat i lípids van disminuir amb el tractament amb BPA, mentre que diverses concentracions d'aminoàcids van augmentar. Aquestes variacions es poden relacionar amb els canvis en els nivells de mRNA dels gens inclosos en les mateixes vies metabòliques. Per exemple, l'alteració en la síntesi d'aminoacil-tRNA està relacionada amb l'augment de les concentracions d'aminoàcids en embrions tractats amb BPA, la qual cosa indicaria la reducció de la demanda energètica per mantenir l'homeòstasi [25]. Els canvis en la concentració d'aminoàcids afecten diverses rutes metabòliques i de biosíntesi. Per altra banda, la ruta potencialment afectada del folat (deficiència de folat) en el desenvolupament d'embrions podria estar associada a una resposta de l'organisme que permeti bloquejar els efectes del BPA i els possibles símptomes depressius relacionats [42]. Addicionalment, el desequilibri observat en les concentracions de l'àcid linoleic també s'ha descrit anteriorment que podria estar relacionada comunament amb diversos trastorns neurològics, inclosa la depressió [25].

L'anàlisi funcional conjunta dels gens i metabòlits va mostrar que els nivells totals de retinoïdes van disminuir amb el tractament amb BPA, mentre que els gens relacionats amb el metabolisme del retinol (*aldh1a2, bco1l, cyp26b1, si:ch1073-13h15.3, rdh12, ugt1a1, ugt5g1*) van aparèixer sobreexpressats. Això, es relaciona amb l'augment general de la degradació dels retinoïdes en embrions de peix zebra degut a exposicions a retinoïdes [43]. A més, Shmarakov *et al.* han atribuït recentment aquest comportament a les respostes transcripcionals i posteriors a la transcripció de l'organisme per activar la biodegradació del BPA [44]. Avui en dia, no es coneix amb precisió les conseqüències de la disrupció dels retinoïdes per BPA, però sí que es coneix que els nivells de retinoïdes exerceixen efectes rellevants sobre el desenvolupament embrionari, el creixement cel·lular, la diferenciació i l'apoptosi [45]. Aquesta pertorbació tan rellevant del metabolisme del retinol també es va observar en l'article IV.

Finalment, l'anàlisi conjunta dels biomarcadors metabolòmics i transcriptòmics va mostrar una disminució de les concentracions de prostaglandines en les mostres d'embrions tractades amb BPA, la qual està relacionada amb la reducció de l'abundància de mRNA del seu enzim productor *ptgs2b*. Aquest efecte es combina amb l'augment de l'expressió de fins a set enzims que catalitzen rutes

metabòliques alternatives, ja sigui la de l'àcid araquidònic o la del metabòlit intermedi PGH2. La interrupció de la via metabòlica de les prostaglandines en els organismes tractats amb BPA podria estar relacionada amb una resposta inflamatòria [46], però encara es desconeix el seu significat fisiològic en el desenvolupament primerenc d'embrions de *D. rerio*. De fet, els efectes de la BPA sobre la síntesi de prostaglandines s'han adreçat únicament en teixits d'animals adults on es descriuen múltiples rols de les prostaglandines amb la reproducció [47]. En aquest treball *s'*ha demostrat que a més que el BPA afecta la funció ovàrica a través de la síntesi de prostaglandines en cèl·lules de granulosa de mamífers [48]. Tot i així, encara no es coneix bé el paper d'aquestes durant el desenvolupament dels embrions de peix zebra.

Podem concloure doncs que les anàlisis metabolòmiques HILIC-MS i trancriptòmiques de RNA-Seq no dirigides han permès aprofundir en els mecanismes de disrupció metabòlica del BPA sobre embrions de *D. rerio*. Aquests resultats contrasten amb d'altres obtinguts a partir de les metodologies dirigides habitualment utilitzades en el camp de la òmica, les quals presenten una capacitat de cobertura més limitada. Cal remarcar que una comprensió més sòlida de la toxicitat del BPA s'obté solament a partir de la integració dels dos nivells òmics, la qual cosa permet caracteritzar completament els efectes del BPA més enllà de la simple detecció de les rutes afectades. En aquest treball, s'ha demostrat que per fer una bona interpretació dels canvis en el metabolisme del retinol es requereix la integració d'ambdues anàlisis òmiques, la qual ha permès confirmar la degradació dels retinoïdes en comptes d'una inhibició de la seva síntesi.

