UNIVERSITAT DE BARCELONA

Desenvolupament i aplicació de metodologies òmiques, analítiques i quimiomètriques en estudis ambientals

Mireia Farrés Rodríguez

DESENVOLUPAMENT I APLICACIÓ DE METODOLOGIES ÒMIQUES, ANALÍTIQUES I QUIMIOMÈTRIQUES EN ESTUDIS AMBIENTALS

Mireia Farrés Rodríguez

UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

FACULTAT DE QUÍMICA

DEPARTAMENT DE QUÍMICA ANALÍTICA

Programa de doctorat:

QUÍMICA ANALÍTICA DEL MEDI AMBIENT

DESENVOLUPAMENT I APLICACIÓ DE METODOLOGIES ÒMIQUES, ANALÍTIQUES I QUIMIOMÈTRIQUES EN ESTUDIS AMBIENTALS

Memòria presentada per Mireia Farrés Rodríguez per optar al grau de

Doctora per la Universitat de Barcelona

Director

Dr. Romà Tauler Ferré

Professor d'Investigació del Departament de Química Ambiental de l'Institut de Diagnòstic Ambiental i Estudis de l'Aigua. CSIC.

Tutora

Dra. Anna de Juan Capdevila

Professora titular del Departament de Química Analítica. Universitat de Barcelona

Als meus pares,

"I, a vegades, contra tot pronòstic
una gran bestiesa capgira tot allò que crèiem lògic,
tot fent evident,
que per un moment,
ens en sortim"

***Captatio Benevolentiae*, Manel**

En primer lloc m'agradaria dedicar el meu agraïment al meu director de Tesi, el Dr. Romà Tauler, per haver confiat en mi, per haver-me donat l'oportunitat de fer aquesta Tesi i per dirigir-me el treball. Gràcies per tot el que m'has ensenyat i per la paciència durant tots aquests anys. De la mateixa manera, també vull agrair a la Dra. Anna de Juan haver acceptat ser la meva tutora durant la Tesi i per ajudar-me en qualsevol moment. Aquest treball no hagués estat de cap manera possible sense l'ajuda del Dr. Benjamí Piña, la Dra. Belen Martrat i el Dr. Joan Grimalt de l'IDAEA, la Dra. Marta Villagrasa de l'ICRA, el Dr. Stefan Tsakovski de Sofia University, la Dra. Olga Jauregui del CCiTUB, i la Dra. Roser Chaler, la Dori Fanjul i la Maria Comesaña del Servei d'Espectrometria de Masses del CSIC.

Gràcies Marta Alier i Stefan, heu estat amb els que he passat gran part d'aquesta Tesi, gràcies per ajudar-me, ensenyar-me i respondre a totes les preguntes sempre. Què dir dels actuals companys de laboratori?! Mil gràcies per les ajudes, els riures (que no faltin mai!), els ànims, per estar allà, per les bones estones compartides i perquè vosaltres heu fet que tot sigui més divertit, de veritat! En ordre de fondo sud a fondo norte... gràcies Xiscu, Cristian, Núria, Carma, Meritxell, Elena, Eva i Elba, i al Víctor de la UB, sou uns cracks!!! També gràcies als companys de l'altre despatx, thank you Xin, Amrita, Joaquim, Igor i una altra vegada, merci Stefan. També vull donar les gràcies a tots amb els que he compartit despatx/laboratori durant una temporada, cadascun d'ells m'ha ensenyat coses noves, ja siguin quimiomètriques o perspectives de la vida en general... Thank you Clecio, Ewelina, Hadi, Jessica, Michele, Elis, Mariana, Alejandro, Mahsa, Yahya... entre molts d'altres. Sara, moltes i moltes gràcies pels riures, els congressos compartits, pels runnings neoyorkins, pels ànims, les ajudes i en general per les estones viscudes durant tota la Tesi. Sílvia, mil gràcies per ser-hi, escoltar, congressos, riures, consells, opinar, festes, escalada, muntanya, vòlei, bici, vins, sopars, triatlons, merci per ser 'tot terreny' i poder compartir moltíssimes coses amb tu. Tot molt necessari per a fer una Tesi. Cristal, mi amiga rutera y también todoterreno, gracias por compartir muchos momentos, charlas en el CSIC, montaña, ferratas, carreras de orientación, esquí, voley, viajes, y un largo etc. Todo esencial para mi. Marta Fort, merci per tot, per compartir la música, pels dinars al CSIC, vòlei, concerts, xerrades i consells, mil gràcies, un plaer haver compartit tot aquest viatge amb tu. Maria, aussie girl, merci por todos los momentos, voley, los momentos bici al volver de inglés y por lo mucho que me ayudaste y enseñaste en el laboratorio de masas.

També vull agrair a tots aquells de fora del CSIC i que de manera indirecta m'han ajudat o animat durant tots els anys de la Tesi. Als que vam començar sent un grup de vòlei... Jose, Surdo, Sílvia again, etc. els riures amb vosaltres i les sortides de tranquis han esdevingut imprescindibles per a

acabar la Tesi. David, per tots els moments compartits, pel vòlei, merci per proposar-nos/'obligar-nos' a fer la triatló, perquè diguem el que diguem al final tot surt bé! :) Mariana, viatgera! gracies per les estones compartides, per Galicia i per Madrid. Elisabet, gràcies per ser-hi sempre sigui des de Castellví, London o LA, impossible resumir tots els moments/facetimes/londontrips que han contribuït a que aquesta Tesi tirés endavant, milions de gràcies. Gemma, sempre hi has sigut, merci per poder comptar amb tu en qualsevol minut, acompanyar-me i animar-me, pels sopars, pels vins, pels moments esquiant... ja ho saps, milions de gràcies. Pau, gràcies pels salts des de les roques a Menorca, pels entrenaments, per ser-hi. Mariona, encara recordo el primer assaig de presentació que et vaig fer aguantar... de fet n'has vist uns quants, gràcies per ser-hi, animar-me, ajudar-me en qualsevol moment. Rous, des de Castellví, Califòrnia, Austràlia, NZ, merci perquè sempre animes, merci per tot, per ser-hi, per més nits d'hivern. Marta Bonastre, milions de gràcies per poder compartir moltes coses amb tu, sobretot moments de muntanya, esquí, bici... un llarg etcètera, tot això ha estat i és bàsic per mi, mil gràcies pel que hem fet i per tot el que queda!! ☺ Gràcies a l'actriu més gran, merci Maria, pels escenaris compartits, per les tonteries, pels riures. Moltes gràcies Cristina i Albert per ser-hi, you're smart!! Núria, gràcies per les xerrades, runnings, muntanya i caminades direcció el mar... merci per ser-hi. Carles, per tu tinc llista llarga de moments, runnings, tenis, esquí, Barça, etc. sàpigues que han estat bàsics i imprescindibles per a fer la Tesi. Cousin (Francesc), també gràcies a tu per compartir festivals, Barça i moltes altres coses amb mi. I molta i molta més gent, els cosins aussies, Anna i Jordi, merci per tot, per preguntar, em va quedar pendent el trip to Australia.... Merci Albert Canals, recordo que em deies que feia coses molt rares a la Tesi... i Isi, lo dicho, merci també per tots els moments i els cafès/birres/cokes prechampions, merci nois per ser-hi :)

A més a més de tots els que he nombrat, també vull donar les gràcies a tota la innumerable gent que m'ha acompanyat al principi, al final o durant aquest viatge. A totes aquelles persones amb les que he compartit moments d'aquesta Tesi, ja sigui a la muntanya, al laboratori, en un concert, al CSIC, esquiant, al gimnàs, en un seminari, escalant, als congressos, ballant, en un viatge, compartint escenari al teatre, en bici, en una reunió, corrent, en una festa, en un sopar... i a totes aquelles persones amb les que m'he creuat i sense saber-ho m'han donat forces durant la Tesi.

Finalment, volia donar les gràcies a tota la meva família avis, tiets, cosins. També a la meva germana Marta i al Foued, per donar vida a la petita Eya, perquè amb ella el final de la Tesi ha estat més dolç. I en especial als meus pares, per tota la paciència, per l'educació que m'heu donat, per recolzar-me, per ajudar-me i, en resum, perquè vosaltres també heu fet possible aquesta Tesi.

Resum

En aquesta Tesi s'han dissenyat, desenvolupat i aplicat noves metodologies quimiomètriques d'anàlisi multivariant de dades que permeten l'extracció d'informació química i bioquímica en estudis òmics i ambientals a partir de dades cromatogràfiques.

S'ha estudiat la influència de diferents factors externs (estadi de creixement, tipus de cultiu, varietat i tipus de mostra) sobre el cultiu del blat mitjançant la comparació de les metodologies ANOVA-PCA i ASCA. Aquestes eines quimiomètriques s'han aplicat sobre l'anàlisi dirigida dels perfils TIC LC-MS dels metabòlits secundaris (al·leloquímics) del blat. Aquestes eines han permès interpretar els canvis dels seus metabòlits en relació als factors externs considerats. La concentració dels metabòlits al·leloquímics DIMBOA-Glc, DIMBOA, HMBOA i MBOA ha canviat de forma sistemàtica durant el creixement de la planta del blat en relació als factors estadi de creixement, tipus de mostra i la seva interacció.

S'ha desenvolupat una estratègia experimental no dirigida per a l'estudi i anàlisi d'experiments metabolòmics de mostres de llevat, utilitzant cromatografia líquida amb detecció per espectrometria de masses (LC-MS) i mètodes quimiomètrics d'anàlisi de les dades experimentals obtingudes. Aquesta estratègia ha permès avaluar l'efecte de condicions estressants sobre el cultiu del llevat (canvi de temperatura de cultiu i concentracions creixents de Cu(II)). S'han avaluat els canvis en les àrees dels perfils d'elució cromatogràfica dels metabòlits resolts per MCR-ALS. Els mètodes quimiomètrics de selecció de variables VIP-PLS i SR-PLS han permès la discriminació entre mostres de llevat control i tractades. La identificació temptativa dels metabòlits s'ha fet a partir dels seus espectres de masses resolts per MCR-ALS. S'ha proposat una interpretació biològica dels canvis observats en el metaboloma com a conseqüència dels canvis en les condicions del seu cultiu (canvis de temperatura i concentracions creixents de Cu(II)). S'han observat modificacions en el metaboloma del llevat quan aquest s'ha cultivat a 42ºC (condicions estressants), i s'ha detectat l'inici procés d'estrès oxidatiu en el llevat quan aquest s'ha cultivat a concentracions (subletals) creixents de Cu(II).

S'ha realitzat l'estudi de les correlacions existents entre les concentracions dels compostos orgànics determinades per CG-MS acumulats als sediments marins (IODP-U1318) durant l'època del Miocè i la temperatura superficial del mar (SST) en el mateix període de temps. S'ha aplicat el mètode VIP-PLSR als perfils cromatogràfics TIC GC-MS de mostres de sediments marins per a la

selecció de nous possibles marcadors paleoclimàtics dels canvis de temperatura en el període de temps investigat. Aquest estudi ha permès la diferenciació entre les aportacions dels compostos del grup de lípids d'origen marí i terrestre.

Finalment, s'han estudiat i comparat diferents mètodes quimiomètrics de selecció de variables, VIP-PLS i SR-PLS, en l'anàlisi de diversos conjunts de dades multivariants de naturalesa i complexitat diversa. En particular, aquest estudi s'ha aplicat a la selecció de variables en els següents casos: 1) selecció dels paràmetres fisicoquímics de l'aigua que influeixen més en el seu gust; 2) selecció dels compostos orgànics fòssils dels sediments marins, les concentracions dels quals (obtingudes per GC-MS) es correlacionaven més amb els canvis de temperatura superficial de l'aigua del mar en estudis paleoclimàtics; i 3) selecció dels gens, a partir d'anàlisi de xips d'ADN (*microarrays*), de l'organisme *Daphnia magna* exposat a dosis subletals de SSRI, en relació a la seva reproducció. S'ha demostrat que el mètode SR té una alta sensibilitat i especificitat. El mètode VIP també ha presentat alta sensibilitat i ha estat més rigorós per a la selecció de variables en els cromatogrames TIC, però ha demostrat baixa especificitat en la resta de conjunt de dades.

Índex

Capítol 1

Objectius i estructura de la Tesi

1.1 OBJECTIUS

L'objectiu principal d'aquesta Tesi és dissenyar, desenvolupar i aplicar noves metodologies quimiomètriques d'anàlisi multivariant de dades que permetin l'extracció d'informació química i bioquímica a partir d'estudis òmics i ambientals que utilitzen mètodes analítics cromatogràfics amb detecció per espectrometria de masses (LC-MS).

De manera més precisa, en aquesta Tesi es proposen els següents objectius específics:

- Estimació de la influència de diferents factors externs (estadi de creixement, tipus de cultiu, varietat i tipus de mostra) sobre el cultiu de blat. Desenvolupament, aplicació i comparació de les metodologies ANOVA-PCA (ANOVA-*Principal Component Analysis*) i ASCA (ANOVA-*Simultaneus Component Analysis)* en l'anàlisi dels perfils metabòlics de la planta de blat obtinguts per LC-MS. Avaluació estadística i interpretació de la significació dels canvis observats en les concentracions dels metabòlits al·leloquímics del blat per la influència dels factors externs considerats.

- Desenvolupament d'una estratègia no dirigida per a l'anàlisi de dades LC-MS dels metabòlits del llevat exposat a condicions estressants (canvi de temperatura de cultiu i concentracions creixents de Cu(II)). Aplicació del mètode MCR-ALS a dades LC-MS metabolòmiques de mostres de llevat control i exposades a condicions estressants. Avaluació estadística dels canvis en les àrees dels perfils de concentració resolts per MCR-ALS que permeten la discriminació entre mostres control i mostres tractades. Aplicació dels mètodes quimiomètrics de selecció de variables VIP-PLS i SR-PLS per a la detecció dels metabòlits més significatius que expliquen els canvis observats en el metaboloma del llevat. Identificació temptativa dels metabòlits a partir dels seus espectres de masses resolts per MCR-ALS i de les bases de dades YMDB (*Yeast Metabolome Database*), KEGG (*Kyoto Encyclopedia of Genes and Genomes*) i MassBank. Interpretació biològica dels canvis en el metaboloma del llevat exposat a condicions estressants.

- Identificació i interpretació de les variacions en les concentracions de compostos orgànics d'origen biològic acumulats en sediments marins i obtinguts per cromatografia de gasos

amb detecció per espectrometria de masses (GC-MS) en relació als canvis de temperatura de l'aigua marina superficial al llarg de 8.3 Ma en el Miocè. Proposta de marcadors paleoclimàtics de variació de temperatura mitjançant l'aplicació del mètode quimiomètric VIP-PLS als perfils GC-MS dels compostos orgànics.

- Estudi i comparació dels resultats de l'aplicació dels dos mètodes quimiomètrics de selecció de variables VIP-PLS i SR-PLS en l'anàlisi de conjunts de dades multivariants de diferent naturalesa i complexitat. Aplicació d'aquests mètodes a la detecció de les variables més significatives (marcadors) en estudis òmics i ambientals.

1.2 ESTRUCTURA DE LA TESI

La present memòria està estructurada en set capítols que es descriuen a continuació:

En el primer capítol es presenten els objectius que han motivat a la realització d'aquesta Tesi i es detalla l'estructura i la relació dels treballs científics de la present memòria.

En el segon capítol es fa primer una introducció general sobre la metabolòmica, les metodologies experimentals d'anàlisis dirigida i no dirigida que utilitzen LC-MS i els mètodes quimiomètrics que es poden emprar en aquests casos. Es descriuen els sistemes experimentals estudiats en aquesta Tesi (blat, llevat i sediments marins), es detallen les etapes de treball experimental general en estudis metabolòmics i es descriuen els mètodes quimiomètrics emprats en aquesta Tesi per al pretractament i tractament de les dades cromatogràfiques. Aquests inclouen diversos mètodes de preprocessament de les dades originals, els mètodes d'anàlisi multivariant, de resolució multivariant, de regressió multivariant i els mètodes de selecció de variables.

En el tercer capítol es presenten els resultats de l'avaluació quimiomètrica de les dades cromatogràfiques obtingudes en l'anàlisi dels compostos al·leloquímics de mostres de blat, mitjançant ANOVA-PCA i ASCA. S'avaluen els efectes de diferents factors experimentals (estadi de creixement, tipus de cultiu, varietat i tipus de mostra) i es determinen els metabòlits més importants en cada cas.

En el quart capítol es presenten els resultats de l'aplicació d'una estratègia d'anàlisi no dirigida dels perfils metabòlics LC-MS del llevat quan s'apliquen condicions ambientals estressants en el seu cultiu (canvis de temperatura i concentracions de Cu(II) creixents). Aquest capítol recull els resultats obtinguts de l'anàlisi conjunta per MCR-ALS de les dades LC-MS de mostres de llevat, la selecció i identificació dels metabòlits la concentració dels quals canvia més en relació a les condicions estressants aplicades, i la interpretació biològica dels fenòmens relacionats amb els canvis metabòlics observats.

En el cinquè capítol es presenten detalladament els resultats obtinguts de l'anàlisi quimiomètrica i interpretació de les senyals paleoclimàtiques de compostos orgànics acumulats en sediments marins i la seva relació amb els canvis de temperatura de la superfície de l'aigua del mar fa milers d'anys (Miocè).

En el sisè capítol es presenten detalladament els resultats obtinguts de la comparació dels mètodes VIP i SR en la regressió per mínims quadrats parcials, per a la selecció i interpretació de les variables més importants en estudis de diferent naturalesa: avaluació de la qualitat sensorial de l'aigua, anàlisi de dades químiques paleoclimàtiques i anàlisi de dades de transcriptòmica.

Finalment, en el setè capítol es recullen les conclusions generals més importants de la present Tesi.

1.3 RELACIÓ DELS TREBALLS CIENTÍFICS PRESENTATS EN LA MEMÒRIA

Article 1. ***Chemometric evaluation of different experimental conditions on wheat (Triticum aestivum L.) development using liquid chromatography mass spectrometry (LC–MS) profiles of benzoxazinone derivatives.***

Autors: Mireia Farrés, Marta Villagrasa, Ethel Eljarrat, Damià Barceló i Romà Tauler
Revista: Analytica Chimica Acta 731 (2012) 24-31.

Article 2. ***Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS.***

Autors: Mireia Farrés, Benjamí Piña i Romà Tauler
Revista: Metabolomics 11 (2015) 210-224.

Article 3. ***LC-MS based metabolomics and chemometrics study of the toxic effects of copper on Saccharomyces cerevisiae***

Autors: Mireia Farrés, Benjamí Piña i Romà Tauler (enviat a Metallomics)

Article 4. ***Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7 Ma old (IODP-U1318)***

Autors: Mireia Farrés, Belen Martrat, Ben de Mol, Joan O. Grimalt i Romà Tauler

Revista: Analytica Chimica Acta 879 (2015) 1-9.

Article 5. ***Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation***

Autors: Mireia Farrés, Stefan Platikanov, Stefan Tsakovski i Romà Tauler

Revista: Journal of Chemometrics 29 (2015) 528-536.

Capítol 2

Introducció

El projecte del genoma humà (*The Human Genome Project*) va fixar l'objectiu de determinar la seqüència completa dels nucleòtids del genoma humà per primera vegada a finals de 1980. La combinació dels recursos adequats i un fort lideratge científic va permetre el desenvolupament de tecnologies ràpides de seqüenciació de ADN (àcid desoxiribonucleic). Des d'aleshores, s'han dilucidat genomes de desenes d'organismes, dels quals el més important és el de la seqüència del genoma humà, que es va posar a disposició de la comunitat científica l'any 2000. El coneixement del genoma i de les noves tecnologies que han emergit posteriorment, estan permetent entendre els riscos de malalties i dels efectes tòxics sobre els organismes vius, així com desenvolupar nous enfocaments predictius d'aquests. A partir d'aquest fet els científics han combinat esforços per a desenvolupar un conjunt de noves eines moleculars i bioinformàtiques amb la finalitat d'aprofitar el potencial que presenta la informació obtinguda a partir del genoma seqüenciat (National Research Council US, 2007). Aquestes eines bioinformàtiques són les que actualment fan possible l'obtenció i l'anàlisi de dades biològiques de grans dimensions (*big data*), i estan tenint efectes molt importants en la investigació i comprensió dels fenòmens biològics tant des de del punt de vista de la genòmica, com de la transcriptòmica, de la proteòmica i de la metabolòmica.

L'aplicació de les tecnologies òmiques als estudis toxicològics ha permès estudiar els efectes dels compostos químics contaminants sobre els éssers vius i identificar els seus perills i, per tant, ha permès vigilar la seva exposició a partir del seguiment de les respostes cel·lulars a diferents dosis. Les tecnologies òmiques comprenen diverses plataformes tecnològiques diferents per l'anàlisi del genoma, el transcriptoma, les proteïnes i els metabòlits. Amb l'ús d'aquestes tecnologies, un sol experiment pot generar una gran quantitat d'informació, i el caràcter integral d'aquesta informació és molt major del que generen els experiments tradicionals. Aprofundir en aquests resultats i aconseguir extreure'n el màxim d'informació és moltes vegades complicat, la qual cosa es pot facilitar mitjançant l'aplicació de la Quimiometria. Aquesta disciplina juga un paper important en l'anàlisi i tractament de conjunts molt grans de dades químiques complexes, multi- i megavariants (*big data*).

La revolució genòmica ha desencadenat un gir en l'anàlisi dels sistemes biològics, que històricament s'ha realitzat en àrees diverses de la biologia, incloent l'ecologia, la biologia del desenvolupament i la immunologia. La biologia de sistemes globals és un terme definit per Nicholson i coautors (Nicholson i Wilson, 2003) que pretén integrar la informació biològica

multivariant per entendre millor les interaccions entre els organismes i el medi ambient. La mesura i la modelització de conjunts de dades i d'informació tan diversos representa un desafiament significatiu a nivell d'anàlisi i modelatge. Les investigacions sobre les vies de resposta cel·lular s'han incrementat de forma constant durant les darreres dècades degut al desenvolupament de les tecnologies òmiques, les quals han posat per davant la biologia de sistemes en detriment de la biologia molecular i cel·lular bàsiques (Westerhoff i Palsson, 2004). La biologia de sistemes combina les dades obtingudes de la identificació dels gens reguladors i de les xarxes bioquímiques per a comprendre millor el funcionament global dels sistemes biològics complexos. Els mètodes convencionals per a la creació d'un a xarxa model inclouen la realització d'una sèrie d'experiments per identificar les interaccions específiques i una recerca exhaustiva de la literatura. S'han creat diverses bases de dades de gran escala que contenen els gens reguladors i les xarxes bioquímiques dels diferents organismes, com són KEGG (*Kyoto Ecyclopedia of Genes and Genomes*) (Kanehisa et al., 2012; Ogata et al., 1999), EcoCyc (Keseler et al., 2013), STKE (*Science's Signal Transduction Knowledge Environment*) (Gough, 2002) i YMDB (*Yeast Metabolome Database*) (Jewison et al., 2012). Tot i que aquestes bases de dades són fonts útils de coneixement, encara queden moltes estructures de la xarxa bioquímica per identificar. S'han aplicat esforços en la recerca dels perfils d'expressió, en els quals s'utilitza per exemple l'anàlisi d'agrupacions (*cluster analysis*) per identificar els gens, proteïnes o metabòlits que s'expressen o es sintetitzen en relació a una funció coneguda. Tot i que aquesta anàlisi d'agrupacions dóna una idea de la correlació entre els compostos analitzats i els fenòmens biològics, aquesta anàlisi no revela la causalitat de les relacions de regulació genètica o de la xarxa bioquímica. S'han proposat diversos mètodes per tal de descobrir automàticament les relacions de regulació bioquímica utilitzant només les bases de dades generades a partir de l'anàlisi de perfils genòmics, proteòmics i metabolòmics (Kitano, 2002). Actualment encara que no és possible processar totes les dades disponibles en un sistema automàtic que permeti la detecció conjunta de les regulacions biològiques, aquestes aproximacions proporcionen informació molt útil, ja que permeten generar possibles hipòtesis sobre l'estructura de la xarxa bioquímica i fer-ne les conseqüents interpretacions. Una vegada s'ha comprès l'estructura d'una xarxa bioquímica determinada, s'ha d'investigar la dinàmica que aquesta segueix. Per a l'anàlisi dinàmica d'un sistema cel·lular determinat s'ha de crear un model, on primer és important considerar acuradament el propòsit de la construcció d'aquest model. Tant si es tracta d'adquirir un coneixement profund del comportament del sistema com si s'han de fer prediccions sobre comportaments complexos en resposta a estímuls, primer s'ha de definir el nivell d'abast i d'abstracció del model proposat.

En aquesta Tesi s'han generat grans conjunts de dades químiques multi- i megavariants (*big data*) a partir de la utilització de les tecnologies òmiques per a l'anàlisi de diferents sistemes, com són els perfils dels metabòlits del llevat, dels metabòlits secundaris del blat i els compostos orgànics acumulats als sediments marins durant milers d'anys. La quimiometria esdevé una eina molt important per a l'anàlisi d'aquestes grans quantitats de dades, la qual permet extreure'n el màxim d' informació.

2.1 LA METABOLÒMICA

Els metabòlits són compostos de baix pes molecular (<1000 Da) presents en les cèl·lules, teixits o fluids biològics, els quals són productes i intermediaris dels processos químics o enzimàtics del metabolisme cel·lular, i com a tals, proporcionen una lectura de l'estat de la cèl·lula. La metabolòmica és el camp de la ciència que caracteritza els metabòlits que es troben a les cèl·lules, teixits o biofluids d'un organisme (Nicholson et al., 1999). Hi ha certa controvèrsia en la utilització de les paraules metabolòmica i metabonòmica. Segons Nicholson i Lindon (Nicholson i Lindon, 2008), la diferenciació d'aquest dos termes és més filosòfica que tècnica. El terme metabonòmica va sorgir al 1999 (Nicholson et al., 1999) a partir de l'arrel grega *meta* (canvi) i *nomos* (lleis o regles) i té com a objectiu mesurar la resposta metabòlica global, dinàmica de sistemes vius a estímuls biològics o de manipulació genètica. En aquest cas, l'atenció es centra en la comprensió del canvi sistèmic, per exemple al llarg del temps, en els sistemes multicel·lulars complexos. D'altra banda, la metabolòmica pretén una descripció analítica més global de les mostres biològiques complexes i té com a objectiu caracteritzar i quantificar totes les molècules de baix pes molecular (metabòlits) en una mostra d'aquest tipus. A la pràctica, aquests dos termes s'utilitzen indistintament i sovint els procediments d'anàlisi i modelatge són semblants.

La metabolòmica és capaç de definir la resposta dels sistemes biològics a la influència genètica i ambiental a través de l'anàlisi del conjunt de metabòlits (metaboloma). Els metabòlits, o molècules petites, dins d'una cèl·lula, teixit, òrgan, fluid biològic o de tot l'organisme constitueixen el metaboloma (Miller, 2007). El metaboloma és el producte final de la expressió del genoma i dels efectes de la interacció de l'organisme estudiat amb el medi. De la mateixa manera que les altres ciències òmiques (com la genòmica, la transcriptòmica i la proteòmica), la metabolòmica és capaç de generar grans conjunts de dades molt completes sobre la mostra que s'analitza (Robertson, 2005). Tot i que la metabolòmica es coneix com un complement de les altres ciències òmiques

(genòmica, transcriptòmica i proteomica), aquesta pot oferir solucions a problemes que es troben en les altres ciències òmiques (Bilello, 2005). L'exposició d'un organisme a un factor d'estrès extern pot donar lloc a canvis en l'expressió gènica i a la producció de proteïnes. Tant el genoma com el proteoma estan sotmesos a una varietat de controls homeostàtics i mecanismes de retroalimentació i aquests canvis s'estenen a nivell del metaboloma. Conseqüentment, la metabolòmica és un indicador dels factors estressants externs més sensible que les altres tecnologies òmiques (Nicholson et al., 1999). La metabolòmica, per tant, té un gran potencial com a tècnica analítica sensible i ràpida, ja que en teoria és capaç d'elucidar les relacions que hi ha entre els nivells de concentració dels metabòlits i l'aplicació d'un factor d'estrès extern, com per exemple l'exposició a contaminants químics (Miller, 2007).

La metabolòmica ambiental és una subdisciplina de la metabolòmica relativament jove, però que està creixent de forma molt ràpida (Viant, 2008). En termes generals, les tècniques metabolòmiques s'utilitzen per caracteritzar les interaccions dels organismes amb el seu entorn. Això s'aconsegueix mitjançant la mesura dels múltiples metabòlits endògens en un teixit o fluid biològic dels organismes estudiats, la qual cosa pot proporcionar una bona descripció del seu fenotip metabòlic funcional. Per exemple, depenent dels metabòlits que es puguin detectar, la informació que s'obtingui pot revelar informació sobre l'estat energètic, reproductiu o oxidatiu d'un determinat organisme. Per tant, la metabolòmica és una eina molt apropiada per a estudiar les interaccions d'un organisme amb el medi ambient, tal i com són els cicles estacionals associats a la reproducció, els efectes de l'alimentació sobre el metabolisme i les adaptacions a canvis ambientals provocats per factors naturals o condicions estressants. Una conseqüència de la gran diversitat química dels components del metaboloma és la dificultat de la seva anàlisi exhaustiva a partir d'una tecnologia analítica única. Més avall es descriuen les tecnologies analítiques emprades en estudis òmics.

Els sistemes ambientals i organismes estan diàriament exposats a diferents agents químics i físics els quals poden afectar de forma significativa el seu metabolisme. Aquests agents presenten un rang molt ampli de concentracions les quals produeixen exposicions de diversa intensitat. En aquest context, els toxicòlegs ambientals han intentat durant molt de temps utilitzar les modificacions o els canvis en l'expressió de gens i de metabòlits específics per determinar les conseqüències de l'exposició en sistemes o organismes biològics. Els nous enfocaments metabolòmics permeten l'avaluació més refinada i sensible de l'exposició a través de l'avaluació

dels canvis dels perfils metabòlics en conjunt. Aquesta mesura permet la distinció dels components tòxics individuals, de l'exposició dels sistemes o organismes a mescles complexes i la detecció de les exposicions després que hagi transcorregut un temps determinat. Es tracta d'esbrinar les empremtes (*fingerprints*) de les respostes biològiques als agents externs tòxics.

El terme biomarcador s'ha aplicat àmpliament per descriure les espècies moleculars que reflecteixen els estats biològics com a conseqüència de les exposicions a agents externs, i a les malalties. En termes generals, un biomarcador es pot definir com un indicador d'un estat biològic donat. Els marcadors biològics o biomarcadors han estat i són àmpliament utilitzats en el camp de la química clínica aplicada a la medicina i toxicologia (Boudah et al., 2013), i el seu ús pot estendre's a la toxicologia ambiental. Els biomarcadors són indicadors interns mesurables d'alteracions moleculars i/o cel·lulars que poden aparèixer en un organisme durant o després de l'exposició a un tòxic (Bennett i Waters, 2000). Els biomarcadors permeten mesurar l'exposició a un agent tòxic i l'extensió de la seva resposta tòxica, i també permeten predir la resposta probable. Els biomarcadors han de ser compostos que siguin mesurables quantitativament, preferiblement de forma no invasiva en les mostres biològiques sensibles, i específics (Timbrell, 1998).

2.2 ANÀLISI DIRIGIDA I NO DIRIGIDA

Depenent de l'àmbit d'aplicació de l'estudi dels metabòlits d'interès, existeixen dos enfocaments principals: l'anàlisi dirigida (*target*) i l'anàlisi no dirigida (*untarget o non-target*) o global (Bouhifd et al., 2013; Goodacre et al., 2004; Patti et al., 2012). L'anàlisi dirigida té com a objectiu determinar les abundàncies relatives i les concentracions d'un grup seleccionat (específic) de metabòlits. En aquest cas, abans de portar a terme l'anàlisi dirigida s'ha de tenir la informació exacta sobre l'estructura dels metabòlits a analitzar. Generalment, l'enfocament dirigit consisteix en una anàlisi quantitativa, ja que implica la comparació dels analits amb compostos de referència, els quals poden estar marcats isotòpicament i afegits a la mostra o bé utilitzats com a estàndards externs per generar la corba de calibratge. L'enfocament dirigit està exhaustivament desenvolupat i ha estat àmpliament utilitzat (Canelas et al., 2009; Griffiths et al., 2010; Lu et al., 2008; Weljie et al., 2006). Un desavantatge important de l'enfocament dirigit és que s'han de conèixer els compostos d'interès a priori, per tant, aquest enfocament depèn de la disponibilitat dels compostos coneguts purs i no és aplicable per a la identificació de nous metabòlits. Tenint en compte que les matrius de les mostres ambientals o d'estudis metabòlics en la majoria dels casos són complexes,

els mètodes d'anàlisi tradicionals s'han enfocat específicament a l'anàlisi dirigida. Aquest enfocament té una bona sensibilitat, permet fer identificacions fiables, fer la quantificació dels compostos diana, i s'ha utilitzat de forma reeixida durant varies dècades. Però l'enfocament dirigit tradicional té un inconvenient significatiu, ja que aquells compostos que no hagin estat prèviament seleccionats es perdran. Això significa totes aquelles substàncies desconegudes o no seleccionades, encara que tinguin canvis de concentració elevats o que tinguin una gran influència en la diferenciació entre les mostres tractades d'un determinat sistema biològic o ambiental, no es podran detectar. Pot haver-hi casos en els quals els canvis de concentració dels compostos seleccionats a priori (*target*) no siguin suficients per explicar les diferències observades entre un conjunt de mostres determinades.

Per altra banda, l'anàlisi no dirigida és l'estudi exhaustiu de tots els metabòlits d'una mostra biològica, i representa una anàlisi global dels analits. L'enfocament no dirigit té com a objectiu determinar els canvis en els perfils metabòlics i a partir d'aquests canvis trobar quins són els diferents metabòlits responsables i fer-ne la seva detecció i quantificació relativa (característiques, *features*) de forma tan extensa com sigui possible (Robertson, 2005). Per exemple, en l'anàlisi no dirigida basada en espectrometria de masses (*Mass Spectrometry*, MS), les dades adquirides poden contenir centenars i fins i tot milers de senyals diferents de masses. Aquesta anàlisi global normalment requerirà d'una estratègia eficient d'emmagatzematge de les grans quantitats de dades generades i de la utilització de mètodes quimiomètrics per a la seva compressió, sense pèrdua dels metabòlits estadísticament més significatius. Si bé l'anàlisi no dirigida (*untarget*) és en general menys sensible que l'anàlisi dirigida (*target*), es pot aplicar a un conjunt de compostos més ampli i ofereix la possibilitat de detectar nous compostos biomarcadors inesperats. Una varietat important de l'anàlisi metabolòmica no dirigida és aquella que focalitza l'anàlisi només en aquells metabòlits que manifesten canvis importants en la seva concentració com a conseqüència dels efectes de les pertorbacions (factors estressants) estudiades.

Donada la sensibilitat, l'alt rendiment i els mínims requeriments de la mostra que es necessiten en l'anàlisi no dirigida, aquesta aproximació te un ampli ventall d'aplicacions biològiques. Malgrat la seva aparició relativament recent com a tecnologia global d'obtenció de perfils, la metabolòmica no dirigida està permetent augmentar la comprensió global del metabolisme cel·lular dels diferents organismes (Patti et al., 2012) i de la seva resposta a factors estressants externs.

2.3 QUIMIOMETRIA I METABOLÒMICA

L'anàlisi dels perfils de concentració metabòlica té els seus inicis ja a principis de la dècada de 1950, quan Williams i coautors van utilitzar dades de cromatografia en paper per demostrar que els llindars de gust i els patrons d'excreció variaven enormement d'individu a individu en persones. Els inicis de la metabolòmica estan ben documentats en el treball de Gates i Sweeley (Gates i Sweeley, 1978), els quals il·lustren l'evolució dels estudis sobre els perfils metabòlics. La terminologia de perfils metabòlics va ser introduïda inicialment per Horning i Horning al 1971 (Horning i Horning, 1971). El treball més recent de van der Greef i Smilde (Greef i Smilde, 2005) descriu més específicament l'evolució dels estudis de metabolòmica i quimiometria. Un avenç analític instrumental important ha estat el desenvolupament de tècniques de ionització suaus en el camp de l'espectrometria de masses (MS). A l'any 1983 Van der Greef i coautors van conjuntar aquesta metodologia analítica amb la metodologia quimiomètrica de reconeixement de patrons (*pattern recognition*) (van der Greef et al., 1983). En aquest treball s'estudiaven les diferències de gènere segons els perfils resultants d'espectrometria de masses de les mostres d'orina d'homes i de dones a partir d'una anàlisi de components principals (*Principal Component Analysis*, PCA). Quan van aparèixer les primeres interfícies per a la combinació de la cromatografia líquida amb l'espectrometria de masses (LC-MS), es van començar a mesurar els perfils metabòlics de plantes mitjançant LC-MS (Games et al., 1984). Els avenços tecnològics dels instruments analítics van proporcionar eines útils per a l'obtenció de perfils metabòlics de microorganismes i de fluids corporals mitjançant l'aplicació l'espectrometria de masses en combinació amb les metodologies de reconeixement de patrons durant la dècada dels 1980, tal i com es descriu en la revisió de Tas i van der Greef (Tas i van der Greef, 1994). Més tard, les millores en les tècniques de separació van seguir oferint un ampli ventall de possibilitats per al desenvolupament de la metabolòmica però, simultàniament, l'ús de la quimiometria va esdevenir cada vegada més important.

Nicholson va donar un impuls inicial molt important en l'aplicació de la ressonància magnètica nuclear (RMN) a l'estudi de la composició metabòlica multicomponent dels biofluids, de les cèl·lules i dels teixits (Nicholson et al., 1983; Nicholson i Wilson, 1989). Uns experiments relativament simples sobre biofluids permetien generar quantitats substancials de dades metabòliques que donaven informació detallada dels processos bioquímics en el conjunt d'organismes, i permetien també fer la investigació de les diferències entre espècies en termes de marcadors toxicològics (Nicholson et al., 1999).

Des de començaments del segle XXI, els avenços en les tecnologies instrumentals d'anàlisi i de tractament de dades han permès que la metabolòmica es desenvolupi i creixi ràpidament. Les plataformes experimentals que s'apliquen en els estudis actuals de metabolòmica són diverses, incloent la ressonància magnètica nuclear (RMN) (Mark A. Warne, 2000; Savorani et al., 2013; Smith et al., 2009), la cromatografia de líquids amb detecció per espectrometria de masses (LC-MS)(Allen et al., 2003; De Vos et al., 2007; Ducruix et al., 2008; Dunn et al., 2008), la cromatografia de gasos amb detecció per espectrometria de masses (GC-MS) (Gullberg et al., 2004; Koek et al., 2006; Lisec et al., 2006), l'electroforesi capil·lar amb detecció per espectrometria de masses (CE-MS) (Hirayama et al., 2009; Nevedomskaya et al., 2010; Soga et al., 2003) i l'espectroscòpia infraroja (Ellis i Goodacre, 2006). Un altre dels factors que contribueix en el ràpid creixement de la metabolòmica és l'ampli camp d'aplicacions que té. Aquestes aplicacions abasten un gran nombre d'àrees de recerca com la nutrició (Gibbons et al., 2015; Orešič, 2009), el descobriment de fàrmacs (Kell i Goodacre, 2014; Robertson i Frevert, 2013), l'estudi de malalties (Kaddurah-Daouk i Krishnan, 2008; Mamas et al., 2011), la biologia de les plantes (Tolstikov i Fiehn, 2002; Tolstikov et al., 2003), dels mamífers (Dunn et al., 2011), dels microorganismes (Winder et al., 2011) i del medi ambient en general (Viant i Sommer, 2013; Viant, 2008, 2009). La selecció de la plataforma analítica adequada pels estudis metabolòmics és important i en gran mesura ve determinada pel sistema biològic estudiat i per les qüestions científiques formulades.

Les tecnologies analítiques utilitzades en estudis metabolòmics generen grans quantitats de dades. Tenint en compte que l'objectiu primordial de la metabolòmica és detectar els canvis en les respostes dels sistemes relacionades amb la seva variació genètica i amb els canvis ambientals, és important separar la variació biològica d'interès de les fonts de variabilitat natural, les quals poden emmascarar els resultats biològics buscats com a conseqüència d'aquests canvis. Per tant, el processament i l'anàlisi de les dades és un pas essencial en la capacitat de processar i interpretar correctament les dades metabolòmiques.

Els mètodes de projecció i regressió lineal són un conjunt de mètodes eficients i útils per a l'anàlisi i modelatge de dades complexes, tals com les que es presenten en els estudis metabolòmics. Un dels mètodes més àmpliament utilitzat per a l'exploració de dades en metabolòmica és l'anàlisi de components principals (*Principal Component Analysis*, PCA) (Smilde et al., 2004). Com que molts dels metabòlits detectats estan bioquímicament relacionats, la variació que s'observa en les

seves concentracions estarà correlacionada. Aquesta propietat és utilitzada en el mètode PCA per construir noves variables que són combinacions de les variables originals i que expressen una mateix font de variació. Per tant, un número limitat d'aquestes variables o components principals seran capaços de descriure la variabilitat original de les dades. La inspecció visual dels valors dels components principals podrà revelar informació important sobre la naturalesa de les mostres (per exemple la seva agrupació) i de les variables originals, estiguin aquestes correlacionades o no. El nombre de treballs que utilitzen el PCA com a eina exploratòria en estudis metabolòmics és molt extens. Per exemple, un de recent és el de Fernández-Varela i coautors (Fernández-Varela et al., 2015) que utilitza el PCA per comparar els perfils metabòlics obtinguts mitjançant GC-MS de grups de poliquets marins exposats i no exposats al petroli cru, i així poder esbrinar possibles patrons metabòlics comuns. Un altre exemple és l'estudi de Collins i coautors, on s'utilitza el PCA per comparar els canvis en els espectres obtinguts per RMN d'una proteïna vinculada a un conjunt d'ARN (àcid ribonucleic) degenerats per tal de definir l'especificitat de les nucleobases (Collins et al., 2015).

La tècnica d'anàlisi de la variància (*ANalisis Of VAriance*, ANOVA) s'utilitza freqüentment per a la diferenciació de conjunts de dades diversos, per exemple, per avaluar la importància dels efectes d'un determinat tractament i avaluar la seva significació estadística (valor *p*, *p-value*). No obstant això, aquesta anàlisi *per se* no proporciona informació sobre quin és el factor causant de l'efecte observat. En els estudis metabolòmics no dirigits (*untarget*) es pretén tractar tot el conjunt de dades obtingut a partir de l'anàlisi instrumental (per exemple a partir dels cromatogrames) de les mostres. I per tal de millorar la interpretació dels resultats de l'ANOVA d'aquests grans conjunts de dades, s'aplica un PCA als components de variància individuals prèviament separats. Aquest ha estat el cas del treball de Stanimirova i coautors, on s'apliquen els avantatges de l'ANOVA i del PCA conjuntament. D'aquesta manera es va facilitar l'avaluació dels efectes d'un determinat tractament contra el càncer a partir d' espectres de fluorescència de raig X (Stanimirova et al., 2011). En un altre treball més recent de Stanimirova i coautors, s'aplica el procediment ASCA (ANOVA-*Simultanoeus Component Analysis*, vegeu secció 2.5.5), a un conjunt de dades adquirides per cromatografia líquida d'alta resolució (HPLC) amb detecció de díodes en línia (*Diode Array Detector*, DAD). L'objectiu del treball va ser detectar la influència de diferents factors (productivitat de la temporada de cultiu, grau de qualitat i procés de pasteurització) considerats en el cultiu i tractament de rooibos mitjançant la variació del contingut fenòlic de la planta. Un treball recent on també s'ha utilitzat la metodologia ASCA és el de Ly-Verdú i

coautors, on es determinen els canvis en els perfils metabòlics mitjançant cromatografia de gasos bidimensional acoblada a un espectròmetre de masses de temps de vol (GCxGC-ToF-MS). La finalitat de l'estudi va ser investigar les fonts de variabilitat dels perfils metabòlics dels extractes de teixit de fetge de ratolins que van estar sotmesos a diferents factors experimentals (temps, dieta, individu) (Ly-Verdu et al., 2015).

Tal i com s'ha esmentat anteriorment, l'augment del nombre de tècniques analítiques ha crescut molt i el seu ràpid desenvolupament tecnològic ha possibilitat la resolució de nous problemes químics i bioquímics. Aquestes noves aproximacions analítiques generen grans conjunts de dades (*big data*), l'anàlisi de les quals permet, cada vegada més freqüentment, proporcionar millors interpretacions dels fenòmens estudiats. En molts d'aquests estudis s'ha de tractar conjunts de dades, que contenen un gran nombre de variables en comparació amb el nombre de mostres. En aquests casos, la selecció d'un grup menor de variables que concentren el màxim d'informació i siguin importants per al problema biològic que s'estudia és important, i pot fer possible l'explicació del fenomen i la seva interpretació biològica. Per tant, els procediments per a la selecció de variables són importants i moltes vegades necessaris per tal d'eliminar les característiques no desitjades, com per exemple el soroll o l'error instrumental o experimental que no estan correlacionats amb el fenomen estudiat. Aquests procediments també permeten fer una anàlisi més concisa sobre la relació que existeix entre un nombre reduït de variables explicatives i la resposta biològica estudiada.

Els models basats en variables latents, com l'anàlisi de components principals (*Principal Component Analysis*, PCA) i la regressió per mínims quadrats parcials (*Partial Least Squares*, PLS) són eines molt útils per fer les representacions de les variacions dels perfils metabòlics. Aquests models serveixen com a eines de diagnòstic per a la detecció de biomarcadors i per trobar patrons de variació comuns en dades complexes. Per tal de relacionar diferents variables experimentals mesurades amb un vector resposta, freqüentment s'utilitza la regressió per mínims quadrats parcials (*Partial Least Squares Regression*, PLSR) (Mehmood et al., 2012). Aquest és un dels mètodes supervisats més populars en quimiometria per construir models de predicció inversa que permetin trobar el millor subespai lineal comú de les variables explicatives (**X**) i de les variables a predir (**y**). Smolinska i coautors (Smolinska et al., 2012), per exemple, han utilitzat PCA i PLS per a l'anàlisi de perfils de dades de RMN (ressonància magnètica nuclear) en la metabolòmica del líquid cefaloraquidi (*CerebroSpinal Fluid*, CSF) per a la possible detecció de

biomarcadors de l'esclerosi múltiple. El mètode PLSR conjuntament amb altres metodologies s'estan utilitzant àmpliament en el camp de la metabolòmica per a la selecció de variables rellevants.

2.4 SISTEMES EXPERIMENTALS ESTUDIATS EN AQUESTA TESI

2.4.1 EL BLAT (*Triticum aestivum* L.)

El blat (*Triticum aestivum*) és un cereal, forma part del regne *Plantae*, pertany a la divisió *Magnoliophyta*, classe *Liliopsida*, ordre *Poales*, família Gramínies (*Poaceae*) i gènere *Triticum*. Existeixen al voltant de 30 tipus de blat amb suficients diferències genètiques entre ells com per ser considerats espècies o subespècies diferents (Mac Key, 2005).

El blat ha format i segueix formant part del desenvolupament econòmic i cultural de l'home, essent un dels cultius més estesos mundialment. En aquests tipus de cultius extensius, el control de les males herbes juntament amb el de plagues i altres malalties, és un aspecte de gran importància a tenir en compte per tal d'obtenir la màxima producció possible d'aquest cereal. Les males herbes són les plantes que creixen en un indret on no han estat cultivades, i interfereixen negativament en l'activitat agrícola. Aquestes poden competir amb els cultius per l'aigua, la llum, els nutrients del sòl, l'espai i el CO_2. Un altre dels problemes associats a les males herbes és que poden servir d'hostes a malalties i insectes o produir substàncies químiques que puguin ser tòxiques als humans o altres plantes (Janick, 1979).

Des dels seus inicis, l'agricultura s'ha orientat cap al mercat urbà amb l'objectiu d'incrementar el rendiment (volum produït per hectàrea) i la productivitat (volum produït per unitat de treball). L'alimentació d'una població en constant creixement ha portat als professionals del sector agrícola a usar tècniques que proporcionin una gran producció. La necessitat creixent de capital i terreny per guanyar en competitivitat, l'alteració de l'ecosistema amb la incorporació de noves tecnologies i mètodes industrials i l'augment de la dependència en l'ús de solucions tecnològiques, fertilitzants i fitosanitaris ha derivat en l'esgotament i contaminació dels recursos naturals. Tot i això, la contaminació ambiental no ha estat l'única problemàtica envers el control de les males herbes. Des de que es van començar introduir els herbicides hormonals l'any 1946 (Troyer, 2001), les empreses agroquímiques han desenvolupat i introduït al mercat una amplia gamma d'herbicides

selectius. No obstant, la natura ha reaccionat envers a aquests agents externs incorporant les substàncies químiques utilitzades com a herbicides als seus sistemes biològics, i d'aquesta forma apareixen les males herbes resistents a aquests herbicides. La resistència als herbicides és un efecte secundari no desitjat que es produeix després de l'ús reiterat d'un determinat herbicida, pel qual una població d'una determinada mala herba deixa de ser controlada amb la mateixa eficàcia per un herbicida.

Tenint en compte l'ús abusiu de productes químics en els sistemes agrícoles, l'al·lelopatia és una de les alternatives per al control de les males herbes. La Societat Internacional d'Al·lelopatia defineix el terme al·lelopatia com a qualsevol procés on estan involucrats metabòlits secundaris (al·leloquímics) produïts per plantes, microorganismes, virus i fongs que influeixen en el creixement i desenvolupament dels sistemes agrícoles i biològics (excloent els animals), incloent els efectes positius i negatius (Torres et al., 1996). Tots els òrgans vegetals contenen quantitats variables de substàncies potencialment al·lelopàtiques, que poden ésser alliberades al medi ambient per mitjà de l'exsudació de les arrels, lixiviació, volatilització i descomposició dels residus de les plantes en el sòl (Willis, 1993). Aquestes substàncies presenten un interès particular perquè proporcionen una via d'alliberament directa de toxines que poden influir sobre la composició de la població microbiana (Woods, 1960). Tant la producció com l'alliberament dels compostos al·leloquímics, el transport, la transformació en el sòl i l'acceptació per part de la planta receptora, depenen de les condicions mediambientals (Einhellig, 1996).

Actualment, els compostos amb activitat al·lelopàtica es consideren com una font d'estructures base a partir de les quals s'originen potencials herbicides. Entre els productes proposats com a molècules base per al desenvolupament d'herbicides es troben les benzoquinones (Netzly et al., 1988). Dins la gran diversitat de molècules procedents del metabolisme secundari de les plantes, els principals compostos amb propietats al·leloquímiques en les gramínies pertanyen a la família de les benzoxazinones. Es troben per exemple en el blat (*Triticum aestivum*), en el blat de moro (*Zea mays*) o en el centè (*Secale cereale*). Els factors ambientals i climàtics, tals com la quantitat de nutrients al sòl, el tipus de fertilitzant aplicat, la temperatura, la radiació solar i l'ús de productes químics pot influenciar en el contingut de compostos al·leloquímics.

Dins les benzoxazinones es poden distingir: els àcids hidroxàmics, les lactames, els metil derivats i les benzoxazoliniones. Les estructures de totes les benzoxazinones es caracteritzen per tenir un

grup hemiacetal cíclic en combinació amb un d'àcid hidroxàmic cíclic o de lactama cíclica. Inicialment, l'evidència de la secreció dels àcids hidroxàmics des de les plantes vives al sòl només havia estat descrita en el sègol (Pérez i Ormenoñuñez, 1991). Posteriorment Wu i coautors (Wu et al., 2000a; Wu et al., 2000b) van descriure que algunes varietats de blat també eren capaces d'exsudar compostos al·lelopàtics. L'alliberament d'al·leloquímics al sòl no només es veu afectat per les seves propietats físico-químiques, sinó també per l'adsorció, desorció i transport en el sòl i per el propi metabolisme dels compostos al·leloquímics. De la mateixa manera, l'activitat fitotòxica de les substàncies alliberades per les plantes es veu influenciada per factors mediambientals, tals com les propietats fisico-químiques del sòl (Sène et al., 2000), l'activació (Nair et al., 1990) i inactivació microbiològica (Laterra i Bazzalo, 1999; Sène et al., 2000) i el nivell de nutrients del sòl (Harper i Lynch, 1982).

2.4.2 EL LLEVAT (*Saccharomyces cerevisiae*)

Els llevats (*Saccharomyces cerevisiae*) són eucariotes unicel·lulars que pertanyen al grup dels fongs. Tot i que hi ha diversos llevats que són objecte continuat d'estudi en biologia, l'espècie *S.cerevisiae*, coneguda també com a llevat de panificació o de cerveseria, constitueix un dels models de recerca més importants en biologia. *S.cerevisiae* forma part del regne de *Fungi*, pertany al fílum *Ascomycota,* classe *Saccharomycetes,* l'ordre *Saccharomycetales* i de la família *Saccharomycetaceae*. Aquesta classificació es basa en les característiques de les cèl·lules, les ascòspores i les colònies que formen així com també la fisiologia cel·lular.

El llevat és un organisme format per una sola cèl·lula i aquesta cèl·lula té una estructura eucariota, és a dir, en la qual el material genètic està rodejat per una membrana constituint el nucli. Els llevats estan àmpliament dispersats a la natura en una gran varietat d'hàbitats, normalment es troben en les fulles de plantes, en flors, en fruites i al terra. El llevat va ser introduït com a organisme experimental als inicis del s. XX (Richards, 1934; Richards i Haynes, 1932), i des d'aleshores ha estat freqüentment utilitzat com a organisme model en estudis d'investigació d'espècies eucariotes superiors (Castrillo i Oliver, 2006; Sherman, 2002) en el camp de la biologia cel·lular i molecular. Es tracta d'un organisme que reuneix la complexitat d'una cèl·lula eucariota i la facilitat de manipulació d'un organisme unicel·lular. Una altra característica important és que treballar amb el llevat és relativament fàcil, ja que es pot replicar ràpidament en medis de cultiu poc complexos i econòmics, i es pot manipular genèticament. De fet, el llevat va ser el primer

organisme eucariòtic del qual es va seqüenciar el genoma complet l'any 1996 (Goffeau et al., 1996). Per altra banda el llevat és considerat un organisme no patogen (*Generally Regarded As Safe*, GRAS) i pot ser manipulat sense precaucions especials. I a causa del seu historial com a organisme d'interès industrial i tecnològic, es disposa d'un gran nombre de fons de coneixements bioquímics i genètics.

Normalment les cèl·lules haploides de *S.cerevisiae* són esferoides de 4-5 µm de diàmetre, i les diploides són el·lipsoides de 6 x 8 µm. El llevat està recobert d'una paret gruixuda (100 a 200 nm) la qual representa el 20% de la massa cel·lular i està composta principalment de polisacàrids. La membrana plasmàtica és una bicapa lipídica amb múltiples proteïnes inserides. El vacúol és un orgànul molt important de l'organisme *S.cerevisiae* i pot arribar a ocupar un volum substancial de la cèl·lula, on s'hi dóna la degradació no específica de proteïnes per part d'una varietat de proteases. A més, serveix com a lloc d'emmagatzematge d'aminoàcids, d'alguns metalls i ions, i col·labora en el procés d'osmoregulació. Els llevats contenen peroxisomes, on tenen lloc diverses funcions metabòliques, com és l'oxidació de fonts de carboni i nitrogen. En el mitocondri del llevat, que conté casi 86000 parells de bases d'ADN (àcid desoxiribonucleic), es codifiquen diverses proteïnes de la cadena de transport electrònic del mitocondri mateix. El nucli de *S.cerevisiae* està separat del citosol per una doble membrana que, contràriament al que ocorre en molts altres eucariotes, no es desfà durant la mitosi. A més dels cromosomes, en el nucli es poden trobar el plasmidi, ARN (àcid ribonucleic) i ADN lineal de doble cadena i els elements *Ty*, que són equivalents als retrotransposons que es troben en altres tipus de cèl·lules eucariotes (Ariño, 2011).

Les cèl·lules de llevat es multipliquen ràpidament per gemmació, una forma asimetria de reproducció asexual, però la reproducció sexual també es pot donar en determinades condicions. Com qualsevol organisme viu, *S.cerevisiae* presenta uns paràmetres òptims de creixement, en referència a requeriments nutricionals i condicions mediambientals. Quan les cèl·lules de llevat es cultiven en medis rics en fonts de carboni, com la glucosa, i en absència d'oxigen es produeix la fermentació alcohòlica. Aquesta forma de creixement del llevat s'explota per a l'elaboració cervesa, vi i altres begudes alcohòliques.