Malgrat el potencial de la integració dels dos nivells òmics, seria necessari realitzar estudis lipidòmics addicionals que permetessin correlacionar els canvis d'expressió gènica amb les alteracions observades en les concentracions dels lípids a causa de l'exposició a BPA.

4.3.2. Comparativa dels efectes del bisfenol A sobre les rutes metabòliques dels embrions de peix zebra

Les dues publicacions d'aquest capítol han permès concloure que el BPA és un disruptor endocrí que produeix principalment alteracions en el metabolisme dels aminoàcids, dels nucleòtids i dels

nucleòsids (veure **Taula 4.2**). Aquests resultats són concordants amb els descrits en l'article de Wang *et al.*, el qual assenyala els efectes d'aquest contaminant sobre el metabolisme de diversos organismes o mostres biològiques [26]. A la **Taula 4.2** es resumeixen els resultats trobats de l'exposició a BPA en els dos estudis descrits en aquest capítol, és a dir, les rutes metabòliques pertorbades en cada cas.

Taula 4.2. Principals rutes metabòliques afectades per l'exposició d'embrions de peix zebra a BPA.

	Anàlisi metabolòmica		**Anàlisi transcriptòmica**
Rutes metabòliques[a]	Exposició de les 48 a les 120 hpf	Exposició de les 2 a les 120 hpf	Exposició de les 2 a les 120 hpf
Metabolisme dels aminoàcids	**	**	**
Metabolisme de les purines	**	*	**
Metabolisme de les vitamines-cofactors	*	**	**
Metabolisme dels lípids	*	*	**
Metabolisme del retinol	*	*	**
Metabolisme de les prostaglandines/àcid araquidònic		*	**
Metabolisme de la taurina i hipotaurina		**	**
Metabolisme del glutamat		**	**

[a]El valor de significació (valor *p*) de cada ruta metabòlica s'ha obtingut mitjançant el programa Metaboanalyst [7].

* Ruta metabòlica afectada amb un valor $p<0,1$ mitjançant el test exacte de Fisher.

** Ruta metabòlica afectada amb un valor $p<0,01$ mitjançant el test exacte de Fisher.

A partir dels resultats es poden avaluar dos factors diferents. D'una banda, l'exposició a dos finestres de temps diferents dintre d'un mateix nivell òmic. D'altra banda, l'avaluació de la mateixa exposició considerant dos nivells òmics diferents.

En el cas de les dos finestres d'exposició, és interessant el cas del metabolisme de les vitamines i els cofactors. Així, es pot observar que l'exposició de 48 a 120 hpf va produir alteracions en el metabolisme del folat mentre que l'exposició de les 2 a les 120 hpf va causar també canvis en els nivells dels metabòlits i en l'expressió gènica a nivell de mRNA del metabolisme de l'àcid ascòrbic.

En canvi, l'estudi combinat de dos nivells òmics va ajudar a rebel·lar diferències en el metabolisme dels lípids, del retinol i de les prostaglandines. Per exemple, en el cas dels lípids, l'estudi metabolòmic va detectar majoritàriament un desequilibri en el balanç dels glicerofosfolípids. En canvi, l'anàlisi transcriptòmica de RNA-Seq va permetre determinar que el BPA també ocasiona un desajust en el metabolisme de l'àcid esteàric i linoleic.

És a dir, l'ús de les dues aproximacions òmiques ha permès detectar i extreure més informació sobre els efectes adversos del BPA en el metabolisme dels organismes. No obstant, cal tenir en compte que dur a terme estudis considerant diferents finestres i nivells d'exposició i diferents nivells òmics pot no ser una qüestió senzilla (quantitat de mostra, cost de l'anàlisi, etc.). Així, s'haurà de definir clarament l'objectiu de l'estudi en l'etapa inicial del disseny experimental per seleccionar quin és el nivell òmic més adequat o si és imprescindible dur a terme l'estudi multi-nivell.

4.4. REFERÈNCIES

1. Garcia, G. R., Noyes, P. D., Tanguay, R. L., Advancements in zebrafish applications for 21st century toxicology, *Pharmacology & Therapeutics*. 2016, *161*, 11-21.
2. Colborn, T., vom Saal, F. S., Soto, A. M., Developmental effects of endocrine-disrupting chemicals in wildlife and humans, *Environmental Health Perspectives*. 1993, *101*, 378-384.
3. Tabb, M. M., Blumberg, B., New modes of action for endocrine-disrupting chemicals, *Molecular Endocrinology*. 2006, *20*, 475-482.
4. Casals-Casas, C., Desvergne, B., Endocrine disruptors: from endocrine to metabolic disruption, *Annual Review of Physiology*. 2011, *73*, 135-162.
5. Cavill, R., Jennen, D., Kleinjans, J., Briede, J. J., Transcriptomic and metabolomic data integration, *Briefings in Bioinformatics*. 2016, *17*, 891-901.
6. Sprent, P., in: Lovric, M. (Eds.), Fisher Exact Test, *International Encyclopedia of Statistical Science*, Springer, Berlin, Heidelberg. 2011, pp. 524-525.
7. Xia, J., Wishart, D. S., Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis, *Current protocols in Bioinformatics*. 2016, *55*, 14.10.11-14.10.91.
8. Soboń, A., Szewczyk, R., Różalska, S., Długoński, J., Metabolomics of the recovery of the filamentous fungus Cunninghamella echinulata exposed to tributyltin, *International Biodeterioration & Biodegradation*. 2018, *127*, 130-138.
9. Huang, S. S., Benskin, J. P., Chandramouli, B., Butler, H., Helbing, C. C., Cosgrove, J. R., Xenobiotics produce distinct metabolomic responses in zebrafish larvae (Danio rerio), *Environmental science & technology*. 2016, *50*, 6526-6535.
10. Soboń, A., Szewczyk, R., Długoński, J., Tributyltin (TBT) biodegradation induces oxidative stress of Cunninghamella echinulata, *International Biodeterioration & Biodegradation*. 2016, *107*, 92-101.
11. Liang, X., Zhang, L., Natarajan, S. K., Becker, D. F., Proline mechanisms of stress survival, *Antioxidants & Redox signaling*. 2013, *19*, 998-1011.
12. Yoon, C., Yoon, D., Cho, J., Kim, S., Lee, H., Choi, H., Kim, S., 1H-NMR-based metabolomic studies of bisphenol A in zebrafish (Danio rerio), *Journal of Environmental Science and Health, Part B*. 2017, *52*, 282-289.
13. Lu, J., Feng, J., Cai, S., Chen, Z., Metabolomic responses of Haliotis diversicolor to organotin compounds, *Chemosphere*. 2017, *168*, 860-869.
14. Li, M., Wang, J., Lu, Z., Wei, D., Yang, M., Kong, L., NMR-based metabolomics approach to study the toxicity of lambda-cyhalothrin to goldfish (Carassius auratus), *Aquatic toxicology*. 2014, *146*, 82-92.
15. Bernat, P., Długoński, J., Comparative study of fatty acids composition during cortexolone hydroxylation and tributyltin chloride (TBT) degradation in the filamentous fungus Cunninghamella elegans, *International Biodeterioration & Biodegradation*. 2012, *74*, 1-6.
16. Dercová, K., Čertík, M., Mal'ová, A., Sejáková, Z., Effect of chlorophenols on the membrane lipids of bacterial cells, *International Biodeterioration & Biodegradation*. 2004, *54*, 251-254.
17. 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.
18. Bernat, P., Gajewska, E., Szewczyk, R., Słaba, M., Długoński, J., Tributyltin (TBT) induces oxidative stress and modifies lipid profile in the filamentous fungus Cunninghamella elegans, *Environmental Science and Pollution Research International*. 2014, *21*, 4228-4235.
19. Ye, G., Chen, Y., Wang, H., Ye, T., Lin, Y., Huang, Q., Chi, Y., Dong, S., Metabolomics approach reveals metabolic disorders and potential biomarkers associated with the developmental toxicity of tetrabromobisphenol A and tetrachlorobisphenol A, *Scientific Reports*. 2016, *6*, 35257.
20. Andreassen, T. K., Skjoedt, K., Korsgaard, B., Upregulation of estrogen receptor alpha and vitellogenin in eelpout (Zoarces viviparus) by waterborne exposure to 4-tert-octylphenol and 17beta-estradiol, *Comparative Biochemistry and Physiology - Part C: Toxicology & Pharmacology*. 2005, *140*, 340-346.
21. Zhou, J., Zhu, X., Cai, Z., Tributyltin toxicity in abalone (Haliotis diversicolor supertexta) assessed by antioxidant enzyme activity, metabolic response, and histopathology, *Journal of Hazardous Materials*. 2010, *183*, 428-433.
22. Zhang, J., Sun, P., Yang, F., Kong, T., Zhang, R., Tributyltin disrupts feeding and energy metabolism in the goldfish (Carassius auratus), *Chemosphere*. 2016, *152*, 221-228.