La Metabolòmica del llevat

S'ha publicat molts estudis que analitzen el perfil metabòlic del llevat utilitzant diferents tècniques analítiques. Entre aquests estudis, un dels més destacats és el de Castrillo i coautors (Castrillo et al., 2003), en el que desenvolupen un protocol per a l'anàlisi dels metabòlits intracel·lulars del llevat mitjançant l'espectrometria de masses utilitzant infusió directa per elecrospray. Aquest mètode és capaç de detectar metabòlits de diferents categories funcionals (intermediaris glucolítics, nucleòtids, nucleòtids de piridina, aminoàcids i compostos orgànics). En aquest treball els metabòlits són extrets directament seguint les directrius del protocol desenvolupat prèviament per Gonzalez i coautors (Gonzalez et al., 1997) utilitzant etanol a ebullició. Els treball de Villas-Bôas i coautors (Villas-Bôas et al., 2005) compara diversos mètodes d'extracció dels metabòlits intracel·lulars del llevat (aminoàcids, àcids orgànics, àcids grassos, nucleòtids, pèptids, sucres, polialcohols i sucres fosfats) mitjançant la plataforma GC-MS. En aquest cas s'argumenta que aquesta plataforma no és la més adequada per a l'anàlisi de les diferents classes de metabòlits. Per altra banda, en el treball de Canelas i coautors (Canelas et al., 2009) també es comparen les tècniques d'extracció per a l'extracció de 44 metabòlits intracel·lulars del llevat (intermediaris fosforilats, aminoàcids, àcids orgànics i nucleòtids). En aquest estudi s'utilitzen les plataformes GC-MS i LC-MS/MS per a l'anàlisi dirigida dels metabòlits.

Metabolòmica dels cultius de llevat en condicions d'estrès de temperatura

S'han publicat diversos treballs en els quals s'estudien els perfils metabòlics del llevat en resposta a diferents tipus d'estrès sobre el seu cultiu, molts d'ells relacionats amb la limitació de nutrients en el cultiu del llevat. Castrillo i coautors (Castrillo et al., 2007) van dur a terme un primer estudi exhaustiu de la biologia de sistemes sobre el control de la taxa de creixement d'una cèl·lula eucariota. En aquest treball es va estudiar el transcriptoma, el proteoma i el metaboloma del llevat per a veure quins efectes tenia la limitació de nutrients en el cultiu de llevat sobre la seva taxa de creixement. Mitjançant anàlisis estadístiques i tècniques multivariants (PCA, test *t*, ANOVA/ANCOVA) es van determinar els canvis correlacionats amb les quatre condicions de nutrients limitants (glucosa, amoni, fosfat i sulfat). En aquest estudi es conclou que la regulació del flux metabòlic és un procés dinàmic, amb la intervenció del control de la transcripció, la traducció i l'ajust dels nivells dels metabòlits. Aquesta regulació del flux metabòlic sembla ser una de les

fonts d'adaptabilitat intrínseca de la cèl·lula eucariota, que permet la seva adaptació als canvis ambientals a curt i a llarg termini.

Més endavant, en el treball de Boer i coautors (Boer et al., 2010) es van mesurar també, mitjançant una anàlisi dirigida LC-MS/MS, els metabòlits intracel·lulars de cultius de llevat en diferents condicions de nutrients (carboni, nitrogen, sofre, leucina i uracil) limitants. Les concentracions dels metabòlits van resultar ser molt sensibles al tipus de nutrient limitant. La restricció de nitrogen (amoni) i de carboni (glucosa) es va caracteritzar per nivells baixos de concentracions d'aminoàcids intracel·lulars i per nivells alts de concentracions de nucleòtids, mentre que la restricció de fòsfor (fosfat) va donar lloc a una situació inversa. Particularment, les respostes en la concentració de metabòlits van ser protagonitzades per aquells estretament vinculats amb els nutrients limitants. Per exemple la glutamina conduïa a la limitació de nitrogen, l'ATP (trifosfat d'adenosina) a la limitació de fòsfor, i el piruvat a la limitació de carboni.

La temperatura és un paràmetre clau per al creixement i el metabolisme del llevat. La temperatura de creixement òptima del llevat és d'entre 25 i 30ºC. Els canvis de temperatura ambiental són alguns dels desafiaments amb els quals han de fer front tots els organismes vius. Quan el llevat s'exposa a temperatures més altes o més baixes, aquest desenvolupa mecanismes que li permeten adaptar-se als canvis de temperatura. S'han publicat diversos estudis sobre els canvis de temperatura dels cultius de llevat des del punt de vista de biologia de sistemes, que integren a la genòmica, la proteòmica i la metabolòmica (Groušl et al., 2009; Hottiger et al., 1987; Morano et al., 2012; Petti et al., 2011). En el treball de Strassburg i coautors (Strassburg et al., 2010), s'estudien els perfils moleculars que engloben la transcriptòmica i la metabolòmica de mostres de cultiu de llevat exposades a temperatures ambientals subòptimes (10ºC i 37ºC), per tal d'esbrinar els seus mecanismes de resposta molecular. En aquest treball es determina que entre els metabòlits analitzats, la trehalosa és la que dona un canvi de resposta més forta a l'estrès de temperatura, tal i com ja s'havia demostrat en estudis anteriors com el de Hottiger i coautors (Hottiger et al., 1987). En relació també al canvi de temperatura del treball de Strassburg i coautors, en els dos règims de temperatura diferents considerats, els patrons de resposta molecular són molt diferents tant a nivell metabòlic com a nivell de transcripció.

En aquesta Tesi, s'ha portat a terme l'estudi de cultius de llevat sotmesos a condicions d'estrès de temperatura (xoc tèrmic) en l'Article 2. En aquest estudi, s'investiguen els perfils metabòlics de

S.cerevisiae obtinguts mitjançant l'anàlisi per LC-MS a través de tècniques quimiomètriques multivariants.

El llevat i el vi. Metabolòmica dels cultius de llevat en condicions d'estrès de coure

Les primeres evidències de la producció de vi es remunten al VI mil·lenni AC en l'antiga Mesopotàmia. Durant mil·lennis, el procés de vinificació es va considerar una propietat intrínseca del most, fins que l'any 1983 Louis Pasteur va demostrar la seva dependència del llevat. Aquest descobriment marca el començament de l'era biotecnològica (Mortimer i Polsinelli, 1999; Pretorius, 2000). En la fermentació del vi, el llevat pot tenir el seu origen en el raïm o a la bodega (material que entra en contacte amb el most). Les fermentacions espontànies són aquelles que es produeixen de forma natural, és a dir, les realitzen els llevats provinents del raïm i del material de bodega, sense cap tipus d'inoculació externa. *S.cerevisiae* és l'única espècie de la microflora present al most capaç de finalitzar la fermentació vínica. Degut a la baixa representació d'aquesta espècie al most inicial, no és estrany que el procés tardi diversos dies en iniciar-se, i que, de vegades, aquest inici no es produeixi (Querol et al., 1994). Per aquesta raó, i al 1890, Müller-Thungau ja va introduir el concepte d'inocular cultius purs de llevat per iniciar les fermentacions víniques (Pretorius, 2000). A nivell comercial els primers cultius purs de llevats vínics en forma de llevat actiu sec no van aparèixer al mercat fins l'any 1965, i el seu ús no es va generalitzar fins als anys 80. Actualment existeixen dues tendències en l'enologia, aquella que confia en la microflora present al raïm i bodegues per realitzar la fermentació de forma espontània, i aquella que prefereix una fermentació més controlada i segura utilitzant inòculs comercials.

La qualitat del vi depèn en gran mesura de la qualitat del raïm. Per tal d'obtenir vins d'alta qualitat és necessari l'ús de raïms sans en l'etapa de maduració i per aquesta raó s'ha de tenir especial cura en la prevenció d'atacs de paràsits de la vinya. Molts d'aquests paràsits són d'origen animal (insectes o àcars) o d'origen vegetal (fongs paràsits). Els agricultors combaten les malalties del raïm i les plagues d'insectes aplicant pesticides, els quals es poden trobar durant la collita del raïm. La presència de pesticides depèn de les característiques químiques dels ingredients actius, així com dels processos de fotodegradació, termodegradació, co-destil·lació i degradació enzimàtica. Els residus de pesticides dels raïms poden ser transferits al most i això pot influir en la selecció i desenvolupament de les soques de llevat. D'altra banda, els llevats també poden influir en els nivells de pesticides en el vi mitjançant la seva reducció o adsorció en les lies del vi. Durant el

procés de fermentació, els llevats poden causar la desaparició dels residus de pesticides degut a la seva degradació o absorció al final de la fermentació, quan els llevats es depositen en forma de lies.

Els fungicides que contenen coure (Cu(II)) es van introduir ja fa més d'un segle i són àmpliament utilitzats en l'agricultura europea. El sulfat de coure és un producte fitosanitari que s'utilitza per al control dels fongs en els raïms. El ió Cu(II) és el responsable de provocar una acció contra els fongs o bactèries. El principal objectiu del coure és el míldiu provocat pel fong *Plasmopara viticola* ja que el coure pertorba les activitats respiratòries, enzimàtiques i membranàcies d'aquest fong. L'ió cúpric és molt estable ja que no el degrada ni la calor ni la llum, i això té vàries conseqüències. La permanència del Cu(II) és molt alta mentre no estigui lixiviat, doncs si aquest s'acumula és una font de possible contaminació i toxicitat. Alts nivells de Cu(II) en el raïm durant la fermentació poden causar l'alentiment en la fermentació o fins i tot que aquesta s'aturi totalment (Brandolini et al., 2002).

En aquesta Tesi, s'ha portat a terme el cultiu del llevat exposat a diferents concentracions de Cu(II) (control, 1mM, 3mM i 6mM) en l'Article 3. En aquest treball s'estudien els perfils metabòlics de *S.cerevisiae* obtinguts mitjançant l'anàlisi per LC-MS a través de tècniques quimiomètriques multivariants i la utilització de diferents mètodes de selecció de variables.

2.4.3 ESTUDI DE MARCADORS PALEOCLIMÀTICS A PARTIR DELS SEDIMENTS MARINS

L'evolució del clima en el futur és un aspecte de gran preocupació social i és de gran interès en l'agenda internacional de la majoria dels països desenvolupats. L'IPCC (*Intergovernamental Panel on Climate Change*) 2014 (Ipcc, 2014a, b) constata que la influència humana en el sistema climàtic és evident, i que les recents emissions antropogèniques de gasos d'efecte hivernacle són les més altes de la història. L'escalfament del sistema climàtic és inequívoc, i des de la dècada de 1950, molts dels canvis observats no tenen precedents en els darrers mil·lennis. Les emissions de gasos antropogènics d'efecte hivernacle han augmentat des de l'era preindustrial, i han estat imputats en gran mesura al creixement econòmic i demogràfic. La mitjana global de la temperatura de l'atmosfera i dels oceans s'ha incrementat considerablement, les quantitats de neu i de gel han disminuït i el nivell del mar ha pujat. Aquests canvis recents del clima han tingut impactes

generalitzats sobre els sistemes humans i naturals (Ipcc, 2014a, b). Conseqüentment hi ha un gran interès per tal d'estudiar fins a quin punt l'home pot influir en l'estat del clima i com això pot afectar al benestar humà i als ecosistemes. En la recerca de la comprensió del comportament del sistema climàtic, el paleoclima s'ha convertit en una eina d'investigació que focalitza l'estudi del clima en el passat per donar a conèixer els seus cicles i patrons de variabilitat natural, i les causes i conseqüències dels canvis en els seus diferents components.

El clima té una variabilitat intrínseca, és a dir, sempre ha estat canviant des de la formació del nostre planeta fa 4600 milions d'anys (Crowley i North, 1988). Des dels seus inicis el planeta ha experimentat canvis dràstics de clima, començant per un considerable refredament després de les temperatures extremadament altes en el seu inici, el creixement i la retirada del capes de gel, així com els períodes d'altes concentracions de gasos d'efecte hivernacle (*GreenHouse Gas*, GHG) a l'atmosfera fa desenes de milions d'anys (Ipcc, 2007). Més recentment, durant el Quaternari, és a dir, en el període que engloba els darrers 1.8 milions d'anys, els canvis climàtics a llarg termini estan dominats pels cicles glacials-interglacials relacionats amb els canvis en els paràmetres orbitals i les respostes no lineals corresponents, com els canvis en el volum global acumulat del gel i en la circulació termohalina (Hays et al., 1976; Paillard, 2001). Els canvis de la temperatura i de la concentració de gasos d'efecte hivernacle centren una intensa investigació a l'actualitat, especialment els de la concentració dels gasos de nitrogen i carboni, els canvis de productivitat dels oceans després dels cicles glacials-interglacials i l'acoblament dels canvis existents entre la terra i l'oceà de les diferents regions del planeta.

La majoria de registres instrumentals de temperatura i de les altres variables climàtiques, tals com la precipitació o la força del vent, engloben només els dos últims segles (Ipcc, 2007). La reconstrucció dels canvis ambientals i climàtics del període preindustrial requereix de l'ús de mètodes indirectes o de marcadors climàtics (*proxies*). Entre aquests marcadors es troben alguns compostos orgànics (principalment lípids) que tenen orígens biosintètics particulars i que són resistents a la degradació, i que per tant es poden preservar en sediments marins o lacustres i en roques sedimentàries durant milers o fins i tot milions d'anys (Rosell-Melé, 2003). Existeixen diferents organismes que sintetitzen aquestes molècules orgàniques, les variacions de concentració de les quals poden donar informació sobre els esdeveniments ambientals i climàtics del passat.

Els marcadors paleoclimàtics més valuosos són taxonòmicament específics, és a dir, es poden atribuir a un determinat grup definit d'organismes, i a través de la seva composició química o configuració estèrica permeten fer una observació indirecta de l'entorn ambiental on van ser sintetitzats. Per tant, donen informació de la variabilitat del clima passat i d'altres paràmetres ambientals. Tant les alquenones derivades de les algues *Haptopyceae* (Brassell et al., 1986b; Prahl i Wakeham, 1987) com els GDGTs (*Glycerol Dialkyl Glycerol Tetraethers*) derivats d'*Arquea* (Schouten et al., 2002) són marcadors paleoclimàtics utilitzats per a la reconstrucció de les temperatures superficials del mar (*Sea Surface Temperature*, SST) i de les temperatures superficials de llacs (*Lake Surface Temperature*, LST).

Les alequenones són cetones de cadena llarga de 37 a 40 àtoms (C_{37}, C_{38}, C_{39}, C_{40}) i de 2 a 4 dobles enllaços o d'insaturació (per exemple, $C_{32:2}$, $C_{37:3}$, $C_{37:4}$). Les alquenones són compostos ubics en els oceans de tot el món i es troben normalment als sediments marins juntament amb els alquil alquenoats d'alquil. Les alquenones són biosintetitzades per algunes algues de la classe *Prymnesiophyceae* (Marlowe et al., 1990; Marlowe et al., 1984). Els principals productors d'alquenones als oceans són les espècies de coccolitofòrids *Emiliana huxleyi* i *Gephyrocapsae oceanica* (Conte et al., 1995; Volkman et al., 1980). La rellevància particular de les alquenones rau en la relació que existeix entre la proporció de patrons d'insaturació que presenten i la temperatura del medi on aquests compostos han estat sintetitzats. El nombre de dobles enllaços que conté la cadena d'hidrocarbonis de les alquenones està directament correlacionada amb la temperatura de l'aigua on viuen les algues que les sintetitzen (Marlowe et al., 1984; Prahl i Wakeham, 1987). En aigües fredes, es sintetitzen més alquenones $C_{37:3}$ i menys $C_{32:2}$, mentre que en aigües més càlides es produeix el fenomen contrari (Brassell et al., 1986a; Brassell et al., 1986b). La distribució d'alquenones dipositades al fons de les aigües marines són un registre fiable de les temperatures corresponents a les zones de màxima activitat fotosintètica (Bentaleb et al., 1999). Així doncs, les alquenones es consideren com una eina potencial molt útil per estimar les temperatures del passat de les aigües superficials (Brassell et al., 1986b; Prahl et al., 1989). Conseqüentment, Brassell i coautors (Brassell et al., 1986b) van proposar la utilització de l'índex d'insaturació d'alquenones $U^{k'}_{37}$ ($C_{37:2}$-$C_{37:4}$/($C_{37:2}$ + $C_{37:3}$+ $C_{37:4}$)), el qual es simplifica quan l'alquenona $C_{37:4}$ es troba per sota els nivells de detecció ($C_{37:2}$/($C_{37:2}$ + $C_{37:3}$)) (Prahl i Wakeham, 1987).

Prahl i coautors (Prahl i Wakeham, 1987) van cultivar una sola soca d'*Emiliana huxleyi* i la van incubar en un rang de temperatures. Aleshores es va analitzar la distribució relativa d'alquenones

en mostres d'aquest experiment de cultiu, i es va trobar l'existència d'una bona relació lineal entre el nombre d'enllaços dobles de les alquenones i la temperatura de l'aigua. Aquesta relació va ser validada comparant els resultats del laboratori amb les anàlisis d'alquenones en mostres de partícules recuperades de la columna d'aigua de l'oceà, i mesurant la temperatura corresponent *in situ*. La relació trobada a partir d'aquest estudi (Müller et al., 1998; Prahl et al., 1988) és (equació 2.1):

$$U^{k'}_{37} = 0.033 \times SST + 0.044 \quad (r^2 = 0.96;\ n = 370) \qquad \textbf{Equació 2.1}$$

Les etapes de treball (*workflow*) en estudis metabolòmics descrits en la següent secció (2.5) són extrapolables a altres àmbits d'estudi en química analítica. En aquest sentit i en el context d'aquesta Tesi s'ha investigat un grup de dades paleoclimàtiques seguint les mateixes etapes de treball que en els estudis metabolòmics (Article 2 i Article 3). S'han analitzat els compostos orgànics acumulats al testimoni de sediment IODP-U1318C extret de la badia de Porcupine al sud oest d'Irlanda. Aquests sediments marins corresponen a principis i mitjans de l'època del Miocè de fins a 11.7 Ma (milions d'anys). A través de l'anàlisi GC-MS i l'aplicació d'eines quimiomètriques multivariants similars a les que s'han emprat en els altres capítols d'estudis òmics, s'han estudiat quins són els compostos dels cromatogrames TIC (*Total Ion Current*) que covarien més intensament en relació als canvis de temperatura superficial del mar (*Sea Surface Temperature*, SST) observats i poden emprar-se com a marcadors, en aquest cas de tipus paleoclimàtic (vegeu capítol 5).

2.5 ETAPES DE TREBALL (*WORKFLOW*) GENERAL EN ELS ESTUDIS METABOLÒMICS

En tota investigació metabolòmica s'ha de realitzar primer un disseny minuciós i exhaustiu de cada una de les etapes relacionades amb 1) la recollida i emmagatzematge de les mostres i el seu tractament , 2) l'elecció de la tècnica analítica, 3) el preprocessament de les dades i la seva anàlisi estadística, 4) la identificació dels metabòlics més rellevants i, per últim, 5) la interpretació biològica dels resultats de l'estudi. En cada una d'aquestes etapes s'han de prendre decisions que desembocaran a l'èxit o al fracàs del treball realitzat. A la figura 2.1 es presenta de manera esquematitzada les principals etapes d'un estudi de perfils metabolòmics, les quals es descriuen en els següents apartats.

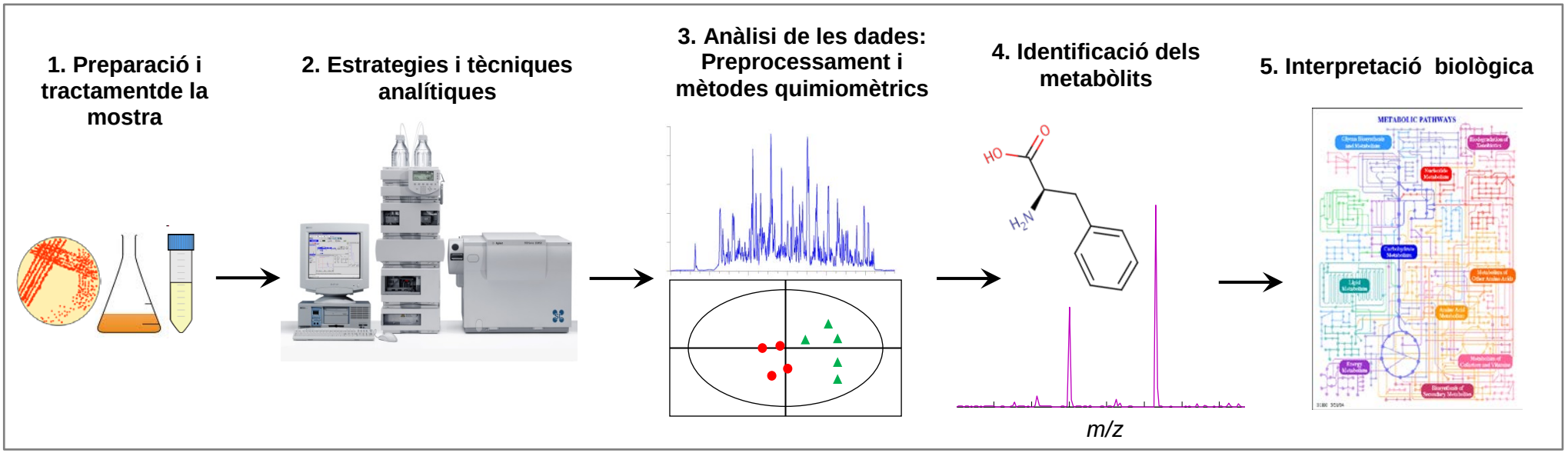


Figura 2.1 Etapes dels estudis metabolòmics

2.5.1 PREPARACIÓ DE LA MOSTRA

La preparació de la mostra és important ja que afectarà al contingut final dels metabòlits de la mostra analitzada, i conseqüentment també afectarà a la interpretació biològica dels resultats obtinguts de l'estudi. El mètode ideal de preparació de la mostra en un estudi metabolòmic no dirigit ha de ser el menys selectiu, senzill i ràpid possible, per evitar així la pèrdua o degradació dels metabòlits durant el tractament de la mostra. L'exactitud i fiabilitat dels resultats venen determinades en gran mesura pels primers passos del tractament de la mostra, és a dir, per un presa de mostra ràpida, per un refredament ràpid de l'activitat metabòlica (també anomenat *quenching*), per la seva separació del medi extracel·lular i per la seva extracció eficient.

Existeixen diversos mètodes àmpliament utilitzats per a l'extracció dels metabòlits intracel·lulars del llevat. Aquesta extracció es pot aconseguir mitjançant elevades temperatures, pHs àcids o bàsics, solvents orgànics, estrès mecànic o per combinació de diversos d'aquests factors. Alguns dels mètodes d'extracció més coneguts utilitzen l'àcid perclòric (Hancock, 1958), l'aigua calenta (Work, 1949) o l'etanol a ebullició (Fuerst i Wagner, 1957; Gonzalez et al., 1997). Posteriorment, s'ha introduït la tècnica de baixes temperatures basada en l'extracció líquida amb dues fases de cloroform i metanol (Koning i Dam, 1992). I, més recentment, s'han proposat dos mètodes basats en cicles de congelació i descongelació en metanol (Prasad Maharjan i Ferenci, 2003) i un altre mètode que utilitza l'acetonitril acídic i metanol (Rabinowitz i Kimball, 2007).

2.5.2 ESTRATÈGIES I TÈCNIQUES ANALÍTIQUES

Donada la complexitat química de les mostres ambientals, és obvi que una sola tècnica analítica no podrà proporcionar una visió global i detallada de tots els metabòlits presents en un sistema biològic determinat. La selecció de la tècnica més apropiada normalment es basa en una solució de compromís, que té en compte velocitat, selectivitat química, sensibilitat instrumental i, per descomptat, la disponibilitat de la instrumentació.

Per a l'anàlisi simultània de la multitud de metabòlits de les mostres, es poden emprar eines com la ressonància magnètica nuclear (RMN) (Wishart, 2008) o l'espectrometria de masses (MS) (Bedair i Sumner, 2008), que són les tècniques més adequades i utilitzades actualment. La RMN és

una tècnica ràpida, robusta i no destructiva. La MS té com punts forts la seva alta sensibilitat i selectivitat, i a més a més, el fet de que és capaç de detectar substàncies "invisibles" per a la RMN. La MS pot emprar-se directament sense la necessitat d'utilitzar abans una tècnica de separació, mitjançant experiments d'infusió directa (Castrillo et al., 2003), però, òbviament, quan s'acobla a les tècniques de separació d'alta resolució com la cromatografia de gasos (*Gas Chromatography*, GC), la cromatografia de líquids (*Liquid Chromatography*, LC) o l'electroforesi capil·lar (*Capillary Electrophoresis*, CE), s'obté una millora substancial de les seves capacitats, i s'incrementa la informació del conjunt de compostos químics (metabòlits) del sistema biològic que s'està estudiant.

L'espectrometria de masses (MS) és una tècnica microanalítica que es pot utilitzar selectivament per detectar i determinar un determinat analit en una mostra. Aquesta tècnica es basa en la identificació de les partícules ionitzades en estat gasós a través de la seva relació massa/càrrega (*m/z*), és a dir, de la massa del ió dividit pel número de càrregues que té. Els espectròmetres de masses, fan ús dels camps magnètics o d'una combinació de camps elèctrics i magnètics per modificar la trajectòria i posició d'aquestes partícules obtenint un espectre *m/z*. L'elevada sensibilitat i selectivitat de l'espectrometria de masses la converteixen en una tècnica ideal per a la detecció de compostos químics en mostres complexes. No obstant això, l'anàlisi d'aquestes mostres requereix generalment la seva separació prèvia per exemple, per un mètode, cromatogràfic.

En aquesta Tesi, l'anàlisi metabolòmica de mostres biològiques i ambientals s'ha realitzat mitjançant la combinació de l'espectrometria de masses i les tècniques de separació, especialment de la cromatografia líquida d'alta resolució (HPLC).

Els resultats que s'obtenen a partir d'un espectròmetre de masses depèn notablement de quines són la seva interfase, font de ionització i analitzador. S'han descrit diverses fonts de ionització (Gelpí, 2002) que es solen classificar segons el grau de fragmentació que provoquen a l'estructura del compost químic. Existeixen tècniques d'ionització fortes com la d'impacte electrònic (*Electronic Impact*, EI), i d'altres més suaus com la d'*electrospray* (ESI), la d'ionització química a pressió atmosfèrica (*Atmospheric Pressure Chemical Ionization*, APCI) o la de desorció/ionització làser assistida per matriu (*Matrix-Assisted Laser Desorption/Ionization*, MALDI). Pel que fa a l'analitzador de masses, hi ha diferents opcions, les més comuns són: el quadropol (Q), la trampa

de ions (*Ion Trap*, IT), el triple quadropol (QQQ), el temps de vol (*Time Of Flight*, TOF), la transformada de Fourier-ressonància d'ió ciclotrònica (*Fourier Transform Ion Cyclotron Resonance mass spectrometry*, FT-ICR), i l'*orbitrap* (OT). Totes aquestes opcions es distingeixen en l'exactitud que ofereixen al determinar la massa molecular dels analits (error entre la massa exacta determinada i el valor teòric), la capacitat per determinar les distribucions isotòpiques (*true isotòpic patern*), el poder d'obtenir patrons de fragmentació en fer acoblaments MS/MS, (MS^n), la velocitat d'escombrat, i el rang de masses que poden mesurar i la resolució.

L'ús d'HPLC-MS en estudis metabolòmics ha experimentat un gran desenvolupament en els últims anys en comparació amb altres tècniques com la cromatografia de gasos, al requerir una preparació menys complexa de les mostres i poder analitzar un ventall més gran d'analits. Les mostres retingudes per l'HPLC entren a l'espectròmetre ionitzades mitjançant un electrospray, tenint en compte que alguns analits s'ionitzaran millor en mode positiu i d'altres en negatiu. No obstant, actualment encara hi ha diverses qüestions que s'han de resoldre abans de la generalització de l'HPLC-MS per a les anàlisis metabolòmiques. Un dels problemes és que no totes les molècules s'ionitzen de la mateixa manera, provocant diferències de sensibilitat notables encara que els components estiguin presents en la mateixa concentració molar. Això complica l'anàlisi de mostres complexes, de les quals no es coneixen els seus constituents. Per altra banda, com que l'anàlisi per MS presenta una alta sensibilitat, en cada anàlisi es poden detectar milers d'ions desconeguts (no identificats). A partir de la mesura dels canvis en les concentracions d'aquests ions es poden diferenciar els grups de mostres analitzades i detectar quins són els possibles biomarcadors que discriminen millor les mostres.

2.5.3 NATURALESA DE LES DADES METABOLÒMIQUES

Les tècniques cromatogràfiques utilitzades tant en les anàlisis metabolòmiques dirigides com en les no dirigides (LC-MS, GC-MS), proporcionen respostes i dades multivariants. La informació que es pot obtenir a partir de l'aplicació dels mètodes quimiomètrics depèn de la naturalesa i estructura de les dades experimentals. Els cromatogrames que s'obtenen estan formats per conjunts de valors numèrics. En aquesta Tesi es tracten dos casos diferents de conjunt de dades: 1) Cromatogames TIC (*Total Ion Current*) que representen en forma de vector de dades. El cromatograma TIC és el resultat de la suma de les intensitats de tot el rang de masses per a cada temps de retenció. 2) Cromatogrames representats en forma de matriu en els quals el temps de

retenció s'ubiquen a la dimensió x de la matriu i el rang de *m/z* analitzades es representen a la dimensió y de la mateixa.

En els experiments de metabolòmica generalment s'analitzen diverses mostres, i això proporciona diferents matrius de dades. La recopilació de totes aquestes matrius de dades, poden proporcionar informació sobre com canvien els perfils metabòlics en cada una de les mostres. Quan aquestes matrius de dades es volen analitzar simultàniament es poden organitzar de diferents maneres, per exemple, en una matriu augmentada $\mathbf{D}_{\text{aug}}$ on les diverses matrius individuals es poden agrupar segons la direcció (o direccions) de l'augmentació.

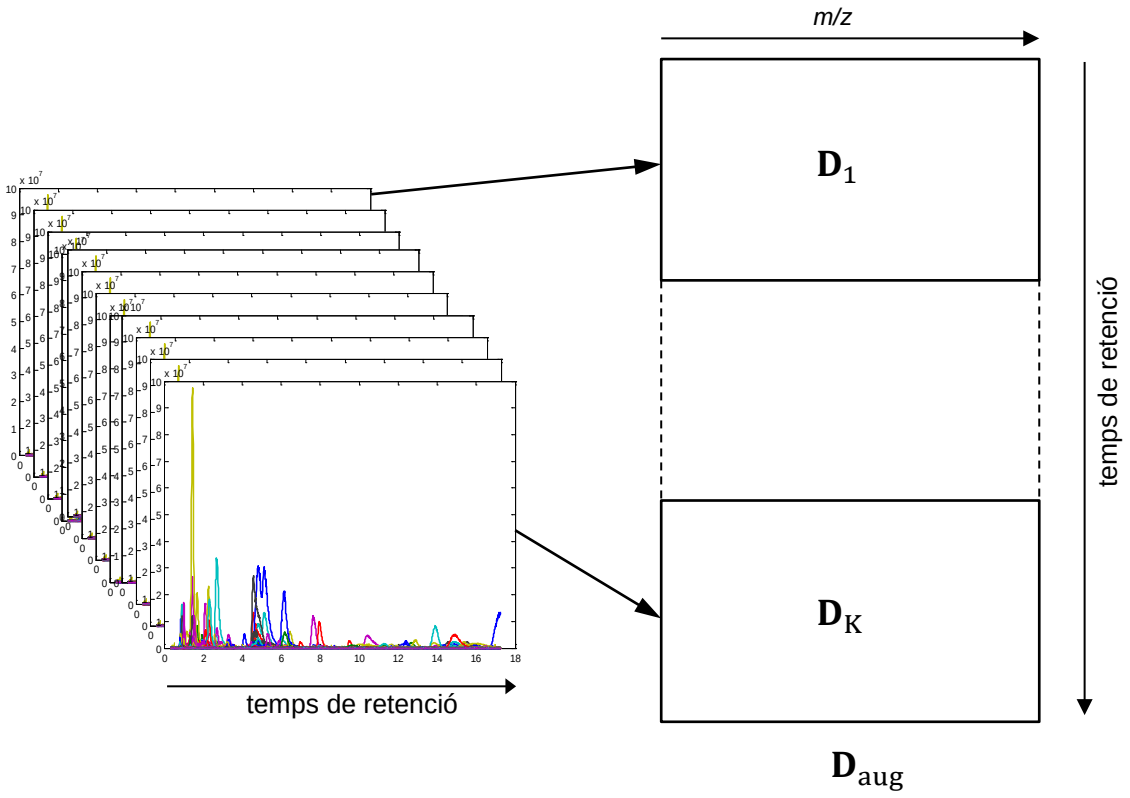


Figura 2.2 Augmentació de matrius de dades en la direcció de les columnes (*m/z*).

La figura 2.2 mostra un conjunt de matrius (cromatogrames) augmentades en direcció de les columnes (*column-wise*). Aquesta augmentació s'utilitza per exemple quan diferents experiments en l'estudi d'un determinat sistema han estat analitzats amb una mateixa tècnica analítica. En aquest cas només cal que les matrius de dades obtingudes en cada experiment tinguin en comú les seves columnes, però el nombre de files de cadascuna d'elles poden diferir. En aquesta Tesi

l'anàlisi per LC-MS de les mostres experimentals proporcionen varies matrius de dades (una per mostra) que es recopilen en una sola matriu augmentada de dades en la direcció de les columnes (vegeu figura 2.2). En aquest arranjament, les diferents matrius de dades s'acoblen aprofitant que tenen la direcció de les columnes en comú (*m/z*).

2.5.4 PREPROCESSAMENT DE LES DADES

El preprocessament de les dades experimentals (per exemple de LC-MS) es una etapa important de l'anàlisi de les dades que es fa per possibilitar o millorar la interpretació dels fenòmens estudiats i també per augmentar la capacitat de predicció d'un determinat model. L'objectiu és eliminar, o almenys reduir, la variació de les dades que pot afectar negativament a l'efectivitat de les interpretacions o prediccions d'un determinat model. Els avenços recents en l'adquisició de dades químiques ha permès un increment considerable de la quantitat de dades i també de la seva complexitat. El senyal que genera un instrument analític es compon d'informació química i estocàstica (variació aleatòria). La part del senyal que es deriva de les substàncies analitzades, idealment, pot ser descrita per un determinat model. La variació aleatòria associada a qualsevol mesura experimental deteriora la qualitat del model, i normalment es produeix per les imperfeccions dels instruments analítics, com per exemple, pel soroll generat pels detectors. També hi ha alguns errors sistemàtics provinents de diverses fonts que si no s'eliminen poden afectar negativament a les prediccions i a la interpretabilitat dels models multivariants.

Els mètodes de preprocessament de dades són part important de l'estratègia de construcció d'un determinat model que permeti la interpretació del fenomen estudiat. És important que aquest preprocessament sigui robust. Quan un preprocessament de dades no es fa de manera adequada, es poden introduir variacions i efectes no desitjats que empitjorin els resultats i conclusions extretes. Per tant, el preprocessament de les dades és un pas crític que pot influir de forma decisiva en els resultats obtinguts i en la seva interpretació. L'ús de tècniques cromatogràfiques acoblades (GC-MS, LC-MS) en l'anàlisi simultània de múltiples mostres donen lloc a múltiples matrius de dades que han de ser preprocessades abans de la seva anàlisi, ja que sovint es troben afectades per les variacions en les seves condicions de mesura experimental. Existeixen un gran ventall de mètodes de preprocessament (Zeaiter i Rutledge, 2009). Els mètodes escollits preprocessar no només dependran del tipus d'informació que es vol obtenir, sinó també del mètode seleccionat posteriorment per a l'anàlisi de les dades. Els mètodes de preprocessament més comuns inclouen

passos com el filtratge o la reducció de soroll del senyal analític mesurat, l'alineament de senyals analítics (pics cromatogràfics) o la seva normalització. A continuació s'expliquen amb més detall els mètodes de preprocessament de dades que han estat més utilitzats en aquesta Tesi.

Correcció de la línia de base i eliminació de soroll

Les variacions de la línia de base, com el soroll de fons o la deriva del senyal instrumental amb el temps (*drift*) es deuen a petits canvis en les condicions experimentals. Existeixen diferents fonts de soroll i de deriva instrumental com són els canvis en la resposta del detector, l'antiguitat dels seus components, la contaminació pels solvents o gasos, i el soroll electrònic. Els mètodes d'eliminació del soroll de fons i de correcció de la línia de base s'apliquen amb l'objectiu de reduir la influència de la variació instrumental no desitjada. Es tracta de separar el senyal procedent d'un compost químic de la mostra analitzada del senyal de fons que ve de la interferència instrumental.

El mètode de mínims quadrats asimètrics (*Asymetric Least Squares*, AsLS) és una eina molt potent i efectiva per a la correcció de la línia de base d'espectres i cromatogrames. Originalment aquest mètode va ser proposat per (Eilers i Boelens, 2005; Eilers, 2003b). El mètode AsLS s'utilitza per al suavitzat de la línia de base i per a la correcció del soroll de fons i intenta millorar les característiques de les dades cromatogràfiques i espectrals.

Alineament

Les fluctuacions naturals i ambientals en les mostres, al laboratori i a la tecnologia analítica (columnes cromatogràfiques) poden influir en les mesures instrumentals de la mostra, tant en la dimensió espectral com en la dimensió temporal. Aquestes diferències es poden resoldre aplicant mètodes d'alineació. En el cas del desajust en la dimensió temporal (per exemple, en els desplaçaments en els temps de retenció cromatogràfics) l'alineació es realitza per assegurar que els pics que són presents en diferents cromatogrames del conjunt de dades analitzat, i que corresponen al mateix component químic, apareguin exactament als mateixos temps de retenció cromatogràfica. I que, per tant, aquests temps de retenció es puguin assignar efectivament al mateix compost químic. L'alineació dels cromatogrames/espectres millora la fiabilitat i la robustesa en la interpretació dels models resultants. El mètode *Parametric Time Warping* (PTW) (Eilers, 2003a) és un mètode d'alineació global, és a dir, alinea tot un cromatograma estirant-lo o

estrenyent-lo en relació a un cromatograma de referència. En canvi, els mètodes com *Correlation Optimized Warping* (COW) (Nielsen et al., 1998; Pierce et al., 2005) o *i*coshift (Savorani et al., 2010) són mètodes que s'apliquen per alineaments locals a vectors de dades (en una direcció).

En aquesta Tesi s'ha utilitzat el mètode COW per corregir els desajustos o canvis en els senyals de les dades en els cromatogrames. Aquest és un mètode de preprocessament de les dades segmentades o a trossos amb l'objectiu d'alinear un vector de dades (mostra) en relació a un vector referència mitjançant la 'deformació' del vector mostra. El vector utilitzat com a referència ha de ser representatiu de les mostres que es volen alinear. L'algorisme funciona mitjançant la divisió del vector de la mostra que s'ha dalinear en segments de mida definida i permetent l'augment o disminució de la longitud d'aquests segments per buscar la correlació òptima amb els segments del vector de referència. D'aquesta manera, els paràmetres que s'han de definir per a l'aplicació del mètode COW d'alineació són: el vector que ha de ser utilitzat com a referència, la longitud del segment (*m*) i el grau de flexibilitat *t* (*slack*). La selecció adequada d'aquests paràmetres es pot realitzar aplicant l'algorisme descrit per (Skov et al., 2006). Quan el nombre de temps de retenció entre el vector a alinear i el vector de referència difereixen, el primer s'interpola linealment amb la finalitat de crear un segment d'igual longitud. En el cas de les dades cromatogràfiques, l'alineació s'efectua en la direcció de l'eix del temps de retenció dels cromatogrames. La relació entre el nombre de punts en el vector de referència i la longitud seleccionada del segment (*m*) determina el nombre de segments amb els que s'ha dividit el cromatograma de la mostra. L'augment o disminució de la longitud màxima del segment de la mostra és controlat pel paràmetre *t* (*slack*).

Normalització

La normalització és una de les etapes de preprocessament més crucial per tal de corregir adequadament la variació sistemàtica entre mostres i les diferents escales de les variables mesurades. Els mètodes de normalització més utilitzats en aquesta Tesi han estat el centrat (*mean-center*) i l'autoescalat de les dades experimentals. El centrat de les dades consisteix en que a cada valor original d'una variable determinada se li resta el valor de la seva mitjana (de tots els valors mesurats de la variable considerada). D'aquesta manera cada variable de la nova matriu centrada té una mitjana igual a zero. El centrat ajusta les diferències en el desplaçament entre les variables, i permet visualitzar les variacions respecte al valor de la seva mitjana, ometent la informació d'aquelles variables que no varien i dels valors constants o *offsets*.

L'escalat és un mètode de preprocessament de dades en el qual cada variable es divideix per un factor determinat, anomenat factor d'escala, que és diferent per a cada variable, i habitualment és la seva desviació estàndard (de tots els valors mesurats d'aquesta variable). L'escalat és important en aquells casos on les variables estan mesurades en escales diferents o en unitats diferents, ja que els models de mínims quadrats es basen en l'ajust de la variació de les dades, això implica que les variables amb més variació són aquelles que resulten més importants. L'objectiu de l'escalat és fer comparables les variables entre sí, ja que d'aquesta manera totes les variables tenen el mateix pes en l'anàlisi i interpretació dels resultats, això vol dir que s'iguala la importància relativa de cada variable. Però, s'ha de tenir en compte que establir la mateixa variància a cada variable pot donar lloc també a problemes, especialment quan les dades contenen variables poc informatives per al model, o són només soroll instrumental, ja que al escalar-les se'ls dona més importància de la que tenen (Gurden et al., 2001). En aquests casos, es recomanable que aquestes variables s'excloguin totalment o que no s'escalin i es divideixin simplement per un número suficientment gran per relativitzar al seva influència en els resultats.

L'autoescalat és un tipus d'escalat de les dades que s'aplica sobre les dades ja centrades. La distribució dels valors de les variables obtinguda en aplicar l'autoescalat és similar a quan només s'escalen, però al mateix temps les dades experimenten la translació a un mateix origen de variació (el zero) tal i com succeeix quan s'aplica el centrat amb la mitjana de les dades.

2.5.5 MÈTODES QUIMIOMÈTRICS

Per a extraure informació química a partir de dades mesurades experimentalment s'utilitzen models matemàtics que descriuen el seu comportament. Un model químic relaciona i descriu el comportament de les variables mesurades experimentalment a partir de les variacions en la composició química de les mostres analitzades. Com que no serà possible l'explicació al cent per cent de tota la variació observada a partir d'aquest model químic, hi haurà també una part residual que descriurà variabilitat desconeguda (soroll) de les dades (Wold, 1995) és a dir,

$$\mathbf{X} = \mathbf{M} + \mathbf{E} = \text{Model químic + Soroll} \qquad \textbf{Equació 2.2}$$

En els estudis de metabolòmica, les observacions i les mostres es caracteritzen per l'ús d'instrumentació com LC-MS o GC-MS, i les mostres mesurades amb aquests instruments

generen conjunts de dades multi- i megavariants. Les tècniques quimiomètriques d'anàlisi i d'exploració multivariant de dades permeten la determinació indirecta de les variables intrínseques (latents) del sistema estudiat, és a dir, de les que causen la variació experimental observada. Les anàlisis de dades multivariants basades en els mètodes de projecció com l'anàlisi de components principals (*Principal Component Analysis*, PCA)(Wold, 1995) i la regressió per mínims quadrats parcials (*Partial Least Squares*, PLS)(Bro et al., 2008) proporcionen una sèrie d'eines eficients i útils per a l'anàlisi i modelatge d'aquest tipus de dades complexes. Aquestes tècniques permeten manipular adequadament matrius grans de dades i comprimeixen la informació multidimensional que contenen, en un nombre reduït de variables latents que expliquen una gran part de la variabilitat de les variables mesurades, així com de les seves relacions. En comprimir la informació present en les dades, la major part de la variància observada es pot analitzar en l'espai de variables latents (de menor dimensió i ortogonal), la qual cosa pot ajudar a entendre millor els fenòmens químics o biològics subjacents. La compressió d'informació estableix també vincles de relació amb les variables originals, que es poden recuperar en qualsevol moment de l'anàlisi. Per tant, la informació continguda en les variables originals no es perd.

La quimiometria proporciona eines potents que permeten extraure informació química interpretable a partir de taules o matrius de dades multivariants, que han estat directament mesurades. En els mètodes quimiomètrics no es postula un model químic de forma explícita a partir de paràmetres fisicoquímics deterministes, o de models mecanicistes explícits. Les tècniques quimiomètriques permeten fer el càlcul i la representació gràfica de les tendències d'associació i d'agrupament més importants que es presenten en les dades, buscant i identificant les possibles variables que influeixen més en l'explicació de la variància de les dades. En quimiometria es poden definir tres categories bàsiques d'anàlisi de dades (Trygg et al., 2006):

(a) Anàlisi exploratòria, que dóna una visió general de les dades per a la detecció tendències, pautes o grups (clústers) (vegeu figura 2.3a).

(b) Anàlisi classificatòria i anàlisi discriminant, el qual classifica les mostres en categories o classes (vegeu figura 2.3b).

(c) Anàlisis de regressió i models de predicció, que són utilitzats quan hi ha una relació quantitativa entre dos blocs de dades (vegeu figura 2.3c).

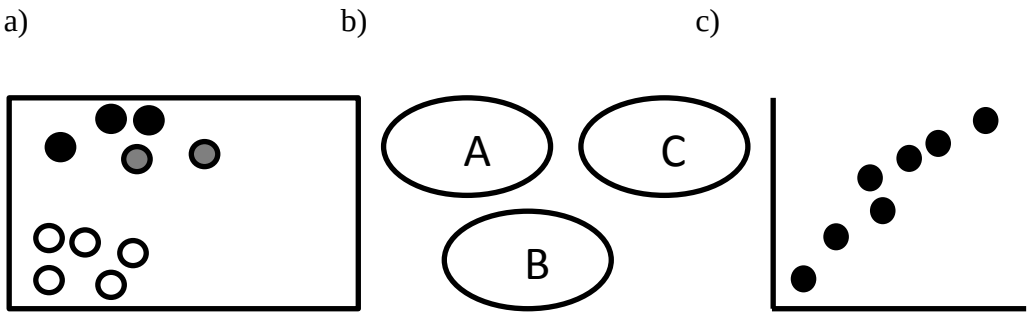


Figura 2.3 Visió generals de les categories bàsiques d'anàlisi quimiomètrica. a) Visió general de les dades, b) classificació de les mostres en categories o classes i c) relació quantitativa entre dos blocs de dades (Trygg et al., 2006).

Anàlisi de Components Principals (PCA)

L'anàlisi per components principals (*Principal Component Analysis*, PCA) (Esbensen i Geladi, 2009; Wold et al., 1987) és un mètode multivariant d'anàlisi de dades dissenyat per extraure i visualitzar la variació sistemàtica de la informació continguda en un conjunt de dades. Aquest mètode és àmpliament emprat en diferents àmbits científics i tecnològics, i en particular en estudis químics, ambientals i d'òmica . El PCA és una eina quimiomètrica que permet la reducció de la complexitat original de les dades, i que projecta els objectes i les variables en un espai de menors dimensions. La hipòtesi inicial en aplicar aquest tipus d'anàlisi és que en el conjunt de dades observades existeix en realitat un nombre limitat i reduït de causes de variació sistemàtica predominant, és a dir, de factors o components que influeixen de forma important sobre la variància observada en les dades. En el mètode PCA, a partir de la combinació lineal de les variables originals, es dedueix un nou conjunt de variables ortogonals no correlacionades (factors o components) que expliquen per a cadascuna d'elles el màxim de la variància continguda en les dades originals. Aquestes noves variables, factors o components principals són ortogonals i per tant el que explica un component no ho explica un altre. Com que en el mètode PCA, a més s'aplica la restricció d'explicació de màxima variància, i es normalitzen les solucions, la solució matemàtica és única. No obstant això, aquestes solucions no tenen significat físic ni interpretació directa. El PCA proporciona uns nous eixos (ortogonals) de representació de les dades, sobre els quals s'expressa el màxim de la variància explicada.

Matemàticament, la descomposició elemental factor a factor en el model PCA ve definida per la següent equació (equació 2.3):

$$\mathbf{x}_{ij} = \sum_{n=1}^{N} \mathbf{t}_{in}\, \mathbf{p}_{jn} + \mathbf{e}_{ij} \qquad \textbf{Equació 2.3}$$

On $\mathbf{x}_{ij}$ és el valor de la variable j en la mostra i, $\mathbf{t}_{in}$ és el *score* del component principal n en la mostra i, $\mathbf{p}_{jn}$ és el *loading* de la variable j en el component principal n , i $\mathbf{e}_{ij}$ és el residual de la part no explicada d'aquest element $\mathbf{x}_{ij}$.

En PCA, la matriu original de dades **X** es factoritza o descompon en el producte de dues matrius ortogonals **T** i $\mathbf{P}^T$ (equació 2.4):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \textbf{Equació 2.4}$$

A l'equació 2.4, la matriu **T** defineix els *scores*, la matriu **P** defineix *els loadings* i **E** representa la matriu residual. El producte de les matriu dels *scores* i dels *loadings* reprodueix la matriu original de dades (equació 2.4) per un nombre determinat de components, n=1,...N.

El mètode més utilitzat actualment per a fer la descomposició de PCA és la descomposició en valors singulars (*Singular Value Decomposition*, SVD) (Golub et al., 1970), doncs és un mètode acurat i ràpid. Independentment del mètode emprat per a calcular els components principals, la determinació del número adequat de components o factors que descriuen la matriu **X** és l'aspecte més crucial en PCA. Per a definir quants components s'han d'incloure a l'anàlisi, s'observa la magnitud dels valors propis o singulars (que són l'arrel quadrada dels propis), ja que aquesta indica la quantitat de variància retinguda per a cada component considerat. Tant els valors propis com la magnitud de variància retinguda per cada component disminueix en afegir més components fins a un nivell petit, i que pràcticament només explica el soroll aleatori de les dades i varia molt poc. El nombre de components N s'escull fins que l'addició d'un nou component ja no proporciona informació addicional rellevant dins del context del problema estudiat.

En el PCA, els components principals es troben ordenats segons la seva quantitat de variància explicada, de manera que el primer component explica més variància que el segon i així

successivament. Per aquest motiu la informació més rellevant del conjunt de dades es troba sempre concentrada en els primers components principals generats. El PCA està dissenyat per a proporcionar una millor visualització de la variància de les dades. El nou conjunt d'eixos ortogonals de coordenades (*loadings*) sobre els que es poden projectar les dades originals (*scores*) proporciona una millor interpretació dels fenòmens que estan causant la variància observada.

Els dos primers components principals defineixen un pla (espai) bidimensional, on totes les mostres es poden visualitzar projectades. Les coordinades de cada una de les mostres projectada en aquest pla són els *scores* **T**. La visualització d'aquests *scores* **T** s'anomena gràfic dels *scores* (*scores plot*). El gràfic dels *scores* és especialment informatiu perquè dona una visió general de totes les mostres analitzades i de com es relacionen les unes amb les altres. A partir d'aquest gràfic es treu informació de l'agrupament de les mostres (clústers), de les tendències i dels valors atípics (*outliers*). El gràfic dels *scores* permet investigar la relació existent entre les mostres, i quan aquesta relació ja està clarament establerta, és possible entendre la raó d'aquesta. En el gràfic dels *loadings* **P** es descriu la influència (pes) de les variables experimentals mesurades en la matriu **X** sobre cadascun dels components descrits pel model PCA. Una característica important és que les direccions del gràfic dels *scores* correspon a les mateixes direccions del gràfic dels *loadings*. Aquest fet és molt útil per a la interpretació de les fonts de variació experimental i de les seves característiques generals, i que es pot fer aleshores a partir de la interpretació simultània dels gràfics dels *scores* i dels *loadings*.

Com que PCA utilitza restriccions d'ortogonalitat, de màxima variància explicada i de normalització dels *loadings*, proporciona solucions úniques. No obstant això, les solucions PCA no són directament interpretables des d'un punt de vista físic, ja que els perfils (vectors) que s'obtenen pels factors de la descomposició (vegeu equació 2.4), **T**, *scores*, i **P**, *loadings*, per a cada component són en realitat una combinació lineal dels perfils vertaders dels components que causen realment la variància observada. En general, els perfils dels factors químics reals no compleixen les restriccions imposades pel model PCA, d'ortogonalitat ni de màxima variància. De tota manera, els resultats obtinguts per PCA són extremadament útils ja que permeten descriure la complexitat del sistema estudiat (nombre de components) i possibiliten una interpretació de com són les fonts de variació descrites per aquests components (a partir dels gràfics de *scores* i *loadings*), i que òbviament estaran relacionades amb les fonts reals de variació

Abans d'analitzar un conjunt de dades mitjançant PCA, i tenint en compte la naturalesa de les dades, s'ha de decidir si cal fer un determinat preprocessament d'aquestes (vegeu secció 2.5.4).

En aquesta Tesi s'ha aplicat el mètode PCA en l'Article 1, en el qual s'avaluen 4 condicions ambientals diferents sobre el desenvolupament del blat. L'estudi s'ha fet a partir de la determinació LC-MS dels canvis dels perfils d'alguns metabòlits del blat obtinguts i de la seva avaluació quimiomètrica. Els quatre factors estudiats han estat: factor estadi de creixement (escala BBCH, 5 estadis), factor varietat (Astron, Ritmo, Stakado), factor tipus de mostra (brots, arrels) i factor tipus de cultiu (convencional, orgànic). En l'Article 2 s'estudien per PCA els perfils metabòlics de *S.cerevisiae* obtinguts per LC-MS; i en l'Article 4 s'analitzen per PCA les els perfils de diversos compostos orgànics fòssils acumulats obtinguts per CG-MS de mostres de sediments marins de fins a 11.7 Ma.

Anàlisi de la variància – Anàlisi de components principals (ANOVA –PCA) i Anàlisi Simultània de Components (ASCA)

L'anàlisi de la variància (*ANalysis Of VAriance,* ANOVA) és una tècnica estadística emprada molt potent que permet determinar i quantificar els efectes de diferents factors experimentals en els resultats observats d'un experiment. Per tal de fer aquesta avaluació de forma adequada, cal que els factors variïn de forma sistemàtica seguint el que s'anomena un disseny estadístic d'experiments (*Statistical Experimental Design or Design of Experiments,* DOE) (Lundstedt et al., 1998). ANOVA és una tècnica àmpliament emprada dins de l'estadística univariant per avaluar l'efecte de factors experimentals a partir de dades obtingudes seguint un disseny estadístic (Kerr i Churchill, 2007; Kerr et al., 2000). Existeixen diverses variants d'aquesta tècnica segons el nombre de possible factors estudiats, que reben diferents denominacions, ANOVA d'una entrada o via (*one-way*), ANOVA de dues entrades o vies (*two-way*), i ANOVA multi entrada o multivia (*N-way*). De tota manera en tots aquestes variants, l'efecte dels factors s'avalua a partir d'una resposta única o univariant, i està per tant limitada per aquesta restricció. En els darrers anys s'han proposat diverses extensions d'ANOVA per dades multivariants (Hoefsloot et al., 2009; Jansen et al., 2008; Jansen et al., 2005a; Jansen et al., 2005b; Smilde et al., 2005; Smilde et al., 2012; Zwanenburg et al., 2011).

Els conjunts de dades multivariants consisteixen en la mesura de múltiples respostes o variables sobre cadascuna de les mostres analitzades, i aquestes en general són conseqüència de la influència dels mateixos factors i dels seus efectes. En la modelització multivariant de dades, aquestes respostes múltiples es modelen conjuntament per tal de trobar la relació entre elles i amb els efectes corresponents. Entre les adaptacions d'ANOVA a dades multivariants destaquen els mètodes ANOVA-PCA i ASCA (ANOVA-*Simultaenous Component Analysis*). Aquestes són dues tècniques d'anàlisis multivariants que combinen les avantatges estadístiques de l'ANOVA per separar fonts de variància, i els avantatges del PCA per tal d'eliminar la covariància entre les variables i així poder explicar el màxim de la variància comuna. Aquest enfocament consisteix en la implementació multivariant de l'ANOVA, en el qual la matriu de dades experimentals es descompon en diverses matrius additives que caracteritzen cada un dels factors del disseny experimental i de l'error residual (Harrington et al., 2005). En particular, les dades metabolòmiques d'aquesta Tesi han estat analitzades utilitzant aquestes metodologies i la seva adaptació a l'anàlisi simultània de components a múltiples nivells (MSCA, *Multilevel Simultaneous Component Analysis*) (Jansen et al., 2005a).

En aquesta Tesi, el model ANOVA emprat ha estat un model de quatre factors que inclou també totes les seves possibles interaccions, el qual es pot escriure matemàticament de la següent manera:

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{X}_{\mathrm{a}} + \mathbf{X}_{\mathrm{b}} + \mathbf{X}_{\mathrm{c}} + \mathbf{X}_{\mathrm{d}} + \mathbf{X}_{(\mathrm{ab})} + \mathbf{X}_{(\mathrm{ac})} + \mathbf{X}_{(\mathrm{ad})} + \mathbf{X}_{(\mathrm{bc})} + \mathbf{X}_{(\mathrm{bd})} + \mathbf{X}_{(\mathrm{abc})} + \mathbf{X}_{(\mathrm{abd})} + \mathbf{X}_{(\mathrm{acd})} + \mathbf{X}_{(\mathrm{bcd})} + \mathbf{X}_{(abcd)} + \mathbf{E} \qquad \textbf{Equació 2.5}$$

A l'equació 2.5, $\mathbf{X}$ és la matriu experimental (formada per exemple pels cromatogrames en mode de suma total de ions, *Total Ion Current*, TIC), $\bar{\mathbf{X}}$ és la matriu de mitjanes de la matriu experimental, en la qual totes les files corresponen a la mitjana del totes les mostres (cromatogrames TIC) del conjunt de dades $\mathbf{X}$. Les matrius $\mathbf{X}_{\mathrm{a}}$, $\mathbf{X}_{\mathrm{b}}$, $\mathbf{X}_{\mathrm{c}}$ i $\mathbf{X}_{\mathrm{d}}$ contenen els efectes individuals dels diferents factors considerats. Les matrius $\mathbf{X}_{(\mathrm{ab})}$, $\mathbf{X}_{(\mathrm{ac}),}$ $\mathbf{X}_{(\mathrm{ad})}$ ... $\mathbf{X}_{(abcd)}$ són les matrius d'interacció dels diferents factors i $\mathbf{E}$ és la matriu residual que representa la variació què no està recollida per les matrius de factors i les seves interaccions. En dades experimentals de tipus òmic o ambiental, aquesta matriu residual conté la variació de tipus natural individual no comuna entre les mostres replicades, a part de la variació del soroll pròpiament degut a la mesura experimental.

Les files de la matriu de mitjanes $\bar{\mathbf{X}}$ contenen les mitjanes dels cromatogrames TIC, i per tant en aquest cas, el rang matemàtic (nombre de files i/o columnes de la matriu linealment independents) d'aquesta matriu de mitjanes és igual a la unitat. Les files de les matrius que representen la variació dels factors considerats estan formades per la mitjanes d'aquests factors en els seus diferents nivells. Per tant el rang matemàtic de cada una d'aquestes matriu de factors serà igual al nombre de nivells dels factors menys un, ja que el resultat serà el nombre de files linealment independents de la matriu. Les files de les matrius d'interacció contenen les mitjanes de les files de la matriu de dades original (cromatogrames TIC de les mostres) que tenen els mateixos nivells dels factors considerats.

Les dades que s'han analitzat en aquesta Tesi s'han obtingut a partir de dissenys experimentals equilibrats (*balanced*), això vol dir que cada nivell dels factors (model de quatre factors) té exactament el mateix nombre d'observacions replicades. En els dissenys estadístics equilibrats, s'utilitza l'anomenada suma de quadrats de tipus I per avaluar els efectes dels diferents factors en l'ANOVA. Aquesta suma de quadrats tipus I consisteix en l'assignació seqüencial (una darrera l'altra) de la part de la variància explicada a cadascun dels efectes principals considerats; posteriorment es consideren les interaccions bidireccionals i així successivament tot incrementant l'ordre de les interaccions. Degut al disseny emprat, la suma de quadrats corresponents a cada un dels factors, a les seves interaccions i a la variància residual és ortogonal (no superposada). Aleshores, la suma total de quadrats de tipus I és equivalent a la suma de quadrats de les matrius dels factors (Iacobucci, 1995; Shaw i Mitchell-Olds, 1993).

L'ANOVA-PCA permet la comparació de les mitjanes dels efectes de cada factor en relació a l'error experimental residual (Harrington et al., 2005). L'ANOVA-PCA proporciona una figura visual (gràfic dels *scores*) en la qual es pot observar si les mitjanes, és a dir, els nivells d'un factor, difereixen significativament respecte la reproductibilitat de la mesura i també permet comparar als factors entre sí. En aquest mètode, el PCA s'aplica separadament a cada una de les matrius dels factors conjuntament amb la matriu residual, per exemple a $\mathbf{X}_a + \mathbf{E}$. Per tal que un factor esdevingui significant, en els resultats del PCA, el seu efecte ha de ser la principal font de variació en comparació amb la de la matriu residual. Aleshores, el primer component principal (PC1) ha de caracteritzar aquest efecte del factor amb un valor molt per sobre del segon component principal (PC2), el qual reflecteix principalment la variació residual de tipus natural i aleatòria (Harrington et al., 2005). La hipòtesi nul·la, que es caracteritza perquè els efectes dels factors considerats no

mostren diferències estadísticament significatives en els resultats experimentals, és rebutjada quan els *scores* d'un factor determinat s'agrupen clarament en relació als nivells d'aquest factor sobre l'eix del PC1, el qual és ortogonal als residuals que es trobaran sobre l'eix del PC2.

En la metodologia ASCA, a diferència de l'ANOVA-PCA, es fa una anàlisi PCA sobre cada matriu de factors individualment ($\mathbf{X}_a$, $\mathbf{X}_b$, $\mathbf{X}_c$... $\mathbf{X}_{(abcd)}$) (Jansen et al., 2005a). El model matemàtic d'ASCA corresponent a l'ANOVA de l'Equació 2.5 per a un disseny experimental de quatre factors és el següent (equació 2.6):

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{T}_a\mathbf{P}_a^T + \mathbf{T}_b\mathbf{P}_b^T + \mathbf{T}_c\mathbf{P}_c^T + \mathbf{T}_d\mathbf{P}_d^T + \mathbf{T}_{(ab)}\mathbf{P}_{(ab)}^T + \mathbf{T}_{(ac)}\mathbf{P}_{(ac)}^T + \mathbf{T}_{(ad)}\mathbf{P}_{(ad)}^T + \mathbf{T}_{(bc)}\mathbf{P}_{(bc)}^T + \mathbf{T}_{(bd)}\mathbf{P}_{(bd)}^T + \mathbf{T}_{(abc)}\mathbf{P}_{(abc)}^T + \mathbf{T}_{(abd)}\mathbf{P}_{(abd)}^T + \mathbf{T}_{(acd)}\mathbf{P}_{(acd)}^T + \mathbf{T}_{(bcd)}\mathbf{P}_{(bcd)}^T + \mathbf{T}_{(abcd)}\mathbf{P}_{(abcd)}^T + \mathbf{E} \quad \textbf{Equació 2.6}$$

A l'equació 2.6 $\mathbf{T}_a, \mathbf{T}_b, \mathbf{T}_c \ldots \mathbf{T}_{(abcd)}$ són les matrius dels *scores* de cada submodel i $\mathbf{P}_a^T, \mathbf{P}_b^T, \mathbf{P}_c^T \ldots \mathbf{P}_{(abcd)}^T$ són les matrius dels *loadings* d'aquest mateix submodel.

Tenint en compte que el PCA s'aplica directament a les matrius de cada factor, i a diferència del mètode ANOVA-PCA, no es possible determinar la variabilitat natural en el gràfic dels *scores* de l'ASCA. No obstant aquesta informació és molt important per avaluar la magnitud de les diferències entre els nivells dels factors en comparació amb la variació natural. Per a determinar l'estimació de la variabilitat dels replicats de les mostres sobre els components principals del PCA es projecta la matriu d'efecte més la matriu residual $\mathbf{X}_k + \mathbf{E}$ als *loadings* $\mathbf{P}_k$ del model PCA per $\mathbf{X}_k$. Les projeccions resultants $\mathbf{Y}_k$ contenen la variació de les rèpliques en el subespai del component principal del factor *k.* Aquesta estimació es pot determinar mitjançant el model de l'equació 2.7 (Zwanenburg et al., 2011).

$$\mathbf{Y}_k = (\mathbf{X}_k + \mathbf{E})\mathbf{P}_k = \mathbf{T}_k + \mathbf{E}\mathbf{P}_k \quad \textbf{Equació 2.7}$$

En aquest model la projecció $\mathbf{Y}_k$ descriu la variació entre les mostres replicades per a un factor determinat $\mathbf{X}_k$. On $\mathbf{P}$ són els *loadings* i $\mathbf{T}_k$ són els *scores* del model PCA.

La diferència més important entre l'ANOVA-PCA i l'ASCA rau en com es comprova la significació dels factors. En l'ASCA, l'aplicació de l'anàlisi simultània de components (SCA) es fa directament sobre les matrius de cada factor. Per tant, per tal de provar estadísticament els resultats, s'aplica un test de permutació (Hoefsloot et al., 2009; Vis et al., 2007) que permet determinar la significació estadística dels efectes de tots els factors i de les seves interaccions. La hipòtesi nul·la (H_0) del test de permutació es basa en que el factor considerat no produeix efectes. El test s'aplica de manera que es permuten totes les files de les matrius de dades originals de tots els factors (en aquesta Tesi s'apliquen 10000 permutacions) i es recalculen la suma de quadrats de tipus I (dades equilibrades). La determinació de la significació de cada factor es fa mitjançant un histograma, que es construeix a partir de la suma de quadrats de totes les matrius permutades per un factor determinat. Quan la suma de quadrats de les matrius permutades és igual a la suma de quadrats de la matriu original s'accepta la hipòtesi nul·la, això vol dir que el factor considerat no té efectes sobre les dades experimentals. El valor *p* de la significació estadística (*p*-value) 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. El valor *p* és el nivell de significació més petit pel que la mostra obtinguda obligaria a rebutjar la hipòtesi nul·la (H_0). Si el valor *p* és menor que el nivell de significació prefixat es rebutja la H_0, en canvi, si és major aquesta hipòtesi s'accepta.

En aquesta Tesi les metodologies ANOVA-PCA i ASCA s'han aplicat en l'Article 1, en el qual s'avaluen 4 condicions ambientals diferents sobre el desenvolupament del blat. L'estudi s'ha fet a partir dels cromatogrames TIC (*Total Ion Current*) obtinguts mitjançant l'anàlisi LC-MS dels canvis dels perfils d'alguns metabòlits del blat obtinguts. Els quatre factors estudiats han estat: factor estadi de creixement (escala BBCH, 5 estadis), factor varietat (Astron, Ritmo, Stakado), factor tipus de mostra (brots, arrels) i factor tipus de cultiu (convencional, orgànic).

Resolució Multivariant de Corbes per Mínims Quadrats Alternats (MCR-ALS)

La resolució multivariant de corbes (*Multivariate Curve Resolution*, MCR) (Tauler, 1995; Tauler i Barceló, 1993; Tauler et al., 1993; Tauler et al., 1995) és un mètode quimiomètric que s'aplica a conjunts de dades agrupades en matrius o taules de dades amb la finalitzat de descriure la variació observada a partir d'un model bilineal amb un nombre reduït de components definits a partir de perfils que tenen significat físic directe i natural. El mètode MCR ha estat freqüentment utilitzat

per obtenir informació qualitativa i quantitativa en anàlisis de mescles complexes analitzades a partir de mètodes espectromètrics (de Juan i Tauler, 2003; de Juan i Tauler, 2006) , com és també el cas de les dades òmiques (Pérez et al., 2009). MCR es basa en un model de descomposició bilineal (vegeu equació 2.8) similar al del mètode PCA.

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad \textbf{Equació 2.8}$$

on $\mathbf{C}$ és la matriu de concentracions, que descriu els perfils de composició dels components presents a les mostres analitzades (files de la matriu) i $\mathbf{S}^{\mathrm{T}}$ és la matriu d'espectres (o respostes a les diferents variables mesurades en les columnes de la matriu) associats als diferents components resolts. Mentre que el nombre de files en les matrius $\mathbf{D}$ i $\mathbf{C}$ és el mateix, el nombre de columnes en $\mathbf{D}$ i $\mathbf{S}^{\mathrm{T}}$ coincideix. Com en el PCA, el nombre de components escollit (nombre de columnes en $\mathbf{C}$ i nombre de files en $\mathbf{S}^{\mathrm{T}}$) és el mínim necessari per descriure de forma suficient la variància en la matriu de dades original, $\mathbf{D}$.

L'objectiu principal dels mètodes de resolució multivariant és recuperar un conjunt de perfils dels components resolts ($\mathbf{C}$ i $\mathbf{S}^{\mathrm{T}}$), els quals permetin fer directament la seva interpretació física. Les solucions buscades per MCR es sotmeten, per tant, a una sèrie de restriccions naturals que proporcionen significat físic als components resolts. Aquestes restriccions són propietats naturals dels perfils dels components reals (per exemple les restriccions de no-negativitat). En MCR no s'apliquen restriccions d'ortogonalitat com en PCA, ja que les solucions que s'obtindrien tindrien valors negatius i no tindrien una significació física directa. De tota manera, mentre que el PCA proporciona solucions úniques, les solucions MCR no són úniques en general i porten associades una cert grau ambigüitat (Tauler et al., 1995). Tot i així, en el cas de dades LC-MS aquesta ambigüitat queda fortament reduïda degut a l'elevada selectivitat de les mesures d'espectrometria de masses.

En el mètode MCR-ALS, el model MCR de l'equació 2.8 es resol a través d'una optimització iterativa per mínims quadrats alternats (*Alternating Least Squares*, ALS) de $\mathbf{C}$ i/o $\mathbf{S}^{\mathrm{T}}$. El mètode MCR-ALS ha estat àmpliament descrit i aplicat en treballs anteriors (de Juan et al., 2014; de Juan i Tauler, 2006b; Jaumot et al., 2005; Tauler, 1995; Tauler i Barceló, 1993; Tauler et al., 1993). En aquesta Tesi només es descriu breument la seva adaptació a l'anàlisi de dades metabolòmiques obtingudes mitjançant LC-MS.

El mètode MCR-ALS descompon la matriu de dades **D** mitjançant un algorisme de mínims quadrats alternats (ALS) que minimitza les dues equacions següents (equació 2.9 i 2.10) sota restriccions adients:

$$\min_{\mathbf{C},restriccions}\left\|\widehat{\mathbf{D}}_{\mathrm{PCA}}-\widehat{\mathbf{C}}\widehat{\mathbf{S}}^{\mathrm{T}}\right\| \qquad \textbf{Equació 2.9}$$

$$\min_{\mathbf{S^T},restriccions}\left\|\widehat{\mathbf{D}}_{\mathrm{PCA}}-\widehat{\mathbf{C}}\widehat{\mathbf{S}}^{\mathrm{T}}\right\| \qquad \textbf{Equació 2.10}$$

on $\widehat{\mathbf{D}}_{\mathrm{PCA}}$ és la matriu de dades reproduïda per PCA pel nombre de components considerat. $\widehat{\mathbf{C}}$ i $\widehat{\mathbf{S}}^{\mathrm{T}}$ són respectivament les estimacions de les matrius de factors **C** i $\mathbf{S}^{\mathrm{T}}$ (concentracions i espectres) de l'equació 2.8 obtinguts durant l'optimització per ALS. En el context d'aquesta Tesi, les dues matrius obtingudes per MCR-ALS, **C** i $\mathbf{S}^{\mathrm{T}}$, reben el nom respectivament de perfils d'elució cromatogràfica i d'espectres de masses dels metabòlits resolts.

El procediment MCR-ALS comença amb la selecció del nombre de components i amb unes estimacions inicials les quals es modifiquen posteriorment de manera iterativa sota l'acció de diferents restriccions. L'estimació del nombre de components (compostos químics) importants es fa com en els mètodes PCA o SVD (Golub et al., 1970), de manera que s'expliqui una part important de la variància en la matriu de dades original, **D**. La descomposició bilineal s'inicia a partir d'una estimació inicial de **C** o de $\mathbf{S}^{\mathrm{T}}$. Per obtenir les estimacions inicials d'una d'aquestes dues matrius, **C** o $\mathbf{S}^{\mathrm{T}}$, es pot fer a partir de les columnes o de les files que siguin més 'pures' de la matriu de dades original, **D**, és a dir, que siguin més diferents entre sí i que no continguin només soroll experimental. Una forma força emprada a la bibliografia de fer aquesta estimació es per exemple a partir d'un procediment similar al proposat en el mètode SIMPLISMA (*SIMPLe-to-use Iterative Self-Modeling Analysis*)(Windig et al., 2005; Windig i Stephenson, 1992). Quan ja es disposa d'aquesta estimació inicial, per exemple de la matriu $\mathbf{S}^{\mathrm{T}}$, s'inicia el procediment d'optimització iteratiu per mínims quadrats alternats (ALS).

En el cas de no aplicar cap restricció durant el procés ALS, les solucions de les equacions 2.9 i 2.10 es descriuen per l'equació 2.11 quan es disposa d'una estimació inicial de la matriu $\mathbf{S}^{\mathrm{T}}$, o per l'equació 2.12 si l'estimació inicial és de la matriu **C**.

$$\hat{\mathbf{C}} = \hat{\mathbf{D}}_{\mathrm{PCA}}(\hat{\mathbf{S}}^{\mathrm{T}})^{+} \qquad \textbf{Equació 2.11}$$

$$\hat{\mathbf{S}}^{\mathrm{T}} = (\hat{\mathbf{C}})^{+}\,\hat{\mathbf{D}}_{\mathrm{PCA}'} \qquad \textbf{Equació 2.12}$$

On $\left(\hat{\mathbf{S}}^{\mathrm{T}}\right)^{+}$ i $(\hat{\mathbf{C}})^{+}$ i són les estimacions de les pseudoinverses de les matrius $\mathbf{S}^{\mathrm{T}}$ i $\mathbf{C}$ respectivament (Golub i Loan, 1996). Les solucions obtingudes en aquest cas són òptimes des del punt de vista de mínims quadrats, però no ho són des d'un punt de vista físic, ja que les concentracions o els espectres poden contenir valors negatius o no complir altres propietats conegudes dels perfils de $\mathbf{C}$ i $\mathbf{S}^{\mathrm{T}}$.

Com ja s'ha esmentat anteriorment, les restriccions són propietats que donen significat físic a les solucions obtingudes del model bilineal proposat en l'equació 2.8, que es compleixen en els perfils reals, i que actuen dirigint el procediment d'optimització iteratiu cap a la solució amb significat físic i interpretable. L'aplicació de restriccions redueix les ambigüitats en el càlcul de la solució final. A partir del coneixement físic del sistema es poden restringir les solucions, de manera que compleixin un determinat nombre de condicions conegudes. En el context de les dades cromatogràfiques de dades òmiques, cal especialment esmentar les restriccions de no-negativitat i de normalització dels espectres de masses ($\mathbf{S}^{\mathrm{T}}$).

La restricció de no-negativitat és la més general en les optimitzacions per mínims quadrats alternats (ALS). Les concentracions dels components químics han de ser sempre valors positius o zero, i per tant, s'apliquen als perfils de concentració ($\mathbf{C}$). De la mateixa manera, en espectrometria de masses, els espectres també són positius. Hi ha diferents formes d'aplicar la restricció de no-negativitat, per una banda mitjançant la substitució directa dels valors negatius per zero durant els diferents passos iteratius, per penalització de la funció a optimitzar o utilitzant algoritmes rigorosos de mínims quadrats no negatius (*non negative least squares*, nnls) (Lawson i Hanson, 1995) i les seva variants més ràpides, com és la de mínims quadrats ràpids no negatius (*fast non-negative least squares*, fnnls) (Bro i De Jong, 1997), tal i com s'expressa en les equacions 2.13 i 2.14.

$$\hat{\mathbf{C}} = fnnls\,(\hat{\mathbf{D}}_{\mathrm{PCA}}, \hat{\mathbf{S}}^{\mathrm{T}}) \qquad \textbf{Equació 2.13}$$

$$\hat{\mathbf{S}}^{\mathrm{T}} = fnnls\,(\hat{\mathbf{D}}_{\mathrm{PCA}}, \hat{\mathbf{C}}^{\mathrm{T}}) \qquad \textbf{Equació 2.14}$$

La restricció de normalització s'aplica per eliminar l'ambigüitat d'escala en els perfils resolts i també per facilitar l'evolució de l'algorisme de mínims quadrats alternats (ALS) durant l'optimització. Es pot aplicar sobre els perfils de concentració (**C**) o sobre els perfils d'espectres ($\mathbf{S}^{\mathrm{T}}$) (Tauler et al., 1995). La restricció de normalització sobre els perfils d'espectres es pot fer de diverses maneres, per exemple dividint tots els valors de l'espectre d'un component determinat per la seva àrea o alçada. D'aquesta manera es forcen els perfils de tots els components a tenir la mateixa àrea o alçada igual a la unitat.

L'avaluació de la qualitat del model MCR-ALS es fa utilitzant dos paràmetres, que són el percentatge de falta d'ajust (*lack of fit*, lof %, vegeu equació 2.15) i el percentatge de variància explicada (R^2 %, vegeu equació 2.16).

$$lof\% = 100\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{i,j}^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{i,j}^2}}, \qquad e_{i,j} = d_{i,j} - \hat{d}_{i,j} \qquad \textbf{Equació 2.15}$$

$$R^2\% = 100\,\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{i,j}^2 - \sum_{i=1}^{n}\sum_{j=1}^{m} e_{i,j}^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{i,j}^2} \qquad \textbf{Equació 2.16}$$

on $d_{i,j}$ és un valor experimental en la matriu de dades per a una variable j i una mostra i; $\hat{d}_{i,j}$ és el valor corresponent calculat utilitzant MCR-ALS (equació 2.8).

El paràmetre *lack of fit* indica la falta d'ajust del model MCR. Per defecte no existeix un valor òptim d'aquest paràmetre, doncs depèn de la qualitat de les dades originals i de la relació senyal soroll de les mesures. Per a trobar el valor adequat d'aquest paràmetre es construeixen models amb un nombre creixent de components i s'observa com aquest increment del nombre de components afecta a l'ajust del model. Si la manca d'ajust millora voldrà dir que el model anterior no tenia la suficient informació per descriure millor el sistema, d'altra banda, si la manca d'ajust empitjora o no varia de forma significativa, vol dir que l'addició de components al model no és necessària. A

més, per escollir el nombre de components i assegurar la qualitat d'un determinat model MCR cal avaluar la qualitat i versemblança dels perfils $\mathbf{C}$ i $\mathbf{S}^{\mathrm{T}}$ obtinguts en el context del problema que es vol resoldre. Com ja s'ha dit al començament, l'objectiu final dels mètodes MCR és la recuperació de perfils $\mathbf{C}$ i $\mathbf{S}^{\mathrm{T}}$ amb significat físic, iguals o el més propers possibles als perfils vertaders dels components que han causat la variació observada experimentalment.

Resolució Multivariant de Corbes aplicada a Matrius Augmentades

El model bilinear MCR abans descrit per una taula o matriu de dades es pot generalitzar a l'anàlisi simultània de múltiples matrius de dades arranjades en matrius augmentades, ja sigui en la direcció de les seves files o de les seves columnes, o de totes dues. L'extensió del model MCR aplicada a l'anàlisi de matrius augmentades en la direcció de les columnes (*column-wise*) es pot expressar en forma matricial de la següent manera (equació 2.17):

$$\mathbf{D}_{\mathrm{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_{\mathrm{K}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_{\mathrm{K}} \end{bmatrix} \mathbf{S}^{\mathrm{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_{\mathrm{K}} \end{bmatrix} \qquad \textbf{Equació 2.17}$$

O de forma més resumida (equació 2.18):

$$\mathbf{D}_{\mathrm{aug}} = \mathbf{C}_{\mathrm{aug}}\mathbf{S}^{\mathrm{T}} + \mathbf{E}_{\mathrm{aug}} \qquad \textbf{Equació 2.18}$$

on la matriu augmentada $\mathbf{D}_{\mathrm{aug}}$, formada per la concatenació de diferents matrius experimentals, es descompon en el producte de la matriu augmentada de perfils d'elució/concentració $\mathbf{C}_{\mathrm{aug}}$ amb la d'espectres $\mathbf{S}^{\mathrm{T}}$, deixant a la matriu residual $\mathbf{E}_{\mathrm{aug}}$, la variància no explicada (residuals) no explicada pel model MCR.

En el cas de dades LC-MS metabolòmiques, cada matriu $\mathbf{D}_{\mathrm{k}}$ conté l'anàlisi d'una mostra mitjançant aquesta tècnica. Generalment, les diferents matrius de dades $\mathbf{D}_{\mathrm{k}}$ es refereixen a l'estudi d'un mateix sistema metabolòmic sota condicions diferents. Per exemple, es refereixen a l'anàlisi LC-MS de diferents mostres extretes d'un mateix organisme (p.e. llevat), que han estat sotmeses a condicions

diferents (temperatura, salinitat, exposició a un determinat compost químic, contaminant ambiental, etc.) i que es comparen amb les corresponents al mateix tipus d'anàlisi per LC-MS de mostres control on no s'han aplicat aquestes condicions. La matriu augmentada $\mathbf{C}_{\mathrm{aug}}$ contindrà els perfils d'elució i concentració dels diferents constituents resolts en les diferents mostres analitzades cromatogràficament (perfils de concentració en les mostres tractades i en les mostres de control). En aquesta matriu augmentada $\mathbf{C}_{\mathrm{aug}}$, els perfils de concentració, $\mathbf{C}_{\mathrm{k}}$, apareixeran seguint el mateix ordre que en les matrius de dades originals $\mathbf{D}_{\mathrm{k}}$ individuals, abans de ser concatenades per donar la matriu experimental augmentada $\mathbf{D}_{\mathrm{aug}}$. Cal remarcar que en el model descrit per l'equació 2.17, els perfils d'elució cromatogràfica (pics cromatogràfics) d'un mateix component resolt en les diferents matrius $\mathbf{C}_{\mathrm{k}}$ poden estar desplaçats i diferir en posició (temps de retenció) i també en la seva forma. D'aquesta manera s'obvien els problemes relacionats amb la dificultat d'alineament cromatogràfic entre diferents mostres $\mathbf{D}_{\mathrm{k}}$ (diferents anàlisis cromatogràfiques). Aquest aspecte dona una gran flexibilitat en la modelització bilineal MCR dels perfils cromatogràfics i contrasta amb altres aproximacions proposades on s'han d'aplicar procediments d'alineament i de modelització dels pics cromatogràfics, els quals necessiten d'ajustos complexos i sotmesos a incerteses.

Per altra banda, i en contraposició als perfils de concentració en la matriu augmentada $\mathbf{C}_{\mathrm{aug}}$, en el model de l'equació 2.17, la matriu espectral $\mathbf{S}^{\mathrm{T}}$ conté els espectres de masses els constituents resolts que ara sí, són els mateixos per a totes les matrius originals individuals. És a dir, es considera que l'espectre MS del mateix constituent o metabòlit en les diferents mostres analitzades (matrius $\mathbf{D}_{\mathrm{k}}$) és sempre el mateix, siguin quines siguin les condicions experimentals de les diferents mostres, controls o tractades. Aquesta restricció dona una gran potencialitat a l'anàlisi conjunta de diverses matrius de dades.

En aquesta Tesi s'han fet dos treballs de metabolòmica basats en l'anàlisi de dades per LC-MS no dirigida (*untarget*), en els que s' estudien les diferències entre els perfils metabòlics en grups diferents de mostres (p.e. mostres control i mostres tractades) (vegeu Article 2 i Article 3). Les estratègies no dirigides requereixen un processament extensiu ja que els cromatogrames LC-MS generats contenen grans conjunts de dades multivariants. Com a conseqüència, un dels primers passos necessaris abans del processament de dades pròpiament és el de permetre la reducció de les dimensions dels conjunt de dades originals sense pèrdua d'informació. En aquesta Tesi, s'han utilitzat dues estratègies per a la construcció de les matrius de dades a partir dels perfils cromatogràfics (LC-MS), l'estratègia de compressió de *binning* (contenidors d'igual mida), i

l'estratègia de les regions d'interès, ROI. Aquestes dues estratègies es descriuen amb més detall a continuació.

Compressió de les dades mitjançant *binning* i aplicació de MCR-ALS en finestres de temps

El procediment de compressió *binning* (per contenidors d'igual mida) transforma les dades originals (*raw data*) en una matriu de dades on els espectres MS als diferents temps de retenció es troben en les seves files, i els cromatogrames als diferents valors *m/z* ajustats en les seves columnes. La conversió d'espectres originals d'alta resolució (mesurats a valors diferents de *m/z* no equidistants) en aquesta nova matriu requereix l'establiment d'un nou eix de valors *m/z* separats en parts equidistants, amb una mida específica (*bin*) prèviament decidida. Aquest procediment comporta un cert grau de pèrdua de resolució espectral en comparació a la resolució instrumental original de les dades. En aquest procediment de *binning,* la compressió de dades es realitza només en la dimensió *m/z.* En l'Article 2 el procediment de *binning* i l'anàlisi MCR-ALS s'ha aplicat a matrius augmentades en la direcció de les columnes (vegeu equació 2.17 i figura 2.2). En la majoria dels casos, com que les dimensions de les matrius individuals poden ser excessivament grans per les capacitats d'emmagatzematge i de càlcul disponibles, aquestes matrius augmentades s'han de subdividir en submatrius que contenen diferents finestres cromatogràfiques. La matriu augmentada corresponent a una mateixa finestra de temps cromatogràfic, j, per a totes les mostres, dóna una matriu augmentada $\mathbf{D}^{j}_{aug}$, la qual es descompon ara de manera similar utilitzant el model bilineal de l'equació 2.8 (vegeu equació 2.19 i 2.20).

$$\mathbf{D}^{j}_{aug} = \begin{bmatrix} \mathbf{D}^{j}_{1} \\ \mathbf{D}^{j}_{2} \\ \mathbf{D}^{j}_{3} \\ \vdots \\ \mathbf{D}^{j}_{K} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^{j}_{1} \\ \mathbf{C}^{j}_{2} \\ \mathbf{C}^{j}_{3} \\ \vdots \\ \mathbf{C}^{j}_{K} \end{bmatrix} \mathbf{S}^{j,T} + \begin{bmatrix} \mathbf{E}^{j}_{1} \\ \mathbf{E}^{j}_{2} \\ \mathbf{E}^{j}_{3} \\ \vdots \\ \mathbf{E}^{j}_{K} \end{bmatrix} \quad \text{per } j\text{=I, II, ..., J finestres de temps cromatogràfic;}$$

i per k=1,2,... mostres analitzades **Equació 2.19**

O de forma més compacta:

$$\mathbf{D}^{j}_{aug} = \mathbf{C}^{j}_{aug}\mathbf{S}^{j,T} + \mathbf{E}^{j}_{aug} \qquad \textbf{Equació 2.20}$$

Cada una de les submatrius ($\mathbf{D}^j_1, \mathbf{D}^j_2, \mathbf{D}^j_3 \ldots \mathbf{D}^j_K$) de $\mathbf{D}^j_{aug}$ té un nombre de files igual al nombre de temps d'elució considerat per a la determinada finestra cromatogràfica seleccionada, j, tot i que aquest nombre de files (temps de retenció) pugui ser diferent per a les diferents submatrius considerades $\mathbf{D}^j_k$. En canvi, el nombre de columnes de les submatrius ($\mathbf{D}^j_1, \mathbf{D}^j_2, \mathbf{D}^j_3 \ldots \mathbf{D}^j_K$) de la matriu augmentada $\mathbf{D}^j_{aug}$ és sempre igual al nombre de valors *m/z* considerats, que en el cas d'aquest procediment de *binning* sempre és el mateix. $\mathbf{C}^j_{aug}$ conté els perfils d'elució cromatogràfica dels diferents constituents en totes les submatrius $\mathbf{C}^j_k$. $\mathbf{S}^{j,T}$ té els espectres de masses purs d'aquests constituents resolts en la finestra de temps cromatogràfica considerada j, i la matriu $\mathbf{E}^j_{aug}$ conté tota la variància no explicada pel model descrit per $\mathbf{C}^j_{aug}$ i $\mathbf{S}^{j,T}$ (Tauler, 1995).

En aquesta Tesi la metodologia de *binning* s'ha aplicat a l'Article 2 en el qual el llevat s'ha cultivat a dues temperatures diferents: en condicions òptimes a 30ºC i en condicions estressants a 42ºC, per a la determinació dels canvis en els seus perfils metabòlics. Aquesta és una anàlisi no dirigida (*untarget*) on els perfils metabòlics són adquirits mitjançant LC-MS i posteriorment analitzats mitjançant MCR-ALS.

Compressió de les dades mitjançant ROI

Aquest mètode ha estat recentment introduït per (Gorrochategui et al., 2015) i s'ha utilitzat inicialment en aquesta Tesi per a facilitar el tractament de les dades obtingudes per LC-MS d'alta resolució. I es basa en la compressió de les dades MS a partir de la recerca de les regions d'interès (*Regions Of Interest*, ROI) en els cromatogrames LC-MS sense pèrdua de resolució espectral. Les regions d'interès contenen les dades *m/z* més rellevants, és a dir, aquelles que tenen una intensitat significativa més alta que un llindar preseleccionat de la relació senyal/soroll (S*ignal to Noise Ratio*, SNR) i que es troben dins d'un rang d'error de lectura dels valors de *m/z* (*m/z error*) similar al de la resolució instrumental de l'equip MS emprat. Els ROIs es busquen espectre per espectre, i les regions on aquests són comuns es combinen per obtenir un determinat nombre final de ROIs, 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 (*m/z error*). A partir d'aquest procediment, finalment s'obté una matriu de dades per a cada mostra analitzada amb els espectres MS a les files (a cada temps de retenció cromatogràfica) i els cromatogrames a les columnes (a cada valor *m/z* dels ROIs escollits).

En aquesta Tesi, aquest procediment de preselecció dels valors de ROI s'ha aplicat en l'Article 3 en el qual es comparen diferents grups de mostres (controls versus tractats). Com que el nombre de ROIs pot variar entre les mostres, l'anàlisi simultània d'aquestes ha de considerar tots els valors obtinguts en les diferents mostres. En el cas d'utilitzar l'estratègia descrita en aquest apartat, s'unifiquen tots els valors dels ROIs de les diferents mostres i, per tant, es consideren tant els valors *m/z* ROI comuns com els no comuns entre mostres.

L'anàlisi simultània MCR-ALS de múltiples mostres requereix la construcció de matrius de dades augmentades en la direcció de les columnes (*column-wise*) (vegeu secció 2.5.3 i figura 2.2). L'estratègia d'anàlisi MCR-ALS de matrius de dades augmentada mitjançant *binning* o ROI és pràcticament la mateixa. La diferència es troba en que en les dades comprimides per *binning* el mètode MCR-ALS s'aplica generalment només a una regió particular del cromatograma (finestra de temps) i el nombre de components resolts és menor (una desena o menys). En canvi en l'estratègia que fa servir el ROI, MCR-ALS s'aplica a tot el cromatograma i, per tant, el nombre de components serà també molt superior (un centenar o més). La matriu augmentada de les mostres comprimides mitjançant el ROI, $\mathbf{D}_{\text{aug}}$, es descompon utilitzant el model bilineal de l'equació 2.8 (vegeu equació 2.21 i 2.22).

$$\mathbf{D}_{\text{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S}^{\mathrm{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} \quad \text{k=1,2,... mostres analitzades} \qquad \textbf{Equació 2.21}$$

O de forma més compacta:

$$\mathbf{D}_{\text{aug}} = \mathbf{C}_{\text{aug}}\mathbf{S}^{\mathrm{T}} + \mathbf{E}_{\text{aug}} \qquad \textbf{Equació 2.22}$$

Cada una de les mostres ($\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3 \ldots \mathbf{D}_K$) de $\mathbf{D}_{\text{aug}}$ té un nombre de files igual al nombre de temps d'elució de cada cromatograma, aquest nombre de files (temps de retenció) pot ser diferent per a les diferents mostres. En canvi, el nombre de columnes de les mostres ($\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3 \ldots \mathbf{D}_K$) de la matriu augmentada $\mathbf{D}_{\text{aug}}$ és sempre igual al nombre de valors *m/z* ROI considerats. $\mathbf{C}_{\text{aug}}$ conté els perfils d'elució cromatogràfica dels diferents constituents en totes les mostres, $\mathbf{C}_k$. $\mathbf{S}^{\mathrm{T}}$ té els

espectres de masses purs d'aquests constituents resolts. La matriu $\mathbf{E}_{\text{aug}}$ conté tota la variància no explicada pel model descrit per $\mathbf{C}_{\text{aug}}$ i $\mathbf{S}^{\text{T}}$ (Tauler, 1995).

En aquesta Tesi la metodologia ROI-MCR-ALS s'ha aplicat a l'Article 3 en el qual el llevat s'ha cultivat a diferents concentracions (control, 1 mM, 3 mM i 6 mM) de sulfat de coure ($CuSO_4$) a 30ºC per a veure si l'ió Cu(II) té un efecte significatiu sobre el creixement del llevat a nivell metabòlic. Aquesta és una anàlisi no dirigida (*untarget*) on els perfils metabòlics són adquirits mitjançant LC-MS i posteriorment analitzats mitjançant MCR-ALS.

Anàlisi de Regressió per Mínims Quadrats Parcials (PLSR)

L'anàlisi de regressió per mínims quadrats parcials (*Partial Least Squares Regresion*, PLSR) (Wold, 1966; Wold et al., 1993; Wold et al., 2001) és un mètode de regressió lineal multivariant que permet trobar models de correlació entre un conjunt de variables predictores, agrupades en una matriu o taula de dades, **X**, i un vector (o matriu) resposta de variables dependents **y** (o **Y**)**,** per un mateix conjunt de mostres.

Tradicionalment, la modelització de tipus lineal de **y** = f(**X**) es realitza a partir d'un mètode de regressió linear múltiple (*Multiple Linear Regression*, MLR), el qual funciona adequadament sempre i quan la relació sigui lineal i el nombre de variables predictores en **X** estiguin poc correlacionades entre si i el seu nombre sigui inferior al nombre de mostres. Existeixen diverses variants d'aquest mètode que permeten escollir un conjunt reduït de les variables originals que amb les que es realitza la regressió multilineal amb èxit, per exemple amb la regressió tipus escalonada (*stepwise*) (Draper i Smith, 1998). Però en molts casos, els instruments de mesura proporcionen un gran nombre de variables **X** que estan fortament correlacionades, a partir de les quals és difícil fer una preselecció inicial sense perdre precisió en les estimacions corresponents. Les variables mesurades freqüentment estan correlacionades, contenen soroll i són incompletes. En estudis òmics i de química ambiental les variables **X** són nombroses i fortament correlacionades entre elles (col·lineals), i contenen soroll experimental. Conseqüentment, el mètode MLR no és generalment adequat i el mètode PLSR proporciona una eina alternativa més adequada per estudiar aquest tipus de problemes.

La regressió PLS és un mètode de regressió inversa en el qual el model calculat relaciona les variables latents (*Latent Variables*, LV) de **X** (matriu de variables predictores) amb **y** (variables a predir) amb la finalitat de maximitzar la covariància entre **X** i **y** de forma òptima amb un nombre mínim de variables latents. L'equació general del mètode de regressió inversa és la següent (equació 2.23):

$$\mathbf{y} = \mathbf{bX} \qquad \textbf{Equació 2.23}$$

on **X** és la taula de dades que agrupa les variables predictores, **y** és el vector resposta de variables dependents i **b** és un vector que conté els coeficients de regressió calculats durant el calibratge (construcció) del model.

Per a cada LV, s'obté un nou vector de coeficients de pes (*weights),* **w,** que descriu la importància de les variables en cada variable latent en a la predicció del vector **y**. La matriu que agrupa els coeficients de pes, **W**, reflecteix la covariància entre **X** i **y** i s'utilitza per al càlcul del vector de regressió (**b**) (vegeu equació 2.24 i 2.25).

$$\mathbf{b} = \mathbf{X}^{+}\mathbf{y} \qquad \textbf{Equació 2.24}$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathbf{T}}\mathbf{W})^{-\mathbf{1}}\mathbf{Q}^{\mathbf{T}} \qquad \textbf{Equació 2.25}$$

El nombre de variables latents, LV, és sempre menor que el nombre de variables originals, i aquestes són ortogonals. Les fórmules matricials que descriuen el procés matemàtic del PLS estan descrites a continuació (equació 2.26 i 2.27):

$$\mathbf{y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \qquad \textbf{Equació 2.26}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \textbf{Equació 2.27}$$

On ara **T** i **U** són les matrius *de scores* de les variables latents (LV) de totes les observacions **X** i **y** respectivament, sobre les mostres, i **P** i **Q** són les matrius *loadings* de les variables en aquestes

variables latents, LV, i **E** i **F** són les matrius que contenen les variàncies no explicades (errors), en **X** i **y** respectivament, pel model PLS.

La construcció d'un model PLS consisteix en buscar el nombre de variables latents necessari per predir i explicar suficientment bé la variable **y** (o la matriu **Y,** en el cas que aquesta sigui una magnitud multivariable també, en la variant del mètode anomenada PLS2). Un mètode força comú per a la determinació del nombre de variables latents és el de la validació creuada (*Cross Validation*, CV) en el calibratge intern del model PLS. La validació creuada és una forma pràctica i fiable de comprovar la significació predictiva d'un determinat model (Wold et al., 1993; Wold et al., 2001). La validació creuada es porta a terme subdividint el conjunt de dades original en un nombre de subgrups sobre els quals es desenvolupa un mateix nombre de models PLSR paral·lels (Shao, 1993). El nombre òptim de variables latents seleccionades per a construir el model PLS ha de donar la variància residual de **y** més petita en la validació creuada tal i com es descriu a l'equació 2.28:

$$\%\ \text{variància residual de } \mathbf{y} = \frac{\sum_i (\hat{\mathbf{y}}_i - \mathbf{y}_i)^2}{\sum_i \mathbf{y}_i^2} 100 \qquad \textbf{Equació 2.28}$$

On $\hat{\mathbf{y}}_i$ és el valor predit i $\mathbf{y}_i$ és el valor mesurat.

En aquesta Tesi, el mètode PLSR s'ha aplicat en l'Article 3, en el que s'ha cultivat llevat a diferents concentracions (control, 1 mM, 3 mM i 6 mM) de sulfat de coure ($CuSO_4$) a 30ºC i s'ha avaluat si el Cu(II) té un efecte en el creixement del llevat a nivell metabòlic. En aquest cas l'anàlisi PLSR s'ha aplicat als perfils metabòlics del llevat adquirits mitjançant LC-MS (**X**) en relació a les diferents concentracions de Cu(II) (**y**).

Per altra banda, el PLSR s'utilitza també en l'Article 4 en el que es determinen i identifiquen quins són els compostos orgànics acumulats en les mostres de sediments marins analitzats mitjançant GC-MS, que covarien més intensament en relació als canvis de temperatura superficial del mar (S*ea Surface Temperature*, SST) observats. El PLSR relaciona la matriu formada pels cromatogrames TIC (*Total Ion Current*), **X,** i les SST associades a cada mostra (**y**).

Anàlisi Discriminant (PLS-DA)

L'anàlisi discriminant per mínims quadrats parcials (*Partial Least Squares – Discriminant Analysis*, PLS-DA) (Barker i Rayens, 2003) és una variant del mètode PLSR en la que la variable dependent **y** és un conjunt de variables binaries que descriuen categories o classes de les mostres analitzades en **X**. El PLS-DA fa l'estimació més eficient de les combinacions lineals dels valors originals i independents de la matriu **X** (LV), els quals es correlacionen de manera òptima amb els canvis observats en la variable dependent, **y**. El PLS-DA construeix un model que maximitza la covariància entre la **X** i la **y** amb un nombre mínim de LV. Per a cada variable latent, LV, es calcula un vector de coeficients de pes (**w**), el qual mostra quines són les variables de la matriu **X** que combinen millor per formar el vector de *scores*, **T.**

En aquesta Tesi aqueta metodologia s'ha aplicat en l'Article 2 en el qual el llevat s'ha cultivat a dues temperatures diferents: en condicions òptimes a 30ºC i en condicions estressants a 42ºC. Els perfils metabòlics adquirits mitjançant LC-MS (**X**) han estat relacionats amb la temperatura de cultiu del llevat (**y**) a través de l'aplicació del PLS-DA.

Per altra banda, en l'Article 3, en el qual el llevat s'ha cultivat a diferents concentracions de sulfat de coure ($CuSO_4$) a 30ºC (per a veure si el Cu(II) té un efecte en el creixement del llevat a nivell metabòlic), s'ha aplicat el mètode PLS-DA a les mostres de llevat sotmeses a concentracions de Cu(II) (**y**, control versus 6 mM) analitzades per LC-MS.

Selecció de variables

En molts casos d'anàlisis de dades multivariants, el nombre de mostres (objectes) és molt petit comparat amb el nombre total de variables analitzades; aquesta és precisament una característica molt habitual de les dades òmiques especialment en estudis no dirigits. No obstant això, moltes de les variables són en molts casos poc rellevants per al problema estudiat, ja que la variació que presenten no està relacionada amb la resposta o efecte investigat (variable dependent **y**), sinó amb altres fenòmens i efectes que són presents i interfereixen en el procés de mesura. És per això que en moltes ocasions el nombre de variables es pot reduir de forma significativa i es poden buscar aquelles variables que són més significatives pel problema que es vol estudiar i així minimitzar la

pèrdua d'informació. Cal recordar però, que aquest és un coneixement que s'ha d'adquirir durant l'estudi, i que no està disponible en les etapes inicials de l'aproximació no dirigida.

La selecció de variables permet la reducció de la complexitat de les dades i millorar la seva possible interpretabilitat química i/o biològica. Els mètodes de selecció de variables tenen per objectiu la selecció d'un conjunt més petit de variables que estiguin relacionades amb la variable resposta i, per tant, siguin necessàries per a construir un bon model explicatiu o predictiu de la variable **y**. O bé, en altres ocasions, per aconseguir interpretar i conèixer quines de les variables **X** estan realment correlacionades (lligades) a les mateixes fonts de variació que les de la variable **y** .

En aquesta Tesi s'han utilitzat els mètodes de selecció de variables importants en la projecció (*Variable Importance in Projection*, VIP) i dels quocients de selectivitat (*Selectivity Ratio*, SR).

Variables importants en la projecció (VIP)

El mètode de selecció de les variables importants en projecció, VIP, va ser inicialment proposat per (Wold et al., 1993). Aquest mètode dona una mesura de la influència individual de cada una de les variables de **X** en la construcció del model PLS per predir **y**. Els VIP, o coeficients VIP (VIP *scores*), es calculen a partir de la suma ponderada dels quadrats dels coeficients de pes (**w**, *weights*), obtinguts en la construcció del model PLS.

Els valors dels VIP són una mesura útil per tal de seleccionar quines variables contribueixen millor en l'explicació de la variància de la resposta **y**. Per a un model PLS sempre hi haurà un valor VIP per a cada variable, que es poden agrupar en un vector VIP.

El valor de VIP de la variable j es calcula a partir de l'expressió (equació 2.29):

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} \mathbf{w}_{jf}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}} \qquad \textbf{Equació 2.29}$$

En la qual $\mathbf{w}_{jf}$ és el valor del weight per la variable j en la variable latent f. SSY_f és la suma de quadrats de la variància explicada per la variable latent f. J és el nombre total de variables en la

matriu de dades **X**. SSY_{total} és la suma de quadrats total de la variable dependent **y** i F el nombre total de variables latents considerades. VIP_j és una mesura de la contribució de cada variable en funció de la variància explicada per cada variable latent del model PLS, en la qual $\mathbf{w}_{jf}^2$ representa la importància de la variable j en la variable latent f.

Les sumes de quadrats de l'equació 2.29 es calculen segons les expressions:

$$SSY_f = \mathbf{b}_f^2 \mathbf{t}_f' \mathbf{t}_f \qquad \textbf{Equació 2.30}$$

$$SSY_{total} = \mathbf{b}^2 \mathbf{T}' \mathbf{T} \qquad \textbf{Equació 2.31}$$

En la qual **T** són els *scores* de la matriu **X** i **b** és el vector de coeficients del model PLS.

Atès que la mitjana dels valors dels VIP al quadrat serà igual a 1 (Chong i Jun, 2005), es pot escollir aquest valor de la unitat o de més gran que la unitat (Chong i Jun, 2005) com a criteri per a la selecció de les variables més importants. El valor de 'més gran que la unitat' no té una justificació estadística i és sensible a la presència d'informació present a la matriu **X** que no estigui realment relacionada amb la resposta **y**. Per tant es recomana emprar aquest criteri només de forma inicial i orientativa.

En aquesta Tesi, aquesta metodologia s'ha aplicat en els Articles 2 i 3 en el qual s'estudien els perfils metabòlics de *S.cerevisiae* obtinguts mitjançant l'anàlisi per LC-MS a través de tècniques quimiomètriques multivariants. També s'ha aplicat en l'Article 4 on s'analitzen les senyals climàtics a partir de compostos orgànics fòssils acumulats als sediments marins de fins a 11.7 Ma i en l'Article 5 on es comparen els mètodes VIP i SR per a la selecció de variables.

Quocient de selectivitat (SR)

El quocient de selectivitat (*Selectivity Ratio,* SR) (Rajalahti et al., 2009a; Rajalahti et al., 2009b) es un paràmetre que es pot emprar per a la visualització de les variables **X** que contribueixen millor a l'explicació de la variància de la variable resposta **y**.

Kvalheim i coautors van desenvolupar un procediment anomenat *target rotation* o *target projection* (TP) per tal de simplificar la interpretació de les variables latents dels models de regressió. TP genera un únic component predictiu mitjançant la projecció de la descomposició de les variables latents sobre la variable resposta (Kvalheim i Karstang, 1989). El valor del quocient SR per a cada variable es defineix com el quocient entre la variància explicada i la variància residual (no explicada) en el vector TP.

En el càlcul del quocient SR s'utilitza tant la capacitat predictiva, a partir del vector de regressió $\mathbf{b}_{\mathrm{PLS}}$, com la capacitat explicativa (covariància). Els *scores* TP s'obtenen mitjançant la projecció de les files de la matriu de variables **X** sobre els coeficients de regressió **b** normalitzats (vegeu equació 2.32).

$$\mathbf{t}_{\mathrm{TP}} = \mathbf{X}\mathbf{b}_{\mathrm{PLS}}/\|\mathbf{b}_{\mathrm{PLS}}\| \qquad \textbf{Equació 2.32}$$

El vector de projecció, *loadings* $\mathbf{p}_{\mathrm{TP}}$, s'obté aleshores projectant les columnes de la matriu **X** sobre el vector dels *scores*, $\mathbf{t}_{\mathrm{TP}}$ (vegeu equació 2.33).

$$\mathbf{p}_{\mathrm{TP}} = \mathbf{X}\mathbf{t}_{\mathrm{TP}}/(\mathbf{t}'_{\mathrm{TP}}\,\mathbf{t}_{\mathrm{TP}}) \qquad \textbf{Equació 2.33}$$

El quocient entre la suma de quadrats de la variància explicada ($\mathrm{SS}_{\mathrm{explicada,i}}$) i la suma de quadrats variància residual ($\mathrm{SS}_{\mathrm{residual,i}}$) resulta ser el vector quocient de selectivitat, SR, que per a cada variable i es defineix a partir de les equacions 2.34, 2.35 i 2.36.

$$\mathrm{SS}_{\mathrm{i,explicada}} = \|\mathbf{t}_{\mathrm{TPi}}\mathbf{p}_{\mathrm{TPi}}{}'\|^2 \qquad \textbf{Equació 2.34}$$

$$\mathrm{SS}_{\mathrm{i,residual}} = \|\mathbf{e}_{\mathrm{TPi}}\|^2 \qquad \textbf{Equació 2.35}$$

$$\mathrm{SR}_{\mathrm{i}} = \mathrm{SS}_{\mathrm{i,explicada}}/\mathrm{SS}_{\mathrm{i,residual}} \qquad \textbf{Equació 2.36}$$

S'aplica un test *F* per tal de definir el criteri a partir del qual una determinada variable es considera que te una capacitat discriminant més elevada, i és seleccionada com a possible variable 'marcadora' del fenomen estudiat. Amb la finalitat de determinar quines variables tenen una major

capacitat discriminatòria en relació al vector resposta **y** i rebutjar la hipòtesi nul·la (quan la variància explicada i la residual és la mateixa), el valor de la *F* calculada (F_{calc}), el qual correspon al valor de SR_i de l'equació 2.36, ha d'excedir un valor crític escollit a partir de la distribució de l'estadístic *F* (F_{crit}) per uns determinats graus de llibertat i un nivell de significació del test prèviament decidit (equació 2.37).

$$F_{calc} = SR_i > F_{crit} = F(\alpha, N-2, N-3) \qquad \textbf{Equació 2.37}$$

En l'equació 2.37, N és el número de mostres i α el nivell de significació. En els treballs d'aquesta Tesi, s'ha escollit el criteri del test *F* a un nivell de significació del 0.05%.

En aquesta Tesi, el mètode de selecció de variables i de possibles marcadors SR s'ha aplicat en l'Article 3, en el qual s'estudien els perfils metabòlics de *S.cerevisiae* obtinguts mitjançant l'anàlisi per LC-MS a través de tècniques quimiomètriques multivariants. El mètode SR també s'ha aplicat també a l'Article 5 en el qual es comparen els mètodes VIP i SR per a la selecció de variables i la seva possible interpretació.

2.5.6 IDENTIFICACIÓ DELS METABÒLITS

En estudis metabolòmics, els senyals (p.e. les àrees dels pics cromatogràfics) dels metabòlits més rellevants (estadísticament significatius) la concentració dels quals canvia per efecte de l'estímul investigat, són utilitzats per la discriminació entre grups de mostres (p.e. mostres control i mostres tractades). Els metabòlits associats a aquests senyals s'identifiquen per tal d'associar la variació observada en els seus canvis de concentració a una explicació biològica coherent, i esbrinar-ne l'efecte del tractament estudiat. Per fer tot això és necessari realitzar la identificació temptativa dels metabòlits a partir dels seus espectres MS. Quan s'utilitza MS d'alta resolució (*High Resolution Mass Spectrometry*, HRMS) com a eina per a la identificació de compostos, les assignacions de metabòlits s'obtenen mitjançant la combinació de la seva massa exacta, de la seva distribució isotòpica, dels possibles patrons de fragmentació i de qualsevol altra informació MS o cromatogràfica disponible.

El càlcul de les combinacions químiques que s'ajusten a una certa massa exacta, és el primer pas per obtenir un conjunt de candidats per a la identificació del metabòlit detectat. Aquest conjunt

possible de candidats es redueix significativament en el cas de disposar d'un espectròmetre de masses que pugui proporcionar un valor molt exacte de la massa molecular. Per a aplicacions generals en el camp de la metabolòmica animal o vegetal, la majoria de metalls es poden excloure (excepte el sodi o el potassi que produeixen adductes comuns amb els productes orgànics ionitzats en els espectres de masses). Per tant, els elements centrals presents en els compostos analitzats seran el carboni, l'hidrogen, l'oxigen, el nitrogen, el fòsfor i el sofre. De tota manera, qualsevol altre element del qual es tingui la mínima evidència de la seva presència a les mostres analitzades ha de considerar-se també en el càlcul de la composició elemental.

Fins l'any 2007, la comunitat científica en estudis metabolòmics encara no havia arribat a un consens total sobre els procediments de treball en metabolòmica (Fiehn et al., 2007). En aquest mateix any, el grup de treball d'anàlisi química (*Chemical Analysis Working Group*, CAWG) de la iniciativa de normalització d'estudis metabolòmics (*Metabolomics Standards Initiative*) va redactar unes normes per a les anàlisis químiques en metabolòmica descrites al treball de Sumner i coautors (Sumner et al., 2007). Tot i així, després d'uns anys des de la publicació d'aquest treball, l'aplicació d'aquestes normes en publicacions científiques encara és limitada (Creek et al., 2014; Salek et al., 2013). L'ús d'un conjunt de normes bàsiques comunes per a la notació de metabòlits és essencial per aconseguir que la comunitat científica pugui avaluar i interpretar les dades i resultats amb confiança (Creek et al., 2014).

Un dels mètodes més comuns per a la identificació dels compostos mitjançant espectrometria de masses, és la comparació amb altres espectres de compostos patrons autèntics del mateix compost mitjançant biblioteques d'espectres de massa, com és per exemple MassBank (Horai et al., 2010). MassBank és una base de dades d'espectres de masses d'alta resolució de metabòlits i compostos químics petits (<3000 Da), en estudis biològics i de ciències de la salut. En aquest aquesta base de dades, la identificació dels metabòlits depèn de les biblioteques que contenen les referències de les dades ESI-MS^n. S'hi poden buscar els compostos químics que tenen les relacions *m/z* obtingudes experimentalment, així com també obtenir els seus espectres MS complets. .