23. Shmarakov, I. O., Retinoid-xenobiotic interactions: the Ying and the Yang, *Hepatobiliary Surgery and Nutrition*. 2015, *4*, 243-267.
24. Edwards, S. W., Preston, R. J., Systems biology and mode of action based risk assessment, *Toxicological Sciences*. 2008, *106*, 312-318.
25. Huang, S. S. Y., Benskin, J. P., Veldhoen, N., Chandramouli, B., Butler, H., Helbing, C. C., Cosgrove, J. R., A multi-omic approach to elucidate low-dose effects of xenobiotics in zebrafish (Danio rerio) larvae, *Aquatic Toxicology*. 2017, *182*, 102-112.
26. Wang, M., Rang, O., Liu, F., Xia, W., Li, Y., Zhang, Y., Lu, S., Xu, S., A systematic review of metabolomics biomarkers for Bisphenol A exposure, *Metabolomics*. 2018, *14*, 45.
27. Nagato, E. G., Simpson, A. J., Simpson, M. J., Metabolomics reveals energetic impairments in Daphnia magna exposed to diazinon, malathion and bisphenol-A, *Aquatic Toxicology*. 2016, *170*, 175-186.
28. Chen, M., Zhou, K., Chen, X., Qiao, S., Hu, Y., Xu, B., Xu, B., Han, X., Tang, R., Mao, Z., Dong, C., Wu, D., Wang, Y., Wang, S., Zhou, Z., Xia, Y., Wang, X., Metabolomic analysis reveals metabolic changes caused by bisphenol A in rats, *Toxicological Sciences:an Official Journal of the Society of Toxicology*. 2014, *138*, 256-267.
29. Boucher, J. G., Gagné, R., Rowan-Carroll, A., Boudreau, A., Yauk, C. L., Atlas, E., Bisphenol A and bisphenol S induce distinct transcriptional profiles in differentiating human primary Preadipocytes, *PLoS One*. 2016, *11*, e0163318.
30. Villeneuve, D. L., Garcia-Reyero, N., Escalon, B. L., Jensen, K. M., Cavallin, J. E., Makynen, E. A., Durhan, E. J., Kahl, M. D., Thomas, L. M., Perkins, E. J., Ankley, G. T., Ecotoxicogenomics to support ecological risk assessment: a case study with bisphenol A in fish, *Environmental Science & Technology*. 2012, *46*, 51-59.
31. Mouriec, K., Lareyre, J. J., Tong, S. K., Page, Y. L., Vaillant, C., Pellegrini, E., Pakdel, F., Chung, B. C., Kah, O., Anglade, I., Early regulation of brain aromatase (cyp19a1b) by estrogen receptors during zebrafish development, *Developmental Dynamics*. 2009, *238*, 2641-2651.
32. Kishida, M., McLellan, M., Miranda, J. A., Callard, G. V., Estrogen and xenoestrogens upregulate the brain aromatase isoform (P450aromB) and perturb markers of early development in zebrafish (Danio rerio), *Comparative Biochemistry and Physiology-Part B: Biochemistry & Molecular biology*. 2001, *129*, 261-268.
33. Chen, J., Saili, K. S., Liu, Y., Li, L., Zhao, Y., Jia, Y., Bai, C., Tanguay, R. L., Dong, Q., Huang, C., Developmental bisphenol A exposure impairs sperm function and reproduction in zebrafish, *Chemosphere*. 2017, *169*, 262-270.
34. Lam, S. H., Hlaing, M. M., Zhang, X., Yan, C., Duan, Z., Zhu, L., Ung, C. Y., Mathavan, S., Ong, C. N., Gong, Z., Toxicogenomic and phenotypic analyses of bisphenol-A early-life exposure toxicity in zebrafish, *PLoS One*. 2011, *6*, e28273.
35. Lam, H.-M., Ho, S.-M., Chen, J., Medvedovic, M., Tam, N. N. C., Bisphenol A disrupts HNF4α-regulated gene networks linking to prostate preneoplasia and immune disruption in noble rats, *Endocrinology*. 2016, *157*, 207-219.
36. Hassan, Z. K., Elobeid, M. A., Virk, P., Omer, S. A., ElAmin, M., Daghestani, M. H., AlOlayan, E. M., Bisphenol A induces hepatotoxicity through oxidative stress in rat model, *Oxidative Medicine and Cellular Longevity*. 2012, *2012*, 194829.
37. Pelayo, S., Oliveira, E., Thienpont, B., Babin, P. J., Raldua, D., Andre, M., Pina, B., Triiodothyronine-induced changes in the zebrafish transcriptome during the eleutheroembryonic stage: implications for bisphenol A developmental toxicity, *Aquatic Toxicology*. 2012, *110-111*, 114-122.
38. Chen, Y., Reese, D. H., A screen for disruptors of the retinol (vitamin A) signaling pathway, *Birth Defects Research Part B: Developmental and Reproductive Toxicology*. 2013, *98*, 276-282.
39. Lakind, J. S., Goodman, M., Mattison, D. R., Bisphenol A and indicators of obesity, glucose metabolism/type 2 diabetes and cardiovascular disease: a systematic review of epidemiologic research, *Critical Reviews in Toxicology*. 2014, *44*, 121-150.
40. Porreca, I., Ulloa-Severino, L., Almeida, P., Cuomo, D., Nardone, A., Falco, G., Mallardo, M., Ambrosino, C., Molecular targets of developmental exposure to bisphenol A in diabesity: a focus on endoderm-derived organs, *Obesity Reviews*. 2017, *18*, 99-108.
41. Rochester, J. R., Bisphenol A and human health: A review of the literature, *Reproductive Toxicology*. 2013, *42*, 132-155.
42. Geng, Y., Gao, R., Chen, X., Liu, X., Liao, X., Li, Y., Liu, S., Ding, Y., Wang, Y., He, J., Folate deficiency impairs decidualization and alters methylation patterns of the genome in mice, *MHR: Basic Science of Reproductive Medicine*. 2015, *21*, 844-856.