Com ja s'ha esmentat anteriorment, *S.cerevisiae* és un dels organismes més intensament estudiats a la bibliografia (vegeu secció 2.4.2) i s'ha emprat com a model en diversos treballs d'aquesta Tesi (vegeu Article 2 i Article 3). És un dels organismes del qual se'n té una descripció més completa. Jewison i coautors (Jewison et al., 2012) van desenvolupar *The Yeast Metabolome Database*

(YMDB), la qual recull informació química, física i biològica de prop de 2000 metabòlits d'aquest organisme. La YMDB és una base de dades que cobreix la majoria del metaboloma del llevat i que conté els metabòlits ja descrits en llibres, articles científics i altres bases de dades electròniques. Per altra banda, la *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa et al., 2012; Ogata et al., 1999) és un recurs bioinformàtic molt poderós que compila moltes dades sobre els compostos químics i sobre la seva reactivitat, per a la comprensió de la seva funció biològica des d'una perspectiva genòmica. YMDB junt amb KEGG i MassBank han estat les eines bàsiques emprades en aquesta Tesi per a la identificació dels metabòlits del llevat detectats en aquesta Tesi.

2.5.7 INTERPRETACIÓ BIOLÒGICA

En general les tècniques quimiomètriques permeten processar un gran nombre de dades biològiques. De tota manera, la recollida i el processament d'aquestes dades no és encara suficient per entendre la dinàmica dels sistemes biològics. A més a més dels passos crucials de selecció i identificació dels metabòlits responsables dels processos biològics estudiats, és necessari l'anàlisi del conjunt d'ells i de conèixer la seva funció en un context biològic. Per a aquest propòsit, és essencial comprendre i interpretar els resultats obtinguts en termes dels mecanismes bioquímics subjacents i de les seves relacions fenotípiques i fisiològiques.

En la biologia de sistemes, la comprensió a nivell de sistemes exigeix un canvi en la noció del "què es busca" en biologia (Kitano, 2002). Mentre que la comprensió dels gens, les proteïnes, els metabòlits i de les seves funcions per separat segueix essent important, l'atenció es centra ara en la comprensió de l'estructura i la dinàmica del sistema format per tots ells en conjunt. Com que un sistema no és només un conjunt de gens, proteïnes i metabòlits, les propietats d'aquest sistema no es poden entendre simplement dibuixant diagrames de les seves interconnexions. Tot i que el diagrama representa un primer pas important, és un anàleg a un full de ruta estàtic, mentre que el que realment es busca conèixer són les xarxes dinàmiques entre les vies bioquímiques, per què aquestes xarxes emergeixen i com es poden controlar.

La identificació precisa i la quantificació relativa d'un elevat nombre de metabòlits en un grup de mostres fan possible l'estudi inicial de les xarxes metabòliques dinàmiques implicades. Aquesta nova metodologia és la que condueix a observacions inabastables mitjançant la utilització dels mètodes clàssics. L'anàlisi de les tipologies d'aquestes xarxes i el seu control, pel que fa a les

pertorbacions ambientals o genètiques específiques, permeten la investigació de les interaccions dinàmiques entre xarxes metabòliques i el descobriment de noves correlacions amb vies caracteritzades bioquímicament així com amb noves vies fins ara desconegudes.

Per a comprendre de forma global la biologia dels sistemes estudiats en aquesta Tesi, s'han de conèixer prèviament les funcions cel·lulars, les vies metabòliques i les reaccions enzimàtiques. La visualització d'aquestes rutes metabòliques pot ser molt rellevant per a la comprensió dels canvis observats en els metabòlits. La base de dades KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa et al., 2012; Ogata et al., 1999) és especialment útil per a integrar els conjunts de dades generades experimentalment. A KEGG, els metabòlits es poden situar sobre els mapes de rutes metabòliques de l'organisme estudiat. I amb la informació obtinguda es formulen noves hipòtesis dels possibles efectes o alteracions observats en les mostres, i profunditzar en els processos bioquímics.

En aquesta Tesi s'ha fet un primer intent d'integració dels resultats obtinguts a partir de la utilització de les bases de dades tipus KEGG i exploració de les rutes metabòliques alterades per efecte de les pertorbacions ambientals estudiades. En el capítol 4 d'aquesta Tesi es mostren dos exemples d'aquest tipus d'aproximació en l'estudi de mostres de llevat sotmeses a estrès tèrmic (Article 2) i a concentracions creixents de Cu(II) (Article 3).

Capítol 3

Avaluació quimiomètrica de diferents factors experimentals sobre el creixement de mostres de blat (*Triticum aestivum* L.) a partir dels perfils LC-MS dels derivats de les benzoxazinones

3.1 INTRODUCCIÓ

Les plantes sintetitzen una gran varietat de metabòlits secundaris que exerceixen rols importants en les interaccions complexes entre els organismes que viuen en el mateix sistema ambiental. En els darrers anys hi ha hagut un interès creixent en les perspectives d'explotació de l'al·lelopatia com a estratègia alternativa per al control de les males herbes, així com també per al control d'insectes i malalties de les plantes (Jabran et al., 2015; Schulz et al., 2013).

Diverses plantes expressen el fenomen al·lelopàtic a través de l'exsudació dels compostos anomenats al·leloquímics. La ubicació d'aquests compostos, la seva interacció amb els microorganismes de la rizosfera i les transformacions bioquímiques i fotoquímiques són factors a tenir en compte en el fenomen al·lelopàtic. A més a més, la producció d'al·leloquímics està influenciada per factors abiòtics com la temperatura, la llum, les característiques del sòl i la pluja, i per factors biòtics com per exemple el cicle de vida de la planta, la competència i els agents patògens, que resulten en una distribució desigual dels al·leloquímics en relació a les condicions ambientals. Aquest escenari complex requereix un enfocament apropiat per a l'avaluació d'aquests compostos. És per això que l'ús de les tecnologies òmiques pot ser molt útil en aquest camp per tal de tenir una visió més completa dels processos complexos que influeixen en la síntesi dels metabòlits de les plantes (Urano et al., 2010).

Villagrassa i altres coautors a través de diferents treballs (Villagrasa et al., 2006a; Villagrasa et al., 2007; Villagrasa et al., 2006b; Villagrasa et al., 2008) van determinar mitjançant una anàlisi cromatogràfica dirigida, el contingut de diversos compostos al·lelopàtics en plantes de blat (*Triticum aestivum* L.) i dels seus productes de degradació en els sòls agrícoles, en diferents varietats de blat cultivades sota diferents sistemes de cultiu. L'objectiu principal de l'estudi de Villagrassa i coautors (Villagrasa et al., 2006a; Villagrasa et al., 2007; Villagrasa et al., 2006b; Villagrasa et al., 2008) era el desenvolupament d'un mètode analític per a la determinació dels derivats de les benzoxazinones en plantes, mitjançant la combinació de l'extracció líquid-sòlid i de l'anàlisi per cromatografia líquida acoblada a espectrometria de masses (LC-MS). Les dades obtingudes a partir d'aquests estudis representaven un bon exemple per a l'estudi i comparació de diversos mètodes quimiomètrics d'anàlisi multivariant en l'avaluació dels efectes que podien tenir

diversos factors experimentals sobre el cultiu del blat en relació a la síntesi i abundància de benzoxazinones (metabòlits secundaris).

Es van cultivar tres varietats de blat (*T. Aestivum* L.) anomenades Astron, Ritmo i Stakado (factor varietat: Astron, Ritmo i Stakado) en dos espais iguals però diferenciats un per l'agricultura orgànica i l'altre per la convencional (factor tipus de cultiu: orgànic i convencional). Les mostres es van recol·lectar en 5 estadis de creixement diferents (factor estadi de creixement: 1r estadi, 2n estadi, 3r estadi, 4t estadi i 5è estadi) d'acord amb l'escala BBCH (*Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie*) (Hess et al., 1997; Zadoks et al., 1974), la qual identifica els estadis de desenvolupament fenològic d'una planta. La presa de mostra es va dur a terme separant els brots de les arrels (factor tipus de mostra: brot i arrel). L'anàlisi LC-MS (Villagrasa et al., 2006a; Villagrasa et al., 2007; Villagrasa et al., 2006b; Villagrasa et al., 2008) es va dur a terme per a la detecció simultània dels compostos glucosats 2-*β*-D-glucopyranosyloxy-4-hydroxy-1,4-benzoxazin-3-one (DIBOA-Glc) i 2-*β*-D-glucopyranosyloxy-4-hydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA-Glc;), les aglucones 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) i 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA), les benzoxazolinones benzoxazolin-2-one (BOA) i 6-methoxybenzoxazolin-2-one (MBOA), i les lactames 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (HBOA) i 2-hydroxy-7-methoxy-1,4-benzoxazin-3-one (HMBOA). Aquests compostos són àcids hidroxàmics del grup de les benzoxazinones amb propietats al·leloquímiques (vegeu secció 2.4.1 de la introducció de la Tesi, capítol 2) i afavoreixen la resistència de la planta a plagues i malalties (Cambier et al., 1999; Fomsgaard et al., 2004; Niemeyer, 1988; Sicker et al., 2000; Søltoft et al., 2008).

Les anàlisis de les dades experimentals presentades en aquest capítol es van realitzar a partir d'una anàlisi metabolòmica dirigida (*target*). L'aproximació d'anàlisi dirigida, tal i com es descriu a la secció 2.2 de la introducció de la Tesi (capítol 2), centra els esforços en la determinació de les concentracions d'un nombre reduït de compostos químics ja coneguts sobre les mostres analitzades on s'han variat un grup de paràmetres o variables experimentals. En particular s'estudien els perfils cromatogràfics TIC (*Total Ion Current*) obtinguts per l'anàlis LC-MS dels derivats de les benzoxazinones del blat (*Triticum aestivum* L.) per a l'avaluació quimiomètrica de diferents factors experimentals en el desenvolupament del blat. En particular, es van utilitzar la combinació del mètode estadístic d'anàlisi de la variància (*ANalysis Of VAriance*, ANOVA) amb el mètode multivariant d'anàlisi de components principals (*Principal Component Analysis*, PCA)

per una banda, ANOVA-PCA, i del mètode d'ANOVA amb el mètode d'anàlisi simultani de components (*Simultaneous Component Analysis*, SCA), ASCA. Aquests mètodes estan descrits a la secció 2.5.5 de la introducció de la Tesi (capítol 2).

Els perfils cromatogràfics TIC van ser investigats amb les metodologies descrites, ANOVA-PCA i ASCA, per a determinar la influència dels diferents factors de cultiu (estadi de creixement, tipus de mostra, varietat, tipus de cultiu) en la síntesi dels metabòlits secundaris seleccionats (DIBOA-Glc, DIMBOA-Glc, DIBOA, DIMBOA, BOA, MBOA, HBOA, HMBOA). Aquests algorismes han estat estudiats a fons i programats durant aquesta Tesi, la qual cosa ha permès entendre millor el model subjacent emprat i, per tant, entendre i interpretar millor els resultats de la seva aplicació. Actualment, l'algorisme ASCA es troba implementat a la PLS Toolboox de Matlab versió 7.8.

Analytica Chimica Acta 731 (2012) 24–31

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Chemometric evaluation of different experimental conditions on wheat (*Triticum aestivum* L.) development using liquid chromatography mass spectrometry (LC–MS) profiles of benzoxazinone derivatives

Mireia Farrés[a], Marta Villagrasa[b], Ethel Eljarrat[a], Damià Barceló[a,b], Romà Tauler[a,*]

[a] Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Catalan Institute for Water Research (ICRA), Parc Científic i Tecnològic de la Universitat de Girona, Edifici H20, Emili Grahit 100, 17003 Girona, Spain

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 13 April 2012
Accepted 16 April 2012
Available online 21 April 2012

Keywords:
Chemometrics
Triticum aestivum L.
Benzoxazinone derivatives
LC–MS
ANOVA–PCA
ASCA

ABSTRACT

Different chemometric techniques have been used to evaluate the effect of distinct experimental conditions and factors on *Triticum aestivum* L. plant development. The study was conducted using three wheat varieties, Astron, Ritmo and Stakado. These varieties were grown under organic and conventional cultivation systems. Samples were collected at five growth stages. Shoots and roots of each plant at these stages were analysed. Three replicates of each analysed sample were performed to improve representativeness and to allow for the evaluation of natural variability and interaction effects. All samples were analysed using Liquid Chromatography Mass–Spectrometry (LC–MS), and the Total Ion Current (TIC) profiles of benzoxazinone derivatives obtained for each sample were investigated. Qualitative and quantitative assessments of these TIC profiles and of their changes in the analysed samples were carried out using different chemometric techniques. Estimation of main effects, and of their possible interaction, was performed by means of Analysis of Variance combined to Principal Component Analysis (ANOVA–PCA) and of Analysis of Variance combined to Simultaneous Component Analysis (ASCA).

1. Introduction

In recent years, there has been an increasing interest to exploiting allelopathy as a biological strategy to minimize the perceived hazardous impacts from herbicides and insecticides in agriculture production. Allelopathy has been defined as any direct or indirect effect (stimulatory or inhibitory) caused by a plant, including microorganisms, on another plant through production of secondary metabolites released into the environment [1,2]. The impact of allelopathy can be exploited for pest and weed control [3,4]. It has been proven that slight changes in the metabolism of plants can be explained by perturbations imposed on them (i.e. plants react to any change in their surroundings), and everything the plant does can be followed by looking at changes in the low molecular weight chemicals (metabolites). There is an enormous diversity of allelochemicals in nature [5]. Among them, benzoxazinones (hyrdoxamic acids, lactams, benzoxazolinones and methyl derivatives of hydroxamic acids) are a group of secondary metabolites implicated on natural plant resistance. Benzoxazinones have been widely studied during the last decade and several analytical methods using HPLC for these compounds have been developed [6,7]. Production of allelochemicals in living plants is affected by abiotic and biotic factors [8].

Understanding the complexity of the influence of different factor levels in a plant growth using an experimental design strategy requires the use of multivariate data analysis methods. The influence of the factors of interest should be separated from each other to draw sensible conclusions from data analysis results. A widely used method for mutivariate data analysis is Principal Component Analysis (PCA). It gives a simplified lower-dimensional representation of the variation that is presented in a dataset. The scores and loadings obtained by PCA can be visualised and interpreted in the context of the problem under study. However, this approach generally does not take the underlying experimental design into account. Thus, the different sources of variation are confounded in the PCA model, and this can seriously hamper the interpretation of the principal components [9]. A well recognised methodology for the analysis of data from designed experiments is Analysis of Variance (ANOVA), which focuses on the separation of the different sources of data variation. Therefore, combining ANOVA with PCA (ANOVA–PCA) will incorporate the pre-knowledge on the data structure into the model. This methodology has been already proposed [10] and it consists in using the experimental design to separate the variation of the experimental hypothesis from other

* Corresponding author. Tel.: +34 93 400 61 40; fax: +34 93 204 59 04.
E-mail address: Roma.Tauler@idaea.csic.es (R. Tauler).

0003-2670/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2012.04.017

potentially confounding sources of variation, previously to the application of PCA. In the same direction, ANOVA–Simultaneous Component Analysis (ASCA) [9,11] has been developed to increase the interpretability of the changes occurred using a multivariate dataset in terms of an appropriate experimental design. By application of ANOVA, the different factors are modelled separately and the significance of each factor and interactions among factors can be investigated [12]. The application of PCA or SCA on the ANOVA model enhances the interpretation of ANOVA results. A comparison of the two ANOVA based methods was done [13,14]. In both cases the influence of each experimental factor could be assessed through PCA scores. And the corresponding PCA loadings can be used to investigate which metabolites exert the largest influence of each factor level.

In the present work, three varieties of wheat (*Triticum aestivum* L.), Astron, Ritmo and Stakado were grown under conventional and organic cultivation, to assess the effect of the different cultivations on the benzoxazinone derivatives. Samples were collected at five growth stages and shoots and roots of each plant were analysed. Four different factors were considered in the experiments: growth stage (1–5), wheat variety (Astron, Ritmo, Stakado), type of cultivation (organic and conventional) and tissue sample (shoots or roots). Three replicates of all these experiments were performed to improve representativeness and allow the evaluation of natural variability and interaction effects. The metabolic profile of benzoxazinone derivatives acquired through Liquid Chromatography–Mass Spectrometry (LC–MS) was used to generate Total Ion Current chromatograms (TICs). Appropriate data pretreatments were performed first, including baseline correction and peak alignment to compensate for minor shifts in retention times. The resulting preprocessed TIC chromatograms were investigated with both ANOVA–PCA and ASCA methodologies; considering the effects of the five different investigated possible factors in order to evaluate the influence of them. This work compared both methods using multivariate datasets with four factors some of them tested at more than two levels.

2. Experimental

2.1. Experimental set-up

Three different wheat (*T. aestivum* L.) varieties, Astron, Ritmo and Stakado, were grown in Lleida (Catalunya, NW of Spain) in triplicate, under conventional and organic cultivation. Shoots and roots of the plants (tissue sample) were collected at five different growth stages. The stages were defined by the BBCH scale, a system for a uniform coding of phenologically similar stages of plant species [6,15].

Wheat samples were lyophilised and extracted by Pressurised Liquid Extraction (PLE) a technique for extracting benzoxazinone derivatives using an ASE 200 (Dionex, Idstein, Germany), the solvent composition was 100% acidified MeOH (1% HOAc). The organic extracts were concentrated to dryness by rotary evaporation and redissolved in 2 mL of acidified water (1% HOAc) then were filtered using 1 μm 25 mm syringe-driven filter units. Purification was performed via LiChrolut RP C_{18} Solid Phase Extraction (SPE) cartridges. Analytes were eluted using 5 mL of acidified MeOH (1% HOAc). Then, extracts were concentrated to dryness and reconstituted with 2.5 mL of [MeOH/H_2O (0.05% HOAc) (60:40)]. Finally 25 μL of 100 ng μL^{-1} of the internal standard (2-MeO-HBOA) were added to the extract for internal standard quantification. A clear explanation of sample preparation is given in [6,16].

2.2. Instrumental Analysis

Chromatographic analyses were performed on a HP 1100 LC–MS. A Synergi Max-RP 80A (C-12 TMS) LC column (250 × 4.6 mm, Phenomenex) with a solvent flow rate of 1 mL min^{-1} was used. The sample injection volume was set at 50 μL. Acidified H_2O (0.05% HOAc) and MeOH were used as the elution solvents A and B, respectively. The solvent gradient adopted was as follow: 0–2 min, 100–70% A; 2–19 min, 70–40% A; 19–21 min, 40–5% A; 21–23 min, 5–5% A; 23–25 min, 5–70% A; 25–30 min, 70–100% A. MS analyses were carried out in a LC–MSD HP 1100 and these were performed in selected ion monitoring (SIM) mode. Mass selective detector equipped with atmospheric pressure ionization source was used with electrospray interface, and it was operated in negative ion mode. The selected analytes and the corresponding ions were the following: 2-β-D-glucopyranosyloxy-4-hydroxy-1,4-benzoxazin-3-one (DIBOA-Glc; 134, 342); 2-β-D-glucopyranosyloxy-4-hydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA-Glc; 164, 372); 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (HBOA; 164, 108); 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA; 134, 78); 2-hydroxy-7-methoxy-1,4-benzoxazin-3-one (HMBOA; 194, –); 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA; 164, 149); benzoxazolin-2-one (BOA; 134, –) and 6-methoxybenzoxazolin-2-one (MBOA; 164, 149). The mass spectrometer was interfaced to a computer workstation running Chemstation software (A.08.03) for data acquisition and processing. Further information about the instrumental analysis is given in [6,16].

2.3. Data arrangement and preprocessing

The resulting Total Ion Current (TIC) chromatograms were the ion sum of the selected analytes (DIBOA-Glc, DIMBOA-Glc, HBOA, DIBOA, DIMBOA, BOA and MBOA; see above). TIC chromatograms were exported one by one through Chemstation software (A.08.03) to 'csv' format files, and imported directly to Matlab version 7.4, where they were preprocessed and analysed using different chemometric methods. A total number of 180 TIC chromatograms were exported and a total number of 640 retention times were selected, ranging from 5.28 to 18 min. The rest of retention times were eliminated because they did not contain any further information. Therefore, the whole data set finally was arranged in a data matrix of 180 rows (equal to the number of samples) and 640 columns (equal to the number of retention time indices).

Every chromatographic run gave a set of chromatographic peaks corresponding to the analysed ions obtained in the fragmentation of the analytes of interest (benzoxazinone derivatives). The aim of the data pre-processing step was to remove the variance unrelated to changes in chemical composition, for a further improved processing step using multivariate data analysis methods afterwards. TIC chromatograms were individually baseline corrected to zero by subtracting the minimum value of each chromatogram from all values of the same chromatogram.

Several techniques have been proposed to correct misalignments or retention time shifts of chromatographic data [17–19]. Possible chromatographic retention time misalignments can be produced by changes in chromatographic columns during their use (ageing), by minor changes in mobile phase composition, by instrumental drifts and also by possible interactions between analytes [18,20–25]. Correlation Optimised Warping (COW) method [18] was selected for peak alignment, this algorithm required two input parameters: the segment length *m* and the slack size *t* [18,20,26]. The length *m* and the slack *t* parameters were selected using a routine that optimises automatically the segment length and the slack size for COW [23]. The optimised values of these two parameters

were 40 and 1 respectively. The selected reference chromatogram was the one with a greater number of representative peaks and it was chosen on a trial error basis by visual inspection of the profiles after preprocessing. It resulted to be the chromatogram of a sample TIC chromatogram of the first growth stage, Ritmo wheat variety, organic cultivation and root tissue.

2.4. Data analysis

ANOVA–PCA and ASCA are two multivariate data analysis approaches that combine statistical advantages of ANOVA to separate the variance sources, and advantages of PCA for eliminating covariation among variables and explain maximum variance. These two multivariate approaches were applied to the analysis of the TIC chromatographic data matrix. The raw data matrix is decomposed into the sum of four different data matrices characterising separately the variance of each one of the investigated factors, plus a residual matrix containing the unexplained variance [10,27]. In particular, data used in this work are analysed using the Multilevel Simultaneous Component Analysis (MSCA) approach [9].

The four factor balanced ANOVA general linear additive model including all possible interactions can be written as follows:

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{X}_a + \mathbf{X}_b + \mathbf{X}_c + \mathbf{X}_d + \mathbf{X}_{(ab)} + \mathbf{X}_{(ac)} + \mathbf{X}_{(ad)} + \mathbf{X}_{(bc)} + \mathbf{X}_{(bd)} + \mathbf{X}_{(abc)} + \mathbf{X}_{(abd)} + \mathbf{X}_{(acd)} + \mathbf{X}_{(bcd)} + \mathbf{X}_{(abcd)} + \mathbf{E} \quad (1)$$

In Eq. (1), $\mathbf{X}$ is the raw experimental Total Ion Current (TIC) data matrix, $\overline{\boldsymbol{X}}$ is the grand mean data matrix with all rows equal to the TIC average chromatogram of the whole dataset, $\mathbf{X}_a$ is the effect of growth stage factor, $\mathbf{X}_b$ is the effect of variety factor, $\mathbf{X}_c$ is the effect of cultivation factor, $\mathbf{X_d}$ is the effect of tissue sample factor, $\mathbf{X}_{(ab)}$ is the interaction of growth stage and variety factors, $\mathbf{X}_{(ac)}$ is the interaction of growth stage and cultivation factor and so on; and $\mathbf{E}$ is the residual matrix representing the natural variation among replicates. Eq. (1) is only a theoretical model which in practice is difficult to interpret for all the contributions, especially for those referred to the interactions.

All data matrices in Eq. (1) have the same sizes but differ in their mathematical rank, which follows the degrees of freedom of ANOVA. All rows in the grand mean matrix, $\overline{\boldsymbol{X}}$, are equal to the average TIC chromatogram, therefore its rank is equal to the unity. The different factor effect matrices contain in their rows the average of the considered factors at its different levels. The rank of each factor matrix is equal to minus one. The interaction matrices have in their rows the average of the samples that are characterised by the same levels of the considered interaction factors.

The proposed four-way factor design is a balanced case, Eq. (1); it means that each factor level had exactly the same number of observations. In balanced designs, type I sum of squares approach is used to test different factors in an ANOVA. It consists in allocating sequentially the part of the explained variance to the main effects; then the two-way interactions and then increasingly higher-order interactions if present. The sum of squares corresponding to each of the factors, to their interaction, and to the residual variance is orthogonal. Thus, the total type I sum of squares is equivalent to the sum of all of the squared factor matrices [28,29]. In this work main attention is given to balanced designs, although in biological studies it is worth extending ANOVA also to unbalanced data [30]. In the ANOVA-PCA methodology, PCA is applied separately to each one of the factor matrices combined with the residual matrix [10], i.e. to $\mathbf{X}_a + \mathbf{E}$. Scores intervals for each sample group can be shown using the convex hull method [31]. The convex hulls are drawn through the extreme objects of the same class around the whole data set in the score plots, thus approximate intervals of the different tested effects can be easily visualised. To test the mentioned significance of the factors, their effect should be the dominant source of variation compared to the residual matrix in PCA results. Therefore, the first principal component (PC1) should mainly characterise this effect over the second principal component (PC2) which will mainly reflect random variation [10]. The null hypothesis, defined as no differences in the experimental results caused by the considered factors, is rejected when the scores of the factor cluster together for its different levels and align along the first principal component axes orthogonal to the residuals in second principal component axes.

In the ASCA methodology, on the contrary, each PCA model is fitted to each factor matrix individually [32]. The ASCA model corresponding to ANOVA equation of a two factor experimental design is given in Eq. (2):

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{T}_a\mathbf{P}_a^T + \mathbf{T}_b\mathbf{P}_b^T + \mathbf{T}_c\mathbf{P}_c^T + \mathbf{T}_d\mathbf{P}_d^T + \mathbf{T}_{(ab)}\mathbf{P}_{(ab)}^T + \mathbf{T}_{(ac)}\mathbf{P}_{(ac)}^T + \mathbf{T}_{(ad)}\mathbf{P}_{(ad)}^T + \mathbf{T}_{(bc)}\mathbf{P}_{(bc)}^T + \mathbf{T}_{(bd)}\mathbf{P}_{(bd)}^T\mathbf{T}_{(abc)}\mathbf{P}_{(abc)}^T + \mathbf{T}_{(abd)}\mathbf{P}_{(abd)}^T + \mathbf{T}_{(acd)}\mathbf{P}_{(acd)}^T + \mathbf{T}_{(bcd)}\mathbf{P}_{(bcd)}^T + \mathbf{T}_{(abcd)}\mathbf{P}_{(abcd)}^T + \mathbf{E} \quad (2)$$

Where component scores of each submodel are given by the matrices indicated by $\mathbf{T}_a$, $\mathbf{T}_b$, $\mathbf{T}_{(ab)}$ and the component loadings are given by matrices $\mathbf{P}_a$, $\mathbf{P}_b$, $\mathbf{P}_{(ab)}$. $\mathbf{E}$ is a matrix in which the residuals of all submodels of the ASCA model are collected ($\mathbf{E} = \mathbf{E}_a + \mathbf{E}_b + \mathbf{E}_{(ab)}$).

As a consequence of applying PCA directly to factor matrices, it is not possible to appreciate the natural variability in the ASCA scores plot. Hence, the estimation of the replicates variation in the principal component subspace of a factor is given by Eq. (3) [14].

$$\mathbf{Y}_k = (\mathbf{X}_k + \mathbf{E})\mathbf{P} = \mathbf{T}_k + \mathbf{E}\mathbf{P}_k \quad (3)$$

In this model the projection $\mathbf{Y}_k$ describes the variation among replicates for a given factor $\mathbf{X}_k$, where $\mathbf{P}_k$ are the loadings and $\mathbf{T}_k$ are the scores of the PCA model for $\mathbf{X}_k$.

The difference between these two methods (ANOVA–PCA and ASCA) is in the way that the significance of factors is tested. As mentioned in ASCA, the SCA step is applied on the factor matrices directly and no assessment of the significance of the results obtained is then made. However, as proposed by [33] and [27], a permutation test can be implemented to assess the statistical significance of the effects of all factors and of their interactions.

Assuming the general factor ANOVA model for balanced data, a permutation test can be set to check the statistical significance of the effects of all factors and of their interactions [33]. The null hypothesis (H_0) assumes that there is no effect of the factor. A permutation test is preformed by randomly permuting the original data matrix (i.e. 10,000 permutations) and recalculating the type I sum of squares of the factors. A histogram of the type I sum of squares of the original and of the permuted data of each factor was constructed, to evaluate the significance of the factor effect. When the randomisations give a type I sum of squares that is equally large as the original type I sum of squares, the null hypothesis is accepted, and the different levels of the factor are considered not to differ. The *p*-value is then calculated dividing the number of cases on which the type I sum of squares are larger than the original type I sum of squares by the number of performed permutations.

All calculations were operated in Matlab software version 7.4. The four factor ANOVA model and the permutation test were achieved by a self-made Matlab algorithm. PCA was performed using Matlab PLS Toolbox version 5.8.

3. Results and discussion

3.1. Total variance

A heat map of the intensities for the TIC chromatograms grouped by the considered factors was initially used to display the overall differences between the TIC chromatograms. The heat map

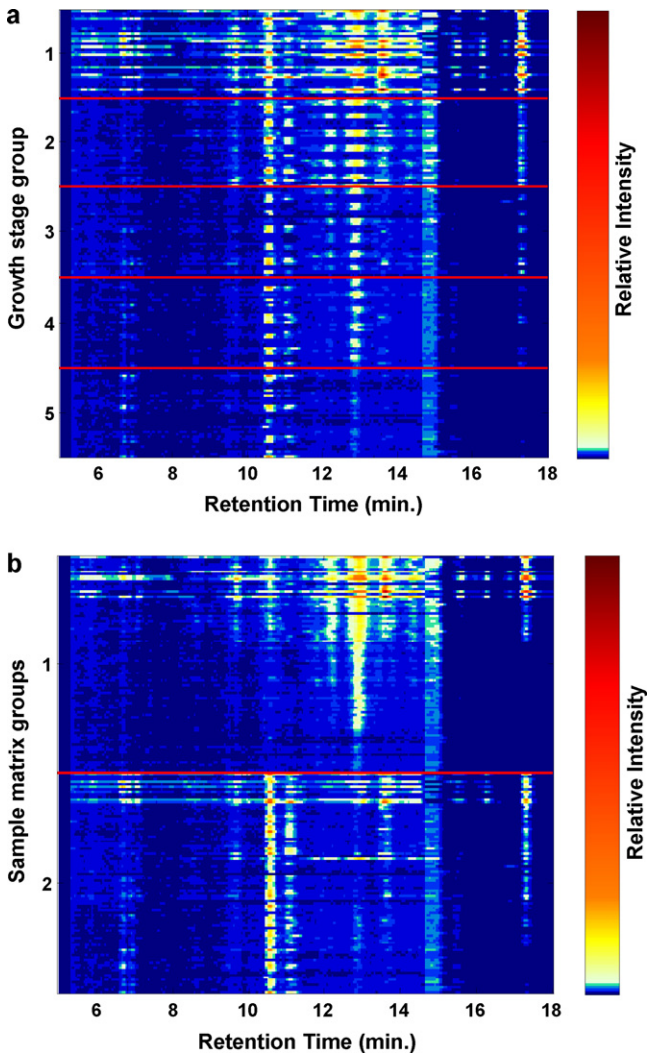


Fig. 1. Heat map image of TIC chromatogram intensities. TIC samples grouped by (in *Y*-axis) (a) growth stage levels and (b) tissue sample levels. *X*-axis is retention times of TIC chromatograms.

(Fig. 1) illustrates the intensity differences among the factor levels. In Fig. 1a it can be appreciated that each level of the growth stage factor has a distinct response in the peak intensity. Similarly the tissue sample factor is shown in Fig. 1b, where depending on the considered tissue sample (shoots or roots) the predominance of the peaks changes. For the remaining factors, the three plant varieties (Astron, Ritmo, Stakado) and the two types of cultivation (conventional and organic), did not show detectable factor effects (corresponding figures are not given).

3.2. ANOVA–PCA

PCA scores obtained from the growth stage data matrix versus residual data matrix is shown in Suppl. Fig. A. The separated effect of the stage factor versus residual data matrix ($\mathbf{X}_a + \mathbf{E}$) is represented; growth factor resulted to be important for the characterisation of the observed data variance. Principal component 1 (PC1) captured 84.77% of total data variance and separated samples of the first growth stage (except for a small group of first growth stage samples) from the samples of other growth stages, which were highly overlapped among them. Therefore, the variance explained by PC1 was describing mostly the growth of the wheat plant, especially in its initial development (first growth stage).

Scores plot resulting from the PCA of the tissue sample factor versus residual data matrices ($\mathbf{X}_d + \mathbf{E}$) is provided in Suppl. Fig. B in supplementary material. This figure shows a first grouping (shoots) along PC1 axis (70.68% of the variance) and another one (roots) along PC2 axis (21.07% of the variance). Consequently in this case, the results from PCA for the tissue sample factor were not conclusive.

When PCA was applied to the wheat variety factor versus residual data matrices ($\mathbf{X}_b + \mathbf{E}$), only a slight differentiation from Astron to Ritmo and Stakado varieties along PC1 (78.95% of variance) was appreciated (see Suppl. Fig. C). However responses to this factor were influenced by a high natural random variability. Finally, it was not possible to recognise any pattern in PCA of the type of cultivation factor versus residual data matrices ($\mathbf{X}_c + \mathbf{E}$), in this case wheat responses were also highly influenced by natural variability (see Suppl. Fig. D).

When comparing the different factor data matrices with the residual data matrix, the same pattern for all the considered factors was observed. Samples belonging to the first growth stage were highly variable among them; consequently, interpreting the score plots became rather difficult. When samples from the first growth stage were not considered, resulting scores changed significantly. Growth stage samples from the second to the fifth levels were distinguished more clearly in PC1 (55.33% of the variance) versus PC2 (32.79% of the variance) score plots (see Fig. 2a) than when all growth stage levels were included. In Fig. 2b, the scores obtained for tissue sample factor are shown. In this case also, when the samples from the first growth stage level were not considered, shoots and roots sample groups were clearly distinguished along PC1, which accounted already for a 80.99% of the data variance.

ANOVA–PCA was applied to the different interaction matrices versus residual matrices, but growth stage factor was the only one that gave significant interactions with other factors. A clustering along PC1 (explained more than 75% of data variance) was observed in all cases and this incremented in the later growth stages. Tissue sample factor did not interact with wheat variety and type of cultivation factors and no differentiation could be observed of the different level groups, with scores of all PCs overlapped (figures are not shown for brevity).

3.3. ANOVA Simultaneous Component Analysis (ASCA)

ASCA results for the permutation test when applied to the growth stage factor are presented in Fig. 3. Vertical lines in the sub-plots of Fig. 3 indicated the sum of squares for the experimental data. As it can be seen, all permutations gave a lower sum of squares than the original data (black line), at a significance level of $p < 0.0001$. The same result was obtained for the tissue sample factor and for the 'growth stage x tissue sample' interaction. Thus, the null hypothesis could be rejected at this significance level for the growth stage and tissue sample factors and for the 'growth stage x tissue sample' factor interaction. The null hypothesis for the remaining tested factors (wheat variety and type of cultivation) was accepted (with significance levels of $p = 0.61$, $p = 0.22$ respectively) and therefore they were not considered relevant, since most of the randomizations gave a sum of squares larger than the experimental sum of squares (figures for these factors are not given for brevity).

The effect of the growth stage factor (data matrix $\mathbf{X}_a$) obtained by ASCA is shown in Fig. 4a. First principal component (PC1) explained 96.70% of data variation. The variation was found both, in the presence or in the absence of the metabolite ions. In every sample a specific group of metabolites was synthesised depending on its growth stage. The number of detected ions was lower in the latest growth stages. These results were consistent with those previously obtained [16], where a general decrease of the total amount of

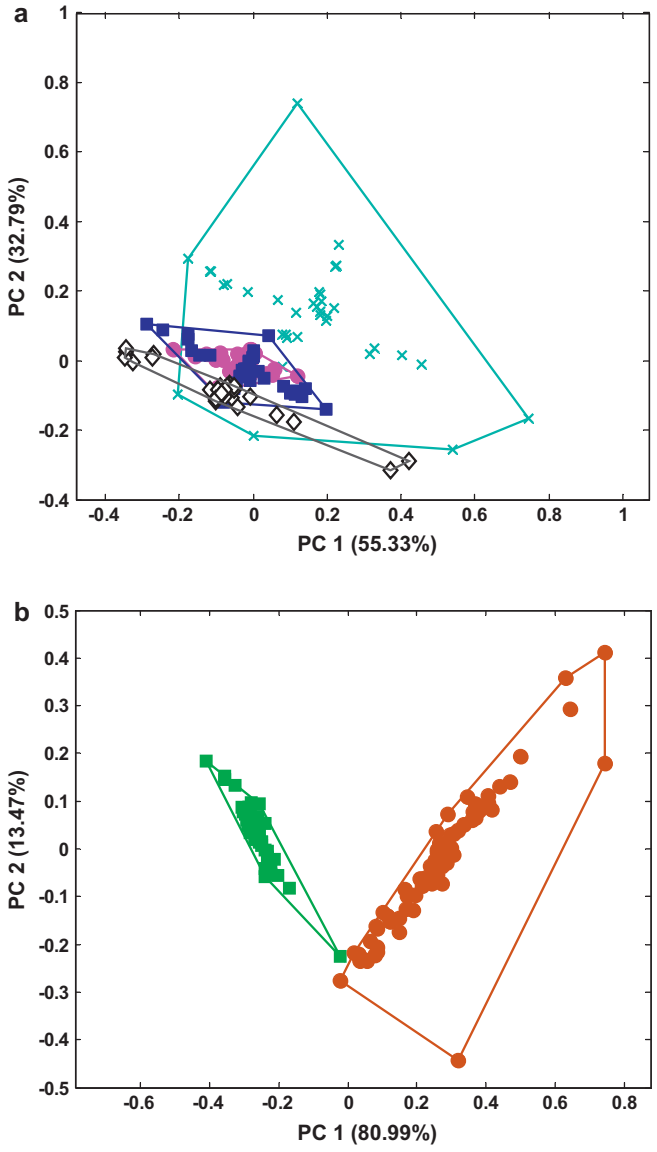


Fig. 2. ANOVA–PCA results: (a) PCA scores plot for growth stage factor versus residual matrix without considering samples from the first growth stage. Different symbols indicate different levels of the factor:cyan crosses are second growth stage, magenta solid circles are third growth stage, blue squares are fourth growth stage and white diamonds are fifth growth stage. (b) PCA scores plot for tissue sample factor versus residual matrix without considering samples from the first growth stage. The two symbols indicate the two levels of this factor: green squares are shoots and brown solid circles are roots. Convex hulls are drawn around each factor level and plotted with the same colour as the level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

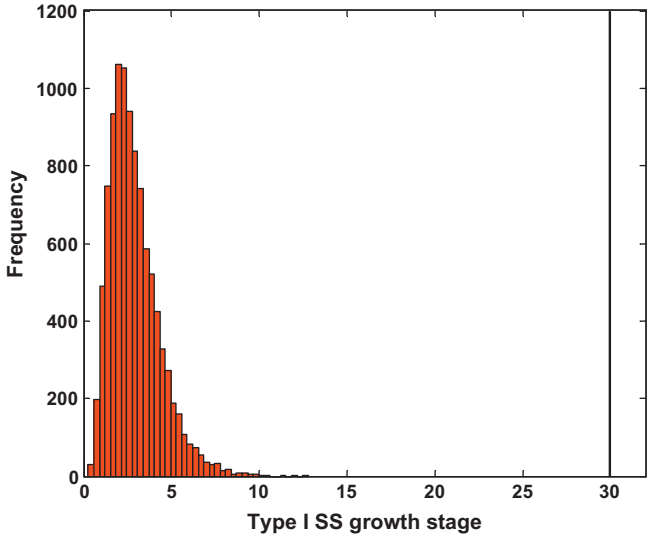


Fig. 3. Distribution plot of sum of square (SS) values obtained after 10,000 permutations of the growth stage factor, indicated by red boxes. Black vertical line represents the sum of squares resulting from the true experimental data for the growth stage factor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

allelochemical content in wheat plants was detected when plant goes from the first to the second growth stage. Although the permutation test verified the significance of this factor (Fig. 3), visual inspection of Fig. 4a does not indicate a clear trend. The observed wide variation within samples of the first growth stage is due to the high amount of ions present in this stage, but not all of these metabolites are present with the same intensity in all samples. This is probably a consequence of the natural variability. In the later growth stages, the same phenomenon also occurs, although to a lesser extent.

Changes in metabolites composition between shoot and root samples are shown in Fig. 4b (data matrix $\mathbf{X}_{d}$). In this plot, PC1 collected all the corresponding variance (100%). Sample scores 37–180 corresponding to second to fifth growth stages were clearly separated in two clusters as it was also obtained by ANOVA–PCA (Fig. 2b). Samples corresponding to the first growth stage were highly variable and shoot and root clusters were not distinguished (samples 1–36 in Fig. 4b). These results were in agreement with those previously obtained by Morgensen and coauthors [34], who also observed the differences in secondary metabolite concentrations according to what type of wheat plant tissue sample (shoot or root) was analysed.

First and second components resulting from the 'growth stage x tissue sample' interaction matrix ($\mathbf{X}_{(ad)}$) accounted for 92.90% and 6.45% of the variance, respectively. PC1 separated the samples from third to the fifth growth stages in two clusters, one corresponding to shoots and the second corresponding to roots. PC2 had the same pattern but only with the samples from the second and fifth stages. Once again the replicates from the first level of the growth stage factor appeared to be influenced by a wide natural variability; hence they were not effectively clustered (figures of interaction matrices are not shown).

It is concluded that the natural variability of the first level of the growth stage factor was strongly hiding the importance of the other considered factors. Therefore, when Simultaneous Component Analysis was performed on factor matrices without considering the first stage growth samples; clustering of scores improved significantly. In Fig. 5, a clear separation of Stakado variety from Astron and Ritmo varieties is shown. Astron and Ritmo varieties did not differentiate in the plot since they were overlapped. On the other side, both types of wheat cultivation could not be distinguished and therefore the type of cultivation factor is not found to be significant in any case (figure not shown).

3.4. Interpreting ASCA loadings

The relative importance of different variables (allelochemicals) was determined by analysing loading plots obtained after ASCA. These plots resembled TIC chromatographic plots and could be interpreted looking at the retention indices, their MS spectra and identifying at what chemical compound corresponded. Retention times that contributed mostly to a particular component were associated with large negative or positive coefficients in the corresponding loading plot. It was confirmed that the larger loadings

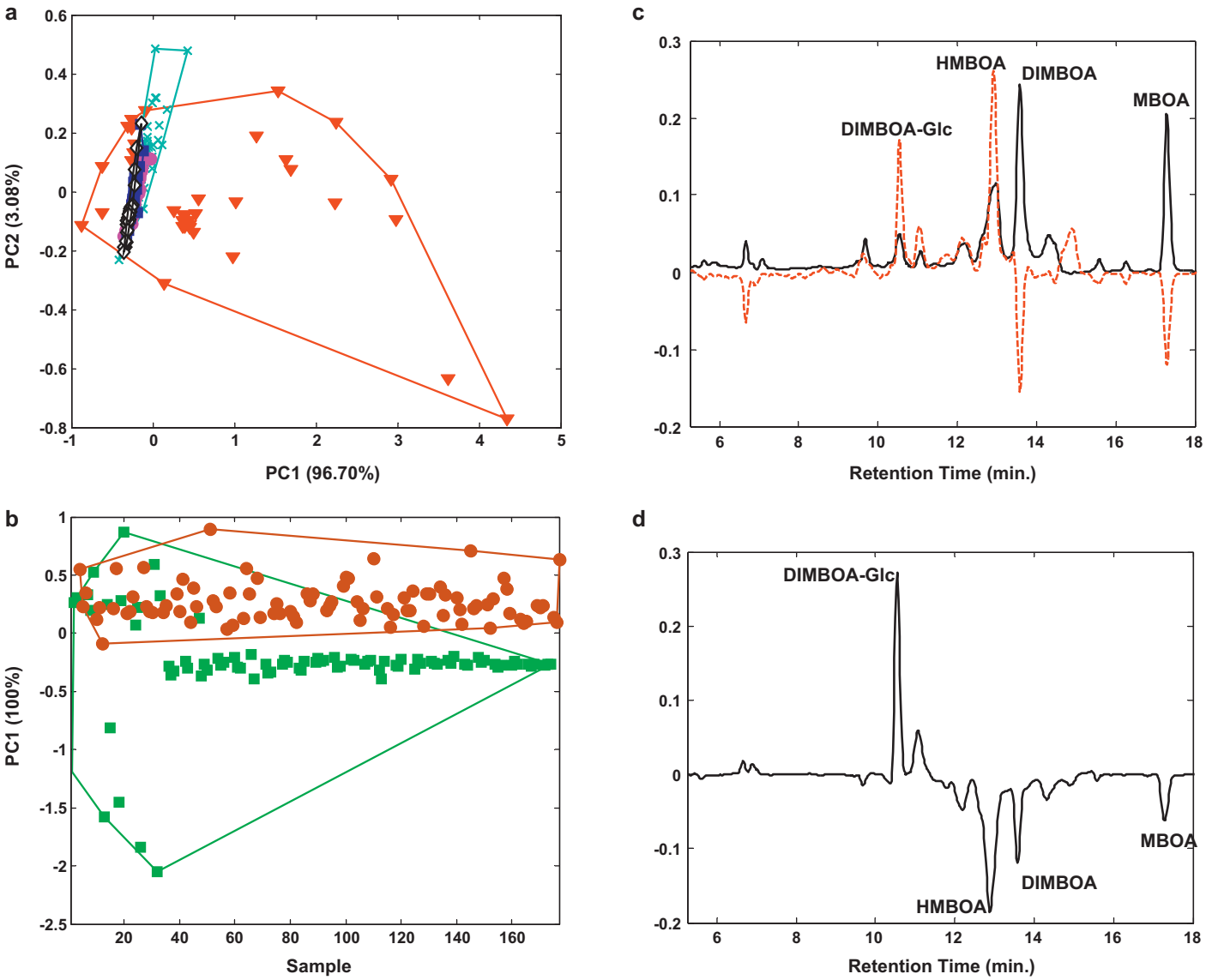


Fig. 4. ASCA results: (a) SCA scores plot of the growth stage factor matrix including projections of $\mathbf{Y}_k$ (see Eq. 3). Symbols indicate the different levels of the growth stage factor: red triangles are first growth stage, cyan crosses are second growth stage, magenta solid circles are third growth stage, blue squares are fourth growth stage and white diamond are fifth growth stage. (b) SCA scores plot of the tissue sample factor matrix including projections of $\mathbf{Y}_k$ (see Eq. 3). Symbols indicate the levels of the sample matrix factor: green squares are shoots and brown solid circles are roots. In the two cases (Fig. 4a and b). Convex hulls are drawn around each factor level and plotted with the same colour as the level. (c) SCA loadings plot of the growth stage factor matrix. Black line are loadings on PC1 and dotted red line are loadings on PC2. (d) SCA loadings plot on PC1 of tissue sample factor matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

corresponded to some of the metabolites investigated in previous studies of the same samples [6,15,16,35]. The metabolites that resulted more significant in this study were the cyclic benzoxazinone derivatives: DIMBOA-Glc, DMBOA, HMBOA and MBOA (DIBOA-Glc and BOA were not significant). All these hydroxamic acids found in this study were secondary metabolites related to host plant resistance to pests and diseases [36–40]. These compounds have been shown to be implicated in the resistance to insects [36,41], fungi [40,42,43] and allelopathic activity against weeds [44,45].

In Fig. 4c the loadings for PCA of the growth stage factor are plotted. Large positive loadings on PC1 were the most important metabolites during the first stage of wheat growth. DIMBOA and MBOA were confirmed to be important markers of this first growing stage. Higher positive loadings on PC2 are important during second growth stage. The main identified markers having higher positive PC2 loadings were the DIMBOA-Glc and HMBOA analytes. High negative loadings of DIMBOA and MBOA on PC2 are markers that differentiate later growth stages.

A general conclusion that emerged from loadings plot of the tissue sample factor (Fig. 4d) was that DIMBOA-Glc was at relatively higher concentrations in root samples, whereas HMBOA, DIMBOA and, less significant, MBOA were at relatively higher concentrations in shoot samples. These results were also in agreement with previous studies [15,16].

Interpreting the loadings of 'growth stage x tissue sample' interaction on PC1; DIMBOA, MBOA and, to a lesser extent, HMBOA can be used as potential markers of shoot samples at the first growth stage and they were differentiated from root samples at the same growth stage. On PC1 this trend was changed in the later growth stages, where the analytes that were markers of shoot samples at the first growth stage (DIMBOA, MBOA, and HMBOA) changed to be significant for the root samples from the third to the fifth growth stages. On PC2 DIMBOA-Glc, DIMBOA and MBOA had large positive

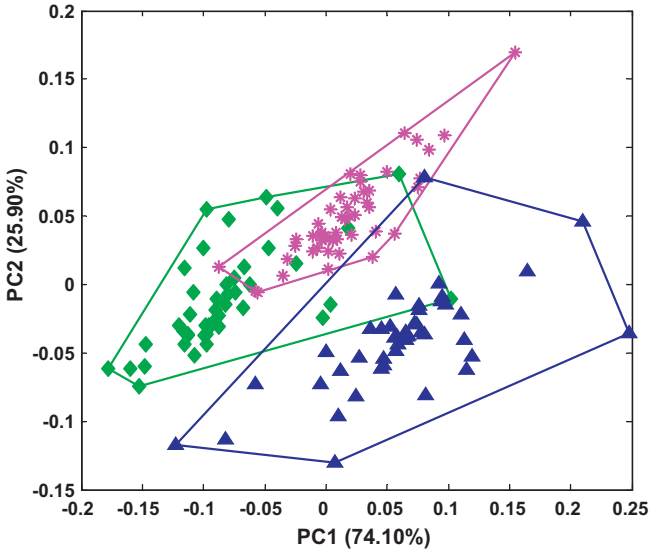


Fig. 5. ASCA results: SCA scores plot obtained without considering the first growth stage level for the wheat variety factor including projections of $\mathbf{Y}_k$ (see Eq. 3). Symbols indicate the levels of the wheat variety factor: green diamonds are Astron variety, magenta asterisks are Ritmo variety and blue triangles are Stakado variety. Convex hulls are drawn around each factor level and plotted with the same colour as the level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

coefficients for shoot samples at the second growth stage, whereas HMBOA was correlated to the root samples at the same stage of growth (figures are not shown for simplicity).

DIMBOA and its decomposition product MBOA have already been documented to inhibit root growth and also seed germination of wild oats (i.e. *Avena fatua*) [45] and other weeds [46,47]. Søltoft and coauthors [40] revealed correlations between the susceptibility to Fusarium Head Blight (a fungal disease) and the concentrations of some benzoxazinoids, DIMBOA-Glc was among them. Huang and coauthors [48] reported that HMBOA had lower phytotoxicity on *Lolium rigidum* compared to DIMBOA and DIBOA (this allelochemical was not significant in the studied data).

Application of ASCA was easier to interpret and more reliable than ANOVA–PCA for the determination of what were the more influencing factors in wheat culture. This was probably due to the high variability existing within samples from the same level group (natural variability), which is common in biological systems. Sample clusters were more difficult to distinguish in score plots resulting from ANOVA–PCA than from ASCA. On the contrary, ASCA allowed a better determination of the significant factors and model building through a statistically sounder validation procedure. Visual interpretation of the natural variability could be then also achieved by projection of residuals in ASCA score plots (see Eq. 3).

Furthermore, multivariate ANOVA–PCA and ASCA methods allow a direct analysis and interpretation of the whole TIC chromatogram in contrast to traditional univariate analysis of individual chromatographic peak areas [16,34].

4. Conclusions

ANOVA–PCA and ASCA enabled the separated analysis and interpretation of what were the effects induced by the different investigated factors in the designed study, giving more easily interpretable information compared to straightforward PCA analysis. Of the tested methods, the ASCA method resulted to be the easiest and most reliable method to use for the analysis of the investigated four-factor ANOVA model.

Concentration of secondary metabolites (benzoxazinone derivatives) showed a systematic trend during the sampling growth period (growth stage factors: 1–5). In the early growth stages there was much variability among replicates, regardless of the level of the other factors. After the initial growth of the wheat plant, the influence of the considered factors was more clearly manifested in the concentration of most of the investigated allelochemicals, and this resulted to be very pronounced for instance in the differentiation between shoots and roots samples.

Acknowledgements

Mireia Farrés acknowledges a PhD grant FI-AGAUR from Generalitat de Catalunya. Funding is acknowledged from Ministerio de Ciencia e Inovación, Spain. CTQ2009-11572 Project.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aca.2012.04.017.

References

[1] R.H. Whittaker, P.P. Feeny, Science 171 (1971) 757.
[2] E.L. Rice, Allelopathy, Academic Press, Orlando, 1984.
[3] R.K. Kohli, J. Crop Prod. 1 (1998) 169.
[4] H. Wu, M. An, D.L. Liu, J. Pratley, D. Lemerle, in: R.S. Zeng, A.U. Mallik, S.M. Luo (Eds.), Allelopathy in Sustainable Agriculture and Forestry, Springer, New York, 2008, p. 235.
[5] R.A. Dixon, Nature 411 (2001) 843.
[6] M. Villagrasa, M. Guillamón, E. Eljarrat, D. Barceló, J. Agric. Food Chem. 54 (2006) 1001.
[7] E. Eljarrat, D. Barceló, TrAC Trends Anal. Chem. 20 (2001) 584.
[8] P.A. Hedin, Crit. Rev. Plant Sci. 9 (1990) 371.
[9] J.J. Jansen, H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman, A.K. Smilde, Anal. Chim. Acta 530 (2005) 173.
[10] P.D.B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Anal. Chim. Acta 544 (2005) 118.
[11] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M.E. Timmerman, Bioinformatics 21 (2005) 3043.
[12] R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh, 1925.
[13] G. Luciano, T. Næs, Food Qual. Prefer. 20 (2009) 167.
[14] G. Zwanenburg, H.C.J. Hoefsloot, J.A. Westerhuis, J.J. Jansen, A.K. Smilde, J. Chemometr. 25 (2011) 561.
[15] M. Villagrasa, M. Guillamón, A. Navarro, E. Eljarrat, D. Barceló, J Chromatogr. A 1179 (2008) 190.
[16] M. Villagrasa, M. Guillamón, A. Labandeira, A. Taberner, E. Eljarrat, D. Barceló, J. Agr. Food Chem. 54 (2006) 1009.
[17] P.H.C. Eilers, Anal. Chem. 76 (2003) 404.
[18] G. Tomasi, F. v. d. Berg, C. Andersson, J. Chemometr. 18 (2004) 231.
[19] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.
[20] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[21] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101.
[22] C. Christin, A.K. Smilde, H.C.J. Hoefsloot, F. Suits, R. Bischoff, P.L. Horvatovich, Anal. Chem. 80 (2008) 7012.
[23] T. Skov, F. v. d. Berg, G. Tomasi, R. Bro, J. Chemometr. 20 (2006) 484.
[24] T.G. Bloemberg, J. Gerretzen, H.J.P. Wouters, J. Gloerich, M. van Dael, H.J.C.T. Wessels, L.P. van den Heuvel, P.H.C. Eilers, L.M.C. Buydens, R. Wehrens, Chemometr. Intell. Lab. Syst. 104 (2010) 65.
[25] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111.
[26] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.
[27] H.C.J. Hoefsloot, D.J. Vis, J.A. Westerhuis, A.K. Smilde, J.J. Jansen, in: D.B. Editors-in-Chief: Stephen, T. Romà, W. Beata (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, p. 453.
[28] D. Iacobucci, in: D.W. Stewart, N.J. Vilcassim (Eds.), AMA Winter Educators Conference: Marketing Theory and Practice, 1995, p. 337.
[29] R.G. Shaw, T. Mitchell-Olds, Ecology 74 (1993) 1638.
[30] I. Stanimirova, K. Michalik, Z. Drzazga, H. Trzeciak, P.D. Wentzell, B. Walczak, Anal. Chim. Acta 689 (2011) 1.
[31] J.A. Fernández Pierna, F. Wahl, O.E. de Noord, D.L. Massart, Chemometr. Intell. Lab. Syst. 63 (2002) 27.
[32] J. Jansen, H. Hoefsloot, J. Greef van der, M. Timmerman, J. West-erhuis, A. Smilde, J. Chemometr. 19 (2005) 469.
[33] D. Vis, J. Westerhuis, A. Smilde, J. van der Greef, B.M.C Bioinformatics 8 (2007) 322.