43. Oliveira, E., Casado, M., Raldúa, D., Soares, A., Barata, C., Piña, B., Retinoic acid receptors' expression and function during zebrafish early development, *Journal of Steroid Biochemistry and Molecular Biology*. 2013, *138*, 143-151.
44. Shmarakov, I. O., Borschovetska, V. L., Blaner, W. S., Hepatic detoxification of bisphenol A is retinoid-dependent, *Toxicological Sciences*. 2017, *157*, 141-155.
45. Bushue, N., Wan, Y. J., Retinoid pathway and cancer therapeutics, *Advanced Drug Delivery Reviews*. 2010, *62*, 1285-1298.
46. Ricciotti, E., FitzGerald, G. A., Prostaglandins and inflammation, *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2011, *31*, 986-1000.
47. Rossitto, M., Ujjan, S., Poulat, F., Boizet-Bonhoure, B., Multiple roles of the prostaglandin D2 signaling pathway in reproduction, *Reproduction*. 2015, *149*, R49-58.
48. OVS, M., Abayalath, N., MAM, I., Sooriyapathirana, S. S., Wijayagunawardane, M. P. B., Kodithuwakku, S., Endocrine disruptor bisphenol-A (BPA) alters the prostaglandins synthesis cascade enzyme gene expression in porcine granulosa cells in vitro, *Wayamba University Sri Lanka International Conference*. 2016.