[34] B.B. Mogensen, T. Krongaard, S.K. Mathiassen, P. Kudsk, J. Agric. Food Chem. 54 (2006) 1023.
[35] M. Villagrasa, M. Guillamón, E. Eljarrat, D. Barceló, J. Chromatogr. A 1157 (2007) 108.
[36] H.M. Niemeyer, Phytochemistry 27 (1988) 3349.
[37] D. Sicker, M. Frey, M. Schulz, A. Gierl, in: W.J. Kwang (Ed.), International Review of Cytology, Academic Press, 2000, p. 319.
[38] V. Cambier, T. Hance, E. de Hoffmann, Phytochem. Analysis 10 (1999) 119.
[39] I.S. Fomsgaard, A.G. Mortensen, S.C.K. Carlsen, Chemosphere 54 (2004) 1025.
[40] M. Søltoft, L.N. Jørgensen, B. Svensmark, I.S. Fomsgaard, Biochem. Syst. Ecol. 36 (2008) 245.
[41] V.H. Argandoña, J.G. Luza, H.M. Niemeyer, L.J. Corcuera, Phytochemistry 19 (1980) 1665.
[42] E. Nakagawa, T. Amano, N. Hirai, H. Iwamura, Phytochemistry 38 (1995) 1349.
[43] Ö. Wahlroos, A.I. Virtanen, Acta Chem. Scand. 12 (1958) 124.
[44] S.K. Mathiassen, P. Kudsk, B.B. Mogensen, J. Agric. Food Chem. 54 (2006) 1058.
[45] F.J. Perez, Phytochemistry 29 (1990) 773.
[46] U. Blum, T.M. Gerig, A.D. Worsham, L.D. Holappa, L.D. King, J. Chem. Ecol. 18 (1992) 2191.
[47] S.V. Copaja, D. Nicol, S.D. Wratten, Phytochemistry 50 (1999) 17.
[48] Z. Huang, T. Haig, H. Wu, M. An, J. Pratley, J. Chem. Ecol. 29 (2003) 2263.

SUPPLEMENTARY MATERIAL

Title: **Chemometric evaluation of different experimental conditions on wheat (*Triticum aestivum* L.) development using liquid chromatography mass spectromety (LC-MS) profiles of benzoxazinone derivatives**

Authors: Mireia Farrés[a], Marta Villagrasa[b], Ethel Eljarrat[a], Damià Barceló[a,b] and Romà Tauler[a,*]

[a] Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain

[b] Catalan Institute for Water Research (ICRA), Parc Científic i Tecnològic de la Universitat de Girona, Edifici H20, Emili Grahit 100, 17003 Girona, Spain

Figure A. ANOVA-PCA results PCA scores plot for growth stage factor versus residual matrix. Different symbols indicate different levels of the factor: red triangles are first growth stage, cyan crosses are second growth stage, magenta solid circles are third growth stage, blue squares are fourth growth stage and white diamonds are fifth growth stage. Convex hulls are drawn around each factor level and plotted with the same colour as the level.

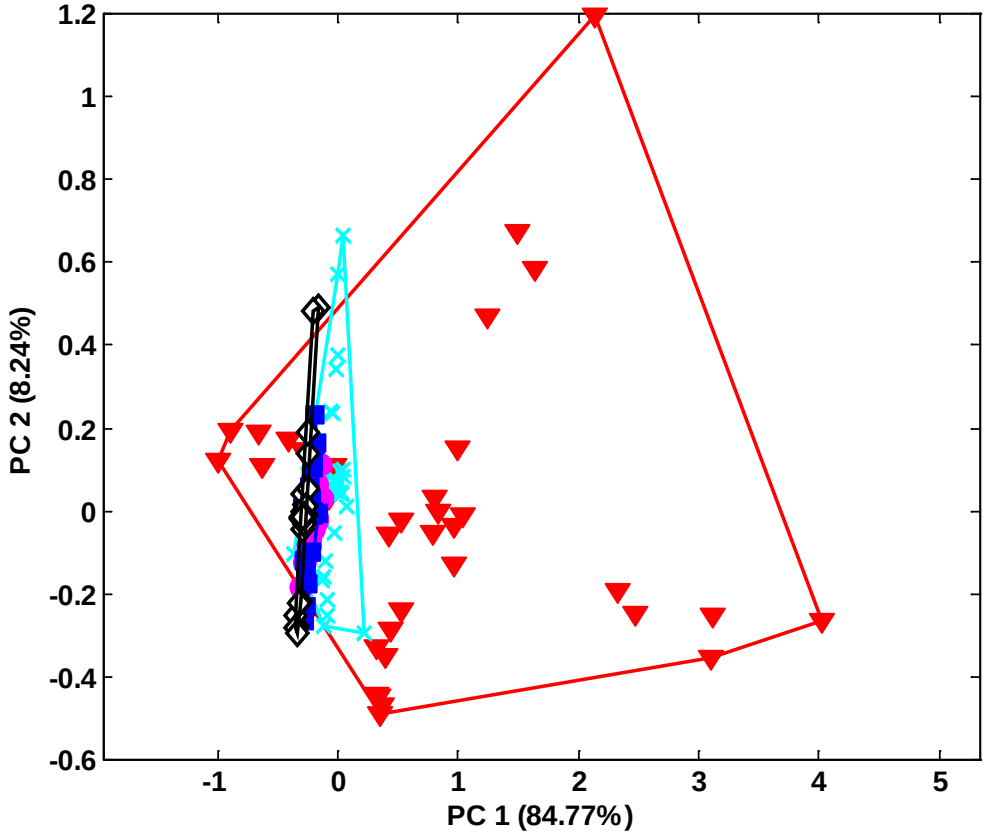


Figure B. ANOVA-PCA results: PCA scores plot for tissue sample factor versus residual matrix. The two symbols indicate the two levels of this factor: green squares are shoots and brown solid circles are roots. Convex hulls are drawn around each factor level and plotted with the same colour as the level.

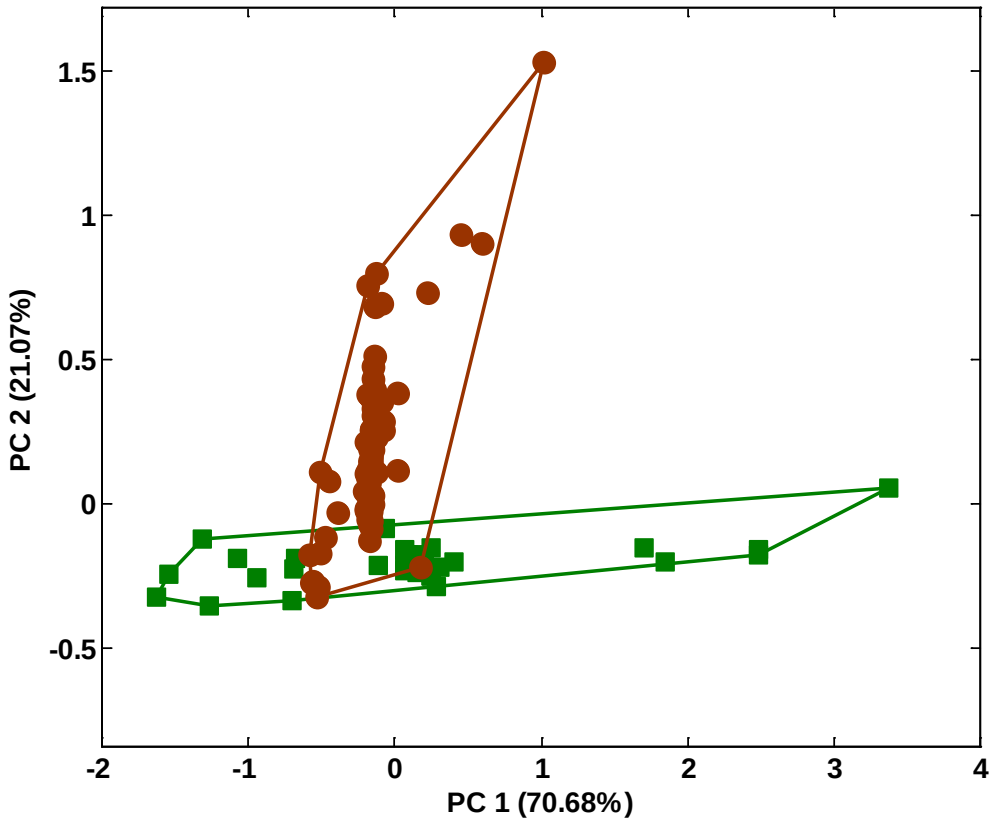


Figure C. ANOVA-PCA results: PCA scores plot for wheat variety factor versus residual matrix. Different symbols indicate the three levels of this factor: green diamonds are Astron, magenta asterisks are Ritmo and blue triangles are Stakado. Convex hulls are drawn around each factor level and plotted with the same colour as the level.

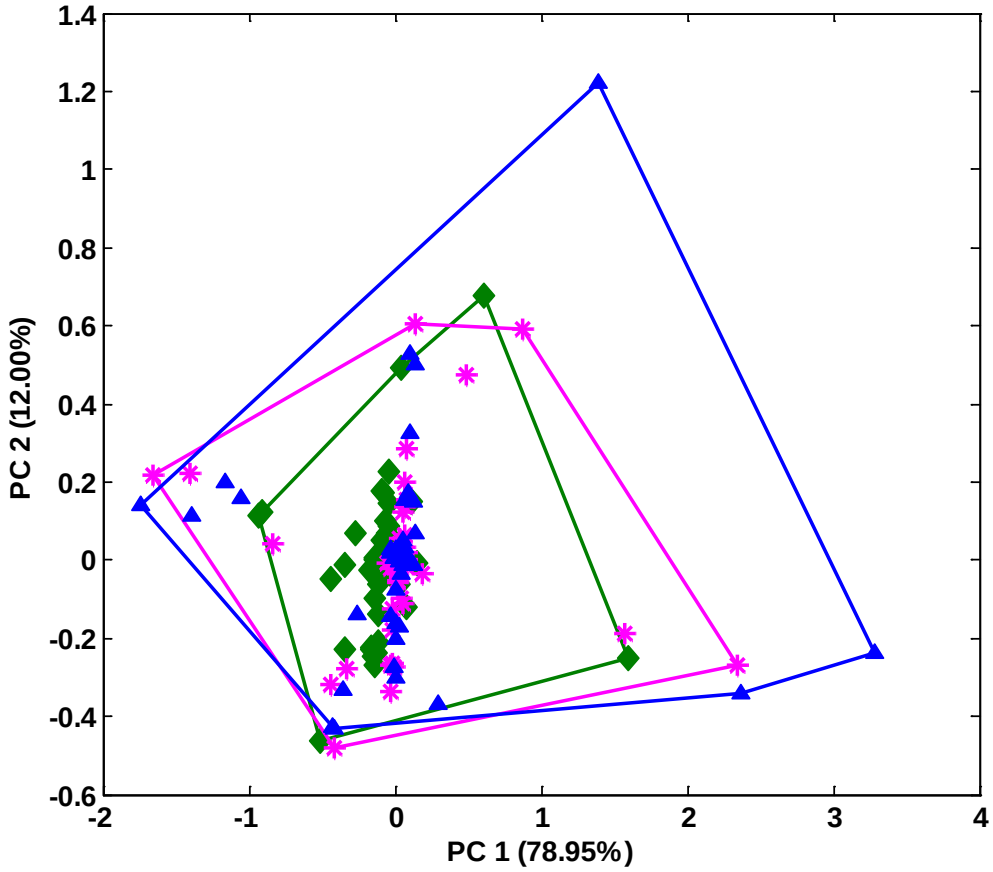
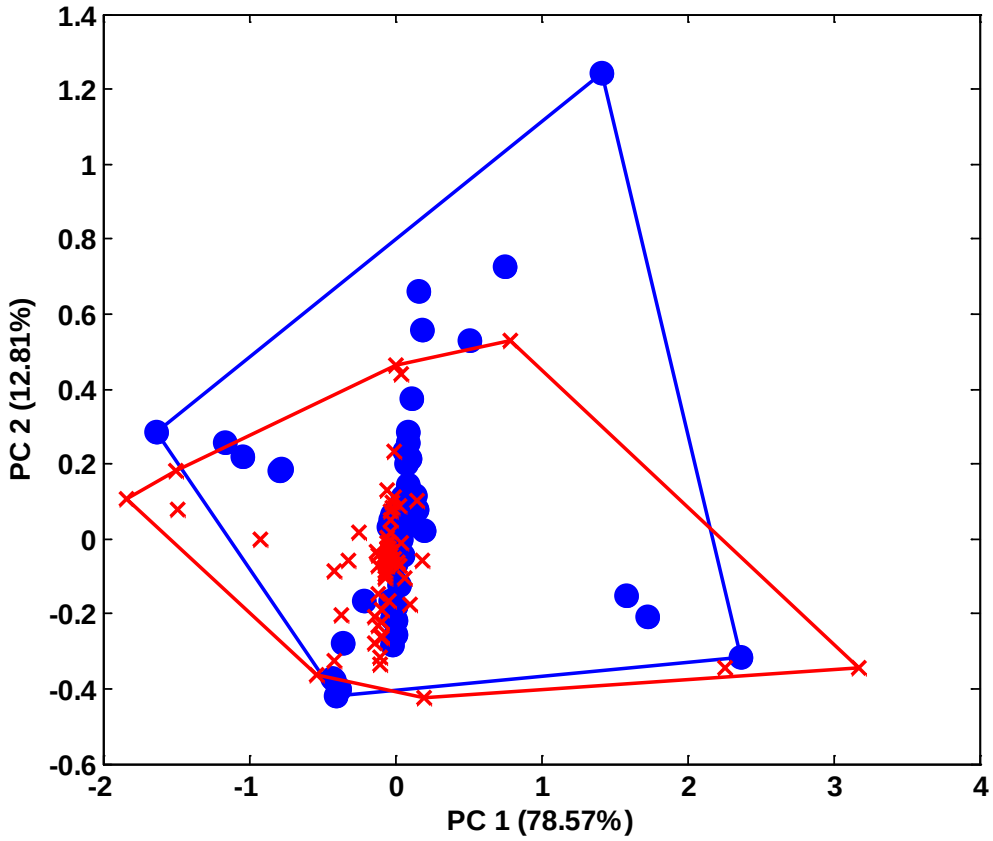


Figure D. ANOVA-PCA results: PCA scores plot for type of cultivation factor versus residual matrix. The two symbols indicate the two levels of this factor: blue solid circles are conventional cultivation and red crosses are organic cultivation. Convex hulls are drawn around each factor level and plotted with the same colour as the level.



3.1.1 DISCUSSIO DELS RESULTATS OBTINGUTS

Els resultats de l'aplicació de l'ANOVA-PCA a cada una de les matrius de dades de cada factor sumades amb la matriu de residuals van permetre observar el mateix patró de comportament per a tots el factors considerats (vegeu Figures A, B, C i D, material suplementari de l'Article 1). Les mostres que pertanyien a la primera etapa de creixement de la planta del blat presentaven una elevada variabilitat entre elles i, conseqüentment, la interpretació del gràfic dels *scores* generat a partir de l'aplicació del PCA a cada una de les matrius així ho mostrava. Tenint en compte aquests resultats, es va aplicar l'ANOVA-PCA a totes les matrius d'efectes però sense considerar les mostres de la primera etapa de creixement. A la figura 2a de l'Article 1 es pot observar que hi ha una millora en la separació dels grups de les etapes de creixement de la planta en relació a la figura A del material suplementari de l'Article 1. Per altra banda, quan es va aplicar l'ANOVA-PCA a la matriu del factor del tipus de mostra hi va haver una millora considerable en la separació del les mostres brots i les mostres arrel (vegeu figura 2b, Article 1). En aquest cas, el primer component principal (PC1) explicava un 80.99% de la variància, i recollia la variació entre el tipus de mostra (brots i arrels). Per altra banda, el mètode ANOVA-PCA també es va aplicar a les diferents matrius d'interacció entre factors, i l'únic factor que interaccionava significativament amb els altres va ser el de l'etapa de creixement de la planta. En tots els casos d'interacció d'aquest factor amb els altres factors (tipus de mostra, varietat i tipus de cultiu) el PC1 explicava més d'un 75% de la variància.

En relació als resultats de l'aplicació del mètode ASCA, l'aplicació del test de permutació permetia rebutjar la hipòtesi nul·la pels factors d'estadi de creixement i del tipus de mostra (brots i arrels), és a dir, aquests dos factors van resultar significatius juntament amb la seva interacció. A la figura 3 de l'Article 1 es presenten els resultats del test de permutació que es va aplicar a la matriu del factor estadi de creixement, on es pot observar la distribució resultant de les sumes de quadrats (*Sum of Squares*, SS) obtingudes a partir de les 10000 permutacions de la matriu original. Totes les SS resultants de les matrius permutades tenien un valor menor a la SS de la matriu original, aquesta última està representada per una línia vertical negre a la figura 3 de l'Article 1. El nivell de significació obtingut de l'aplicació del test de permutació al factor estadi de creixement va ser $p < 0.0001$.

L'aplicació de SCA (*Simultaneous Component Analysis*) a les tres matrius dels factors significatius (estadi de creixement, tipus de mostra i interacció estadi de creixement x tipus de mostra), va permetre detectar quins eren els metabòlits secundaris de la planta del blat més afectats pels factors considerats (a partir dels gràfics dels *loadings* corresponents). L'efecte del factor estat de creixement de la planta es pot observar a la figura 4a de l'Article 1, on es representen els *scores* del SCA. El PC1 explicava el 96.7% de la variància i és en aquest eix on es separen les mostres en relació als diferents nivells d'aquest factor. A la figura 4b del mateix article hi són representats els *scores* que resultaven de l'aplicació del SCA sobre la matriu del factor del tipus de mostra. Es pot veure com PC1 explicava pràcticament tota la variància i separava les mostres dels brots en relació a les mostres de les arrels, tot i que en les mostres corresponents a l'estadi 1 de creixement aquest patró no era tan evident, ja que les mostres presentaven una major variabilitat natural entre elles (mostres 1-36 figura 4b, Article 1). Aquests resultats coincideixen amb els d'un estudi anterior de Mogensen i coautors (Mogensen et al., 2006), on també es van observar diferències en les concentracions dels metabòlits secundaris en la planta de blat en relació al tipus de mostra analitzada (brots o arrels).

El primer i segon components resultants de l'aplicació del SCA a la matriu de la interacció dels factors 'estadi de creixement x tipus de mostra' recollien el 92.90% i el 6.45% de la variància respectivament. El PC1 separava les mostres des de la tercera a la cinquena etapa de creixement en dos grups, un corresponent a les arrels i l'altre corresponents als brots. I en el PC2 es seguia el mateix patró però, en aquest cas, amb les mostres del segon i del cinquè estadi. Un cop més, les rèpliques del primer estadi del factor creixement van estar influenciades per una gran variabilitat natural i aquestes no es van agrupar eficaçment. Per tant, el SCA es va aplicar també a les matrius dels diferents factors sense considerar les mostres del primer estadi de creixement de la planta. L'agrupament dels *scores* va millorar significativament en aquest cas. Tal i com es pot comprovar a la figura 5 de l'Article 1, hi ha una bona separació entre la varietat de blat Stakado de les varietats Astron i Ritmo. En aquest cas els dos primers components principals explicaven el 100% de la variància. Conseqüentment es clou que la variabilitat natural del primer nivell del factor etapa de creixement amaga significativament la importància dels altres factors considerats.

Com a resum, l'aplicació de la metodologia ASCA a dades metabolòmiques en una anàlisi dirigida té dos aspectes pràctics principals. En primer lloc, l'observació dels gràfics dels *scores* del PCA de les matrius dels factors individuals i de les matrius d'interaccions va permetre entendre el

comportament dels metabòlits en resposta als quatre factors considerats durant el cultiu de la planta del blat (estadi de creixement, tipus de mostra, varietat, tipus de cultiu). En segon lloc, un benefici important de l'enfocament ASCA és la possibilitat d'aïllar les fonts de soroll no aleatòries o no desitjades de les altres. El soroll experimental pot causar interferències en els metabòlits analitzats i conseqüentment dificultar la identificació dels pics cromatogràfics d'interès. El mètode ASCA es centra en la variabilitat associada als factors experimentals considerats i millora l'aplicació de PCA en termes d'interpretabilitat i de selecció de possibles característiques o aspectes rellevants subjacents de les dades. En relació a la comparació de les metodologies ANOVA-PCA i ASCA per a l'anàlisi del conjunt de metabòlits secundaris del blat, es clou que el mètode ASCA permet fer una millor interpretació de les fonts de variabilitat. La inspecció visual del gràfic dels *scores* de l'ASCA corrobora la disminució de l'efecte de la variabilitat natural i del soroll. I el test de permutació permet determinar de forma més entenedora i estadísticament fiable quins són els factors significatius.

La importància relativa de les diferents variables (metabòlits al·leloquímics) es va fer a partir de l'anàlisi dels gràfics dels *loadings* obtinguts després de l'aplicació de l'ASCA. L'observació dels gràfics dels *loadings* va permetre la identificació dels metabòlits que resultaven més importants en els patrons de comportament descoberts. Els metabòlits secundaris que actuen com a al·leloquímics més rellevants en relació als factors significatius (estadi de creixement, tipus de mostra i interacció estadi de creixement x tipus de mostra) segons la metodologia ASCA van ser: DIMBOA-Glc, DIMBOA, HMBOA i MBOA. S'ha demostrat que aquets àcids hidroxàmics estan implicats en la resistència de la planta als insectes (Argandoña et al., 1980; Niemeyer, 1988), als fongs (Nakagawa et al., 1995; Søltoft et al., 2008; Wahlroos i Virtanen, 1958) i a l'activitat al·lelopàtica contra les males herbes (Mathiassen et al., 2006; Perez, 1990).

El metabòlit DIMBOA i el seu producte de descomposició MBOA han estat relacionats amb la inhibició al creixement de l'arrel i també de la germinació de llavors de civada silvestre (*Avena fàtua*) (Perez, 1990) i d'altres males herbes (Blum et al., 1992; Copaja et al., 1999). Blum i coautors (Blum et al., 1992) clouen en el seu estudi que MBOA és més potent en la germinació de *Trifolium incarnatum* que el seu precursor DIMBOA. No obstant, un altre estudi (Kato-Noguchi et al., 2010) clou que el DIMBOA va resultar tenir un efecte fitotòxic major en 5 espècies de males herbes en comparació amb els compostos BOA i MBOA. Pel factor estadi de creixement del blat, DIMBOA i MBOA van resultar els metabòlits més significatius durant el primer estadi de

creixement de la planta, mentre que en el segon estadi predominaven DIMBOA-Glc i HMBOA. Els metabòlits que diferenciaven el cinquè estadi de creixement van ser DIMBOA i MBOA.

En relació al factor tipus de mostra, el metabòlit DIMBOA-Glc es trobava en altes concentracions a les arrels mentre que HMBOA, DIMBOA i en menor abundància MBOA eren més presents als brots. Wu i coautors (Wu et al., 2001) van trobar una concentració major del metabòlit DIMBOA a les arrels en comparació amb els brots, tot i així en altres estudis es va trobar un major contingut de DIMBOA en els brots que en les arrels (Mogensen et al., 2006). Aquests resultats diversos es poden explicar tenint en compte que hi ha varis factors externs com ara la quantitat de nutrients al sòl, la temperatura, la precipitació, la radiació, l'estrès i l'ús de pesticides sintètics que poden influir en el contingut d'al·leloquímics (Einhellig, 1996). D'altra banda Huang i coautors van determinar que HMBOA tenia menor fitotoxicitat en *Lolium rigidum* en comparació amb DIMBOA i DIBOA (2,4-dihydroxy-1,4-benzoxazin-3-one). També s'han trobat correlacions significatives entre la susceptibilitat d'una determinada malaltia fúngica (*Fusarium Head Blight*) i les concentracions d'algunes benzoxazinones, entre elles DIMBOA-Glc (Søltoft et al., 2008).

En relació als metabòlits més abundants trobats a partir del estudi de la interacció entre els factors estadi de creixement i tipus de mostra, DIMBOA, MBOA i en menor mesura HMBOA van resultar ser més importants en les mostres de brots del primer estadi de creixement, que es distingien de les mostres d'arrels del mateix estadi. En els estadis de creixement posteriors, els mateixos metabòlits que diferenciaven les mostres dels brots del primer estadi de creixement (DIMBOA, MBOA i HMBOA) van esdevenir significatius per a les mostres de les arrels des del tercer fins al cinquè estadi de creixement.

Capítol 4

Avaluació quimiomètrica dels canvis dels perfils metabòlics LC-MS del llevat (*Saccharomyces cerevisiae*) sotmès a condicions ambientals estressants (canvis de temperatura i de concentracions de Cu(II))

4.1 INTRODUCCIÓ

En la metabolòmica, igual que en les altres ciències òmiques, es generen grans conjunts de dades multivariants d'origen biològic. Aquests estudis proporcionen una visió global de l'estat bioquímic d'un organisme sota condicions específiques. Els enfocaments metabolòmics no dirigits (*untarget*) tenen com a objectiu realitzar l'anàlisi del metaboloma de la forma més àmplia possible per a classificar els fenotips en base a un determinat patró observat. En principi, aquests enfocaments no només ofereixen la caracterització dels canvis en el metaboloma sinó també, a diferència de l'anàlisi dirigida (*target*), permeten la detecció de metabòlits prèviament descartats o desconeguts. L'anàlisi quimiomètrica de les dades obtingudes en estudis metabolòmics és un veritable repte a causa de la seva subtilesa i de la complexitat del problema estudiat. En aquesta aproximació quimiomètica s'ha de posar èmfasi tant en el disseny estadístic planejat dels experiments com en les metodologies quimiomètriques capaces d'analitzar les dades metabolòmiques multi- i megavariants generades en el context de tecnologies instrumentals avançades com la cromatografia de líquids amb detecció per espectrometria de masses (LC-MS).

Una de les necessitats clau en metabolòmica és la capacitat d'analitzar d'una manera eficient un nombre extens de metabòlits cel·lulars, dels quals hi ha una manca de coneixement previ de la seva presència o absència en les mostres analitzades. Actualment no hi ha una tècnica única que cobreixi tot el metaboloma d'un determinat organisme en la seva amplitud i profunditat, en una sola anàlisi. Cobrir una àmplia gamma de metabòlits del metaboloma total en una sola anàlisi és un dels principals reptes de la metabolòmica.

La cromatografia de líquids amb detecció per espectrometria de masses (LC-MS) permet analitzar una gamma àmplia d'espècies, en comparació amb la cromatografia de gasos amb detecció per espectrometria de masses (GC-MS), on els diferents metabòlits s'han de derivatitzar primer per aconseguir la seva volatilitat i separar-se cromatogràficament a la columna. Actualment, la plataforma LC-MS és àmpliament utilitzada en metabolòmica (Tang et al., 2014). Aquesta popularitat es deu a la gran versatilitat que té en la detecció d'una gran varietat de molècules polars (metabòlits) en una sola anàlisi.

Les columnes cromatogràfiques de fase reversa (*Reverse Phase*, RP), entre elles les columnes C18, són tradicionalment i extensa utilitzades en cromatografia líquida i proporcionen una bona separació per a compostos no molt polars i per aquells compostos més dèbilment polars. No obstant això, la principal limitació de les columnes RP és la falta d'una retenció adequada de les molècules altament hidrofíliques, iòniques i polars en la fase estacionària. En aquest cas, i amb l'objectiu de separar les molècules més polars, les quals són abundants en el metaboloma del llevat (organisme estudiat en aquest capítol), és necessari fer servir el mètode de separació conegut com cromatografia líquida d'interacció hidrofílica (*Hydrophilic Interaction LIquid Chromatography*, HILIC), el qual ha crescut en popularitat a causa de la seva alta compatibilitat amb MS (*Mass Spectrometry*), i que permet millorar substancialment la detecció d'analits molt polars com són els metabòlits (Lämmerhofer, 2010).

En aquest capítol es presenta el desenvolupament i aplicació d'un procediment (*workflow*, vegeu secció 2.5 de la introducció de la Tesi, capítol 2) d'anàlisi no dirigida (*untarget*) en dos experiments metabolòmics diferents de l'organisme *Saccharomyces cerevisiae*. L'objectiu final de la metodologia de treball emprada en aquests estudis és aconseguir trobar marcadors (metabòlits) que siguin reflex dels canvis metabòlics produïts com a conseqüència de dues situacions d'estrès aplicades als cultius del llevat. En un primer estudi d'aquest capítol, en l'Article 2, el llevat es va cultivar a dues temperatures diferents: en les seves condicions òptimes de creixement a 30ºC i en les condicions estressants de temperatura més elevada de 42ºC. En un segon estudi, en l'Article 3, les mostres de llevat es van cultivar a diferents concentracions (control, 1 mM, 3 mM i 6 mM) de sulfat de coure ($CuSO_4$), a 30ºC per avaluar quin és l'efecte del Cu(II) en el metaboloma del llevat. El sulfat de coure és un producte fitosanitari que s'utilitza per al control dels fongs en els raïms. L'ió Cu(II) és el responsable de provocar una acció d'eliminació dels fongs i/o bactèries. El principal objectiu del sulfat de coure és el míldiu, una malaltia produïda el fong *Plasmopara vitícola* que afecta a les fulles i als raïms dels ceps, ja que coure pertorba les activitats respiratòries, enzimàtiques i membranàcies dels fongs. L'ió Cu(II) és molt estable ja que no es degrada per la calor ni per la llum i això fa que la persistència ambiental del Cu(II) sigui molt alta mentre no sigui lixiviat. Quan aquest element s'acumula pot ser una font important de contaminació amb conseqüències toxicològiques per al llevat i conseqüentment tenir conseqüències en la fermentació del vi (Avery et al., 1996; Brandolini et al., 2002) (veure secció 2.4.2 de la introducció de la tesi, capítol 2).

Cultiu del llevat

Les cèl·lules de *Saccharomyces cerevisiae* es van fer créixer en un medi ric no selectiu anomenat YEP (*Yeast Extract Peptone*), el qual conté extracte de llevat i peptona bacteriològica. Quan el medi YEP es mescla amb la font de carboni (glucosa), el nou medi s'anomena YPD (*Yeast Peptone Dextrose*). Les cèl·lules de llevat poden ser crescudes tant en medi líquid com en medi sòlid, ambdós preparats amb els mateixos components afegint agar en el medis sòlid per a la seva solidificació. La temperatura òptima de creixement de *S.cerevisiae* és de 30ºC.

Les soques de *S.cerevisiae* es poden mantenir indefinidament en medi YPD amb glicerol al 15% a -80ºC (glicerinats). Per recuperar les cèl·lules de llevat a partir dels estocs, es va rascar lleugerament el glicerinat amb un escuradents estèril i es va estriar en el medi YPD solidificat per l'agar en plaques de Petri . Les plaques es van incubar a 30ºC fins l'aparició de colònies (3 dies aproximadament).

El control del creixement del llevat en medi líquid es va fer mitjançant la mesura de la densitat cèl·lules a través de la densitat òptica (OD) a 600 nm (OD_{600}) amb un espectrofotòmetre. El llevat, al ser un microorganisme microaeròfil, quan es cultiva en líquid es fa créixer en agitació (150 rpm) per tal d'afavorir l'intercanvi gasós i evitar la formació de grumolls. Per a la preparació dels cultius cel·lulars de *S.cerevisiae* primerament es fa fer un precultiu, aquest es va preparar inoculant una colònia crescuda en el medi sòlid al medi líquid YPD. Una petita quantitat d'aquest precultiu en fase exponencial es va inocular a un nou medi, el cultiu resultant (OD_{600} 0.1) era la mostra esperada per procedir l'experiment (vegeu Articles 2 i 3).

Extracció dels metabòlits del llevat

El metabolisme del llevat es va inactivar (*quenching*) refredant les mostres amb gel. El sobrenedant (medi de cultiu) es va eliminar centrifugant les mostres a 4 ºC. Posteriorment, el *pellet* (cèl·lules de llevat) es va netejar amb un tampó fosfat salí (*Phosphat-Buffered Saline*, PBS) que va ajustar el pH de les mostres a 7.4 amb la finalitat de garantir l'estabilitat d'una gran varietat de metabòlits. Els metabòlits intracel·lulars del llevat es van extraure utilitzant el protocol de Gonzalez i coautors (Gonzalez et al., 1997) que es basa en la incubació de les mostres de llevat amb 5 ml d'etanol (75%) a 80 ºC durant 3 min.

Anàlisi per LC-MS

En els treballs d'aquest capítol (Article 2 i 3), les anàlisis de les mostres de llevat es van portar a terme utilitzant una columna cromatogràfica HILIC TSKgel Amide-80 (Tosoh). Aquesta columna és adequada per a l'anàlisi d'un grup gran de metabòlits. En el treball de Tostikov i Fiehn (Tolstikov i Fiehn, 2002) es va utilitzar aquesta columna TSKgel Amide-80 per a la detecció d'oligosacàrids, glucòsids, aminosucres, aminoàcids de les fulles de la planta *Cucurbita màxima.* En aquestes anàlisis es van detectar fins a un total de 60 pics cromatogràfics. En el treball de Bajad i coautors (Bajad et al., 2006) també es va utilitzar la columna TSKgel Amide-80 per a la detecció de metabòlits de l'organisme *Escherichia coli*, que formen part del metabolisme central de carboni, aminoàcids i del metabolisme dels nucleòtids d'aquest organisme. En aquest últim treball es van detectar i identificar fins a 79 metabòlits per espectrometria de masses en tàndem (MS/MS).

Compressió de les dades LC-MS

L'anàlisi LC-MS d'alta resolució en mode d'anàlisi completa d'ions (*full scan*) genera grans matrius de dades, on la mesura de les relacions massa càrrega (*m/z*) pot presentar fins a quatre decimals de precisió. Degut a les enormes dimensions de les matrius de dades generades, és necessària la reducció de les dimensions de les matrius durant la importació de les dades a l'entorn de programació Matlab. En els dos treballs d'aquest capítol (Article 2 i Article 3) es presenten dues estratègies diferents de reducció de les dades. En l'Article 2 la compressió de les dades es va realitzar mitjançant un procediment de *binning* (vegeu secció 2.5.5 de la introducció de la Tesi, capítol 2). El procediment de *binning* consisteix en agrupar els valors *m/z* en petits intervals regulars (contenidors d'igual mida) dins d'un determinat rang de valors de *m/z*. Totes els valors de les intensitats dels senyals agrupats dins d'un d'aquests intervals o contenidors se sumen i s'atribueixen a un determinat valor de *m/z* descriptiu de l'interval. En l'Article 3, la compressió de dades de les mostres de llevat es va fer a partir de la selecció d'un nombre reduït de regions d'interès dels valors *m/z* (*Regions Of interest*, ROI) (Gorrochategui et al., 2015) (vegeu secció 2.5.5 de la introducció de la Tesi, capítol 2). Donada les característiques de les dades de MS, on hi ha una densitat relativament petita de valors d'intensitat elevada comparada amb el gran nombre de valors absents (*scarcity*) o poc diferents del soroll, aquesta segona estratègia (ROI) pot representar un enorme estalvi d'emmagatzematge de la informació MS rellevant sense pèrdua de resolució *m/z* (que sí es perdia en el cas del procediments de *binning*). El procediment ROI permet

la compressió de les dades originals LC-MS sense pèrdua de resolució espectral. Els ROI contenen les dades dels cromatogrames més interessants, és a dir, d'aquells cromatogrames de valors *m/z* que tenen intensitats MS més significants i majors a una determinada relació senyal/soroll llindar. A diferència del procediment de *binning*, en el ROI la representació de les dades LC-MS comprimida és una matriu de dades amb un rang de valors de *m/z* no equidistant.

La compressió de dades mitjançant el procediment ROI presenta dos avantatges importants, el primer és que permet reduir la dimensió de les matrius de dades de forma considerable sense perdre resolució espectral. Contràriament, amb el procediment de *binning* els valors de *m/z* queden assignats a un valor en el que la resolució experimental ha disminuït de forma important. Per altra banda, l'aplicació del procediment ROI evita que els cromatogrames s'hagin de dividir per finestres per a la seva anàlisi i resolució per MCR-ALS, ja que el nombre de *m/z* és considerablement menor

Metabolomics (2015) 11:210–224
DOI 10.1007/s11306-014-0689-z

ORIGINAL ARTICLE

Chemometric evaluation of *Saccharomyces cerevisiae* metabolic profiles using LC–MS

Mireia Farrés · Benjamí Piña · Romà Tauler

Received: 27 March 2014 / Accepted: 7 June 2014 / Published online: 25 June 2014

Abstract A new liquid chromatography mass spectrometry (LC–MS) metabolomics strategy coupled to chemometric evaluation, including variable and biomarker selection, has been assessed as a tool to discriminate between control and stressed *Saccharomyces cerevisiae* yeast samples. Metabolic changes occurring during yeast culture at different temperatures (30 and 42 °C) were analysed and the complex data generated in profiling experiments were evaluated by different chemometric multivariate approaches. Multivariate curve resolution alternating least squares (MCR-ALS) was applied to full spectral scan LC–MS preprocessed data multisets arranged in augmented column-wise data matrices. The results showed that sectioning the MS-chromatograms in different windows and analysing them by MCR-ALS enabled the proper resolution of very complex coeluted chromatographic peaks. The investigation of possible relationships between MCR-ALS resolved chromatographic peak areas and culture temperature was then investigated by partial least squares discriminant analysis (PLS-DA). Selection of most relevant resolved chromatographic peaks associated to yeast culture temperature changes was achieved according to PLS-DA-Variable Importance in Projection scores. A metabolite identification workflow was developed utilizing MCR-ALS resolved pure MS spectra and high-resolution accurate mass measurements to confirm assigned structures based on entries in metabolite databases. A total of 65 metabolites were identified. A preliminary interpretation of these results indicates that the strategy described in this study can be proposed as a general tool to facilitate biomarker identification and modelling in similar untargeted metabolomic studies.

Keywords Metabolic profiling · Untargeted metabolomics · Metabolite identification · *Saccharomyces cerevisiae* · Multivariate curve resolution-alternating least squares · Partial least squares-discriminant analysis · Liquid chromatography–mass spectrometry

1 Introduction

Cell metabolites describe the physical and chemical characteristics of organisms. Metabolomics aims to measure the global, dynamic metabolic response of living complex multicellular systems to biological stimuli or genetic manipulation (Nicholson and Lindon 2008). It determines changes in low molecular weight organic metabolites in complex biological samples. By identifying biochemical compounds whose concentrations have varied due to a biological stimulus, metabolomics allows uncovering new possible targets (biomarkers) for biochemical interpretation of biological changes.

Currently, a range of analytical platforms are used for metabolomic analysis, including direct infusion mass spectrometry (MS) (Højer-Pedersen et al. 2008), gas chromatography coupled to mass spectrometry (GC–MS) (Lu et al. 2008), two-dimensional GC coupled to MS (GC × GC–MS), liquid chromatography coupled to MS (LC–MS) (Bajad et al. 2006), capillary electrophoresis coupled to MS (CE–MS), and proton nuclear magnetic resonance (1H NMR) spectroscopy and Fourier transform infrared (FT-IR) spectroscopy. Complete chromatographic separation of the components of complex biological

M. Farrés · B. Piña · R. Tauler (✉)
Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain
e-mail: Roma.Tauler@idaea.csic.es

samples is often difficult to achieve. Despite the fundamental advantages of metabolomics, so far no metabolomic platform allows for the reliable complete separation, detection and identification of all metabolites. Actually, the analysis of the full metabolomes is a very difficult task due to the large chemical diversity of cellular metabolites (Villas-Bôas et al. 2005; Werf et al. 2007; Garcia et al. 2008; Theodoridis et al. 2012; Xu et al. 2014)

In general, targeted metabolomics approaches are directed to the detection and quantification of specific classes of compounds. In contrast, non-targeted metabolomics aims to study the widest possible range of compounds and enables the identification of most discriminatory metabolites that can be used as biomarkers (Glinski and Weckwerth 2006). A global non-targeted metabolomics in combination with multivariate data analysis aims to the isolation of previously unknown biomarkers specific to a particular biological stimulus.

LC–MS-based approaches are of particular importance for non-targeted metabolomics. Metabolites can be extracted with aqueous alcohol solutions and directly analysed. In principle, LC–MS does not require any prior pretreatment of samples to distinguish between different metabolite groups of interest and it is suitable for the detection of a wide range of metabolite classes. Depending on the type of chromatographic column used for the analysis, various metabolite groups can be reliably analyzed using LC–MS. High mass spectrometry resolution with electrospray ionization (ESI) is the preferred method in terms of universality, high throughput, resolution and sensitivity (Niessen 1999).

When metabolomic profiles are analysed by LC–MS in full spectral scan mode, some drawbacks, like baseline distortion, retention time peak shifting and possible peak shape distortions from one chromatographic run to another, and possible strong peak coelution problems can appear. Different chemometric methods can be used to reduce the effects of these drawbacks, such as baseline correction methods (Eilers 2004), peak alignment methods (Savorani et al. 2010), warping methods (Nielsen et al. 1998), wavelets methods (Walczak et al. 1996) and multivariate curve resolution (Parastar and Akvan 2014). In particular, because of the ubiquitous existence of the large number of overlapping embedded peaks, multivariate curve resolution methods can be very useful and necessary to achieve the goals of metabolomics studies by full spectral scan LC–MS.

Saccharomyces cerevisiae is a budding yeast species, which comprises a group of unicellular fungi belonging to *Ascomycetes* phylum. *S. cerevisiae* has been used as a model for higher eukaryote species in biology because its similar metabolism (Sherman et al. 2002; Castrillo and Oliver, 2006). In this study the LC–MS metabolomics approach is coupled to different chemometric methods, such as MCR-ALS and PLS-DA, to explore the changes observed in the metabolite profiles of *S. cerevisiae* when it is cultivated at different temperatures. In this work, a new strategy using Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) (Tauler 1995; Peré-Trepat et al. 2005) is proposed as a general approach for proper investigation and resolution of complex and extensive LC–MS data sets (in full spectral scan mode), where huge amounts of information can be uncovered, including strongly hidden coeluted and embedded unknown chromatographic peaks. Related approaches have been already proposed in previous works (Pérez et al. 2009; Szymańska et al. 2009; Siano et al. 2011) to solve similar coelution problems in metabolomics, but this work goes a step further and apart from their resolution, metabolites are also identified by their exact mass. In addition, Partial Least Squares-Discriminant Analysis (PLS-DA) (Barker and Rayens 2003) is applied to the MCR-ALS results to investigate what metabolites were more influenced by the temperatures changes on yeast cultures, acting therefore as a possible biomarkers of temperature stimulus on yeast cultures.

2 Experimental

2.1 Chemicals

Pure metabolites threonine, valine, isoleucine, glutamic acid, adenosine monophosphate (AMP), adenosine triphosphate (ATP), 3-phosphoglyceric acid, glucose-1-phosphate, fructose-6-phosphate, fructose-1,6-biphosphate, itaconic acid, succinic acid and citric acid were obtained from Sigma-Aldrich (St. Louis, USA). Stock individual standard solutions (500 μg mL^{-1}) were prepared dissolving accurate amounts of pure standards in acetonitrile:water 1:1. Two standard mixture samples of these compounds were prepared at 10 and 20 μg mL^{-1} concentration levels in acentonitrile:water 1:1. Ethanol, Acetonitrile and HPLC grade water were obtained from Merck (Darmstadt, Germany).

2.2 Culture conditions

Yeast strains W303a were grown in glass cultured tubes overnight at 30 °C and 150 rpm in non-selective medium (yeast extract peptone dextrose, YPD, 5 g L^{-1} yeast extract, 10 g L^{-1} peptone, 20 g L^{-1} glucose). Eight shake flask cultures were performed in 100-mL flasks with 50 mL medium. Samples culture media were inoculated with 50 μL of yeast pre-cultures in YPD medium and incubated for at 30 °C and 150 rpm. After 7 h four flask cultures

were incubated at 30 °C and the other four flask cultivations at 42 °C, in both cases for 1 h.

2.3 Quenching and extraction of metabolites

Four types of samples were investigated (i) one standard mixture at 20 μg mL^{-1} (ii) one standard mixture at 40 μg mL^{-1}; (iii) four yeast samples cultivated at 30 °C; and (iv) four yeast samples cultivated at 42 °C. The same analytical pretreatment was applied in biological and standards samples. Metabolites extraction procedure was performed using three blank (without yeast) samples.

After culture, samples were poured to a 50 mL Falcon tubes and the metabolism of the cultures samples was rapidly inactivated cooling down the mixture on ice. Once cooled down, all tubes were centrifuged at 4,000 rpm for 15 min at 4 °C. The supernatant was removed and yeast pellets remained into the Falcon tube. Pellet was then cleaned up with phosphate buffered saline (PBS). At this point, the two standard mixture samples at 20 and 40 μg mL^{-1} were added. A volume of 25 mL of PBS was poured into each sample to adjust their pH to 7.4. Falcon tubes were centrifuged at 4,000 rpm for 10 min at 4 °C. Again the supernatant was removed. This step was repeated twice. All through the clean-up procedure, samples were kept in cold.

Extraction of yeast metabolites was carried out according to the procedure previously described (Gonzalez et al. 1997). Metabolites extraction was performed into 50 mL Falcon tubes, adding 5 mL of solvent (75 % ethanol) to the cell pellet and further incubation of the suspension for 3 min at 80 °C. After cooling down the mixture on ice, sample volume was concentrated and dried by evaporation using nitrogen gas. The residue was resuspended to a final volume of 0.5 mL with the LC mobile phase (95 % acetonitrile). Prior to pouring the final volume to a vial, it was filtered through 0.2 μm GHP membranes (GHP, Acrodisc Syringe Filters, Pall Life Sciences, USA) to further ensure removal of any residual protein/debris before LC analysis.

2.4 LC–MS

An Accela liquid chromatograph (Thermo Scientific, Hemel Hempstead, UK) equipped with a quaternary pump, a thermostated autosampler and a TSKgel Amide-80 5-μm (100 × 2.0 mm) column purchased from Tosoh (Tokyo, Japan) was used. LC solvents were 0.5 mM ammonium acetate in 90 % acetonitrile at pH 5.5 (solvent A) and 2.5 mM ammonium acetate in 90 % acetonitrile in 60 % acetonitrile at pH 5.5 (solvent B). The gradient elution was as follows: $t = 0$, 5 % B; $t = 8$, 60 % B; $t = 12$, 95 % B; $t = 17.5$, 95 % B; $t = 20$, 5 % B; $t = 30$, 5 % B. Injection volume was 5 μL and flow rate was 0.3 mL/min.

An LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Hemel Hampstead, UK) equipped with an ESI source in positive mode was used to acquire mass spectra profiles in full scan mode. Operation parameters were: source voltatge, 3.5 kV; sheath gas, 40 (arbitrary units); auxiliary gas, 10 (arbitrary units), sweep gas, 5 (arbitrary units); and capillary temperature, 275 °C. The acquired mass range was from 50 to 1000 Da. The mass spectrometer was interfaced to a computer workstation running Xcalibur 2.1 software for data acquisition and processing.

2.5 Data import

Full scan MS spectra of different chromatographic runs were saved in raw mode in Xcalibur software 2.0 (Thermo Xcientific, San Jose, CA) and converted to mzXML by ReAdW software (Seattle Proteome Center 2014) and imported to MATLAB (The Mathworks Inc. Natick, MA, USA) computer environment with the mzxmlread.m function from the Bioinformatiocs Toolbox 3.0.

3 Chemometric data preprocessing and analyses

Chemometric data analysis included different multivariate data analysis methods like Principal Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA) and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) (Jaumot et al. 2005). Matlab R2007a (Mathworks Inc. Natick, MA, USA) and PLS Toolbox 5.8.1 (Eigenvector Research Inc., Wenatchee, WA, USA) were used as computer programming environments for all chemometric analyses.

3.1 Data preprocessing

Each analysed yeast sample produced a raw full scan MS chromatogram which was imported to Matlab and initially binned to their integer mass to facilitate faster processing and chemometric analysis. Each full scan MS chromatogram was stored in a data matrix with dimensions of 3,587 rows (retention times, ranging from 0 to 30 min) and 951 columns (mz intensity values, ranging from 50 to 1,000 Da).

Raw full scan MS chromatograms were size reduced, giving a total number of 546 *mz* values within the mass range between 55 and 600 Da. MS chromatograms were then interpolated to the same retention times giving a total number of 2020 retention times, ranging from 0 to 17 min. Therefore the final size of every data matrix corresponding to a full scan MS chromatogram of a yeast sample was of 2020 rows by 546 columns. Baseline and background contributions were corrected by subtraction of the mean

chromatogram of the blank samples. To have most of data values at reasonable units (between 0 and 2), the intensity scale of all chromatograms was divided by 10^8. The resulting preprocessed data matrices from all yeast samples were then analysed by MCR-ALS.

On the other hand, the eight Total Ion Current (TIC) yeast MS chromatograms were also arranged altogether in a single TIC data matrix (8 rows × 2,020 columns). Before its analysis, chromatographic peaks were appropriately aligned to compensate possible between run retention time shifts. The Correlation Optimized Warping (COW) method (Nielsen et al. 1998; Tomasi et al. 2004) was selected for this purpose. The application of this method required as input parameters, the segment *m,* which is the length of the sections in which the chromatogram is divided, the slack size *t,* which is the maximum chromatographic peak warping allowed and a reference chromatogram (Nielsen et al. 1998). These parameters were selected according to the method proposed in previous works (Skov et al. 2006). In order to improve the application of the peak alignment procedure, the TIC matrix (8 ×°2,020) was divided in two submatrices with dimensions of: 8 ×°1,300 and 8 ×°720, and COW alignment was then performed in each part individually. After alignment both windows were then rejoined. Before PCA, the already aligned TIC chromatograms in TIC data matrix were mean-centred.

3.2 Data arrangement

Two types of data sets were analysed in this work: (i) the eight individual Total Ion Current (TIC) chromatograms arranged in a single TIC data matrix and (ii) the eight full scan MS chromatograms sectioned in different windows and arranged in different column-wise augmented data matrices (see below).

No further arrangement was required for the single TIC data matrix. In contrast, every full scan individual preprocessed MS chromatogram data matrix was divided in ten separate submatrices corresponding to different time windows. This MS chromatogram window subdivision was done manually according to peak shape and peak density, and more specifically, not to miss those chromatographic peaks that could change with temperature. The first time window was discarded for MCR-ALS analysis since it only contained signal background and noise, i.e. all chromatograms from blanks, standards and yeast samples had the same shape profile at that initial time window. Therefore a final number of nine windows (j = I, II,…, IX) were selected from every full scan MS chromatogram data matrix of the ten analyzed samples (4 control yeast samples, k = 1,2,3,4; 4 temperature stressed yeast samples, k = 5,6,7,8 and two standard mixture samples, k = 9,10) (see Fig. 1). Therefore, for each of the ten analyzed samples, k = 1,…, 10, nine time windows were obtained, j = 1, …, 9 giving the individual data submatrices $\mathbf{D}_k^j$. As can be seen in Fig. 1, individual data submatrices ($\mathbf{D}_k^j$) corresponding to the same chromatographic window for the different analyzed samples were arranged in nine column-wise augmented data matrices. The dimensions of these nine column-wise augmented data matrices depended on the dimensions of the selected time windows (selected retention times). Thus, from window I to window IX the augmented matrices dimensions were: 1010 × 546, 1110 × 546, 2210 × 546, 2910 × 546, 2080 × 546, 2760 × 546, 3410 × 546, 3210 × 546 and 4010 × 546. In every case the first dimension refers to the sum of the retention times of the ten included samples (4 control, 4 stressed samples and 2 standards), and the second dimension is equal to the number of *mz* values included in the analysis, which were in all cases 546 *mz* values (from 55 to 600 Da).

3.3 Principal component analysis

Principal component analysis (PCA) was used for initial exploration of the behaviour of yeast samples metabolic profiles according to temperature changes. PCA compresses the information contained in the original variables into a small number of new orthogonal variables (components) built from linear combinations of the original variables explaining most of the measured data variance (Wold et al. 1987; Esbensen and Geladi 2009). Plots of firsts components are usually enough to explore the main sources of variance in the original data. Here PCA was performed on the TIC chromatograms of the eight yeast samples at the two culture temperatures (30 °C and 42 °C).

3.4 Partial least squares-discriminant analysis (PLS-DA)

PLS-DA (Barker and Rayens 2003) is a PLS regression method (Geladi and Kowalski 1986b; Wold et al. 2001) which correlates a set of response variables **y** to a set of predictor variables **X**, where **y** is a set of binary variables of describing the categories of **X**. PLS-DA estimates in an very efficient way the best linear combinations of the independent original **X**-values (called latent variables, LV), which correlate optimally with the observed changes of the dependent variable, **y**. PLS-DA tries to build a model that maximizes the covariance between **X** and **y** with a minimum number of latent variables. For every latent variable, a vector of weight coefficients shows what **X**-variables are best combined to form the **X**-scores vector.

This method was used in this work to investigate what metabolites could be more influenced by temperature

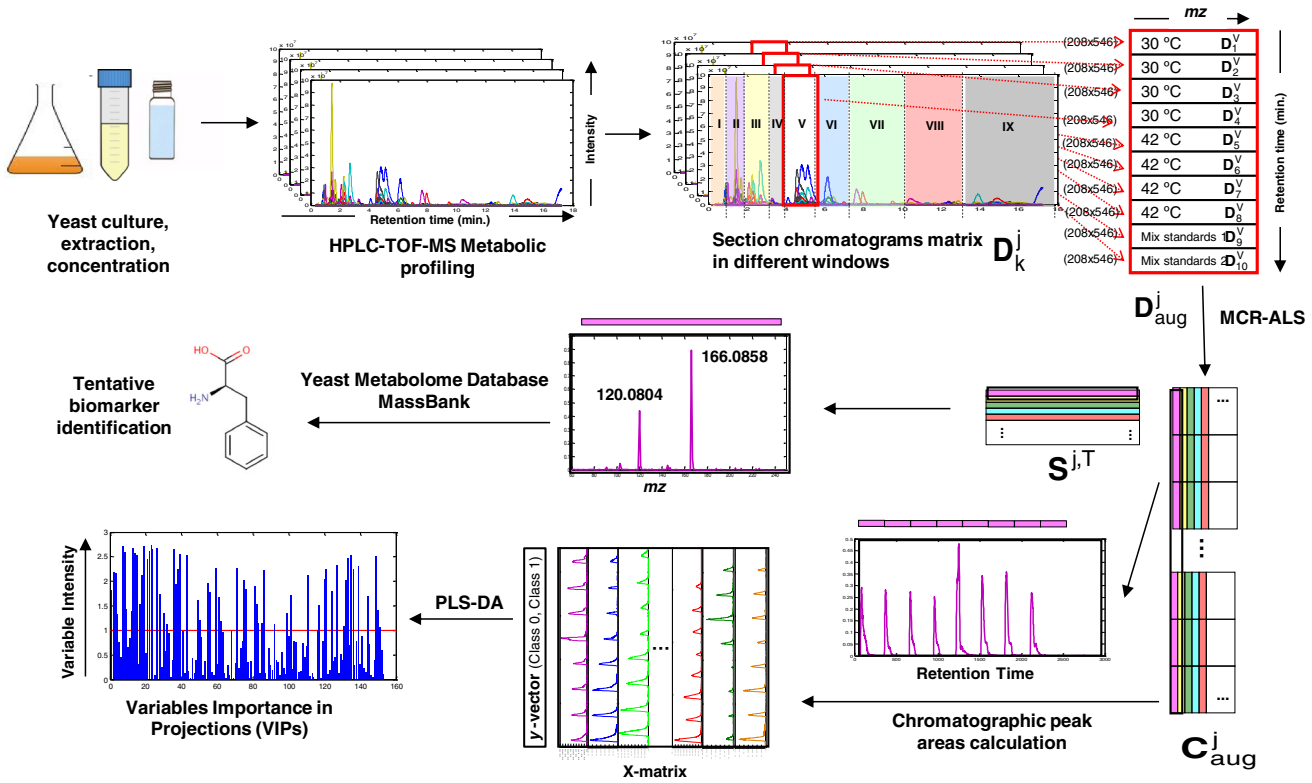


Fig. 1 Schematic representation of the workflow following untargeted (LC–MS) data generation. The workflow involved experimental analysis, data pre-processing and data analysis in order to identify possible biomarkers (yeast metabolites)

changes on yeast cultures. PLS-DA regression was applied to optimally model class variable **y** (samples cultured at 30 °C categorized in class 0 and samples cultured at 42 °C categorized in class 1, which described temperature changes in relation to the observed changes in the predictor variables (**X** matrix) (Wold 1966; Geladi and Kowalski 1986a, b; Wold et al. 2001). In this work, two PLSA-DA analysis were performed. In a first analysis, PLS-DA was applied to the **X** TIC data matrix (with dimensions of 8 × 2020), as in PCA. In a second analysis, PLS-DA was applied to the **X** TIC data matrix (with dimensions of 8 × 91), containing the peak areas of the components resolved by MCR-ALS in the full scan MS chromatographic analysis of the eight yeast samples at the two temperatures (see Sect. 3.5).

To investigate the more influent variables (peak retention times and possible metabolites associated to them) in the PLS-DA model, Variable Importance in Projection (VIP) scores (Wold et al. 1993; Wold 1995; Wold et al. 2001) were calculated. VIP scores (Wold et al. 1993) are a weighted sum of squares of PLS weights for each variable and measure the contribution of each predictor variable to the model. It is frequently used as a parameter for variable selection (Chong and Jun 2005; Rajalahti et al. 2009; Andersen and Bro 2010). For a given model and problem there is one VIP-vector, summarizing the contribution of the selected number of components on the prediction of the *y* variable (Wold et al. 2001). On the other hand, since the average of squared VIP score is equal to 1, the 'greater than one' rule is used as a criterion for variable selection (Chong and Jun 2005).

3.5 Multivariate curve resolution-alternating least squares (MCR-ALS)

The goal of this analysis was to resolve the maximum number of individual elution profiles and pure mass spectral profiles of the possible metabolites extracted from the investigated yeast samples. MCR-ALS is chemometric method which allows for the resolution of multiple components in unknown unresolved mixtures from chromatographic systems, including strongly coeluted, overlapped and embedded peaks.

In the particular case under study, the MCR bilinear model is mathematically described according to:

$$\mathbf{D}_k^j = \mathbf{C}_k^j S^{j,T} + \mathbf{E}_k^j \text{ for } j = \text{I, II, ..., IX windows and } k = 1, 2, \ldots, 10 \text{ samples} \quad (1)$$

Rows of data matrices $\mathbf{D}_k^j$ are the different elution times of the samples chromatographic analysis. Columns of data matrices $\mathbf{D}_k^j$ are the mass spectra recorded at the different elution times. $\mathbf{C}_k^j$ is the matrix of MCR-ALS resolved elution profiles in window j and sample k, and $\mathbf{S}^{j,T}$ is the matrix of their corresponding pure mass spectra. These resolved pure mass spectra can be then used for the identification of the different metabolites. $\mathbf{E}_k^j$ contains the unexplained variance related to background and noise contributions not modelled by $\mathbf{C}_k^j$ and $\mathbf{S}^{j,T}$.

This data analysis can be extended to the simultaneous analysis of the different control and stressed yeast samples and standard metabolite mixture samples (chromatographic runs), which facilitated the resolution of the coeluted metabolites simultaneously present in the yeast samples. Data submatrices $\mathbf{D}_k^j$ corresponding to the same time window are settled one on the top of the other (column-wise augmented matrices In this new data arrangement, the new column (*mz*) vector subspace is the same for all sample matrices. The new column-wise augmented data matrix D_{aug}^j can be decomposed similarly using the bilinear model equation:

$$\mathbf{D}_{aug}^j = \begin{bmatrix} \mathbf{D}_1^j \\ \mathbf{D}_2^j \\ \mathbf{D}_3^j \\ \vdots \\ \mathbf{D}_{10}^j \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1^j \\ \mathbf{C}_2^j \\ \mathbf{C}_3^j \\ \vdots \\ \mathbf{C}_{10}^j \end{bmatrix} \mathbf{S}^{j,T} + \begin{bmatrix} \mathrm{E}_1^j \\ \mathbf{E}_2^j \\ \mathbf{E}_3^j \\ \vdots \\ \mathbf{E}_{10}^j \end{bmatrix}$$

$$= \mathbf{C}_{aug}^j \mathbf{S}^{j,T} + \mathbf{E}_{aug}^j \text{ for } j = \mathrm{I}, \mathrm{II}, \ldots, \mathrm{IX} \text{ windows}$$

The detailed procedure followed in the MCR-ALS analysis of every window, $\mathbf{D}_{aug}^j$ is shown in Fig. 1. Nine chromatographic windows were considered for MCR-ALS analysis. These nine chromatographic windows (j = I, II, …, IX) covered the full investigated time range. Ten submatrices corresponding to the ten samples (k = 1, 2, …, 10) were considered in each augmented data matrix $\mathbf{D}_{aug}^j$ for all studied windows. Every individual data matrix (one window, one sample) had a number of rows equal to the total number of recorded elution times in the considered chromatographic region, although the total number of considered rows (retention times) did not exactly match among the different considered sample submatrices, $\mathbf{D}_k^j$. The number of columns was always equal to the same number of considered *mz*. $\mathbf{C}_{aug}^j$ has the resolved augmented elution profiles of the resolved peaks. $\mathbf{S}^{j,T}$ is the matrix of pure mass spectra of the resolved coeluted compounds, and $\mathbf{E}_{aug}^j$ matrix is the noise and background signal absorption not explained by the model described by $\mathbf{C}_{aug}^j$ and $\mathbf{S}^{j,T}$(Tauler 1995).

Before starting the Alternating Least Squares (ALS) iterative process to solve Eqs. (1) and (2), the number of components is initially estimated by principal component analysis (PCA) (Wold et al. 1987) or more simply, by the singular value decomposition (SVD) (Golub and Loan 1996). In the present study, the applied constraints have been non-negativity and spectra normalization. Non-negativity was applied to chromatographic and mass spectra profiles and normalization constraint was applied to the pure mass spectra profiles to fix their scale during ALS optimization. See (Tauler and Barceló 1993; Tauler 1995; Tauler et al. 1995; de Juan et al. 2009; Tauler et al. 2009) for further details of MCR-ALS method and constraint implementation.

Figures of merit of the MCR-ALS optimization procedure are the percent lack of fit, which is the difference among the input data $\mathbf{D}_{aug}^j$ and the data reproduced from the product obtained by MCR-ALS ($\mathbf{C}_{aug}^j \mathbf{S}^{j,T}$); and the percent of explained variance (R^2).

Due to the high selectivity of pure component mass spectra, rotation ambiguities were practically reduced to a minimum. Only in cases where strongly coeluted peaks have common molecular ions, it is expected to have some degree of rotation ambiguity. Moreover the simultaneous analysis of multiple data matrices, including those from the analysis of the standard mixture samples, reduced more the possible presence of rotation ambiguities associated to MCR-ALS solutions of the augmented window data matrices.

Full scan LC–MS data matrices of control and temperature stressed samples, together with those of the standard mixture samples, were simultaneously analysed by MCR-ALS (see Sect. 3.2 and Eq. 2). First four matrices were the LC–MS full scan data matrices of the control yeast samples (cultured at 30 °C). Second four matrices were the LC–MS full scan data matrices of the stressed yeast samples (cultured at 42 °C) and last two matrices were the full scan LC–MS data matrices of the two standard mixtures. The later were included to check that MCR-ALS method was appropriately applied to separate coeluted chromatographic peaks of the components of the standard mixtures. As it was already mentioned in Sect. 3.2, since the simultaneous resolution of the whole LC–MS full scan chromatogram would give an augmented data matrix of dimensions (20200 × 546), the complete chromatogram obtained for each individual sample was sectioned in nine windows, subdividing then the MCR-ALS analysis in nine MCR-ALS differentiated analysis of the corresponding data submatrices, $\mathbf{D}_k^j$, j = I, II, …, IX (see Sect. 3.2 and Fig. 1).

4 Metabolite identification

Due to the high number of peaks generated from a metabolomic analysis and in order to identify yeast metabolites, the application of MCR-ALS was used to facilitate the resolution of the coeluted and embedded chromatographic peaks. Otherwise, resulting MS chromatograms were too complex to process them at once and for the non-target searching of the individual components. Results from the ALS optimization are shown as: resolved eluted profiles, $\mathbf{C}^{j}_{aug}$ matrices, and pure mass spectra, $\mathbf{S}^{j,T}$ matrices (see Eq. 2). Peak areas of the resolved elution profiles were used to investigate possible temperature effects and mass spectra of the corresponding compounds were used for yeast metabolites identification and confirmation (Fig. 1). Since LTQ-Orbitrap instrument allowed for a very accurate mass measurement of four decimal places, possible metabolites candidates were investigated according to their accurate mass (positively ionized) value. Most abundant measurements of *mz* of all resolved pure mass spectra, $\mathbf{S}^{j,T}$, were used to match a metabolic feature to a single or small number of molecular formula in combination with chemical and biological knowledge. Accurate *mz* data were searched in Yeast Metabolome Database (YMDB) (Jewison et al. 2012). The candidates were checked in full scan MS chromatogram from LTQ-Orbitrap direct data acquisition having full MS accuracy. MZmine 2 framework (Pluskal et al. 2010) was used to search the resolved peaks in the LTQ-Orbitrap original raw data, taking as a reference the peak retention time obtained in the MCR-ALS eluted profiles matrix and the molecular formula matched in YMDB. Mass tolerances of 0.01 millimass units (mmu) were allowed for matching a particular molecular formula when searched in LTQ-Orbitrap full scan MS chromatograms. In this way, it was ensured that considered MCR-ALS resolved elution profiles matched with the ones originally present in raw MS chromatograms. Further confirmation of the MCR-ALS resolved candidates (metabolites) was done by comparison with the metabolite mass spectrum from the MassBank mass public spectral database (Horai et al. 2010).

Although some ion types were expected (protonated peaks, isotope peaks) others were not expected (adduct ions). Initial identifications were refined using previous analyses of the yeast metabolome (Canelas et al. 2009; Beltran et al. 2012). Predicted exact mass for different adducts were calculated using the Mass Spectrometry Adduct Calculator from Metabolomics Fiehn Lab webpage (Huang et al. 1999). Identified metabolites were further functionally and metabolically characterized using the KEGG database (Kanehisa et al. 2012). Section 5.4 gives an example of detailed application of the procedure described above.

5 Results and discussion

5.1 PCA and PLS-DA of MS-TIC chromatograms

When principal component analysis (PCA) was applied to the mean-centered MS TIC data matrix (with 8 samples and 2020 measured chromatographic retention times), three principal components already explained 88.83 % of data variance. In Fig. 2a, scores of the first two components are given. PC1 explains 47.76 % of the data variance and separates the samples in relation to the yeast culture temperatures. Samples grouped in the negative side of PC1 axis were grown at 42 °C and samples grouped on the positive side of PC1 axis are the control samples grown at 30 °C. Variances explained by PC2 and PC3 are related to other unknown variability sources not dependent of temperature.

Partial least-squares discriminant analysis (PLS-DA) was applied to investigate the more relevant variables (peak retention times) related to the discrimination between control and stressed yeast samples. The performance of PLS-DA model was also calculated on the mean-centred MS TIC data matrix. PLS-DA was assessed by using leave-one-out cross-validation method (adequate for a small number of samples as in this study). $\boldsymbol{y}$ vector containing the class labels was also mean-centred. First PLS latent variable (LV1) already accounted for 47.32 % of $\mathbf{X}$ data variance and for 86.66 % of the dependent variable $\boldsymbol{y}$ (low temperature control samples and high temperature stressed yeast samples). This confirms again that the main source of variance in TIC chromatograms was related to temperature changes. In the scores plot of PLS-DA (see Fig. 2b), the two groups of samples (control and temperature stressed) were clearly distinguished. To help the visualization of the more influent variables PLS-DA VIP values were calculated (see Sect. 3.3 and Fig. 2c).

Due to the strong co-elution among multiple chromatographic peaks at the same retention times in all TIC chromatograms, the evaluation of the relative importance of the different variables (peak retention times) on temperature changes was rather difficult and further analysis was performed using MCR-ALS analysis of full scan MS chromatograms. This allowed the improved mathematical resolution of the coeluted peak profiles and the estimation of their corresponding pure mass spectral profiles, and their further identification (see below).

5.2 MCR-ALS of full scan LC–MS chromatograms

A total number of nine column-wise augmented data matrices, each one corresponding to one of the nine windows, were analysed separately by MCR-ALS. All MCR-ALS

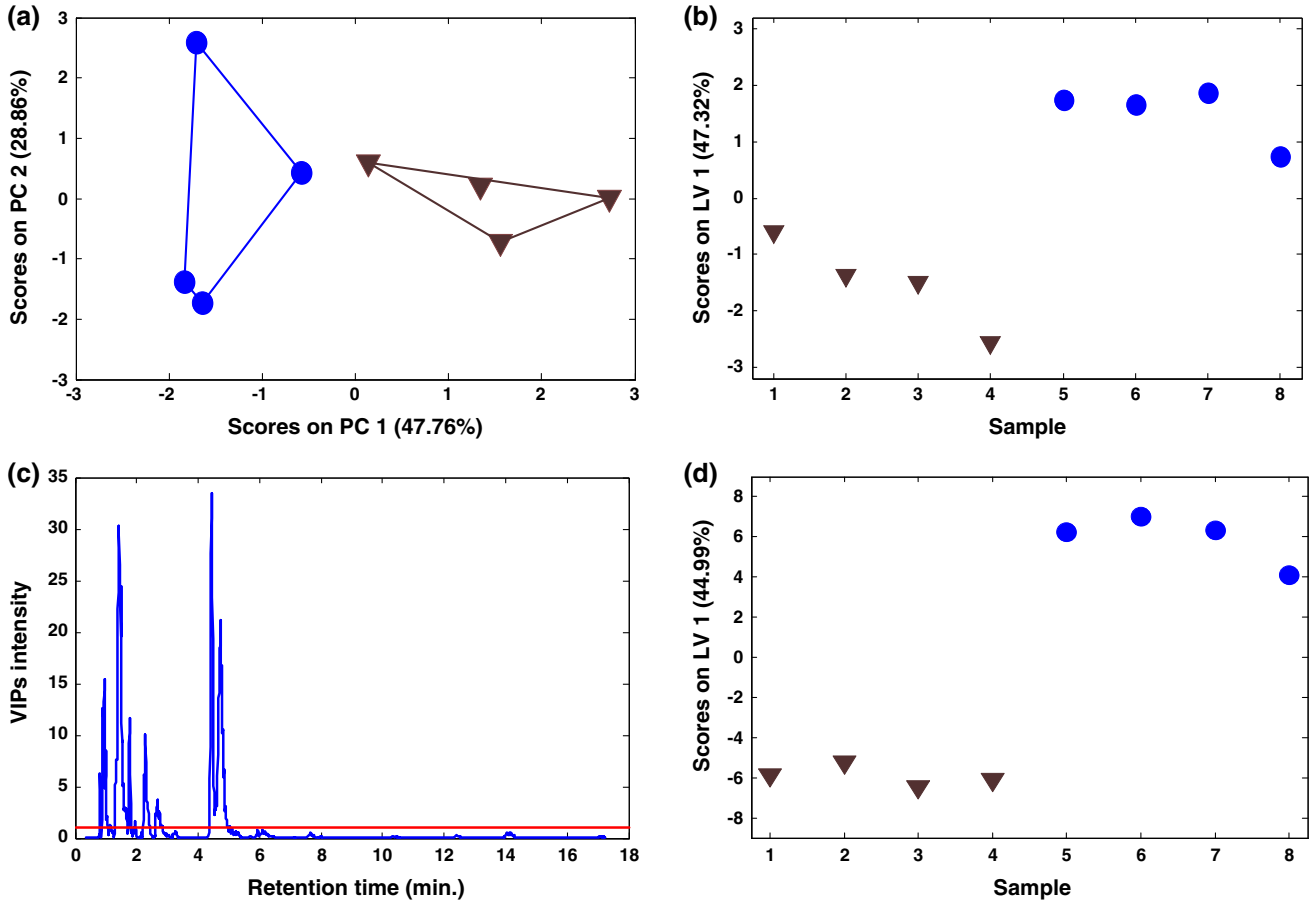


Fig. 2 **a** PCA scores plot for the eight yeast samples (MS TIC chromatograms). Convex hulls are drown around each yeast culture temperature group with the *same color* as the *corresponding symbols*. **b** PLS-DA scores plot for the eight yeast samples at the two temperatures using their MS TIC chromatograms. **c** Variables importance in projection (VIP) plot resulting from PLS-DA analysis of full scan MS TIC yeast chromatograms. *Horizontal red line* shows the threshold value used to select the variables with the most important VIP scores. **d** PLS-DA scores of autoscaled chromatographic peak areas obtained by MCR-ALS analysis of full scan MS chromatographic data of the analysed yeast samples. In **a**, **b** and **d** *blue solid circles* are control samples cultured at 30 °C and *brown triangles* are yeast samples cultured at 42 °C (Color figure online)

explained variance (R^2) percentages were higher than 98 %. A relatively large number of MCR-ALS components (resolved peaks) were needed to explain properly the observed data variance and patterns. Not all MCR-ALS resolved components corresponded to true chromatographic peaks which could be assigned to separate metabolites, since other possible signal contributions such as the background and solvent contributions could also be present.

Figure 3 is an example of MCR-ALS applied to the column-wise augmented data matrix ($\mathbf{D}^{j}_{aug}$) corresponding to time window III (elution times from 1.82 to 2.69). In this example, the two standard mixture samples were omitted from the figure because there was no standard compound eluting in this time window. As it can be seen in Fig. 3, the four coeluted components (h, p, r, d) were successfully separated by MCR-ALS analysis. Elution profiles ($\mathbf{C}^{III}_{aug}$) and pure mass spectra ($\mathbf{S}^{III,T}$) are shown. The four contributions (h, p, r, d) were identified by their mass spectra as explained in Sects. 4 and 5.4 for their further metabolite identification). There were other components (not shown) which were sections of chromatographic peaks corresponding to windows II and IV. There were also some minor noise interferences without chromatographic peak shape and very imprecise spectra, which were finally not shown in the figure for clarity.

As stated before, when resolving the column-wise augmented data matrices of every window j, $\mathbf{D}^{j}_{aug}$ with MCR-ALS, *mz* resolution was restricted to one integer mass value to facilitate its numerical analysis (see Sect. 3.1). This resolution was generally enough to resolve the elution profiles of the coeluted metabolites, and allowed the simultaneous estimation of their full scan MS spectra at this limited resolution. In some cases, however, this was not sufficient, and an optimal resolution could not be

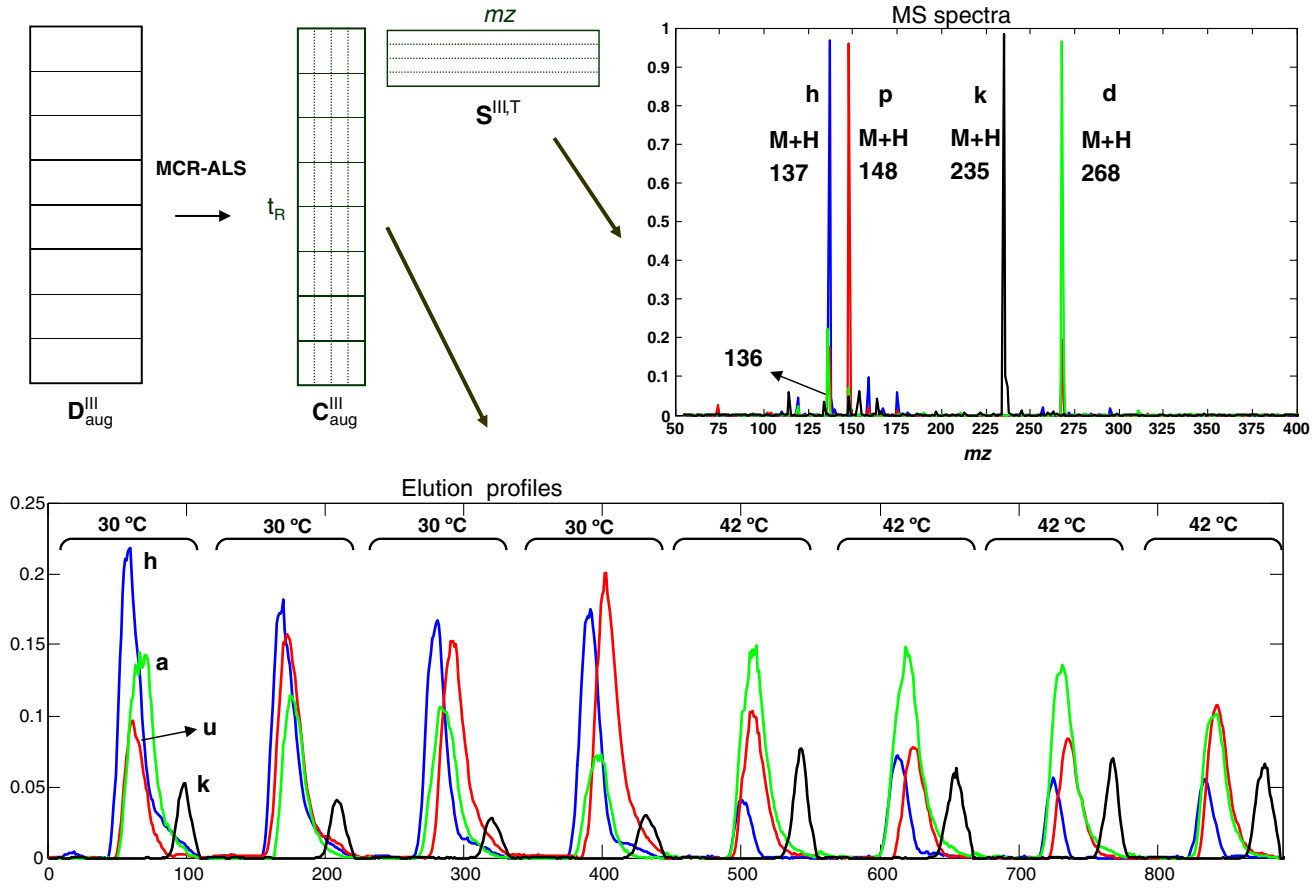


Fig. 3 Example of MCR-ALS resolution (Multivariate curve resolution-alternating least squares) simultaneously applied to the column-wise augmented data matrix ($\mathbf{D}^{III}_{aug}$) corresponding to time window III including the control and the stressed yeast samples. $\mathbf{C}^{III}_{aug}$ is the matrix of MCR-ALS resolved elution profiles. $\mathbf{S}^{III,T}$ is the matrix of MCR-ALS resolved MS pure spectra. Compound labels identification in this example are: *h* hypoxanthine, *p* palmitic acid, *r* arbitol/ribitol, *d* deoxyguanosine

achieved. For instance, when two peaks were overlapped in the same elution profile but giving the same MS, after checking their exact mass in the high resolution raw data these two compounds resulted to have slightly different masses. Therefore, to allow the separation of these two chromatographic peaks as different components higher resolution data should be processed by MCR-ALS.

Once the whole set of column-wise augmented data matrices corresponding to the nine time windows were analysed by MCR-ALS, the areas of all MCR-ALS resolved peaks for each component profile were calculated and arranged in a data table. In total, 91 components were resolved and their peak areas calculated for every one of the eight analysed yeast samples, i.e. a table with a total number of 8 × 91 peak areas was obtained.

Metabolite profiles in all window profiles were strongly overlapped and their proper resolution was strongly facilitated by the proposed MCR-ALS analysis. It might be argued that similar results could have been achieved using LC–MS in single ion monitoring (SIM) mode once their characteristic molecular MS ions were identified (target analysis). However this analysis would have required the exploration of such a large number of possibilities that it would have made this metabolomic study very tedious and impractical, if not unaccurate in many cases. Moreover, for the goals of the present study, the larger the number of possible unknown metabolites (non-target analysis) simultaneously analyzed and resolved the more interesting could be the conclusions derived from the obtained results.

5.3 PLS-DA of chromatographic peak areas

PLS-DA was applied to the peak areas of all 91 MCR-ALS resolved elution profiles, for control and treated yeast samples (X matrix with dimensions of 8 × 91). Prior to PLS-DA model calculation, the peak areas were autoscaled to give equal relevance to their possible change due to the temperature differences in control and treated yeast samples. PLS-DA model was developed to investigate what metabolite peak areas were more important in the

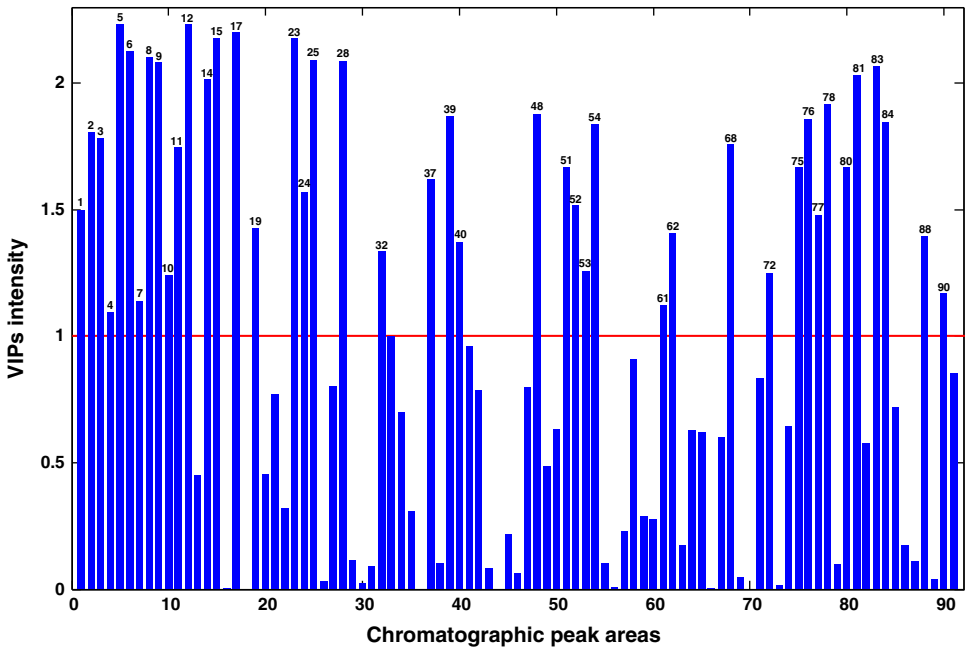


Fig. 4 Variables importance in projection (VIP scores) plot resulting from PLS-DA analysis of the autoscaled chromatographic peak areas obtained by MCR-ALS analysis of yeast samples. *Horizontal red line* shows the threshold value selecting the most important variables (Color figure online)

discrimination between yeast culture samples at 30 and 42 °C and to identify potential metabolites markers of the temperature effects on yeast metabolism. No outlier samples were detected using a leave-one-out cross-validation (adequate strategy for a small number of samples). One PLS-DA component was enough to explain most of the class variance (98.02 % variance) using the 44.99 % of **X** variance related to the changes on chromatographic peak areas of the resolved components, with specificity and sensitivity values equal to 1 for each class.

In Fig. 2d, the projection of the first latent variable (LV1) scores is given; two groups of samples are clearly distinguished. Yeast samples at 30 °C were projected on the negative scores axis whereas yeast samples at 42 °C were projected on the positive scores axis.

VIP values (Eriksson et al. 2006) calculated from PLS-DA model (see Sect. 3.3) revealed what variables (metabolites) were more important to discriminate the effects of temperature changes on yeast metabolism. Variables whose VIP values were higher than 1.0 were considered as potential indicators (Chong and Jun 2005) of these effects (see Fig. 4). Results are listed in Table 2. PLS-DA VIPs (Fig. 4) and the corresponding PLS-DA weights provided a summary of PLS-DA results. In this table chromatographic peaks increasing ('Ups') or decreasing ('Downs') at 42 °C relative to the control samples cultured at 30 °C are given.

5.4 Tentative identification of possible biomarker compounds

Identification of the metabolites corresponding to all 91 resolved chromatographic peaks was attempted. Taking as example the pure MS spectrum of one of the components resolved by MCR-ALS (at window IV, component 12, peak 40 on Table 2), its corresponding metabolite identification was performed in the following way: Its main integer mass value was 166 (molecular mass positively ionized). When this mass value was searched in the YMDB database, possible candidates were L-methionine (R)-S-oxide, D/L-Phenylalanine, 4-Pyridoxolactone, *N*-Formylanthranilic acid and 7-Methylguanine. These were the yeast metabolites that when positively ionized, $[M + H]^+$, give the value of 166 as the most abundant *mz* of its mass spectrum. When these candidates were searched in the raw full scan MS chromatograms, using then high resolution *mz* values through MZmine2 framework (Pluskal et al. 2010), D/L-Phenylalanine was the selected candidate that better matched with the resolved chromatographic peak. Its accurate mass was 166.0858, which when searched in MassBank database (Horai et al. 2010) was confirmed to be the spectrum of D/L-Phenylalanine (with an accurate mass of 166.0868). Some MS pure spectra resolved by MCR-ALS gave additional less intense signal product ions which could be also used for identification and confirmation of possible metabolite candidates. For instance, in the case of the MCR-ALS resolved pure MS spectrum of D/L-Phenylalanine another ion mass was detected at 120.0804, which matched very well with a product ion of metabolite D/L-Phenylalanine at 120.0813 in the MassBank data base. This should be considered a clear additional advantage of the proposed MCR-ALS strategy which is difficult to be achieved using traditional direct identification approaches. MCR-ALS resolution strategy allows the simultaneous identification of both molecular and product ions in the spectrum of every resolved component.

Table 1 Temptative identification of yeast metabolites associated to chromatographic peaks changing their areas when culture temperature changed from 30 to 42 °C

Ups				Downs			
Peak number	C-number	Metabolite	Weight	Peak number	C-number	Metabolite	Weight
3	C06104	Adipic acid	0.1401	1	C00033	Acetic acid	−0.1285
5	C05853	2-Phenylethanol	0.1568	2	C00097	L-Cysteine	−0.141
6	C00077	L-Ornithine	0.1532	4			−0.1098
11	C00864	Pantothenate	0.1388	7	C00116	Glycerol	−0.1121
12	C01571	Capric acid	0.1569	8	C00147	Adenine	−0.1523
14	C00474	Arabitol/ribitol	0.1491	9	C00262	Hypoxanthine	−0.1516
15	C00559	Deoxyadenosine	0.1549	10	C00249	Palmitic acid	−0.1169
17	C06423	Caprylic acid	0.1558	23	C00791	Creatinine	−0.155
19	C00120	Biotin	0.1255	24			−0.1315
25	C00474	Arabitol/ribitol	0.1518	32		LysoPC(18:1(11Z))	−0.1213
28	C01087	2-Hydroxyglutaric acid	0.1517	37	C00902	2-Oxohexanoic acid	−0.1337
40	C00079	L-Phenylalanine	0.1229	39	C00160	Glycolic acid	−0.1437
51			0.1356	48	C00250	Pyridoxal	−0.1439
54	C00049	L-Aspartic acid	0.1423	52	C02794	L-3-Hydroxykynurenine	−0.1293
62	C00041	L-Alanine	0.1247	53	C00082	L-Tyrosine	−0.1176
72			0.1173	61	C00152	L-Asparagine	−0.1112
75	C00114	Choline	0.1356	68			−0.1392
76	C07113	Acetophenone	0.1432	77	C02059	Phylloquinone	−0.1277
80			0.1356	78	C00378	Thiamine	−0.1454
81			0.1496	84	C00192	Hydroxylamine	−0.1428
83			0.151	88	C01346	dUDP	−0.1242
90			0.1135				

Following the same procedure, a total number of 65 metabolites were tentatively identified out of the 91 potentially different chromatographic peaks (Table 2). In most cases (54 out of 65), observed *mz* values differed from the calculated ones by less than 50 ppm (Table 2), and this error was considered acceptable given the methodology used in this work. All metabolites displayed in Table 2 are either bona fide yeast metabolites classified as such in the YMBD (see YMDB numbers in Table 2) or they have been identified in similar yeast metabolome studies (Canelas et al. 2009; Beltran et al. 2012). Identified metabolites include 17 out of the 20 protein amino acids, plus L-Ornithine, although L-Serine and L-Glutamine. Other major components of the yeast metabolome were identified, as nucleosides and their derivatives, vitamins, glycerol and some organic acids and lipids (Table 2). In summary, the proposed methodology correctly identified many of the major components of the yeast metabolome. Further work and improvement of the proposed methodology is persued to allow the complete description of the more important metabolites present in yeast cells.

5.5 Biological interpretation of the changing metabolites

Results obtained in previous Sects. 5.3 and 5.4 showed that some of the MCR-ALS resolved compounds displayed a relative increase or decrease of their abundances (relative chromatographic areas) depending on the culture temperature (42 vs. 30 °C). Tentative identification of metabolites whose abundance increased (Ups) or decreased (Downs) at 42 °C relative to the standard 30 °C culture temperature is displayed in Table 1. In this case, almost 80 % of the peaks showing a significant change of their area due to temperature were tentatively identified. Functional analyses of the altered metabolites did not allow an accurate assessment of the metabolic pathway changes underlying the acclimatisation of the yeast cell to growth at high temperatures. However, it is noticeable the increase of two pentose alcohol, tentatively identified as ribitol and arabitol (Tables 1, 2), coupled to the decrease of glycerol. Yeast is known to maintain their osmotic balance by modifying their internal concentration of glycerol and/or sugar alcohols (Hohmann 2002). Although *S. cerevisiae* is believed

Table 2 Tentative adscription of yeast metabolites to observed mass values for the different chromatographic peaks (not only those showing area changes when culture temperature change)

Peak number	Retention time	Highest mass ion	Proposed metabolite	KEGG C-number	YMDB	Adduct	Adduct m/z	error mz (ppm)	>50 ppm
1	1.38	102.0547	Acetic acid	C00033	YMDB00056	M + ACN + H	102.0549	1.96	
2	1.33/1.69	123.0552	L-Cysteine	C00097	YMDB00046	M + 3ACN + 2H	123.0569	13.81	
3	1.7–2	293.1167	Adipic acid	C06104		2M + H	293.1231	21.83	
5	1.8	267.1452	2-Phenylethanol		YMDB01072	2M + Na	267.1355	36.13	
6	1.48	155.0811	L-Ornithine	C00077	YMDB00353	M + Na	155.0791	12.90	
7	1.9–2	156.0656	Glycerol	C00116	YMDB00283	M + ACN + Na	156.0631	16.02	
8	2.1	136.0619	Adenine	C00147	YMDB00887	M + H	136.0618	0.73	
9	2.26	137.0457	Hypoxanthine		YMDB00555	M + H	137.0458	0.63	
10	2.28–2.38	148.0965	Palmitic acid	C00249	YMDB00069	M + H + K	148.1053	59.42	*
11	1.85–2.5	220.1175	Pantothenate	C00864	YMDB00203	M + H	220.1180	2.16	
12	2.1	211.1119	Caproic acid	C01571	YMDB00677	M + K	211.1095	11.37	
13	2.3	268.1028	Deoxyguanosine	C00330	YMDB00505	M + H	268.1041	4.76	
14	2.59	235.1192	Arabitol/ribitol	C00474	YMDB00591	M + 2ACN + H	235.1288	40.83	
15	1.95	252.1088	Deoxyadenosine	C00559	YMDB00503	M + H	252.1091	1.09	
17	2.1	167.0922	Caprylic acid	C06423	YMDB00676	M + Na	167.1042	71.81	*
19	1.8	245.0949	Biotin	C00120	YMDB00282	M + H	245.0955	2.35	
20	3.6	209.1032	Thymine	C00178	YMDB00885	M + 2ACN + H	209.1033	0.48	
21	4.1	195.0871	Uracil	C00106	YMDB00098	M + 2ACN + H	195.0877	3.08	
23	2.7	114.0659	Creatinine	C00791		M + H	114.0662	2.63	
25	2.6	235.1192	Arabitol/ribitol	C00532	YMDB00591	M + 2ACN + H	235.1288	40.83	
26	3.1–3.3	494.3212	LysoPC(16:1(9Z))		YMDB02210	M + H	494.3241	5.90	
28	2.7	166.072	2-Hydroxyglutaric acid	C01087	YMDB00059	M + NH4	166.0710	6.02	
30	3.3–3.4	192.1588	L-Arginine	C00062	YMDB00592	M + NH4	192.1455	69.22	*
31	2.7	184.0633	L-2-Aminoadipate		YMDB00999	M + NA	184.0580	28.65	
32	2.85	522.3528	LysoPC(18:1(11Z))		YMDB02211	M + H	522.3554	5.01	
34	4.3	162.0576	Cystine	C01420	YMDB00861	M + 2ACN + 2H	162.0457	73.29	*
35	5.9	150.0583	L-Methionine	C00073	YMDB00318	M + H	150.0583	0.00	
37	2.3	148.0965	2-Oxohexanoic acid	C00902	YMDB00388	M + NH4	148.0968	2.03	
38	1.8–2	205.0673	2-Oxobutanoate	C00109	YMDB00071	2 M + H	205.0707	16.58	
39	4.5–4.6	159.0761	Glycolic acid	C00160	YMDB00807	M + 2ACN + H	159.0764	1.89	
40	4.5–4.6	166.0858	L-Phenylalanine	C00079	YMDB00304	M + H	166.0863	3.01	
41	4.7–4.8	205.0972	L-Tryptophan	C00078	YMDB00126	M + H	205.0972	0.12	
48	5–5.3	209.0917	Pyridoxal	C00250	YMDB00392	M + ACN + H	209.0920	1.43	
49		132.102	L-Isoleucine	C00407	YMDB00038	M + H	132.1019	0.76	
49		132.102	L-Leucine	C00123	YMDB00387	M + H	132.1019	0.76	
52	5.9–6	225.0858	L-3-Hydroxykynurenine		YMDB00105	M + H	225.0870	5.26	
53	6.1–6.2	182.0804	L-Tyrosine	C00082	YMDB00364	M + H	182.0812	4.39	
54	7.1	175.0866	L-Aspartic acid	C00049	YMDB00896	M + ACN + H	175.0713	87.39	*
56		116.0705	L-Proline	C00148	YMDB00378	M + H	116.0706	0.86	
57	6.2	118.0863	Histamine	C00388	YMDB01556	M + 3ACN + 2H	118.0869	5.08	
58	6	118.0858	L-Valine	C00183	YMDB00152	M + H	118.0863	4.23	
59	6	72.0805	2-Nonanone		YMDB01383	M + 2H	72.0752	74.11	*
60	6.6–6.7	148.0605	L-Glutamate	C00025	YMDB00271	M + H	148.0605	0.00	
61	7.6–7.7	150.0778	L-Asparagine	C00152	YMDB00226	M + NH_4	150.0873	63.30	*
62	7.6	90.0546	L-Alanine	C00041	YMDB00154	M + H	90.0550	4.44	
64	9.1	147.0764	L-Glutamine	C00064	YMDB00002	M + H	147.0764	0.00	

Table 2 continued

Peak number	Retention time	Highest mass ion	Proposed metabolite	KEGG C-number	YMDB	Adduct	Adduct m/z	error mz (ppm)	>50 ppm
64	9.1	147.0764	L-Serine	C00065	YMDB00112	M + ACN + H	147.0764	0.00	
65	7	218.1384	Ergosterol	C01694	YMDB00543	M + H + K	218.1548	75.25	*
67	8.8	120.065	L-Threonine	C00188	YMDB00214	M + H	120.0655	4.16	
69	11–11.2	337.1677	Pyridoxamine	C00534	YMDB00889	2M + H	337.1871	57.53	*
70	9.5	162.1122	Octadecanoic acid	C01530	YMDB00682	M + H + K	162.1210	54.08	*
71	10.2	258.1088	Glycerophosphocholine		YMDB00309	M + H	258.1101	5.04	
73	9.7–9.8	246.0982	Deoxyuridine	C00526	YMDB00508	M + NH4	246.1084	41.54	
74	11.3–11.4	401.1649	Cortisol	C00735		M + K	401.1725	18.84	
75	11.7–11.8	146.1175	Choline	C00114	YMDB00227	M + ACN + H	146.1413	162.86	*
76	9.5	121.0648	Acetophenone	C07113	YMDB01629	M + H	121.0648	0.00	
77	10–11.3	151.123	Phylloquinone	C02059	YMDB01526	M + 3H	151.1239	5.80	
78	12.7–12.9	154.0972	Thiamine	C00378	YMDB00220	M + ACN + 2H	154.0767	133.05	*
82	12–12.1	130.0974	Methyl-3-ethyl-butanoate		YMDB01749	M + H	130.0988	11.00	
84	13.8–14.2	116.0819	Hydroxylamine	C00192		M + 2ACN + H	116.0818	0.86	
86	14.9–15	309.1377	Adenosine	C00212	YMDB00058	M + ACN + H	309.1306	22.89	
87	15.5–15.6	131.118	lignoceric acid	C08320	YMDB00684	M + 2H + Na	131.1231	38.89	
88	14.5–15	145.1079	dUDP	C01346	YMDB00746	M + H + 2Na	145.1020	40.96	
89	14.2–14.3	181.0968	Limonene	C06078	YMDB01727	M + 2Na–H	181.0964	2.21	
91	13.3	137.1069	2-Methyl-5-propylpyrazine		YMDB01504	M + H	137.1073	3.10	

to only use glycerol for this protective function, the observed changes may reflect a re-equilibrium in osmolite concentrations as a response to the high growth temperature. Similarly, decrease of two structural lipids/organic acids (palmitic acid and the lysophospholipid tentatively identified as LysoPC(18:1(11Z)) may reflect an alteration on the yeast lipid composition to adapt membrane fluidity to the high temperatures. Finally, it is known that continuous growth at high temperatures induces a switch in yeast metabolism towards respiration from fermentation (Mensonides et al. 2013)). Therefore, it is conceivable that some of the observed changes, like the increase of short/medium-length organic acids adipic, caprylic, and capric acids, or the decrease of fermentation sub-products, like acetate and glycerol, may reflect this metabolic adaptation to temperature changes. Further interpretation of the observed changes will require a more accurate investigation and knowledge of biochemical pathways of yeast growing under different temperatures.

6 Concluding remarks

Direct analysis of preprocessed TIC chromatograms by PLS-DA resulted to be a rather limited strategy to uncover yeast growth metabolite concentration changes under different temperatures. This limited strategy did not allow for the identification of the most important metabolites related to yeast culture growth under different temperatures, due to the strong overlapping of the chromatographic peaks associated for the large number of different coeluted metabolites. The novel MCR-ALS strategy presented here allowed the resolution of coeluted chromatographic peaks, the calculation of their corresponding peak areas and the resolution of their corresponding pure MS spectra.

In this work, a new workflow for metabolite identification in untargeted metabolomics is demonstrated. The resolved pure MS spectra together with the high mass accuracy offered by the LTQ-Orbitrap enabled the identification of the resolved chromatographic peaks. A total number of 65 metabolites out of the 91 total detected were successfully identified. Changes in MCR-ALS chromatographic peak areas of some of the metabolites in control and stressed yeast samples were used to detect possible variations of metabolite concentrations at the two different culture temperatures by means of PLS-DA-VIP scores analysis. Results revealed that the concentrations of 43 metabolites were significantly changed according to the yeast culture temperature (stressing factor). Further research is proposed to complete the biochemical interpretation of the effects of temperature on yeast metabolome and confirm possible biomarkers of these effects.

Preliminary analysis, indicate that some metabolites linked to cell growth were affected by temperature with a consistent pattern of temperature-driven metabolic adaptation, although changes observed are still giving an incomplete description of the temperature effects on yeast growing process. The proposed strategy can simplify considerably the biochemical LC–MS data interpretation and allow the uncovering of new targets for discovery (biomarkers).

Acknowledgments The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement No. 32073.

References

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics, 24*, 728–737.

Bajad, S. U., Lu, W., Kimball, E. H., Yuan, J., Peterson, C., & Rabinowitz, J. D. (2006). Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *Journal of Chromatography A, 1125*, 76–88.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173.

Beltran, A., Suarez, M., Rodríguez, M. A., Vinaixa, M., Samino, S., Arola, L., et al. (2012). Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. *Analytical Chemistry, 84*, 5838–5844.

Canelas, A. B., ten Pierick, A., Ras, C., Seifar, R. M., van Dam, J. C., van Gulik, W. M., et al. (2009). Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics. *Analytical Chemistry, 81*, 7379–7389.

Castrillo, J. I., & Oliver, S. G. (2006). Metabolomics and systems biology in *Saccharomyces cerevisiae*. In A. Brown (Ed.), *Fungal genomics* (pp. 3–18). Berlin, Heidelberg: Springer.

Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*, 103–112.

de Juan, A., Rutan, S. C., & Tauler, R. (2009). 2.19-Two-way data analysis: Multivariate curve resolution—iterative resolution methods. In D. B. Stephen, T. Romà, & W. Beata (Eds.), *Comprehensive chemometrics* (pp. 325–344). Oxford: Elsevier.

Eilers, P. H. C. (2004). Parametric time warping. *Analytical Chemistry, 76*, 404–411.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006). *Multi- and megavariate data analysis, part 1, basic principles and applications*. Umetrics, AB: Umetrics Academy.

Esbensen, K. H., & Geladi, P. (2009). 2.13-Principal component analysis: Concept, geometrical interpretation, mathematical background, algorithms, history, practice. In D. B. Stephen, T. Romà, & W. Beata (Eds.), *Comprehensive chemometrics* (pp. 211–226). Oxford: Elsevier.

Garcia, D. E., Baidoo, E. E., Benke, P. I., Pingitore, F., Tang, Y. J., Villa, S., et al. (2008). Separation and mass spectrometry in microbial metabolomics. *Current Opinion in Microbiology, 11*, 233–239.

Geladi, P., & Kowalski, B. R. (1986a). An example of 2-block predictive partial least-squares regression with simulated data. *Analytica Chimica Acta, 185*, 19–32.

Geladi, P., & Kowalski, B. R. (1986b). Partial least-squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17.

Glinski, M., & Weckwerth, W. (2006). The role of mass spectrometry in plant systems biology. *Mass Spectrometry Reviews, 25*, 173–214.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore: Johns Hopkins University Press.

Gonzalez, B., François, J., & Renaud, M. (1997). A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. *Yeast, 13*, 1347–1355.

Hohmann, S. (2002). Osmotic stress signaling and osmoadaptation in yeasts. *Microbiology and Molecular Biology Reviews, 66*, 300–372.

Højer-Pedersen, J., Smedsgaard, J., & Nielsen, J. (2008). The yeast metabolome addressed by electrospray ionization mass spectrometry: Initiation of a mass spectral library and its applications for metabolic footprinting by direct infusion mass spectrometry. *Metabolomics, 4*, 393–405.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*, 703–714.

Huang, N., Siegel, M. M., Kruppa, G. H., & Laukien, F. H. (1999). Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data. *Journal of the American Society for Mass Spectrometry, 10*, 1166–1173. http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator/.

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. *Chemometrics and Intelligent Laboratory Systems, 76*, 101–110.

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. (2012). YMDB: The yeast metabolome database. *Nucleic Acids Research, 40*, D815–D820. http://www.ymdb.ca/.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research, 40*, D109–D114. http://www.genome.jp/kegg/kegg2.html.

Lu, H., Liang, Y., Dunn, W. B., Shen, H., & Kell, D. B. (2008). Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. *TrAC, Trends in Analytical Chemistry, 27*, 215–227.

Mensonides, F. I. C., Hellingwerf, K. J., de Mattos, M. J. T., & Brul, S. (2013). Multiphasic adaptation of the transcriptome of *Saccharomyces cerevisiae* to heat stress. *Food Research International, 54*, 1103–1112.

Nicholson, J. K., & Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature, 455*, 1054–1056.

Nielsen, N.-P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*, 17–35.

Niessen, W. M. A. (1999). State-of-the-art in liquid chromatography–mass spectrometry. *Journal of Chromatography A, 856*, 179–197.

Parastar, H., & Akvan, N. (2014). Multivariate curve resolution based chromatographic peak alignment combined with parallel factor analysis to exploit second-order advantage in complex

chromatographic measurements. *Analytica Chimica Acta, 816*, 18–27.

Peré-Trepat, E., Lacorte, S., & Tauler, R. (2005). Solving liquid chromatography mass spectrometry coelution problems in the analysis of environmental samples by multivariate curve resolution. *Journal of Chromatography A, 1096*, 111–122.

Pérez, I. S. N., Culzoni, M. A. J., Siano, G. G., García, M. A. D. G., Goicoechea, H. C. C., & Galera, M. A. M. N. (2009). Detection of unintended stress effects based on a metabonomic study in tomato fruits after treatment with Carbofuran pesticide. Capabilities of MCR-ALS applied to LC-MS three-way data arrays. *Analytical Chemistry, 81*, 8335–8346.

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics, 11*, 395.

Rajalahti, T., Arneberg, R., Kroksveen, A. C., Berle, M., Myhr, K.-M., & Kvalheim, O. M. (2009). Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry, 81*, 2581–2590.

Savorani, F., Tomasi, G., & Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance, 202*, 190–202.

Seattle Proteome Center (SPC)—Proteomic Tools. (2014). *Institute for systems biology.* http://tools.proteomecenter.org/software.php.

Sherman, F., Christine, G., & Gerald, R. F. (2002). Getting started with yeast. In G. Christine & R. F. Gerald (Eds.), *Methods in enzymology* (pp. 3–41). Waltham, MA: Academic Press.

Siano, G. G., Pérez, I. S., García, M. D. G., Galera, M. M., & Goicoechea, H. C. (2011). 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, 85*, 264–275.

Skov, T., Berg, F. V. D., Tomasi, G., & Bro, R. (2006). Automated alignment of chromatographic data. *Journal of Chemometrics, 20*, 484–497.

Szymańska, E., Markuszewski, M. J., Vander Heyden, Y., & Kaliszan, R. (2009). Efficient recovery of electrophoretic profiles of nucleoside metabolites from urine samples by multivariate curve resolution. *Electrophoresis, 30*, 3573–3581.

Tauler, R. (1995). Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems, 30*, 133–146.

Tauler, R., & Barceló, D. (1993). Multivariate curve resolution applied to liquid chromatography—diode array detection. *TrAC, Trends in Analytical Chemistry, 12*, 319–327.

Tauler, R., Maeder, M., & de Juan, A. (2009). 2.24-Multiset data analysis: Extended multivariate curve resolution. In D. B. Stephen, T. Romà, & W. Beata (Eds.), *Comprehensive chemometrics* (pp. 473–505). Oxford: Elsevier.

Tauler, R., Smilde, A., & Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics, 9*, 31–58.

Theodoridis, G. A., Gika, H. G., Want, E. J., & Wilson, I. D. (2012). Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Analytica Chimica Acta, 711*, 7–16.

Tomasi, G., Berg, Fvd, & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241.

Villas-Bôas, S. G., Rasmussen, S., & Lane, G. A. (2005). Metabolomics or metabolite profiles? *Trends in Biotechnology, 23*, 385–386.

Walczak, B., van der Boagert, B., & Massart, D. L. (1996). Application of wavelet packet transform in pattern recognition of near-IR data. *Analytical Chemistry, 68*, 1742–1747.

Werf, M. J. V. D., Overkamp, K. M., Muilwijk, B., Coulier, L., & Hankemeier, T. (2007). Microbial metabolomics: Toward a platform with full metabolome coverage. *Analytical Biochemistry, 370*, 17–25.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.

Wold, S. (1995). PLS for multivariate linear modeling. In H. van de Waterbeemd (Ed.), *QSAR: Chemometric methods in molecular design, methods and principles in medicinal chemistry* (Vol. 2, pp. 195–218). Weinheim: Verlag Chemie.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*, 37–52.

Wold, S., Johansson, A., & Cochi, M. (Eds.). (1993). *PLS-partial least squares projections to latent structures*. Leiden: ESCOM Science Publishers.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.

Xu, Y.-J., Wang, C., Ho, W. E., & Ong, C. N. (2014). Recent developments and applications of metabolomics in microbiological investigations. *TrAC, Trends in Analytical Chemistry, 56*, 37–48.

LC-MS based metabolomics and chemometrics study of the toxic effects of copper on *Saccharomyces cerevisiae*

Mireia Farrés[a], Benjamí Piña[a], Romà Tauler[a]

[a] Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain.

Abstract

Copper containing fungicides are used to protect vineyards from fungal infections. Higher residues of copper on grapes at toxic concentrations are a potentially toxic and affect the microorganisms living in vineyards, such as *Saccharomyces cerevisiae*. In this study, the response of metabolic profiles of *S.cerevisiae* at different concentrations of copper sulphate (control, 1 mM, 3 mM and 6 mM) analysed by liquid chromatography coupled to mass spectrometry (LC-MS) is evaluated. Data sets were analysed by an untargeted metabolomics approach, based on the application of the multivariate curve resolution-alternating least squares method (MCR-ALS). Peak areas of the MCR-ALS resolved elution profiles in control and in Cu(II)-treated samples were compared using partial least squares regression (PLS), and the intracellular metabolites best contributing to samples discrimination were selected and identified. Fourteen metabolites showed significant concentration changes upon Cu(II) exposure, following a dose-response effect. The observed changes were consistent with the expected effects of Cu(II) toxicity, including oxidative stress and DNA damage. This research confirmed that LC-MS based metabolomics coupled to chemometrics is a powerful approach for discerning metabolomics changes in *S.cerevisiae* and for elucidating modes of toxicity of environmental stressors, including heavy metals like Cu(II).

1. Introduction

Copper (Cu(II)) containing fungicides have been used for more than one century in Europe on agricultural soils, such as vineyard soils. These fungicides have been used to protect crops from fungal infections such as downy mildew by *Plasmopara viticola* since the end of the 19th century. Copper is a potentially toxic metal which, at low concentrations, can act as an essential micronutrient for microbial growth. At toxic concentrations, Cu(II) interacts with cellular nucleic acids and enzyme active sites, although a principal initial site of Cu(II) action is considered to be at plasma membrane. Thus, exposure of fungi and yeasts to elevated Cu(II) concentrations can lead to a rapid decline in membrane integrity [1]. Total Cu(II) concentrations in such soils can affect toxicologically to microorganisms living in vineyards. Cooper-based agrochemicals should not be toxic, under normal circumstances, but intensive and long term use of these fungicides has increased Cu(II) concentrations in soils significantly [2]. The use of copper formulates in biological vineyards has caused high levels in copper residues on grapes causing in some causes slow or stuck fermentations [3]. Several studies have been published about Cu (II) toxicity microorganisms [4-7] and specifically the effects of copper on *Saccharomyces cerevisiae* (yeast) have been the focus of many of these studies [1, 2, 8-11], although we have not encountered metabolomics studies yet.

The development of new omic methodologies and technologies for environmental risk assessment may represent great opprotunities for the identification o emerging risks in the application of fungicides on vineyards. The recent availability of a range of omic technologies provide researchers enormous opportunities to uncover the effects of xenobiotics on many parmameters simultaneously [12]. Omic technologies are valuable tools to measure biochemical changes associated with mode of action at the level of DNA/RNA (transcriptomics), proteins (proteomics) and at the metabolome (metabolomics). They provide the means to identify biomarkers for dose response modelling. These omics repertories might not be sufficiently informative *per se*, since these data are highly multivariate in nature, therefore one must use advanced multivariate data anlaytical tecniques able to cope with the challenges inherent with these very complex analytical data sets. Chemometric methods offer multiple efficient and robust methods for modeling and analysis of complicated chemical/biological data tables, with the goal to produce more interpretable and reliable models capable of handling incomplete, noisy and collinear data structures [13].

Metabolomics enables the detection of possible alterations in the metaboloma of organisms as a result of their exposure to bioactive compounds. However, the develompment of robust metabolomics models for the study of the mode of action (MoA) of bioactive compounds is a complex procedure. The application of bioactive compounds at sublethal doses under experimental

conditions is an the preferable option for metabolomics thus enabling the detection of their primary effects on the metabolism of the biological system, while excluding undesirable secondary effects [14].

Untargeted analytical methods can detect hundreds of metabolites with no or with only a limited prior knowledge of the metabolite composition of samples. Liquid chromatography – mass spectrometry (LC-MS) - based approaches have been shown to be particularly useful for untargeted metabolomics [15]. Recently, several untargeted approaches have been proposed to analyse metabolomic profiles of *S.cerevisiae* using diferent analytical techniques coupled to chemometric evaluation [16-19]. In the present work, we propose and use an untargeted approach to study the metabolomic profiles of a laboratory *S.cerevisiae* strain (*BY4741*) exposed to different sublethal concentrations of copper sulfate ($CuSO_4$). The goal of this approach is to observe what are the main changes in the yeast metabolic profiles and to tentatively identify what metabolites had their concentrations changing more in relation to copper (Cu(II)) exposition. A similar chemometrics strategy to the one applied in our previous work [19] is proposed for the analysis of the LC-MS data and resolution of the metabolic profiles using the multivariate curve resolution-alternating least squares (MCR-ALS) method [20-23]. Determination of biomarkes was based on the use of Student's *t*-test and the aplication of variable selection methods: variable importance in the projection (VIP) [24] and selectivity ratio (SR) [25, 26] in the partial least squares (PLS) regression analysis [24, 27, 28].

2. Materials and Methods

2.1 *S.cerevisiae* culturing and exposure to sublethal concentrations of Cu(II)

Saccharomyces cerevisiae samples were exposed to copper (added as $CuSO_4$) at sublethal concentrations. *BY4741(MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) yeast strain cells were pre-cultured in non-selective YPD (yeast peptone dextrose) medium on an orbital shaker (150 rpm) at 30 ºC overnight to obtain the initial culture (pre-culture). Four flasks of 250 ml of YPD medium were inoculated with the yeast pre-culture to an optical density at 600 nm (OD_{600}) of 0.1. These cultures were then grown for 6h to an OD_{600} of 0.8. Volumes of 75 ml of the yeast culture samples were exposed to increasing concentrations of $CuSO_4$ of 1 mM, 3 mM and 6 mM to for 3.5 hours to an OD_{600} of 2.5, with three replicates at each concentration. Three additional replicates were prepared as control samples without any addition of Cu(II).

2.2 Quenching and *S.cerevisiae* metabolites extraction

Metabolism of yeast culture was rapidly inactivated cooling down the samples on ice. Once cooled down, cells from the late exponential growth phase were harvested by centrifugation (4000 rpm for 15 min at 4 ºC) discarding the supernatant and washed twice with phosphate buffered saline (PBS) to adjust their pH to 7.4. Final cell pellets were kept cold until the extraction.