CAPÍTOL 5. *Conclusions*

Els mètodes analítics basats en espectrometria de masses desenvolupats en aquesta Tesi han permès fer una anàlisi completa i fiable del metaboloma d'organismes biològics model, com el llevat (*Saccharomyces cerevisiae*) i els embrions de peix zebra (*Danio rerio*). En aquest context, l'ús d'estratègies de tractament de dades ha estat molt útil per extreure la informació biològica de les dades metabolòmiques, i d'aquesta manera permetre entendre els efectes dels diferents factors ambientals estudiats sobre aquests organismes model.

A continuació es resumeixen les conclusions referents als mètodes analítics i quimiomètrics presentats en aquesta Tesi, i als efectes dels disruptors endocrins avaluats en embrions de peix zebra:

Metodologies analítiques no dirigides

- Les fases estacionàries HILIC són una gran alternativa en els estudis de metabolòmica no dirigida on s'analitzen compostos d'elevada polaritat i baix pes molecular. En aquests estudis HILIC, els factors experimentals més influents en l'anàlisi dels metabòlits són el tipus de fase estacionària i el modificador orgànic. En concret, les fases estacionàries amida i zwitteriònica són les més recomanables per a l'anàlisi de metabòlits polars. A més, el modificador orgànic que proporciona una millor separació és l'acetonitril, ja que el metanol dificulta la formació de la capa d'aigua a la superfície de la fase estacionària HILIC. Per últim, altres factors experimentals com el pH o la força iònica no han mostrat efectes rellevants en l'anàlisi dels metabòlits investigats.
- S'ha demostrat el potencial de les metodologies analítiques basades en CE-MS per dur a terme estudis de metabolòmica no dirigida, sobretot pel cas de metabòlits molt polars. En concret, les condicions de separació optimitzades amb capil·lars de sílice fosa han permès obtenir perfils metabòlics de mostres complexes mitjançant l'ús d'instrumentació senzilla i baix consum de reactius i mostra. L'ús de tècniques de tractament de dades adequades ajuda a resoldre possibles problemes de reproductibilitat en els temps de migració o en les àrees dels pic, comuns en CE.

Metodologies quimiomètriques

- S'ha demostrat la utilitat de les estratègies de tractament de dades basades en el mètode de resolució MCR-ALS per a l'anàlisi de dades metabolòmiques no dirigides. MCR-ALS és un mètode robust i flexible per al processament de dades complexes com les que es generen en els estudis òmics. A més, MCR-ALS permet l'anàlisi de dades en aquells casos en que les separacions obtingudes s'allunyin del comportament ideal, com ho són freqüentment les que s'observen en les separacions de CE-MS i de LC-MS, on es donen problemes com, pics asimètrics, desplaçaments en els temps dels pics entre mostres i la presència de coelucions molt fortes degudes a l'elevat nombre de compostos. A partir de l'aplicació del mètode MCR-ALS, ha estat possible detectar, resoldre i identificar un gran nombre de metabòlits en una sola anàlisi cromatogràfica o electroforètica, la qual cosa possibilita una millor comprensió de la naturalesa de les mostres biològiques analitzades, moltes vegades de gran complexitat.
- En la resolució per MCR-ALS del metabòlits és imprescindible una etapa prèvia de compressió de les dades en la direcció espectral. Aquest preprocessament es pot dur a terme mitjançant el procediment d'interpolació, d'agrupament en caixetes (*binning*) o cerca de les regions d'interès (ROI). En els dos primers casos d'interpolació i *binning*, sovint es requereix d'una etapa addicional de divisió dels cromatogrames en finestres de temps i/o de massa (*windowing*) per poder dur a terme el posterior anàlisi per MCR-ALS. En canvi, la cerca de ROIs disminueix de forma més efectiva la mida de les dades permetent la seva anàlisi directa per MCR-ALS i mantenint l'exactitud en la resolució espectral de les dades originals (alta resolució de *m/z*) a diferència dels mètodes tradicionals d'interpolació i *binning*.
- L'avaluació dels efectes dels diferents factors ambientals investigats sobre els organismes biològics estudiats és possible de forma senzilla i ràpida mitjançant la utilització de mètodes estadístics univariants i multivariants, els quals permeten maximitzar la quantitat d'informació que es pot extreure a partir de l'anàlisi dels diferents conjunts de dades òmiques. Els mètodes de tractament de dades estadístics univariants, com el test *t*, el test *U* de Mann-Whitney i l'ANOVA, han permès el tractament independent de cadascuna de les variables