Intracellular metabolites were extracted from the *S.cerevisiae* culture using the boiling ethanol protocol as described previously [29]. Metabolites extraction was performed into 15 mL Falcon tubes, adding 5 mL of solvent (75 % ethanol) to the cell pellet and further incubation of the suspension for 3 min. at 80 ºC. After cooling down the mixture on ice, sample volume was concentrated and dried by evaporation using nitrogen gas (N_2). The residue was re-suspended to a final volume of 0.4 mL with the LC mobile phase (75 % acetonitrile). Prior to pouring the final volume to a vial, the solution was filtered through 0.2 μm GHP membranes (GHP, Acrodisc Syringe Filters) to further ensure removal of any residual protein/debris before LC analysis.

2.3 LC-MS analysis

A Waters Acquity UPLC system (Waters) and a TSKgel Amide-80 5-μm, (250 x 2.0 mm) HILIC (Hydrophilic Interaction LIquid Chromatography) column purchased from Tosoh Bioscience were used. LC solvents were 0.5 mM ammonium acetate in 90 % acetonitrile at pH 5.5 (solvent A) and 2.5 mM ammonium acetate in 90 % acetonitrile in 60 % acetonitrile at pH 5.5 (solvent B). The gradient elution was as follows: t = 0, 25 % B; t = 8, 30 % B; t = 12, 60 % B; t = 17, 60 % B; t = 20, 25 % B; t = 27, 25 % B. Injection volume was 5 μL and flow rate was 0.15 mL/min.

A Waters LCT Premier orthogonal accelerated time of flight (TOF) mass spectrometer (Waters), operated in negative electrospray ionization (ESI) mode was used to acquire mass spectra profiles in full scan mode from 100 to 800 Da. The mass spectrometer was interfaced to computer workstation running MassLynx V 4.1 software for data acquisition and processing.

Since the aim was to separate highly polar molecules, such as the metabolites more abundant in yeast metabolome, the use of HILIC columns was specially adequate for their analysis. HILIC columns have grown popularity in recent years due to their high compatibility with mass spectrometry and to their detection capabilities improvement for polar metabolites [30]. In a previous work, this approach was already shown to be adequate for yeast growth metabolome

analysis stressed by increasing temperatures [19]. In this work, the same analytical methodology has been extended to the analysis of yeast metabolome stressed by Cu(II) treatment.

2.4 LC-MS data pretreatment

Full scan MS spectra of the different chromatographic runs were saved in raw mode and were then converted to cdf format by MassLynx V4.1 software and imported to MATLAB R20012b (Mathworks Inc. Natick, MA, USA) computational environment using mzcdfread.m and mzcdf2peak.m functions from the Bioinformatics Toolbox.

Due to the huge size of MS spectra and storage requirements, especially in full scan high *m/z* resolution LC-MS acquisition modes, different data analysis strategies can be proposed. One of them is data binning, which reduces storage and facilitates data analysis steps. However, in this case the resolution power of the raw measurements is lost and several steps are required for its recovery (see for instance strategies used in previous works [16, 19, 31, 32]. Another sounder strategy, is the one based on the use of the regions of interest (ROI), already proposed in some open source software packages for metabolomics (*XCMS* software)[33]. These strategies take advantage of the sparse nature of the raw MS data and consider only intensity data values higher than a preselected threshold value and having peak elution features profiles. They have been recently adapted to the MATLAB environment (see [34]) and they have been used in the present work. The implementation of ROI approaches requires the input of a signal-to-noise (SNR) threshold value, the mass accuracy of the mass spectrometer expressed in ppm, and the minimum number of retention times these signals were repeatedly obtained. In the present work, the values of these parameters were 250 (0.15% of maximum MS signal intensity), 0.05 (*m/z* resolution), and 10 retention times respectively. Each region of interest contained masses with significant intensity with no loss of the original spectral resolution. In this way, every MS spectra provided a data matrix with a number of rows equal to the number of measured retention times (ranging from 0 to 27 min.) in the chromatogram and a number of columns equal to the number of finally selected *m/z* ROI values considering all simultaneously analysed chromatographic runs. Data from 12 samples (3 controls, 3 exposed to 1 mM, 3 exposed to 3 mM and 3 exposed to 6 mM Cu(II) concentrations) were arranged in a single column-wise augmented data matrix, one at the top of each other with their common and uncommon *m/z* values from ROIs (see [34]). Finally, the augmented (from the 12 individual samples) data matrix to be analysed (see below Equation 2) had dimensions of 22662 x 1076.

2.5 LC-MS data analysis by MCR-ALS

As stated in previous work [19], due to the high number of highly overlapped peaks generated in LC-MS metabolomics analysis of yeast samples, the application of MCR-ALS is proposed and used to facilitate their resolution and identification. The goal of MCR-ALS analysis was to resolve directly the maximum number of individual elution profiles and pure mass spectral profiles of all possible metabolites extracted from the investigated yeast samples and to investigate the effect of Cu(II) exposure on them.

MCR-ALS is a powerful chemometrics method able to analyse multicomponent systems with strongly overlapping contributions from complex chemical systems, including chromatographic ones. The mathematical basis of the bilinear model used by MCR in the case of the particular case of the analysis of a single yeast sample (k) is shown in equation 1:

$$\mathbf{D}_k = \mathbf{C}_k\mathbf{S}^T + \mathbf{E}_k \quad \text{for} \quad k = 1, 2, \ldots, 12 \text{ samples} \qquad \textbf{Equation 1}$$

In this equation, the rows of the data matrices $\mathbf{D}_k$ (I x J) have the MS spectra at all retention times (i= 1,…I) in the chromatographic analysis of this yeast sample, and the columns have the corresponding elution profiles at all the measured mass spectra *m/z* channels (j=1,…J). $\mathbf{C}_k$ is the matrix of MCR-ALS resolved elution profiles in yeast sample k, and $\mathbf{S}^T$ is the matrix of their corresponding resolved pure mass spectra. $\mathbf{E}_k$ contains the unexplained variance related to background and noise contributions not modelled by $\mathbf{C}_k$ and $\mathbf{S}^T$.

MCR-ALS can be extended to the simultaneous analysis of multiple yeast samples analysed by LC-MS as it is shown in next equation 2:

$$\mathbf{D}_{aug} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \vdots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} \qquad \textbf{Equation 2}$$

or briefly:

$$\mathbf{D}_{aug} = \mathbf{C}_{aug}\mathbf{S}^T + \mathbf{E}_{aug} \qquad \textbf{Equation 3}$$

The augmented data matrix ($\mathbf{D}_{aug}$) has a number of 22662 rows, equal to the total number of recorded elution times in the simultaneous analysis of the different yeast samples and corresponding chromatographic runs (k = 1, 2, …, 12 samples, three control samples, and three 1mM Cu(II), three 3mM Cu(II) and three 6mM Cu(II) exposed samples). The number of columns

of $\mathbf{D}_{\text{aug}}$ is 1076, which is equal to the number of *m/z* ROI values finally considered. $\mathbf{C}_{\text{aug}}$ is the augmented matrix of the resolved elution profiles in the different chromatographic runs, $\mathbf{S}^{\text{T}}$ is the matrix of pure MS spectra of them and $\mathbf{E}_{\text{aug}}$ matrix is the noise and background signal not explained by the model [20]. One clear advantage of the MCR-ALS of column-wise augmented data matrices as shown in Equations 2 and 3 is that chromatographic peak alignment among chromatographic runs is not necessary because MCR-ALS allows for a complete freedom in the modelling of elution profiles [35] in the different runs.

Although the number of MCR-ALS components is usually estimated by singular value decomposition (SVD) [36], in this work an arbitrary sufficient large number of components was initially proposed. The criterion used to select this number is that it should explain a significant amount of variance and that it should give information about all possible detected chromatographic peaks in the system. In LC-MS data, due to the high selectivity of MS signals, chromatographic peak shape features are rather easily distinguished. Usually, apart from resolving chromatographic elution profiles of the extracted metabolites, other additional components describing solvent and background contributions were also considered.

Once the number of components and the initial estimates have been selected, the bilinear models described by Equations 2 and 3 were solved using a constrained ALS (Alternating Least Squares) iterative approach. In the present study, the applied constraints have been non-negativity for the elution profiles in, $\mathbf{C}_{\text{aug}}$, and for the MS spectra in $\mathbf{S}^{\text{T}}$ as well as their normalization. [20, 22].

MCR-ALS model quality was evaluated by the percent of explained variance (R^2) and the lack of fit (lof) values [37], see Equations 4 and 5 respectively:

$$\text{lof}\,\% = 100\sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} x_{ij}^2}} \quad e_{i,j} = d_{i,j} - \hat{d}_{i,j} \qquad \textbf{Equation 4}$$

$$R^2 = 100\sqrt{\frac{\sum_{i,j} x_{ij}^2 - \sum_{i,j} e_{ij}^2}{\sum_{i,j} x_{ij}^2}} \qquad \textbf{Equation 5}$$

In which $d_{i,j}$ is the experimental value from the data matrix for variable *j* and sample *i*; $\hat{d}_{i,j}$ is the corresponding value calculated using MCR-ALS (Equation 1).

2.6 Detecting metabolite concentration changes due to Cu(II) treatment

To discover what yeast intracellular metabolites concentrations changed because of the Cu(II) exposition, peak areas of the MCR-ALS resolved elution profiles ($\mathbf{C}_{\text{aug}}$) were statistically evaluated. For this purpose, a Student's *t*-test ($p < 0.05$) was applied to the areas of MCR-ALS resolved compounds of control samples compared to samples exposed to 6 mM of Cu(II).

Furthermore, differences between two groups of samples were also analysed by partial least squares – discriminant analysis (PLS-DA). PLS-DA [38] is a PLS regression method (see below) where the set of predictor variables, **X** (yeast metabolites), are correlated to a set of binary variables, **y** (controls and Cu(II) exposed), describing the categories of X. In this work, PLS-DA analysis was performed using the data matrix **X** of the peak areas of the MCR-ALS resolved components, and the **y** data-vector where control samples were categorized as class 0 and 6 mM exposed samples were categorized as class 1. Prior to PLS-DA model calculation, the peak areas were autoscaled to give equal relevance to their possible change due to the exposure to Cu(II). PLS-DA model was assessed using leave-one-out cross-validation due to the small number of samples. PLS-DA model is assessed with the sensitivity parameter which measures the proportion of samples which are correctly identified as exposed to Cu(II), and with the specificity parameter which measures the proportion of controls which are correctly identified as not being exposed to Cu(II).

PLSR is a general multivariate linear regression method [27, 28, 39] which finds the correlation patters between any set of independent predictor variables, grouped in a matrix **X**, and any dependent response vector response, **y**. PLSR estimates in a very efficient way the best linear combinations of the variables **X** which enable a good prediction of **y** variables. These new combinations of variables are called latent variables (LV). PLSR builds a model that maximizes the covariance between **X** and **y** variables with a minimum number of latent variables (LV). In this work, PLSR was also applied to correlate the **X** -matrix with the peak areas of the elution profiles (metabolites concentration) resolved by MCR-ALS and the **y**-vector of Cu(II) concentrations. Both, PLSR and PLS-DA methods have been described in detail elsewhere [28, 38, 39].

To further investigate the more influent variables (metabolites) in the PLSR and PLS-DA models, the variables importance in projection (VIP) method was used [24, 28]. VIP scores are defined as a weighted sum of squares of PLS weights which take into account the amount of explained **y** variance in each extracted LV. This method is frequently used as a parameter for variable selection [40, 41]. Since the average of the squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion for variable selection [42]. This is not a statistically justified limit, and it can be

shown that is very sensitive to the presence of non-relevant information pertaining to **X** [41, 43]. The selectivity Ratio (SR) is another variable selection method frequently used to detect the more important variables of a multivariate data set. The SR value for a variable in particular is defined as the ratio between its explained residual variances of the spectral variables on the target-projected component. In order to ascertain which variables have higher discriminatory abilities, a *F*-value statistic is calculated for every variable to check whether exceeds the critical value from a *F*-distribution assumption [25].

2.7 Tentative identification of metabolites whose concentration change due to Cu(II) exposure

Results from the MCR-ALS optimization are given as resolved eluted profiles in $\mathbf{C}_{\mathrm{aug}}$, and pure mass spectra, $\mathbf{S}^{\mathrm{T}}$ factor matrices (Equation 3). Changes in peak areas from the resolved elution profiles in Cu(II) exposed samples compared to control samples were used to investigate the possible effects of Cu(II) in yeast metabolism. A first evaluation is performed from fold changes using a *t*-test. Additionally, all peak areas resolved by MCR-ALS were autoscaled, and PLS-DA was then applied to identify the most important metabolites responsible for the discrimination of Cu(II) treated and controls samples according to VIP and SR scores. Final selection of metabolites more affected by Cu(II) treatment was done taking into account the three approaches: *t*-test, VIPs and SR.

Pure mass spectra of those components having their concentrations changing considerably by Cu(II) treatment were used for their putative metabolite identification. Since the ROI *m/z* selection procedure (see above) allowed not to lose *m/z* accuracy from raw measured data, metabolite identification was carried out directly from MCR-ALS resolved pure mass spectra in $\mathbf{S}^{\mathrm{T}}$. Since the observed signals in $\mathbf{S}^{\mathrm{T}}$ included the molecular ion and the corresponding charged molecular ion adducts of neutral metabolites, the masses of the most common negative ion adducts were also considered to the elemental composition calculator searching engine. The calculation was conducted 8 different times for each negative ion peak (i.e., allowing for [M-H], [M-2H], [M-3H], [M-H_2O-H], [M+Na-2H, [M+HAc-H], [2M-H], [2M+HAc-H], [3M-H], where HAc = acetic acid. These metabolites were searched in two on-line databases resources, the Yeast Metabolome Database (YMDB) [44] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [45, 46]. The final list of the tentatively identified metabolites was then used to investigate and interpret the most probable metabolic pathways and mechanisms affected by the addition of Cu(II) to the yeast culture samples.

3. Results and discussion

3.1 Yeast growth

Fig. 1 indicates the inhibitory effect of Cu(II) ions in the concentration range 1-6 mM. Concentrations of Cu(II) 6 mM resulted in a slight decreased in the late exponential growth phase (at 8-10h) of yeast. To a lesser extent, the slow growth of *S. cerevisiae* was also observed at 3 mM of Cu(II). Yeast growth inhibition increased with increasing Cu(II), but these differences had no greater significances since the Cu(II) concentrations were at sublethal levels.

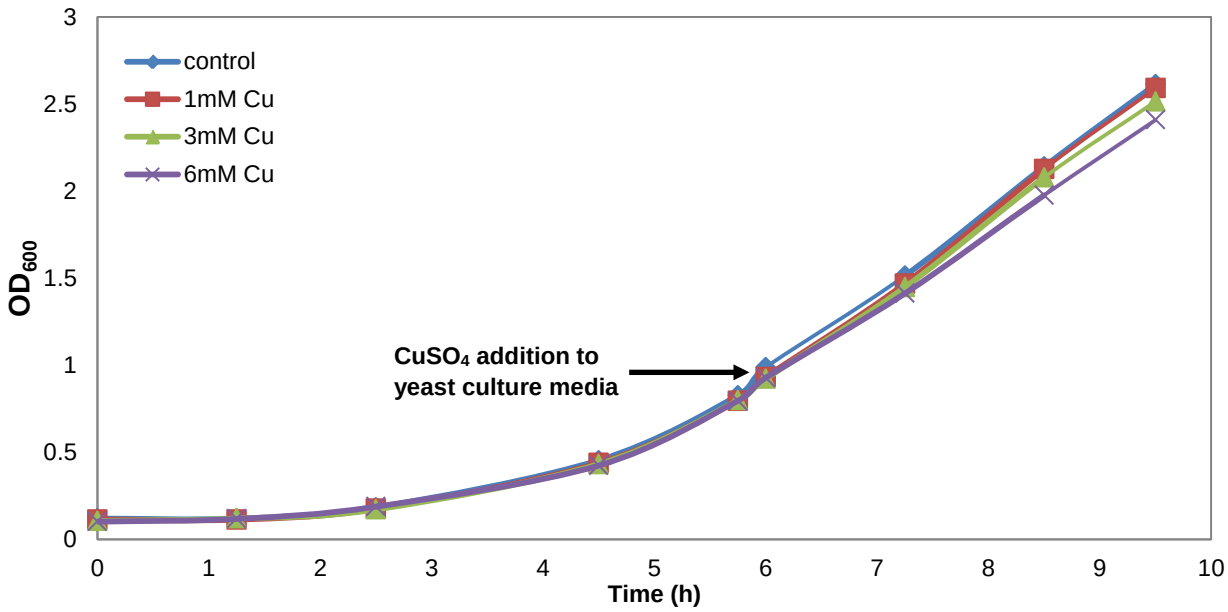


Figure 1 Effect of Cu(II) exposure at different concentrations on the growth of *Saccharomyces cerevisiae*. The growth of biomass was measured by optical density at 600 nm (OD_{600}). Cu(II) ions as $CuSO_4$ were added at the 6^{th} culture at an OD_{600} of 0.8. The results represent the mean of three individual cultivations with standard deviations lower than 5%.

3.2 Metabolomic and chemometric analysis

Differences in metabolic profiles between treated and control sample groups were investigated by PLS-DA. Prior to this PLS-DA, ROI data compression (section 2.4) and MCR-ALS resolution of coeluted peaks (section 2.5) were carried out. Reduction of *m/z* mode dimensions of original raw MS data matrices by ROI strategy and matrix augmentation of the different experiments (3 controls, 3 exposed to 1 mM, 3 exposed to 3 mM and 3 exposed to 6 mM Cu(II) concentrations), gave a data matrix with dimensions of 22662 x 1076 ($\mathbf{D}_{aug}$) using the procedure described in section 2.4. MCR-ALS was applied using a large number of components, such as 100, as to reflect the number of the eluted metabolites plus a number of extra components that explained background, solvent and other non-well defined chromatographic contributions (systematic noise sources). With this high number of components, MCR-ALS explained a total variance (R^2) higher than 99%, and a lack of fit (lof) lower than 9% (Equations 4 and 5). In Figure 2, an example of a small chromatographic region with three resolved components, out of the set of the 100 components included in the global MCR model is given. Elution profiles of these three components resolved for the individual chromatographic runs (above for control and treated samples) included in the considered column-wise augmented data matrix are shown, together with their corresponding pure spectra (left below) and resolution of three elution profiles (right below) for one of the replicates of the Cu(II) 3 mM treated sample.

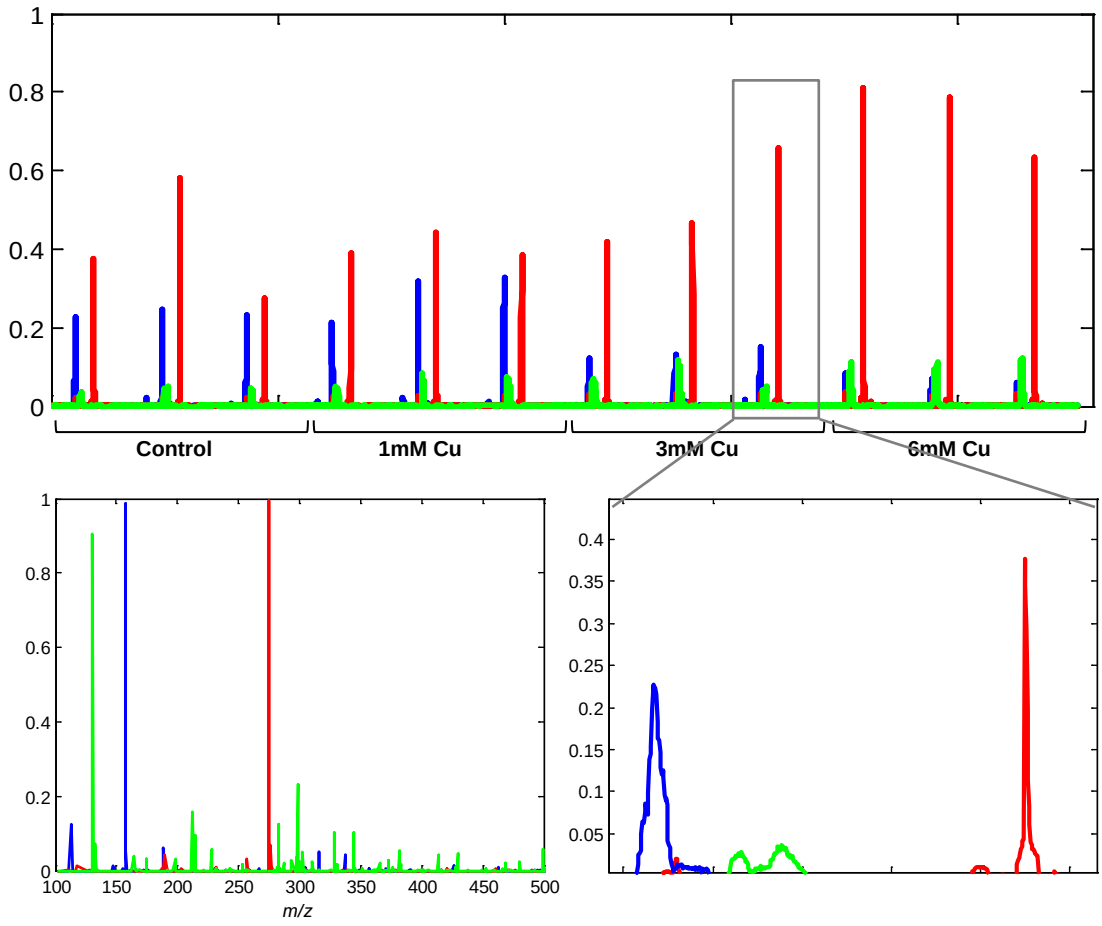


Figure 2 Example of MCR-ALS results of one one chromatographic region with three coeluted components (out of the whole set of 100 components considered in the global MCR-ALS analysis of this region) in the analysis of the column-wise data matrix (see Equation XX) which includes the simultaneous analysis of twelve different yeast samples (three control yeast samples, three 1 mM Cu(II), three 3 mM Cu(II) and three 6 mM Cu(II) exposed yeast samples. a) Resolved elution profiles for these three components in the individual chromatographic runs. b) Expansion of the resolution of these three components in one of the yeast samples exposed at 6 mM Cu(II), and c) pure resolved mass spectra.

As a result of MCR-ALS, the whole set of peak areas for all resolved elution profiles of the different metabolites in the different investigated samples were obtained and collected in a table. Effect of Cu to yeast control and exposed samples was then thoroughly studied by a statistical comparison of the calculated peak areas of all resolved components in these two types of samples. Firstly, a *t*-test was applied to the peak areas of control and 6 mM Cu(II) exposed samples.

Secondly, a PLS-DA analysis was performed to detect what were the peak areas discriminating between control and Cu(II) exposed samples (see procedure in the method section 2.6). A PLS-DA discrimination model with one latent variable (LV), already explained 41.05 % of the total **X** variance (peak areas of MCR-ALS resolved elution profiles) and 96.02 % of the total class variance (**y**), with specificity and sensitivity values equal to 1 for each class.

A new PLSR regression model was then tested to reveal whether metabolic concentration changes of *S. cerevisiae* followed Cu(II) dose application. This PLSR model was applied to the autoscaled areas of all 100 MCR-ALS resolved components according to the four treatment levels, controls (0 mM Cu(II)) and Cu(II) at 1 mM, 3 mM and 6 mM concentration levels). **X** data matrix had dimensions of 12 x 100, and the **y** -vector had the four concentration levels of Cu(II) exposure (0, 1, 3 and 6 mM), with 3 replicates at each level. Using leave-one-out cross-validation, the PLSR model with one-latent variable, already captured a large part of the **y** variance (around 94%). Figure 3 displays the resulting scores of this PLSR model; as it can be seen, the four sample groups (control and Cu(II) at 1mM, 2mM, and 6mM concentration) were separated according Cu(II) dose. These results indicated that the metabolic concentrations changed linearly according to exposure to increasing Cu(II) doses. Control and Cu(II) exposed samples to 1mM dose were closer, with the group of samples exposed to 3mM more separated from them. Samples exposed to 6mM Cu(II) give the group more isolated in Fig.2. These results are consistent with the growth curve shown in Fig.1 where *S.cerevisiae* growth was more inhibited at higher concentrations of Cu(II).

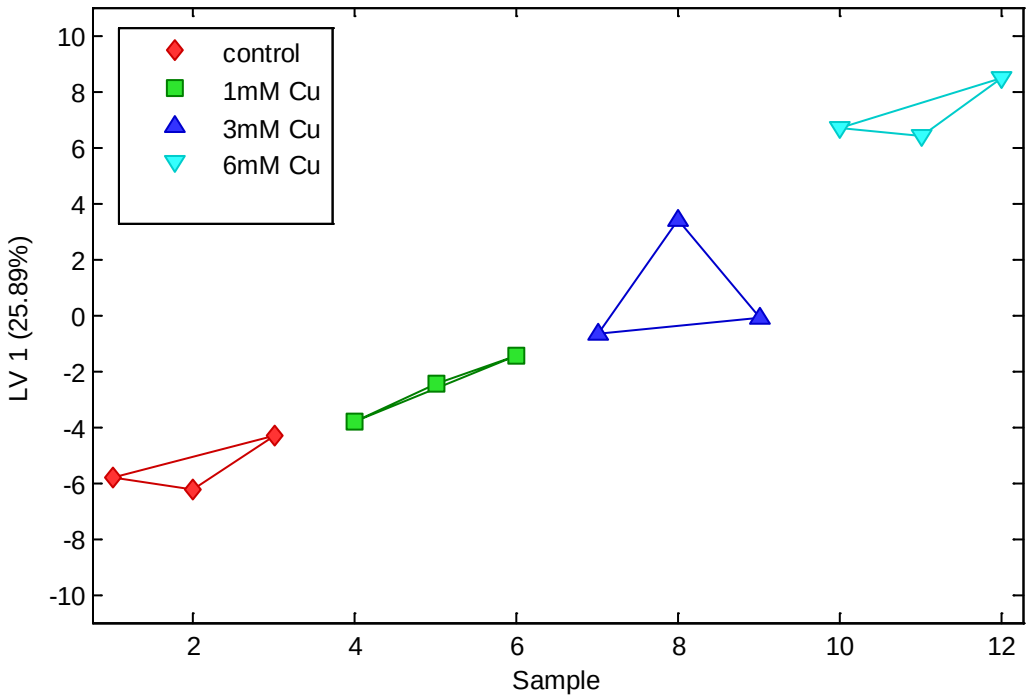


Figure 3. LV1 PLSR Scores plot in the analysis of yeast samples exposed to 0 mM (control), 1 mM, 3 mM and 6 mM of Cu(II).

3.3 Time-course of concentration changes in specific metabolites

Changes in metabolite concentrations (fold-change) are summarized along with their metabolite assignments in Table 1. VIP and SR values (see method section 2.6), calculated from PLS-DA model revealed what variables (metabolites) had the greatest influence on the discrimination between control and 6 mM Cu(II) exposed *S. cerevisiae* metabolic samples. 41 metabolites had VIP scores higher than one. When this threshold was increased to 2, 19 metabolites were selected. A total number of 14 metabolites were coincident using three different analyses: Student's *t*-test ($p < 0.05$), VIP (greater than 2) and SR (*F*-test, 95%). These metabolites were identified as described in section 2.7. Table 1 shows the tentative identification of these metabolites (compound name, molecular formula, adduct, error and KEGG number), their corresponding concentration fold change, and its concentration trend (up = increasing or down = decreasing). Among the different changes observed, glutathione (GSH) concentration was reduced to less than 1% of the levels in control samples in the highest dose group (6mM Cu(II)). Nicotinamide D- ribonucleotide and Nicotinate D-ribonucleoside concentration showed a remarkable increase with a fold-change of 26.5 and 11.6 respectively. Furthermore, concentrations of trehalose and L-Glutamic acid also increased with a fold-change higher than 6.

Table 1 Most relevant metabolites whose concentrations changed due to Cu(II) yeast culture exposition and their KEGG tentative identification.

Met. ID	Compound	Molecular formula	Adduct	Error (ppm)	KEGG C-number	Fold-change	Trend
1	Gultathione	$C_{10}H_{17}N_3O_6S$	$[M-H]^-$	20.48	C00051	105.2	DOWN
2	L-Glutamic acid	$C_5H_9NO_4$	$[M-H]^-$	0.22	C00025	5.1	UP
3	L-Dihydroorotic acid	$C_5H_6N_2O_4$	$[M-H]^-$	7.13	C00337	6.2	DOWN
4	L-Phenylalanine	$C_9H_{11}NO_2$	$[M-H]^-$	9.89	C00079	1.8	UP
5	2-Isopropylmaleic acid	$C_7H_{10}O_4$	$[3M-H]^-$	7.08	C02631	3.3	UP
6	Nicotinate D-ribonucleoside	$C_{11}H_{14}NO_6$	$[M+HAc-H]^-$	32.26	C05841	11.6	UP
7	Adenosine diphosphate ribose	$C_{15}H_{23}N_5O_{14}P_2$	$[M-H_2O-H]^-$	18.64	C00301	3.6	UP
8	L-Leucine	$C_6H_{13}NO_2$	$[M-H]^-$	10.16	C00123	3.1	UP
9	L-Isoleucine	$C_6H_{13}NO_2$	$[M-H]^-$	10.16	C00407	3.1	UP
10	L-Aspartic acid	$C_4H_7NO_4$	$[M-H]^-$	10.55	C00049	5.4	DOWN
11	Nicotinamide D-ribonucleotide	$C_{11}H_{15}N_2O_8P$	$[M-H]^-$	42.18	C00455	26.5	UP
12	Guanosine	$C_{10}H_{13}N_5O_5$	$[M-H]^-$	19.26	C00387	1.7	UP
13	L-Pipecolate	$C_6H_{11}NO_2$	$[M+HAc-H]^-$	10.21	C00408	2.8	UP
14	Trehalose	$C_{12}H_{22}O_{11}$	$[M+HAc-H]^-$	35.68	C01083	6.3	UP

3.4 Biological interpretation

The 14 metabolites whose concentrations varied between control and Cu(II) exposed samples (Table 1) felt into different functional categories according to KEGG (not mutually exclusive): glutathione metabolism (glutathione and L-glutamic acid), amino acid metabolism (L-phenylalanine, 2-Isopropylmaleic acid, L-leucine, L-isoleucine and L-aspartic acid), purine metabolism (guanosine and adenosine diphosphate ribose), nicotinate and nicotinamide metabolism (nicotinate D-ribonucleoside and nicotinamide D- ribonucleotide), starch and sugar metabolism (trehalose), pyrimidine metabolism (L-dihydroorotic acid) and lysine degradation (L-pipecolate).

Exposure to sublethal Cu(II) concentrations involved a complex response that led to yeast cell to acclimate to prevented cell death. Yeast responses to Cu(II) exposure promoted three major changes in the metabolism: defense against reactive oxygen species (ROS), cell protection and DNA repair.

Glutathione was clearly downregulated when yeast was exposed to high concentration of Cu(II) (6 mM). Since free Cu(II) ions are likely to participate in the formation of reactive oxygen species (ROS), Cu(II) and Cu(I) ions will participate actively in oxidation and reduction biochemical reactions. In the presence of reducing agents such as glutathione (GSH), Cu(II) can be reduced to Cu(I), which is then able to catalyze the formation of hydroxyl radicals (OH·) from hydrogen peroxide (H_2O_2) via the Haber-Weiss reaction [47, 48], causing the depletion of glutathione concentrations [49, 50], in agreement with results reported in Table 1. Large amounts of ROS will led to protein denaturation, membrane order alteration and damage of intracellular enzyme activity and consequent reduced metabolism [51]. In *S. cerevisiae*, exposure to low doses of H_2O_2 or oxidative stress by glutathione depletion induces apoptosis [52]. In our case, we observed a near total depletion of glutathione without significant effects on yeast growth. The proposed methodology allowed the detection of an early step of the oxidative stress process, and confirmed glutathione as the first line of defence against oxidants in the cell. The increase of concentration of L-Glutamic acid (see Table 1), a precursor of GSH, may reflect the activation of metabolic pathways leading to the restoration of physiological GSH levels.

Trehalose is capable of reducing oxidant-induced modifications of proteins during exposure of yeast cells to H_2O_2, and it is considered a general stabilizer of starving or stressed yeast cells [53]. Therefore, its increased concentration at higher doses of Cu(II) (see Table 1) may be explained by its capacity to scavenge free radicals and therefore protecting cellular constituents from oxidative damage [54].

The presence of reactive oxygen species (ROS) and of Cu(II) ions themselves in the presence of reducing agents such as glutathione may result in damage to DNA. ROS species and OH· radicals either oxidize bases, generate other radicals that result in crosslinks, or cleave the phosphoester bonds between specific nucleotides in the DNA [55]. DNA damage induces several cellular responses that enable the cell either to eliminate or cope with the damage. DNA ligase is an enzyme important for DNA repair and replication. The first step of DNA ligase-mediated DNA repair involves nucleophilic attack on the α phosphorous adenosine triphosphate (ATP) or nicotinamide adenine dinucleotide (NAD^+), resulting in the release of pyrophosphate or nicotinamide D-ribonucleotide (NMN) (see Table 1) and formation of a covalent intermediate (ligase-adenylate), in which adenosine monophosphate (AMP) is linked via phosphoamide bond to lysine [56, 57]. This or a similar mechanism may account for the increase in NMN, nicotinamide D-ribonucleoside (NAR) and adenosine diphosphate ribose (ADP-ribose) concentrations observed in Table 1 upon exposure to high Cu(II) concentrations. NAD+ is a co-enzyme of pivotal importance in the redox balance of metabolism, as it is continuously interconverted between an oxidized (NAD^+) and

reduced (NADH) state. The increases of NAR and ADP-ribose concentrations (see Table 1) may be explained by the fact that NAD^+ is synthetized from NAR [58] and consumption of NAD^+ involves release of nicotinamide and transfer of the remaining ADP-ribose moiety onto acceptor molecules [59].

4. Conclusions

In this study, a new LC-MS based metabolomics approach combined with chemometrics was used to explore Cu(II) toxicity in *S. cerevisiae* cultures . The exposure to sublethal concentrations of Cu(II) produced significant changes at the metabolic level, even at conditions where yeast growth was not significantly affected. Changes in MCR-ALS chromatographic peak areas provided a useful insight about potential metabolites that changed their concentration following the exposure to different Cu(II) doses. ROI data compression strategy allowed not losing any spectral resolution nor *m/z* accuracy from LC-MS raw data and enabled the tentative identification of the resolved chromatographic peaks. The intracellular metabolites best contributing to samples discrimination were selected by means of VIP-PLS and SR-PLS variable selection methodologies. Fourteen metabolites showed significant concentration changes upon Cu(II) exposure, following a dose-response effect. The observed metabolic changes were consistent with the expected effects of Cu(II) intoxication and with the physiological responses to its presence.

High concentrations of Cu(II) caused increased concentrations of ROS which led to reduced yeast metabolism and DNA damage. An early step of the oxidative stress process was detected, because there was a near total depletion of glutathione without significant effects on yeast growth. Glutathione was confirmed to act as a first line of defence against oxidants in the cell. The concentration increase of L-Glutamic acid reflected the activation of metabolic pathways leading to the restoration of physiological glutathione levels. Yeast cells counteracted the resulting oxidative stress by protecting their cellular constituents by increasing trehalose concentration at higher Cu(II) doses. A DNA repair mechanism activation was observed and explained by the increase of nicotinamide D- ribonucleotide, nicotinamide D-ribonucleoside and adenosine diphosphate ribose concentrations when the higher Cu(II) concentrations were used.

Acknowledgements

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

References

1. Avery, S.V., Howlett N.G., and Radice, S. (1996). *Copper toxicity towards Saccharomyces cerevisiae: dependence on plasma membrane fatty acid composition.* Applied and Environmental Microbiology *62* (11): p. 3960-3966.
2. Komárek, M., Čadková, E., Chrastný, V., Bordas, F. and Bollinger, J.C. (2010). *Contamination of vineyard soils with fungicides: A review of environmental and toxicological aspects.* Environment International *36* (1): p. 138-151.
3. Brandolini, V., Tedeschi, P., Capece, A., Maietti, A., Mazzotta, D., Salzano, G., Paparella, A. and Romano, P. (2002). *Saccharomyces cerevisiae wine strains differing in copper resistance exhibit different capability to reduce copper content in wine.* World Journal of Microbiology and Biotechnology *18* (6): p. 499-503.
4. Flemming, C.A. and J.T. Trevors (1989). *Copper toxicity and chemistry in the environment: a review.* Water, Air, and Soil Pollution *44* (1-2): p. 143-158.
5. Cervantes, C. and F. Gutierrez-Corona (1994). *Copper resistance mechanisms in bacteria and fungi.* FEMS Microbiology Reviews*14* (2): p. 121-137.
6. Giller, K.E., E. Witter, and S.P. McGrath (1998). *Toxicity of heavy metals to microorganisms and microbial processes in agricultural soils: a review.* Soil Biology and Biochemistry *30* (10–11): p. 1389-1414.
7. Dupont, C.L., G. Grass, and Rensing C. (2011) *Copper toxicity and the origin of bacterial resistance-new insights and applications.* Metallomics *2* (11): p. 1109-1118.
8. Sun, X.y., Zhao, Y., Liu, L.l., Jia, B., Zhao, F., Huang, W.d and Zhan, J.c. (2015). *Copper Tolerance and Biosorption of Saccharomyces cerevisiae during Alcoholic Fermentation.* PLoS ONE *10* (6)
9. Azenha, M., Vasconcelos, M.T., and Moradas-Ferreira, P. (2000). *The influence of Cu concentration on ethanolic fermentation by Saccharomyces cerevisiae.* Journal of Bioscience and Bioengineering *90* (2): p. 163-167.
10. Presta, A. and Stillman, M.J. (1997). *Incorporation of copper into the yeast saccharomyces cerevisiae. Identification of Cu(I)-metallothionein in intact yeast cells.* Journal of Inorganic Biochemistry *66* (4): p. 231-240.
11. Welch, J.W., Fogel, S., Cathala, G. and Karin, M (1983), *Industrial yeasts display tandem gene iteration at the CUP1 region.* Molecular and Cellular Biology *3* (8): p. 1353-1361.
12. dos Santos, S., Piculell, L., Medronho, B., Miguel, M.G. and Lindman, B. (2012). *Phase behavior and rheological properties of DNA-cationic polysaccharide mixtures.* Journal of Colloid and Interface Science *383* (1): p. 63-74.
13. Trygg, J., Holmes, E. and Lundstedt, T. (2006). *Chemometrics in Metabonomics.* Journal of Proteome Research *6* (2): p. 469-479.
14. Aliferis, K.A. and Jabaji, S (2011). *Metabolomics – A robust bioanalytical approach for the discovery of the modes-of-action of pesticides: A review.* Pesticide Biochemistry and Physiology, *100* (2): p. 105-117.
15. Dunn, W., Erban, A., Weber, R., Creek, D., Brown, M., Breitling, R., Hankerneier, T., Goodacre, R., Neumann, S., Kopka, J. and Viant M. (2013) *Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics.* Metabolomics, *9* (1): p. 44-66.
16. Ortiz-Villanueva, E., Jaumot, J., Benavente, F., Piña, B., Sanz-Nebot, V. and Tauler, R. (2015). *Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling.* Electrophoresis *36* (18): p. 2324-2335.

17. Puig-Castellví, F., Alfonso, I., Piña, B. and Tauler, R. (2015). *A quantitative 1H NMR approach for evaluating the metabolic response of Saccharomyces cerevisiae to mild heat stress.* Metabolomics 11 (6): p. 1612-1625.

18. Son, H.S., Hwang, G.S., Kim, K.M., Kim, E.Y., van den Berg, F., Park, W.M., Lee, C.H. and Hong, Y.S. (2009). *1H NMR-Based Metabolomic Approach for Understanding the Fermentation Behaviors of Wine Yeast Strains.* Analytical Chemistry *81* (3): p. 1137-1145.

19. Farrés, M., B. Piña, and R. Tauler (2015). *Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS.* Metabolomics, *11* (1): p. 210-224.

20. Tauler, R. (1995). *Multivariate curve resolution applied to second order data*. Chemometrics and Intelligent Laboratory Systems *30* (1): p. 133-146.

21. Tauler, R., Kowalski, B., and Fleming, S. (1993). *Multivariate curve resolution applied to spectral data from multiple runs of an industrial process.* Analytical Chemistry *65* (15): p. 2040-2047.

22. Tauler, R. and Barceló, D. (1993). *Multivariate curve resolution applied to liquid chromatography—diode array detection.* TrAC Trends in Analytical Chemistry *12* (8): p. 319-327.

23. Tauler, R., Smilde, A. and B. Kowalski (1995). *Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution.* Journal of Chemometrics *9* (1): p. 31-58.

24. Wold, S., Johansson, A. and Cochi, M (1993). *PLS-partial least squares projections to latent structures* 3D QSAR in Drug Design, Theory, Methods, and Applications, ed. H. Kubinyi. ESCOM Science Publishers: Leiden. 523-550.

25. Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.M., Ulvik, R.J. and Kvalheim, O.M (2009) *Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles.* Analytical Chemistry *81* (7): p. 2581-2590.

26. Rajalahti, T., Arneberg, R., Kroksveen, A.C., Berle, M., Myhr, K.M. and Kvalheim, O.M. (2009) *Biomarker discovery in mass spectral profiles by means of selectivity ratio plot.* Chemometrics and Intelligent Laboratory Systems *95* (1): p. 35-48.

27. Wold, H. (1966). *Estimation of Principal Components and Related Models by Iterative Least squares,* in *Multivariate Analysis.* Academic Press. p. 391-420.

28. Wold, S., Sjöström, M., and L. Eriksson, L. (2001). *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems *58* (2): p. 109-130.

29. Gonzalez, B., François, J. and Renaud, M. (1997). *A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol.* Yeast *13* (14): p. 1347-1355.

30. Lämmerhofer, M. (2010). *HILIC and mixed-mode chromatography: The rising stars in separation science.* Journal of Separation Science *33* (6-7): p. 679-680.

31. Gorrochategui, E., Casas, J., Porte, C., Lacorte, S. and Tauler, R. (2015). *Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells.* Analytica Chimica Acta *854*: p. 20-33.

32. Navarro-Reig, M., Jaumot, J., García-Reiriz, A. and 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.* Analytical and Bioanalytical Chemistry *407* (29): p. 8835-8847.

33. Tautenhahn, R., Bottcher, C. and Neumann, S. (2008). *Highly sensitive feature detection for high resolution LC/MS.* BMC Bioinformatics *9* (1): p. 504.

34. Gorrochategui, E., Jaumot, J. and Tauler, R. (2015) *A protocol for LC-MS metabolomic data processing using chemometric tools.* Protocol Exchange

35. de Juan, A. and Tauler, R. (2001). *Comparison of three-way resolution methods for non-trilinear chemical data sets.* Journal of Chemometrics *15* (10): p. 749-771.

36. Golub, G.H. and Loan, C.F.V. (1996) *Matrix Computations*. The Johns Hopkins University Press (Baltimore and London).

37. de Juan, A., Jaumot, J. and Tauler, R. (2014). *Multivariate Curve Resolution (MCR). Solving the mixture analysis problem.* Analytical Methods *6* (14): p. 4964-4976.

38. Barker, M. and Rayens, W. (2003). *Partial least squares for discrimination.* Journal of Chemometrics *17* (3): p. 166-173.

39. Geladi, P. and Kowalski, B.R. (1986). *Partial least-squares regression: a tutorial.* Analytica Chimica Acta *185*: p. 1-17.

40. Andersen, C. M. and Bro, R. (2010). *Variable selection in regression—a tutorial.* Journal of Chemometrics, *24* (11-12): p. 728-737.

41. Farrés, M., Platikanov, S., Tsakovski, S. and Tauler, R. (2015). *Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation.* Journal of Chemometrics, 2015: p. 528-536.

42. Chong, I.G. and Jun, C.H. (2005). *Performance of some variable selection methods when multicollinearity is present.* Chemometrics and Intelligent Laboratory Systems *78* (1–2): p. 103-112.

43. Tran, T.N., Afanador, N.L, Buydens, L. and Blanchet, L. (2014). *Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC).* Chemometrics and Intelligent Laboratory Systems *138*: p. 153-160.

44. Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A.C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelinkov, I., Wilson, M. and Wishart, D. (2012) *YMDB: the Yeast Metabolome Database.* Nucleic Acids Research, 2012. *40* (D1): p. D815-D820.

45. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Research *40* (D1): p. D109-D114.

46. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research *27* (1): p. 29-34.

47. Bremner, I., *Manifestations of copper excess* (1998). The American Journal of Clinical Nutrition *67* (5 Suppl): p. 1069S-1073S.

48. Gaetke, L.M. and C.K. Chow, C.K. (2003). *Copper toxicity, oxidative stress, and antioxidant nutrients.* Toxicology *189* (1–2): p. 147-163.

49. Estruch, F. (2000). *Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast.* FEMS Microbiology Reviews *24* (4): p. 469-86.

50. Izawa, S., Y. Inoue, Y. and Kimura, A. (1995). *Oxidative stress response in yeast: effect of glutathione on adaptation to hydrogen peroxide stress in Saccharomyces cerevisiae.* FEBS Letters *368* (1): p. 73-6.

51. Berger, F., Ramírez-Hernández, M.A.H. and Ziegler, M. (2004). *The new life of a centenarian: signalling functions of NAD(P).* Trends in Biochemical Sciences. *29* (3): p. 111-118.

52. Madeo, F., et al., *Oxygen Stress: A Regulator of Apoptosis in Yeast.* The Journal of Cell Biology, 1999. *145*(4): p. 757-767.

53. Benaroudj, N., Lee, D.H. and Goldberg, A.L. (2001). *Trehalose Accumulation during Cellular Stress Protects Cells and Cellular Proteins from Damage by Oxygen Radicals.* Journal of Biological Chemistry *276* (26): p. 24261-24267.

54. de Jesus Pereira, E., Panek, A.D. and Eleutherio, E.C.A. (2003). *Protection against oxidation during dehydration of yeast.* Cell Stress & Chaperones *8* (2): p. 120-124.

55. Linder, M.C. (2012) *The relationship of copper to DNA damage and damage prevention in humans.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *733* (1–2): p. 83-91.

56. Odell, M., Sriskanda, V., Shuman, S. and Nikolov, D.B. (2000). *Crystal Structure of Eukaryotic DNA Ligase–Adenylate Illuminates the Mechanism of Nick Sensing and Strand Joining.* Molecular Cell *6* (5): p. 1183-1193.

57. Sriskanda, V., Moyer, R.W. and Shuman, S. (2001). *NAD+-dependent DNA Ligase Encoded by a Eukaryotic Virus.* Journal of Biological Chemistry *276* (39): p. 36100-36109.

58. de Figueiredo, L.F., Gossmann, T.J, Ziegler, M. and Schuster S. (2011). *Pathway analysis of NAD+ metabolism.* Biochemical Journal *439* (2): p. 341-348.

59. Koch-Nolte, F., Fischer, S., Haag, F. and Ziegler, M. (2011). *Compartmentation of NAD+-dependent signalling.* FEBS Letters *585* (11): p. 1651-1656.

4.1.1 DISCUSSIÓ DELS RESULTATS OBTINGUTS

Canvis metabòlics produïts per l'exposició del llevat a canvis de temperatura (Article 2)

L'aplicació de PCA a la matriu de dades formada pels cromatogrames TIC (*Total Ion Current*) LC-MS de les 8 mostres de llevat (4 cultivades a 30ºC i 4 cultivades a 42ºC) i 2020 temps de retenció, va permetre obtenir tres components principals (*Principal Component*, PC) que explicaven el 88.83% de la variància de les dades. A la figura 2a de l'Article 2 es pot observar que el PC1 (47.76%) agrupava les mostres de llevat en relació a la temperatura de cultiu. La variació observada en els altres dos PCs corresponia a un altra tipus de variabilitat natural que no estava associada a la temperatura de cultiu. Per altra banda, l'anàlisi discriminant PLS-DA de la mateixa matriu de dades (**X**) en relació a la temperatura de cultiu ($\boldsymbol{y}$) va resultar en un model d'una sola variable latent que recollia el 47.32% de la variància de **X** i un 86.66% de la variància de $\boldsymbol{y}$. Aquesta variable latent separava les mostres segons la seva temperatura de cultiu, és a dir, discriminava les mostres control cultivades a 30ºC de les mostres estressades cultivades a 42ºC (vegeu figura 2b, Article 2). Per a la investigació dels principals metabòlits afectats pel canvi de temperatura en el cultiu del llevat es va inspeccionar primer els gràfic dels VIP (vegeu secció 2.5.5 de la introducció de la Tesi, capítol 2), però a causa de que existia una forta coelució entre els pics cromatogràfics de l'anàlisi LC-MS (vegeu figura 4.1), els cromatogrames TIC, els quals contenen la suma de les intensitats de les masses per cada temps de retenció, no van poder ser utilitzats de forma efectiva per a la selecció dels possibles metabòlits marcadors. Conseqüentment, es va realitzar l'anàlisi MCR-ALS de les matrius de dades obtingudes en mode d'anàlisi completa d'ions (*full-scan*) i aconseguir la resolució de tots els pics cromatogràfics estiguessin o no coeluits.

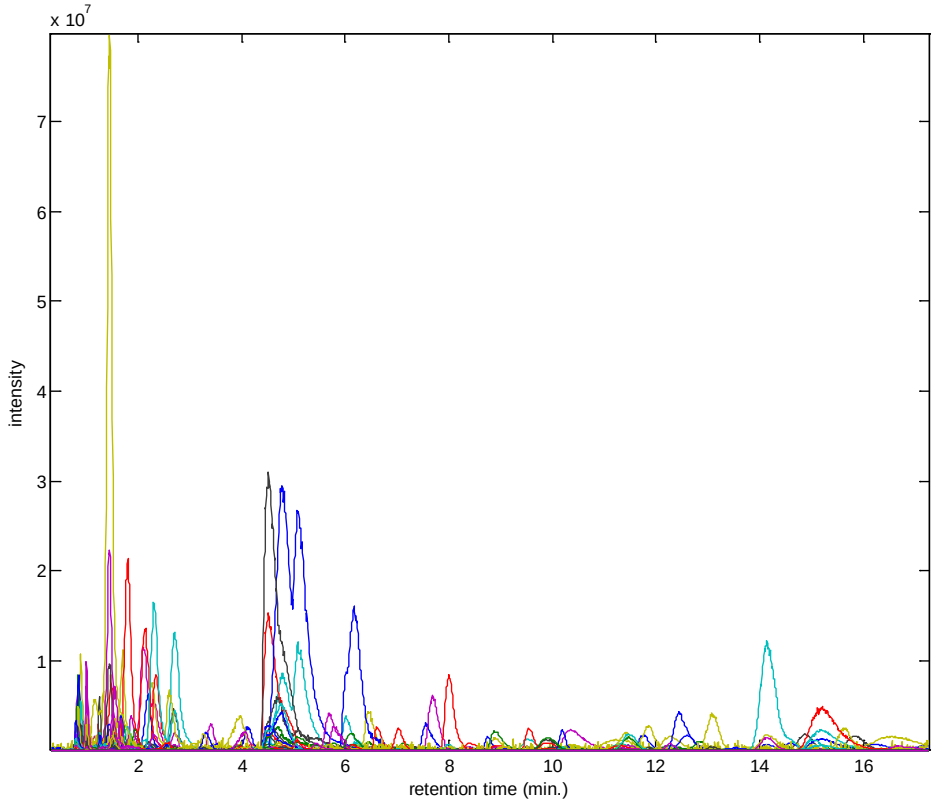


Figura 4.1. Exemple de cromatograma LC-MS original d'una mostra de llevat cultivada a 30 ºC.

Els cromatogrames originals es van dividir en nou finestres i per a cada finestra es va construir una matriu augmentada (vegeu secció 2.5.3 de la introducció de la Tesi, capítol 2). Es van obtenir un total de nou matrius de dades augmentades les dimensions de les quals eren: 1010 x 546, 1110 x 546, 2210 x 546, 2910 x 546, 2080 x 546, 2760 x 546, 3410 x 546, 3210 x 546 i 4010 x 546. En cada cas la primera dimensió correspon als temps d'elució dels cromatogrames (4 controls, 4 mostres estressades i 2 mostres d'estàndards), i la segona dimensió correspon al nombre dels valors *m/z*. El procediment d'anàlisi MCR-ALS per a cadascuna de les matrius augmentades es descriu a la secció 2.5.5 de la introducció de la Tesi (capítol 2). La variància explicada (R^2) en tots els casos va resultar major al 98% i el percentatge de la falta d'ajust (*lack of fit*, lof) menor al 11% (vegeu secció 2.5.5 de la introducció de la Tesi, capítol 2). Tot i això, no tots els components resolts per MCR-ALS van correspondre a veritables pics cromatogràfics (metabòlits), sinó que també es van resoldre altres contribucions a les senyals cromatogràfics, com el soroll de fons, la línia base o les contribucions del mateix dissolvent. Finalment, es van resoldre els perfils d'elució de 91 metabòlits (vegeu taula 2, Article 2) a partir de l'anàlisi simultània de les 8 mostres de llevat

analitzades (4 mostres control i 4 mostres estressades) i 2 mostres d'estàndards, i també es va resoldre els espectres de masses purs de cadascun d'ells.

L'anàlisi PLS-DA va permetre investigar la correlació existent entre la nova matriu de dades obtinguda a partir de les 91 àrees dels perfils resolts de les 8 mostres de llevat (**X** de dimensions 8 x 91) i la temperatura de cultiu de cada una de les mostres de llevat **y** (30ºC i 42ºC). El model amb una sola variable latent explicava el 98.02% de la variància de la variable temperatura, **y,** a partir del 44.99% de la variància de les variables corresponents a les àrees dels perfils d'elució dels metabòlits resolts en la matriu **X,** amb una especificitat (proporció de negatius que s'identifiquen correctament) i sensibilitat (fracció de positius reals que s'identifiquen correctament) de 1 per a cada classe. Aquest model PLS-DA agrupava clarament les mostres en relació a la temperatura de cultiu de les mostres de llevat (vegeu figura 2d, Article 2). Els valors dels VIP de l'anàlisi PLS-DA van permetre estimar quins eren els metabòlits les concentracions dels quals canviaven més segons la temperatura de cultiu. Aquests metabòlits es mostren a la taula 1 de l'Article 2.

A la taula 2 de l'Article 2 es mostren els 65 metabòlits que es van identificar temptativament del conjunt dels 91 pics cromatogràfics resolts mitjançant l'anàlisi per MCR-ALS. Tots els metabòlits identificats d'aquesta taula 2 són bé presents a les bases de dades YMDB (*Yeast Metabolome Database*) o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (vegeu secció 2.5.6 de la introducció de la Tesi, capítol 2) o bé es van identificar en estudis previs similars sobre el metaboloma del llevat (Beltran et al., 2012; Canelas et al., 2009). Els metabòlits identificats incloïen 17 dels 20 aminoàcids de les proteïnes, a més a més de L-ornitina. També es van identificar altres components essencials del metaboloma del llevat, com diversos nucleòsids i els seus derivats, vitamines, glicerol i alguns àcids orgànics i lípids (vegeu taula 2, Article 2).

Els metabòlits que s'indiquen a la taula 1 de l'Article 2, són els que mostraven un augment o disminució relativa de les seves abundàncies quan el llevat es va cultivar a 42ºC respecte el cultiu a temperatura estàndard (30 ºC). En total 22 metabòlits van incrementar la seva concentració relativa quan el llevat es van cultivar a 42 ºC, dels quals 16 es van identificar temptativament. Per altra banda, 21 metabòlits van disminuir la seva concentració relativa quan el llevat es va cultivar a 42 ºC (condicions estressants), dels quals 18 també van ser temptativament identificats (veure taula 1, Article 2).

L'anàlisi funcional dels metabòlits identificats no va permetre una avaluació exacte de tots els canvis metabòlics que es produeixen en l'aclimatació de les cèl·lules de llevat a les altes temperatures. No obstant, cal remarcar l'augment de concentració d'alguns dels metabòlits identificats temptativament, com per exemple el ribitol i l'arabitol, juntament amb la disminució de glicerol (vegeu taula 1, Article 2). Això es pot interpretar de la següent manera, el llevat manté el seu equilibri osmòtic mitjançant la modificació de la concentració interna de glicerols i/o polialcohols (Hohmann, 2002), però es creu que *S.cerevisiae* només utilitza el glicerol en funcions de protecció. Els canvis observats poden aleshores reflectir un re-equilibri de les concentracions d'osmòlits com a resposta a l'increment de la temperatura de cultiu del llevat. De la mateixa manera, la disminució l'àcid palmític i el lisofosfolípid identificat temptativament com a LysoPC(18:1(11Z)), podria reflectir una alteració en la composició dels lípids del llevat per adaptar la fluïdesa de la membrana a les temperatures més elevades del medi de cultiu del llevat. Finalment, es coneix que el creixement continu del llevat a altes temperatures indueix un canvi metabòlic des de la fermentació cap a la respiració (Mensonides et al., 2013). Conseqüentment, és concebible que alguns dels canvis metabòlics observats, com l'increment de concentració dels àcids orgànics adípic, caprílic i càpric de cadena curta/mitjana, o la disminució de concentració dels sub-productes de la fermentació com són l'acetat i el glicerol, reflecteixin aquesta adaptació metabòlica als canvis de temperatura.