procedents de les dades metabolòmiques per a la selecció de potencials biomarcadors entre grups de mostres biològiques. Per contra, els mètodes de tractament de dades estadístics multivariants, com el procediment ASCA i el rMANOVA, són especialment útils en l'anàlisi dels conjunts de dades complexes basats en dissenys experimentals per a l'avaluació global de la significació dels factors investigats sobre el conjunt dels metabòlits estudiats i detectar aquells que es troben més afectats pels canvis i factors ambientals estressants considerats.

- Les estratègies de fusió de dades proposades en aquesta Tesi han facilitat la comprensió global dels efectes dels factors ambientals estressants investigats sobre els processos biològics dels organismes estudiats. Aquestes estratègies de fusió de dades metabolòmiques no dirigides s'han realitzat a partir de l'anàlisi simultània de les mateixes mostres amb els mètodes HILIC-MS i CE-MS i la metodologia ROIMCR. Aquests procediments de fusió han permès incrementar el nombre de metabòlits afectats detectats i obtenir informació més precisa sobre els canvis en els perfils metabòlics dels organismes biològics, especialment quan s'empra l'estratègia de fusió de dades de nivell baix, és a dir, quan es fusionen directament les dades experimentals obtingudes per les dues tècniques, HILIC-MS i CE-MS.

Efectes de compostos diruptors endocrins (EDCs) en embrions de peix zebra

- S'ha demostrat que l'exposició d'embrions de peix zebra a BPA, PFOS i TBT des de les 48 a les 120 hores posteriors a la fertilització (hpf), causa efectes significatius en el seu metaboloma. Els dos primers compostos químics (BPA i TBT) han causat una major disrupció metabòlica que el PFOS. L'anàlisi metabolòmica no dirigida ha permès definir que els embrions *D. rerio* responen, en general, de manera similar als tres EDCs. El desequilibri en el metabolisme dels aminoàcids i en la biosíntesi de proteïnes demostra que en els tres casos es produeixen efectes tòxics, estrès oxidatiu i alteracions en la proliferació cel·lular. A més, s'han detectat efectes específics sobre vies de senyalització particulars, com per exemple els canvis en el metabolisme dels lípids per l'exposició a BPA i TBT. Finalment, s'ha determinat que el BPA produeix també efectes pertorbadors en el metabolisme del retinol, el qual està estretament relacionat amb el procés de fototransducció (visió) dels embrions.

- Els embrions de peix zebra exposats a BPA des de les 2 a les 120 hpf també presenten canvis significatius i similars als trobats en els embrions exposats a aquest disruptor endocrí de les 48 a les 120 hpf. Cal destacar que l'anàlisi conjunta o fusionada de les dades metabolòmiques i transcriptòmiques no dirigides ha permès confirmar que el BPA interromp el metabolisme del retinol, de les prostaglandines, l'homeòstasi lipídica i altera significativament els nivells de colesterol durant el temps d'exposició estudiat. A més, gràcies al canvis en l'expressió gènica a nivell de mRNA s'ha confirmat l'activitat estrogènica i androgènica del BPA.
- L'anàlisi conjunta (fusionada) de les dades obtingudes a partir de les dues aproximacions òmiques (metabolòmica i trasncriptòmica) permet entendre i interpretar millor (de forma més global) la toxicitat del BPA sobre aquests organismes biològics. La comprensió dels efectes adversos del BPA en les rutes metabòliques dels embrions de peix zebra és possible únicament a partir de la integració funcional dels canvis observats en els metabòlits i en l'expressió gènica a nivell de mRNA.