Canvis metabòlics produïts per l'exposició del llevat a concentracions elevades de Cu(II) (Article 3)

El creixement de les diferents mostres de llevat (3 control, 3 mostres exposades a 1 mM Cu(II), 3 mostres exposades a 3 mM Cu(II) i 3 mostres exposades a 6 mM Cu(II)) mesurat a partir de l'absorbància dels cultius de llevat (OD_{600}) es mostra a la figura 1 de l'Article 3. L'addició del sulfat de coure ($CuSO_4$) va donar lloc a una lleugera disminució del valor de l'absorbància de les mostres de llevat exposades a 6 mM de Cu(II), i en menor mesura, de les mostres de llevat exposades a 3 mM de Cu(II). La reducció del creixement del llevat en la seva fase exponencial més tardana es va atribuir a l'exposició al Cu(II), i aquesta disminució es va accentuar lleugerament a concentracions més elevades de Cu(II).

Abans de l'anàlisi de les matrius de dades obtingudes per LC-MS de les mostres de llevat amb Cu(II) mitjançant MCR-ALS, es va procedir a la recerca de les regions cromatogràfiques on

existien senyals amb intensitats MS elevades (*Regions Of Interest*, ROI). Aquest procediment es descriu a la secció 2.5.5 de la introducció de la Tesi (capítol 2). L'estratègia ROI va permetre obtenir una matriu de dades de dimensions més reduïdes amb la mateixa resolució experimental. En total, l'anàlisi simultània de les 12 mostres de llevat prèviament indicades implicava el processament d'una matriu augmentada amb 22662 files, corresponents als temps d'elució dels cromatogrames, i 1076 columnes, corresponents al nombre de valors diferents *m/z* finalment seleccionats mitjançant aquesta estratègia ROI.

Per a seleccionar el nombre òptim de components en l'anàlisi MCR-ALS, es va fer una estimació inicial de pocs components i l'anàlisi es va repetir de forma consecutiva incrementant el número de components. Després de cada anàlisi MCR-ALS, la incorporació d'un nou component es va considerar apropiada només si resultava amb una disminució de la falta d'ajust (*lack of fit*, lof) i un increment de la variància explicada (R^2) (vegeu secció 2.5.5 de la introducció de la Tesi, capítol 2). Conseqüentment, si l'addició d'un component extra causava un valor superior o igual a la falta d'ajust i una menor variància explicada, aquest component es rebutjava automàticament. L'anàlisi MCR-ALS de la matriu augmentada de dades va resultar en 100 components, la majoria dels quals estaven relacionats amb els diferents metabòlits eluïts durant la cromatografia. Alguns dels components resolts servien per explicar el senyal de fons, la línia base, la contribució dels dissolvents i altres contribucions cromatogràfiques no definides. La variància explicada (R^2) va ser major al 99% i el percentatge de la falta d'ajust (*lack of fit*, lof) menor al 9%.

La selecció final dels metabòlits la concentració dels quals canviava en relació a l'exposició de Cu(II) es va fer tenint en compte els tres enfocs següents: un el test estadístic *t* de *Student* i els dos mètodes de selecció de variables VIP i SR. L'anàlisi estadística mitjançant els tres enfocs esmentats es va aplicar a les àrees dels components finalment resolts per MCR-ALS. Aquestes àrees es van agrupar en una matriu **X** de 6 files (3 files corresponents als tres controls de les mostres de llevat i 3 files corresponents a tres mostres tractades a 6 mM de Cu(II)) i 100 columnes corresponents als components resolts per MCR-ALS. El model PLS-DA es va aplicar a la matriu **X** (de dimensions 6 x 100) en relació a la concentració de coure (control i 6 mM de Cu(II), **y**). En aquest model una sola variable latent va explicar el 96.02% de la variància de **y** amb el 41.05 % de la variància de **X**, amb una especificitat (proporció de negatius que s'identifiquen correctament) i una sensibilitat (fracció de positius reals que s'identifiquen correctament) de 1 per a cada classe.

També es va aplicar un model de regressió PLSR per relacionar les àrees resoltes per MCR-ALS en totes les mostres analitzades (matriu **X** de dimensions 12 x 100) als quatre nivells d'exposició al Cu(II) considerats, **y** (0 mM, 1 mM, 3 mM i 6 mM). Una sola variable latent va explicar el 94% de la variància del vector de concentracions de Cu(II) (**y**) amb un 25.89% de la variància de **X** (vegeu figura 3, Article 3). A la mateixa figura 3 de l'Article 3, les mostres de llevat control i les exposades a 1 mM de Cu(II) estan més properes, en canvi les mostres de llevat exposades a 3 mM de Cu(II) estan lleugerament més allunyades d'aquest primer grup. Finalment, les mostres de llevat exposades a 6 mM de Cu(II) pertanyen al grup més aïllat de la resta de mostres. Aquest fet és congruent amb la corba de creixement de les mostres de llevat en presència de Cu(II) de la figura 1 de l'Article 3, en la qual s'observa com el creixement l'organisme *S.cerevisiae* va ser lleugerament més inhibit a concentracions majors d'aquest ió. Els resultats exposats va confirmar, per tant, que els perfils metabòlics de llevat canviaven linealment d'acord amb l'increment de les dosis d'exposició a Cu(II).

Un total de 14 metabòlits van ser detectats com a rellevants mitjançant el test *t* de *Student* ($p < 0.05$), els VIP (valor llindar de 2) i el SR (test *F*, 95%). La taula 1 de l'Article 3 mostra el resultat de la identificació temptativa dels 14 metabòlits, el canvi en la concentració (*fold-change*) i la tendència del canvi de concentració (creixent o decreixent).

Els metabòlits identificats pertanyen a diferents categories funcionals d'acord amb la base de dades KEGG (no mutualment excloents): el metabolisme del glutatió (glutatió i L-àcid glutàmic), el metabolisme dels aminoàcids (L-fenilalanina, àcid 2-isopropilmaleic, L-leucina, L-isoleucina i L-àcid aspàrtic), el metabolisme de les purines (guanosina i adenosina difosfat ribosa), el metabolisme del nicotinat i nicotinamida (nicotinat D-ribonucleòsid i nicotinamida D-ribonucleòtid), el metabolisme del midó i el sucre (trehalosa), el metabolisme de la pirimidina (L-àcid dihidrooròtic) i la degradació de la lisina (L-àcid pipecòlic) (vegeu taula 1, Article 3).

L'exposició dels cultius de llevat a concentracions subletals de Cu(II) va desencadenar una resposta complexa en el metabolisme del llevat per tal que aquest pogués aclimatar-se a les noves condicions de Cu(II) en el medi, i d'aquesta manera impedir la mort cel·lular. L'anàlisi del metaboloma va permetre investigar tres principals canvis en el metabolisme a conseqüència de l'adaptació de l'organisme a l'increment de concentracions de Cu(II) en el medi de cultiu:

1) Defensa del llevat contra les espècies reactives de l'oxigen (*Reactive Oxidative Species*, ROS). Els ions de Cu lliures són propensos a participar en la formació de ROS, els ions Cu(I) i Cu(II) participen activament en l'oxidació i reducció de reaccions bioquímiques. En presència d'agents superòxids o reductors tals com el glutatió (GSH), el Cu(II) es pot reduir a Cu(I) i aquest és capaç de catalitzar la formació de radicals hidroxil (OH·) a partir de peròxid d'hidrogen (H_2O_2) mitjançant la reacció Haber-Weiss (Bremner, 1998; Gaetke i Chow, 2003) de manera que causa l'esgotament del glutatió (Estruch, 2000; Izawa et al., 1995) d'acord amb els resultats observats a la taula 1 de l'Article 3. Grans quantitats de ROS en el medi provoquen la desnaturalització de proteïnes, alteracions en la membrana cel·lular i el dany a l'activitat enzimàtica intracel·lular amb la conseqüent reducció del metabolisme. L-àcid glutàmic és un metabòlit precursor del glutatió, i l'increment de la seva concentració (vegeu taula 1, Article 3) podria reflectir l'activació de les rutes metabòliques que condueixen a la restauració dels nivells fisiològics de glutatió en l'organisme *S.cerevisiae*.

2) Protecció cel·lular del llevat. La trehalosa és capaç de reduir les modificacions oxidants induïdes en les proteïnes durant l'exposició de les cèl·lules de llevat a H_2O_2 (Benaroudj et al., 2001). El metabòlit trehalosa es considera un estabilitzador quan les cèl·lules estan sotmeses a condicions estressants, conseqüentment, l'augment de la seva concentració a dosis més altes de Cu(II) (vegeu taula 1, Article 3) es podria explicar per la seva capacitat d'eliminar els radicals lliures i, així, protegir els constituents cel·lulars del dany oxidatiu (de Jesus Pereira et al., 2003).

3) Reparació de l'ADN del llevat. La presència de ROS i d'ions de coure en presència també d'agents reductors com el glutatió produeix dany a l'ADN. Les espècies ROS i els radicals OH· oxiden les bases de l'ADN, generen altres radicals que resulten en reticulacions, o trenquen enllaços fosfoèster entre els nucleòtids específics de l'ADN (Linder, 2012). Els dany a l'ADN indueix varies respostes cel·lulars que permeten eliminar-lo o reparar-lo. L'ADN lligasa és un enzim important per a la reparació i replicació de l'ADN i el primer pas per a la reparació implica un atac nucleofílic als metabòlits α fosfat adenosina trifosfat (ATP) o nicotamida adenina dinucleòtid (NAD^+), resultant en l'alliberament de pirofosfat o nicotinamida D-ribonucleotid (NMN) i la formació d'un intermediari covalent (lligasa-adenilat), en el qual l'AMP (adenosina monofosfat) es vincula a la lisina mitjançant l'enllaç fosfoamida (Odell et al., 2000; Sriskanda et al., 2001). Aquest mecanisme, o un de similar, podria explicar l'increment de concentracions de NMN, nicotinamida D-ribonuleosid (NAR) i adenosina difosfat ribosa (ADP-ribosa) en relació a

l'increment de la concentració Cu(II) en el medi de cultiu del llevat. NAD^+ és un coenzim fonamental en l'equilibri redox del metabolisme de llevat, ja que és contínuament interconvertit entre estats oxidat (NAD^+) i reduït (NADH). L'increment de les concentracions de NAR i ADP-ribosa s'expliquen pel fet que NAD+ es sintetitza a partir del metabòlit NAR (de Figueiredo et al., 2011) i el consum de NAD implica l'alliberament de nicotinamida i la transferència de la fracció restant d'ADP-ribosa a l'acceptor de molècules (Koch-Nolte et al., 2011).

Capítol 5

Anàlisi quimiomètrica de senyals paleoclimàtics a partir dels perfils GC-MS de compostos orgànics acumulats als sediments marins

5.1 INTRODUCCIÓ

En aquest capítol es proposa fer l'anàlisi quimiomètrica de dades climàtiques obtingudes a partir de l'anàlisi no dirigit per GC-MS de mostres de sediments marins. Existeixen diversos estudis anteriors on es mostra l'aplicació dels mètodes d'anàlisi multivariant de dades paleoclimàtiques (Brassell et al., 1986a; Conte i Eglinton, 1993; Poynter et al., 1989). En aquests estudis s'investiguen possibles marcadors climàtics en relació al canvi de temperatura superficial del mar (*Sea Surface Temperature*, SST). No obstant, en aquests casos les anàlisis es basen en un nombre limitat de compostos marcadors. En aquest capítol, l'estudi que es presenta és molt més general i no es limita a un nombre determinat de compostos, sinó que s'extén a tots els compostos que es detecten en els cromatogrames TIC (*Total Ion Current*) de l'anàlisi GC-MS de les mostres considerades. A diferència dels altres capítols d'aquesta Tesi, les dades analitzades en aquest capítol, no són canvis de concentració de metabòlits d'organismes biològics (blat o llevat), sinó dades de canvis de concentracions de compostos orgànics acumulats als sediments marins al llarg dels anys i que provenen d'organismes biològics. L'estratègia seguida per a l'anàlisi quimiomètrica emprada en aquests cas és molt similar a la que s'ha seguit en els estudis metabolòmics realitzats en els altres capítols d'aquesta Tesi (Article 1, Article 2 i Article 3). A la figura 5.1 es mostra l'esquema de treball seguit en aquest capítol i que manté un clar paral·lelisme amb el de la figura 2.1 de la introducció de la Tesi (capítol 2), la qual representa les etapes de treball dels estudis metabolòmics emprades en aquesta Tesi. De manera similar com s'ha fet en els Articles 2 i 3 del capítol 4, en l'article presentat en aquest capítol (Article 4) s'ha aplicat una estratègia d'anàlisi no dirigida de determinació dels canvis de concentració dels compostos orgànics acumulats en les mostres de sediments marins, i que responen als canvis de temperatura observats a principis i mitjans de l'època del Miocè de fins a 11.7 Ma (milions d'anys).

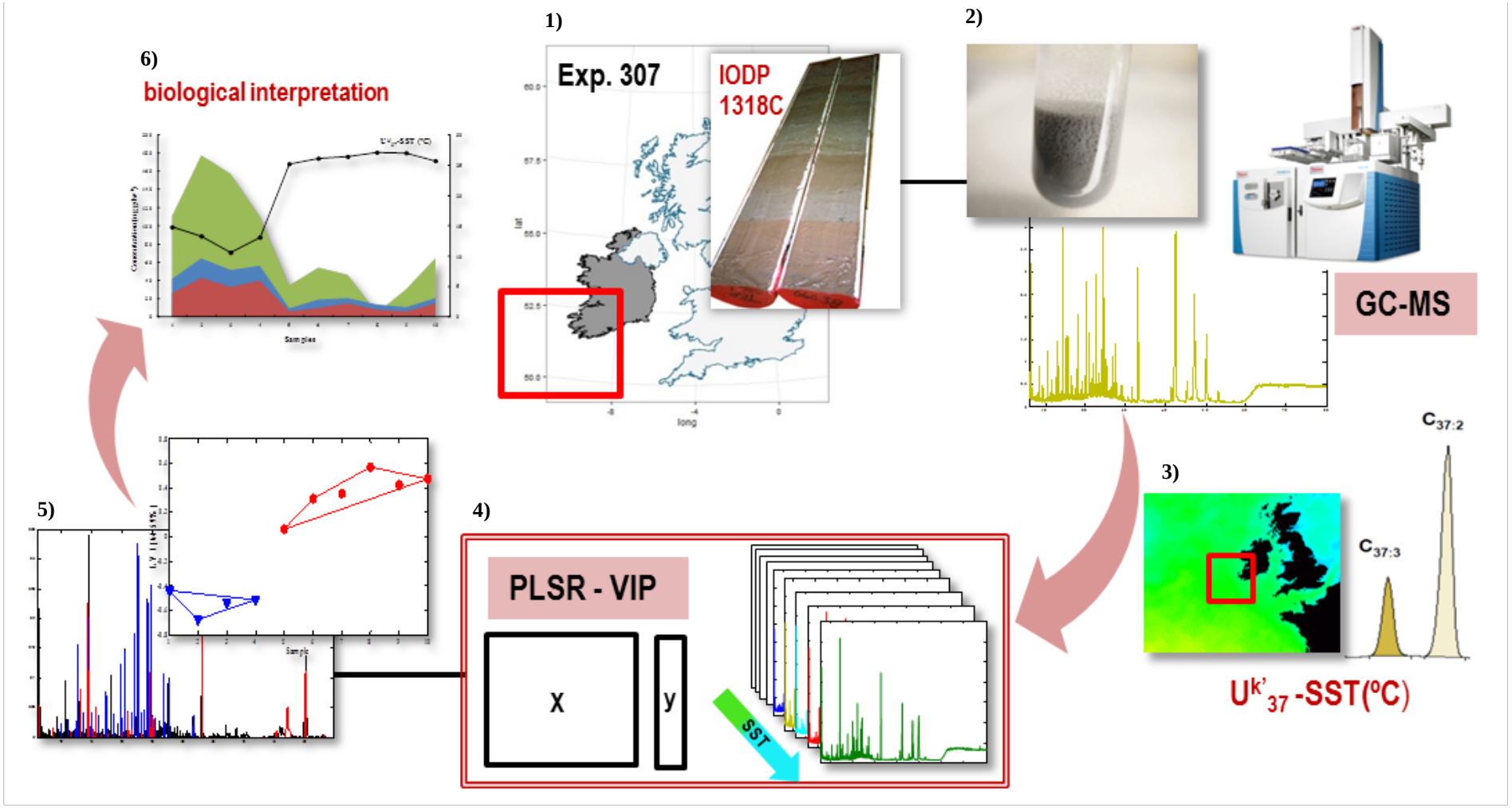


Figura 5.1 Esquema de treball de l'Article 4

L'objectiu de l'estudi realitzat en aquest capítol és determinar i identificar quins són els compostos orgànics acumulats en les mostres de sediments marins que covarien més intensament en relació als canvis de temperatura superficial de l'aigua de mar (SST) observats. Aquest objectiu és similar al que es proposa en els estudis de metabolòmica descrits en els capítols 3 i 4 d'aquesta Tesi.

La figura 5.1 resumeix les etapes de treball portades a terme en aquest capítol i en el treball corresponent, Article 4. Les mostres estudiades provenen del testimoni de sediment IODP-U1318C extret de la badia de Porcupine al sud oest d'Irlanda (vegeu l'etapa 1 de la figura 5.1) durant l'Expedició 307 del programa integrat de perforació oceànica (*Integrated Ocean Drilling Program*, IODP). Aquest és un testimoni de sediment que correspon a sediments que es van acumular des de principis i mitjans de l'època del Miocè durant 8.3 Ma (vegeu taula 1, Article 4). La determinació de la temperatura superficial de l'aigua (SST) associada a cada capa de sediment es va calcular aplicant l'índex $U^{k'}_{37}$, que és la relació del grau d'insaturació de les alquenones detectades a les diferents mostres (vegeu secció 2.4.3 de la introducció de la Tesi, capítol 2). D'aquest testimoni se'n van extraure deu mostres que van ser analitzades posteriorment.

L'etapa 2 de la de figura 5.1 correspon als procediments per a l'extracció dels compostos fòssils de mostres de sediments marins, els quals han estat adequadament descrits en treballs previs (Villanueva i Grimalt, 1996; Villanueva et al., 1997) i en el treball que es presenta en aquest capítol (Article 4). Les mostres de sediment es van assecar per congelació i els compostos orgànics es van extraure per sonificació mitjançant diclorometà (CH_2Cl_2). Els extractes van ser hidrolitzats amb hidròxid de potassi (KOH, 10%) en metanol (MeOH). L'extracció amb hexà (C_6H_{14}) va generar una fracció enriquida amb compostos neutres, que va ser netejada amb aigua ultrapura i derivatitzada amb bis(trimethylsilyl)trifluoroacetamide (BSTFA). La fracció àcida es va extraure amb una solució KOH/MeOH després de l'acidificació amb àcid clorhídric (HCl, 25%). L'emulsió es va rentar amb aigua ultrapura i la fase orgànica es va extraure amb hexà. La metilació dels àcids grassos lliures es va realitzar amb trifluorur de bor (BF_3).

Les estratègies tècniques i analítiques utilitzades per l'anàlisi de les fraccions neutra i àcida també estan representades a l'etapa 2 de la figura 5.1. L'anàlisi es va portar a terme mitjançant cromatografia de gasos amb detecció per espectrometria de masses (GC-MS) en mode d'anàlisi completa de ions.

Previ al tractament de dades generades per GC-MS, a l'etapa 3 de la figura 5.1, es pot observar que es va fer el càlcul de les SST a partir de la quantificació de pics cromatogràfics. Tal i com s'ha esmentat anteriorment, les SST associades a les diferents mostres de sediment es van calcular mitjançant l'índex d'insaturació d'alquenones $U^{k'}_{37}$ ($C_{37:2}/(C_{37:2} + C_{37:3})$). L'índex es va quantificar a partir dels cromatogrames TIC (*Total Ion Current*) de les fraccions saponificades ($U^{k'}_{37}$ = heptatriaconta-15E,22E-dien-2-one/(heptatriaconta-15E,22E-dien-2-one + heptatriaconta-8E,15E,22E-trien-2-one)). La temperatura anual mitjana de la superfície del mar (SST) es va calcular mitjançant el calibratge global que utilitza l'equació $U^{k'}_{37}$ = 0.33 x SST + 0.044 (r^2=0.96; N=370) (Müller et al., 1998). Les temperatures associades a cada una de les mostres es situen entre els 10.6 ºC i 27.1 ºC (vegeu taula 1, Article 4).

L'etapa de treball 4 de la figura 5.1 correspon a l'arranjament dels cromatogrames TIC en forma de matriu i l'anàlisi per PLSR en relació a les SST. En aquest estudi, la selecció de variables més importants (vegeu etapa 5 figura 5.1) es va portar a terme mitjançant el mètode de selecció de variables importants en la projecció (*Variable Importance in Projection*, VIP) descrit a la secció 2.5.5 de la introducció de la Tesi (capítol 2). Finalment, l'última etapa de treball de la figura 5.1 (etapa 6) correspon a la interpretació biològica dels resultats obtinguts a partir dels *scores* i VIP del PLSR.

Analytica Chimica Acta 879 (2015) 1–9

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7 Ma old (IODP-U1318)

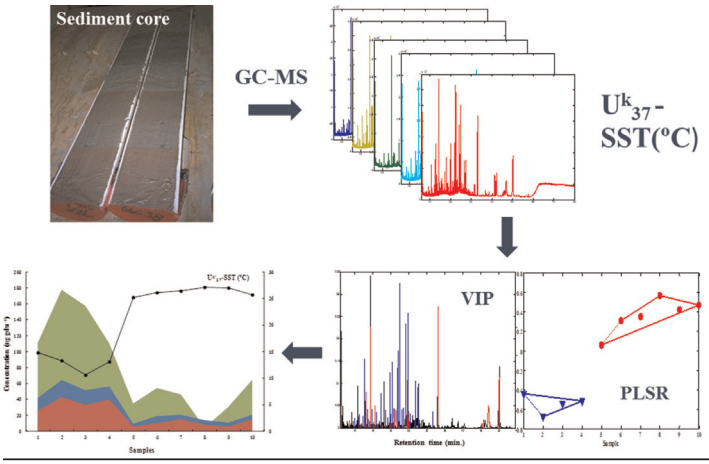
Mireia Farrés [a], Belen Martrat [a], Ben de Mol [b,1], Joan O. Grimalt [a], Romà Tauler [a,*]

[a] Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Jordi Girona 18, E-08034 Barcelona, Spain
[b] Barcelona Center for Subsurface Imaging (Barcelona-CSI-CSIC), Passeig Marítim de la Barceloneta 37, E-08003 Barcelona, Spain

HIGHLIGHTS

- Organic compounds from old marine sediments were analysed and identified by GC–MS.
- Total ion current chromatograms were correlated to sea surface temperatures by PLSR.
- Sea surface temperature proxies were selected using VIP PLSR scores.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 February 2015
Received in revised form 13 April 2015
Accepted 23 April 2015
Available online 25 April 2015

Keywords:
Sea surface temperature
Climate marker
Early-middle Miocene
Chromatographic profiles
Partial least squares regression
Variable importance in projection

ABSTRACT

This study focuses on the extraction of climate signals and processes using a combined approach which includes the analysis of a high number of lipid molecules in marine sediments, and the chemometric analysis of the acquired data. Neutral and acidic fractions of marine sediments from site IODP-U1318 (south-west of the UK, Porcupine Seabight) were quantified by GC–MS. The alkenone unsaturation index, $U^{k'}_{37}$, was estimated from the composition of C_{37} alkenones and it was then used for the estimation of sea surface temperatures (SST) for reference. Principal component analysis (PCA), explained 77.45% of the total data variance, and differentiated neutral fraction GC–MS total ion current (TIC) profiles according to SST values of the different sediment sections. GC–MS TIC chromatograms were correlated to sea surface temperatures (SST) by partial least squares regression (PLSR). The compounds more robustly in line with SST values at each sediment section explained 93% of the SST variance and they were identified using the variable importance in projection (VIP) scores method. The proposed approach enables an objective identification of organic compounds sensitive to SST variability throughout complete chromatographic profiles. As a result of this multivariate unbiased approach, lipid composition of sediments was differentiated between compounds of marine (long chain *n*-alkanes, long chain *n*-alkan-1-ols) and

* Corresponding author.
E-mail address: roma.tauler@idaea.csic.es (R. Tauler).
[1] Present address: Senergy AS Norway, C/O Aker Brygge Business Centre, P.O. Box 1433, 0115 Oslo, Norway.

http://dx.doi.org/10.1016/j.aca.2015.04.051

terrestrial (short chain *n*-alkan1-ols, alkenols, cholesterol, squalene) origin, whose concentrations were directly and inversely correlated to SST, respectively.

1. Introduction

Climate change is the result of complex interactions between many concurrent variables and parameters [1,2]. Marine sediments are interesting archives in which a wealth of information related with the past climate changes on earth is stored. They offer the opportunity to focus on a wide diversity of planet sites and therefore provide specific information on how climate changed in those sites

About 1000 fossil species can be identified in regular marine sections. These fossil remains may provide crucial information for climate change studies, e.g. $\delta^{18}O$ or $\delta^{13}C$ [3,4] or Mg/Ca in foraminifera [5]. However the same sediment section may contain about 1000 million of organic molecules which store an even huge amount of information on the past geochemical processes, including those related to climate change. Only a tiny fraction of this information is currently used [6].

To take advantage of the past climate knowledge using these stored compounds, the structural composition of these molecules should be known, their biological source must also be known and their changes related to climate conditions must be identified in the precursor organisms. Also, these compounds have to be relatively stable to diagenesis and the studied molecules have to be in sufficient concentrations in the samples to allow their determination. This last requirement involves that robust analytical methods affording the analysis of trace compounds in a large number of sediment samples must be available. Chromatographic methods are needed to fulfil these analytical requirements, and among these, gas chromatography (GC) coupled to mass spectrometry (GC–MS) is the instrumental technique affording the quantitative determination for larger numbers of molecules in one single analytical run. The use of this technique involves a restriction for the study of GC-amenable molecules in the paleoclimatic studies. However, even in these conditions the number of compounds resulting from the gas chromatograms is very large and multivariate data analysis is needed for handling the databases generated in the paleoceanographic studies [7].

No single proxy is free from ambiguity when describing climate parameters and their temporal changes. Hence, multiproxy strategies are recommended for precise and accurate reconstruction of the past. In this respect, lipids are of great utility in providing paleoclimatic information, specifically as a potential tool for estimating past sea surface temperatures (SST) [6,8,9]. Several geochemical SST proxies based on organic molecules are available. The most extensively used encompasses the study of the relative abundance of di- and tri-unsaturated long-chain ketones (alkenones) synthesized by coccolithophores [10,11]. Alkenones are ubiquitous biomarkers in the oceans, and the relative composition of the triunsaturated alkenone vs. the diunsaturated alkenone is linearly related to the sea surface water temperature (SST) over geological time-scales [10–13]. In this study, we have used GC profiles for correlating fossil compounds to alkenone-based reconstructed SST values in a sedimentary marine record recovered from the continental margin southwest of the UK (Porcupine Seabight). Handling and analysis of this multimolecular data set demand chemometric methods to study their occurrence in relation to SST changes.

Multivariate data analysis has proved useful for studying sources of heavy minerals in sediments from the Gulf of California and the Orinoco-Guayana Shelf [14], or for evaluating the origin of microfossils in samples from the Niger Delta region and the North Atlantic Ocean [15]. Multivariate data analysis is also helpful in paleoclimate studies, e.g. for the identification of correlations between molecular and isotopic composition and other parameters as descriptors of climatic processes [16], for the clustering and covariance analysis of biomarkers (*n*-alkanes, alkan-1,15-diols, alkenones, methyl esters, dinosterol) covary downhole in sediments [17], or for the determination of class correlation between SST and alkyl alkenoates or alkenones [18]. While these previous studies are based on the analysis of a rather limited number of compounds, in the present study the chemometric approach has been extended to the whole chromatographic profiles arranged in a single data matrix.

The present study focuses on the application of advanced multivariate data analysis methods for the untargeted investigation of concentration changes in fossil organic compounds accumulated in marine sediments up to 11.7 Ma. The aim of the study is to determine and identify what organic compounds in sediment marine samples covary strongly with observed SST changes over years. Principal component analysis (PCA; [19,20]) and partial least squares regression (PLSR; [21–23]) chemometric methods have been used to understand the relationships of the studied compounds with SST changes.

2. Material and methods

2.1. Study area

Sediments from site IODP-U1318C were retrieved during the Integrated Ocean Drilling Program (IODP) Expedition 307 at the Challenger Mound in the Porcupine Seabight (51°26.16′N, 11°33.0′W, 420.9 m below seafloor; [24]). The area is of interest due to its mounds, which are deep water coral banks. The majority of these coral banks is constructed by the framework builder *Lophelia petusa*,*Madrepora oculata* and associated fauna, with biodiversity comparable to tropical coral reef settings [25]. Cold water corals tolerate a wide range of environmental factors, such as SST from 4 to 12 °C and salinity from 32 to 36 psu (practical salinity units) [26]. The IODP-U1318C strata analysed are characterized by greenish-gray silty clay down to fine sand with carbonate contents from the early-middle Miocene, up to 11.7 Ma (tie points at 80.90 m, 1.357 Ma and 86.25 m, 10.216 Ma; Table 1) [27].

2.2. Methodology

The procedures and equipment for extracting, isolating and quantifying the fossil compounds have been described by [28,29]. Sample collection was limited to only two extract fractions (acidic and non-acidic compounds). Sediment samples were freeze-dried and extracted via sonication using CH_2Cl_2. The extracts were hydrolyzed with 10% KOH in MeOH. Extraction with hexane yielded a fraction enriched in neutral compounds, which was cleaned with ultrapure water and derivatized with bis(trimethylsilyl)trifluoroacetamide. The concentration of each compound in the neutral fraction was determined using *n*-hexatriacontane as internal standard. The acid fraction was extracted from the KOH/MeOH solution after acidification with HCl (25%). The emulsion was washed with ultra pure water to dissolve salts, and the organic phase extracted with hexane. Methylation of free fatty acids was performed with BF_3 in MeOH (10%) which was

Table 1
Reconstructed SST vs. core depth/age of samples from site IODP-U1318C, in the Porcupine Seabight (51°26.16′N, 11°33.0′W, 420.9 m below seafloor).

Sample	IODP code	Depth (m)[a]	Age (Ma)[b]	SST (°C)
1	1318 C 3 2 115-116	82.15	3.427	14.8
2	1318 C 3 3 070-071	83.30	5.331	13.3
3	1318 C 3 3 077-078	83.37	5.447	10.6
4	1318 C 3 3 147-148	83.97	6.441	13.1
5	1318 C 3 4 064-065	84.64	7.550	25.2
6	1318 C 3 4 126-127	85.26	8.577	26.1
7	1318 C 3 5 006-007	85.56	9.073	26.4
8	1318 C 3 5 035-036	85.85	9.554	27.1
9	1318 C 3 5 076-077	86.26	10.233	27
10	1318 C 3 6 015-016	87.15	11.706	25.7

[a] Miocene unconformity—erosion around 85.41 m (Kano et al. [27]).
[b] Tie points at 80.90 m, 1.357 Ma and 86.25 m, 10.216 Ma (Kano et al. [27]).

removed later with a saturated solution of NaCl in ultra pure water. The methyl esters were recovered with hexane. The concentration of each compound in the acid fraction was determined using nonadecanoic acid as internal standard. Each purified extract fraction was diluted in toluene and analyzed using GC–MS. He was the carrier gas. Oven temperature was programmed from 90 °C (held 1 min) to 170 °C at 20 °C/min, then to 280 °C at 6 °C/min (held 25 min), and finally to 315 °C (held 12 min.) at 10 °C/min. Mass spectra were acquired in the electron ionization mode (70 eV) scanning from *m*/*z* 42 to 700 with a cycle time of 1 s. Data were acquired and initially analyzed with Xcalibur software.

The alkenone unsaturation index ($U^{k'}_{37}$) was quantified from the total ion current chromatograms (TICs) of the saponified fractions [($U^{k'}_{37}$ = heptatriaconta-15E,22E-dien-2-one/heptatriaconta-15E,22E-dien-2-one + heptatriaconta-8E,15E,22E-trien-2-one), $C_{37:2}/(C_{37:2}+C_{37:3})$]. A global core top calibration, which uses the equation $U^{k'}_{37} = 0.033 \times SST + 0.044$ ($r^2 = 0.96$; $n = 370$; [30]) provided the annual mean SST for reference (Table 1).

2.3. Data analysis

2.3.1. Data arrangement and pre-processing

Total ion current (TIC) chromatograms were obtained by summing the signal intensities of all MS detected ions between *m*/*z* 42 and 700. GC–MS data were converted from raw to Xcalibur data NetCDF (.cdf) format using the Xcalibur file converter (version 2.0.7; Thermo Fisher Scientific Inc.). NetCDF files were imported to Matlab version 7.4 through NetCDF reader (version 1; http://www.mathworks.com/matlabcentral/fileexchange/15177-netcdf-reader) and further pre-processed using chemometric methods. A total number of 20 TIC chromatographic profiles (10 neutral sediment fractions and 10 acidic sediment fractions) was exported. This process involved the measurement of the MS intensity values at 2940 retention times, between 6 and 55 min. Other retention times were not considered because they did not contain any information of interest. Both the neutral and acidic data sets were interpolated to have the same retention times and rearranged in two different data matrices, each with 10 rows (equal to the number of samples) and 2940 columns (equal to the number of retention times), as shown in Fig. 1. Each sample chromatogram provided a set of chromatographic peaks corresponding to the different *m*/*z* ions, within the *m*/*z* range of 42 and 700 amu, from which TIC was calculated. Prior to integration of digitalized chromatograms in the database three pre-treatment methods were followed: (i) baseline correction, (ii) peak alignment and (iii) mean-centring (Fig. 1).

Baseline offsets were observed in GC–MS data, due to different effects such as column bleed, non-linearity in detectors and other instrumental artefacts. These baseline offsets may distort compound quantification by shifting upwards biased peak areas [31]. Chromatograms baseline correction was achieved by decreasing the contributions from noise and background signal. The asymmetric least squares (AsLS) method is a powerful system for chromatogram background baseline correction, and it was applied here as originally proposed [32,33]. According to the AsLS

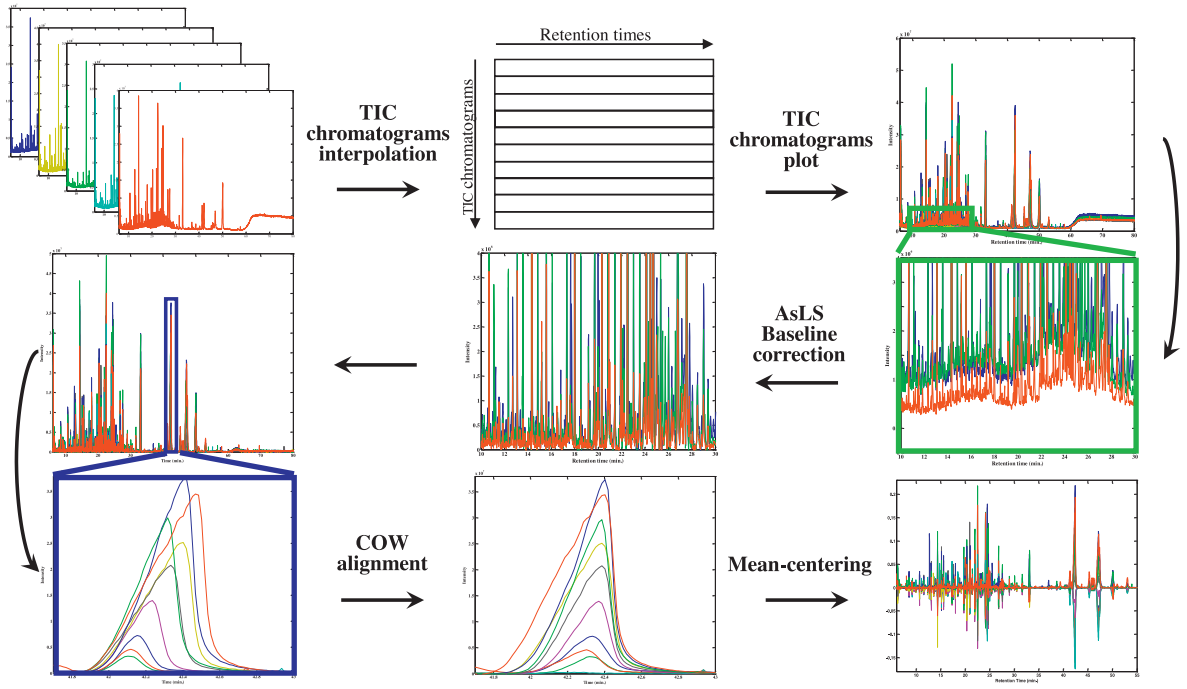


Fig. 1. GC–MS data pre-processing steps performed in the IODP-U1318C samples analyzed in this study.

algorithm, two parameters should be optimized, p for peak asymmetry and λ for smoothness, tuning them until offset was minimized (by visual inspection).

TIC chromatographic profiles should be aligned appropriately before chemometric analysis to compensate for possible drifts in retention times between different chromatographic runs. Misalignment between injections can be produced by changes in chromatographic columns during use (ageing), as well as pressure, temperature and flow rate fluctuations and interaction between analytes, all of which may contribute to changes in the retention time of a given compound [34–36]. Chemometric data analysis of TIC chromatograms benefits from adequate peak alignment pre-treatment procedures. In the present study, the correlation optimized warping (COW) algorithm [34,37] was selected for peak alignment. The method is useful for analysis of TIC chromatographic shifted data. COW requires the selection of the segment length m, the slack size t and a reference chromatogram. Here, we used a routine for the automatic parameter selection (m, t and reference chromatogram) [38] for the COW alignment. Finally, both chromatographic data matrices (neutral and acidic fractions) and the SST vector were mean-centred. This step is required prior to chemometric analysis in order to focus on data variance and reduce constant offset contributions. For mean-centring, the mean of each data column (retention time or variable) is subtracted from all values in that column. This enhances the differences among samples, reduces the number of principal components (PCs) and provides better prediction models.

2.3.2. Principal component analysis

PCA method [19,20] is closely related to the empirical orthogonal functions (EOF) method, very commonly used in climate studies [39]. It is one of most used multivariate exploratory data analysis method in chemometrics. It reduces the number of variables to a smaller number of fundamental orthogonal components which are linear combinations of the original variables. Each of these principal components are obtained in order to describe the maximum data variance. Here, the two pre-processed TIC data matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$ for the neutral and acidic extracted sample fractions, respectively, were analysed by PCA. Separately, each of these two matrices comprised 10 samples measured at 2940 chromatographic retention times (TIC chromatograms). In PCA the equation $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$ is solved, where $\mathbf{X}$ is either of the two pre-processed chromatographic data matrices ($\mathbf{X}_1$ or $\mathbf{X}_2$), $\mathbf{T}$ is the score matrix, $\mathbf{P}$ is the loading matrix and $\mathbf{E}$ is the residuals matrix. The $\mathbf{T}$ columns are orthogonal and map the main sample patterns in the principal components. The rows of $\mathbf{P}^T$ are orthonormal and map the main chromatographic patterns (variables) in the principal components. PCA was used to explore differences between samples and relate them to their associated SST values (either in the neutral, $\mathbf{X}_1$, or in the acidic fractions, $\mathbf{X}_2$).

2.3.3. Partial least squares regression

PLSR [21] is a multivariate linear regression method used to find correlation models between predictor variables ($\mathbf{X}$ data matrix) and the response (usually arranged in a $\mathbf{y}$ vector) measured on the same set of samples. In this study, $\mathbf{X}$ showed the TIC chromatograms and $\mathbf{y}$ lists the SST values associated with them. PLSR is particularly useful in cases of complex databases such as the IODP-U1318C chromatograms (Section 2.3.1), where the $\mathbf{X}$ matrix contains a large number of correlated variables. In these cases, PLSR provides a reliable estimate of the best linear combinations of the independent original $\mathbf{X}$ values (named as latent variables, LVs), optimally correlated with the changes observed in the dependent

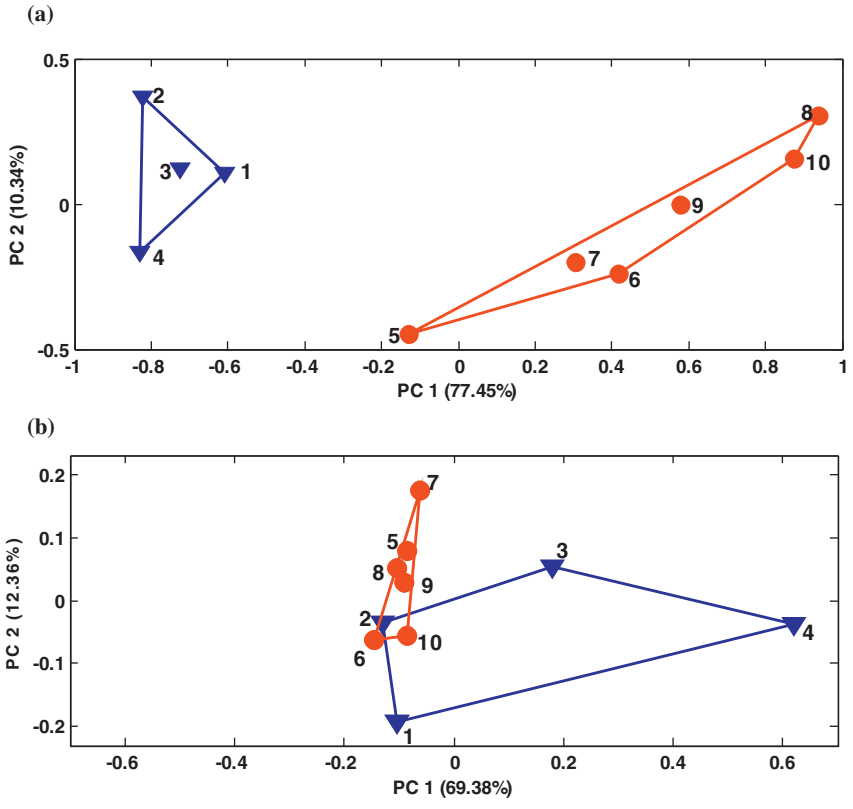


Fig. 2. PCA scores plot for (a) neutral and (b) acidic fraction samples. Blue triangles are samples associated to low SST (<15 °C) values and red solid circles are samples associated to high SST (>25 °C) values. Convex hulls are drawn around each SST group with the same color as the corresponding symbols. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variable, **y**. PLSR builds a model that maximises the covariance between **X** and **y** with a minimum number of latent variables. For every latent variable, a vector of weight coefficients is obtained and shows which **X**-variables combine best with **y**, SST values. The PLSR method is described in detail elsewhere [22,23].

The variable importance in projection (VIP) method [23,40,41] was used for the identification of the compounds associated with chromatographic peaks in the **X** matrix that could be considered potential climate markers. The PLSR–VIP method is used to measure the contribution of each predictor variable to the model. Prior to PLSR analysis, the alkenone chromatographic peaks used for SST estimation (i.e. $C_{37:2}$ and $C_{37:3}$) were deleted from the **X** variables data set. In summary, PLSR was used in this work to build a linear model which optimally correlated TIC chromatographic profiles of the fossil organic compounds in IODP-U1318C sediment samples with SST values associated to the samples. Some of the predictor (**X**) variables, i.e. chromatographic peaks with the highest weight coefficients and VIP scores, were selected as good indicators of the changes in SST (**y**) values. Correlation direction, positive or negative sign, between the abundance of these compounds and SST values can then be assessed considering the signs of weight PLSR coefficients.

3. Results and discussion

3.1. Data pre-processing

AsLS baseline correction method applied to TIC chromatographic experimental data (10×2940 dimensions) improved the content and appearance of all peak signals (smoothness of baselines) using optimal values for λ and p of 10^5 and 10^{-3} respectively (see Section 2.3.1 for the meaning of these parameters). COW alignment was then applied to the previously baseline corrected chromatograms and yielded optimized values for the length m and the slack size t of 35 and 14, respectively (see Section 2.3.1 for the meaning of these parameters). Both COW correction and mean-centring (see Fig. 1) enhanced the chromatographic signal content and reduced significantly undesired chromatographic variations not caused by chemical changes.

3.2. PCA of GC–MS TIC chromatograms

PCA scores plot of the neutral fraction data matrix is shown in Fig. 2a. PC1 accounted for 77.45% of the total chromatographic data variance. Two groups of samples clustered along this PC1. The first, with negative PC1 scores, was associated to low SST (<15 °C; Table 1). The second, with positive PC1 scores, was linked to high SSTs (>25 °C; Table 1). Hence, this PC1 separated the samples analysed in two paleo-SST groups. The most recent ones, between 3.4 and 6.5 Ma, are those with lower SST as values displayed by PCA.

These results highlight that a significant amount of observed concentration changes of compounds present in the sediments were related to SST. In contrast, PCA applied to the acidic sample fractions did not reveal any systematic pattern of samples distribution according to SST (Fig. 2b). The samples associated with high SST values were having the same PC1 scores as samples 1 and 2 which were at low SST values. PCA scores from samples at low SST were very weakly clustered along the positive side of PC1 axis. Since PCA can only allow a qualitative visualization of the more important variance sources, PLSR was further used to provide a more quantitative evaluation and interpretation of the obtained results.

3.3. PLSR of GC–MS TIC chromatograms and SST values

3.3.1. Neutral fraction PLSR results

PLSR was applied to the 10 TIC chromatograms from neutral fraction samples and SST values (**X** and ***y*** variables, respectively). Using leave-one-out cross-validation, the first PLS latent variable (LV1), accounted for 64.59% of the **X** data variance and for as much as 93.01% of the dependent variable ***y*** (SST) variance (Fig. 3). PLSR results confirmed therefore the presence of a strong correlation pattern between chromatographic peak areas of the neutral compounds and SST changes in the different sediment sections.

In the PLSR **T** scores plot of the neutral sample fractions (Fig. 3) separation of the samples increased compared to previous PCA. The 'greater than one rule' VIP scores is used as a criterion for initial variable selection [42]. Fig. 4 shows the retention times of possible SST markers associated with the largest VIP scores which are differentiated as follows: (i) chromatographic peaks shown in red are those directly (positively) correlated with SST; and (ii) chromatographic peaks in blue are those inversely (negatively) correlated with SST; (iii) black chromatographic peaks are those not found to be either correlated to be relevant in PLSR analysis.

Table 2 includes the relevant chromatographic peaks for the PLSR analysis. Once retention times of selected chromatographic peaks were recognized in the original chromatogram, the corresponding measured mass spectra were retrieved from the original experimental data and identified by comparison to library data and synthetic standards. The correlation coefficients of the concentrations of these compounds in the samples and the SST data are also shown in Table 2. Most of the VIP chromatographic peaks gave a highly significant p value and only a minor portion of the chromatographic peaks had a correlation coefficient <0.6. However, these were also considered because their abundance did indeed change in relation to SST.

From VIP score values, the relative abundance of cholesterol and other sterols was found to be relatively high when the SSTs values were also high (>25 °C). Sterols in general are quite stable, and cholesterol in particular is synthesized by most phytoplankton species. In deep sea waters, crustacean zooplankton produce a significant amount of C_{27} derived steroid compounds (>90%) through modification of other types of sterols [43,44]. Squalene also strongly correlated with SST (Table 2). This compound and other cyclic (tri) terpenoids are precursors of sterols. High

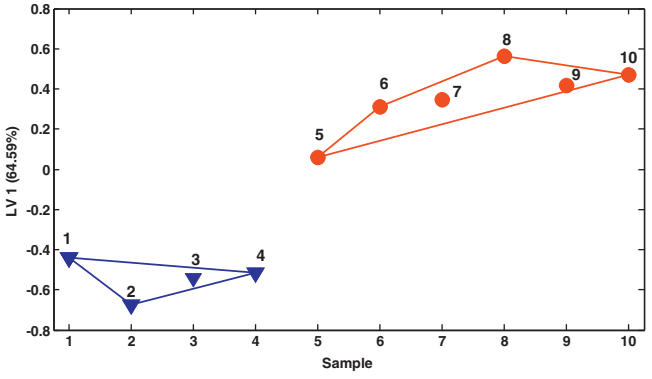


Fig. 3. Scores plot for LV1 in the PLSR analysis of neutral fraction samples. Blue triangles and lines are samples associated to low SST (<15 °C) values and red solid circles and lines are samples associated to high SST (>25 °C) values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

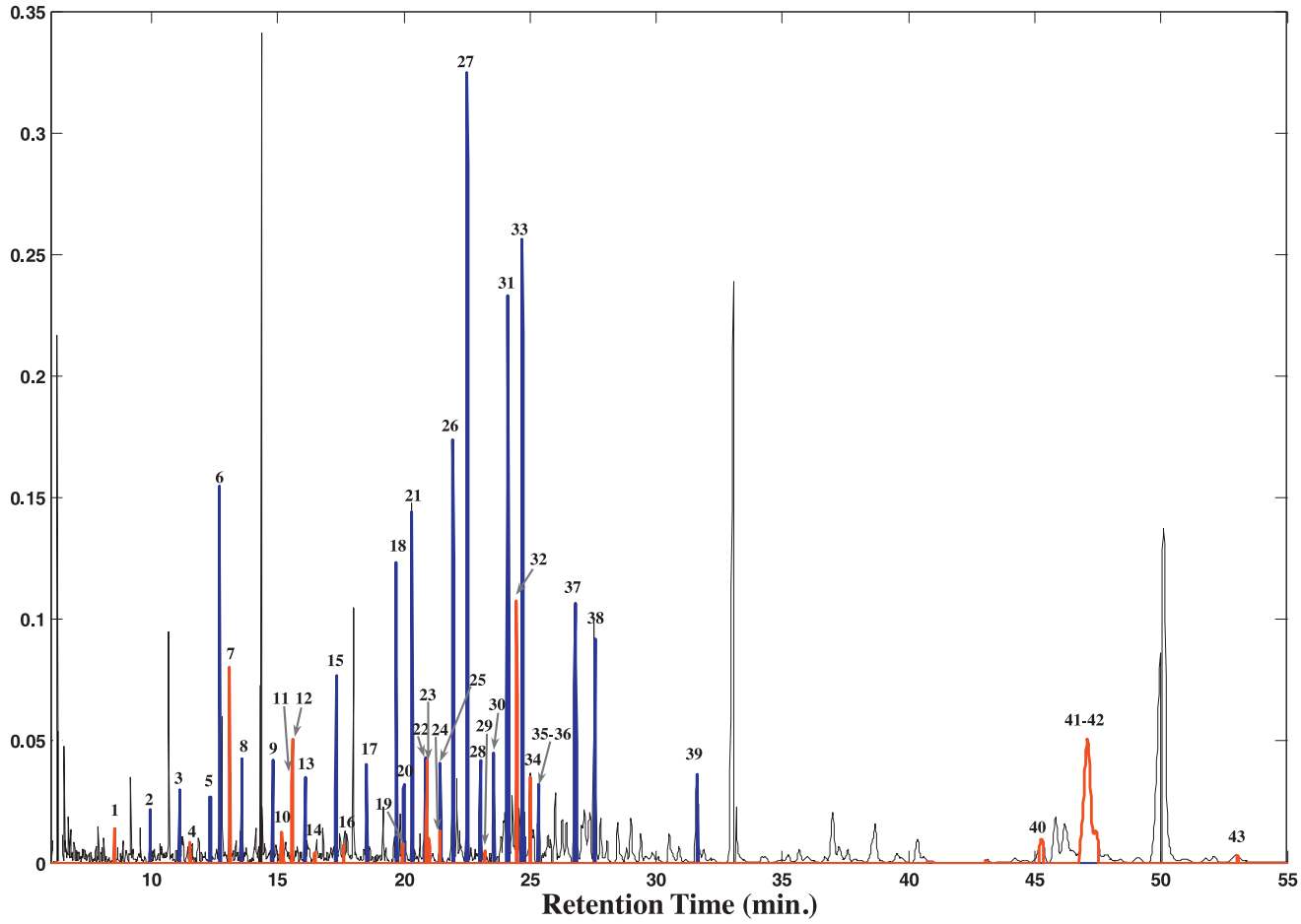


Fig. 4. Representative TIC chromatogram sample against time with red peaks positively correlated to SST and blue peaks inversely correlated to SST. Peaks are numbered according to VIP numbers from Table 2. Black chromatographic peaks are not relevant (VIP score <1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

abundance of this lipid reflects a high amount of phytoplankton and zooplankton inputs [45,46]. The short chain *n*-alkane-1-ols (<C_{20}, octadecan-ol and eicosanol) were higher when SST values were also high, but to a lesser extent than for cholesterol and squalene (Table 2). These compounds derive from a variety of sources such as marine plankton and/or bacteria [47]. Fig. 5a shows the concentration profiles of: heptatriacontadien-2-ol and non-atriacontatrien-3-ol, which bear a strong correlation with SST (Table 2). These alkenols could originate as a consequence of non-selective bacterial reduction of alkenones in anoxic sediments or by biosynthesis of the algae producing the alkenones [48,49]. The concentrations of these alkenones and the previously described compounds in sediments reflect increased marine productivity during higher SST values in site IODP-U1318C (Fig. 5a).

Long chain *n*-alkanes (from tetracosane to tritriacontane) had inverse correlations (negative) with the SST values (Table 2). As an example, Fig. 5b plots three of the long chain *n*-alkanes—pentacosane, heptacosane and hentriacontane—which were predominantly and inversely correlated with SST, with higher abundances at the lowest SST values. Long chain odd carbon numbered *n*-alkanes are major lipids of the epicuticular wax of vascular plants and common components of eolian dust [50–52]. They are well preserved in sediments over time [53]. In particular, plant wax components are removed from the leaf surface by rain or wind and transported by eolian, fluvial and/or ice mobilizations [54]. High abundances of organic dust components were also found to be inversely correlated with SST in the North Atlantic and the Pacific Ocean during glacial periods [55,56].

Some of the short chain *n*-alkanes such as heneicosane, eicosane and nonadecane showed high abundances at the lowest SST values (Table 2). They originate from a variety of sources, such as marine algae and bacteria [45,57]. The abundance of tetracosan-1-ol, pentacosan-1-ol, hexacosan-1-ol, heptacosan-1-ol octacosan-1-ol, triacontan-1-ol and dotriacontan-1-ol were relatively high at the lowest SST values. Long chain *n*-alkan-1-ols are ubiquitous constituents of vascular plants waxes, together with long chain *n*-alkanes [50]. The 9-octadecen-1-ol showed inverse correlation with SST. The 29-methyl hopane abundance was slightly higher in samples with SST <15 °C, which suggests an increase in bacterial activity [58].

The compounds from the neutral fractions found to be correlated to SST have either a marine or a terrestrial source. One interesting result of this approach is that it provides a differentiation between marine and terrestrial inputs as a consequence of the unbiased chemometrics data handling method used in this work. The group of marine lipids increase at higher SST indicating a high productivity and higher biomass production. The group of terrestrial compounds indicate higher inputs at lower temperatures. Climatic processes at lower temperatures have transported higher fluxes of terrigenous materials at IODP U1318C site which is consistent with previous observations in the north Atlantic in glacial periods [55].

Table 2
Assignment of VIP chromatographic peaks from the PLSR analysis of TIC chromatographic profiles matrix. Correlation coefficient and *p* value for each compound vs. temperature are annotated (UI, unidentified compound).

VIP	Retention time (min)	M^+ (*m*/*z*)	Compound	Correlation coefficient	*p* value
1	8.52		UI	0.68	0.0293
2	9.95	268	Nonadecane	−0.96	0
3	11.1	282	Eicosane	−0.95	0
4	11.52		UI	0.66	0.0366
5	12.32	296	Heneicosane	−0.93	0.0001
6	12.7	340	(*Z*)-9-Octadecen-1-ol	−0.25	0.4869*
7	13.1	342	Octadecan-1-ol	0.63	0.0505*
8	13.55	310	Docosane	−0.93	0.0001
90	14.82	324	Tricosane	−0.96	0
10	15.17		UI	0.69	0.0261
11	15.23		UI	0.69	0.028
12	15.57	370	Eicosan-1-ol	0.66	0.0385
13	16.1	338	Tetracosane	−0.89	0.0006
14	16.47		UI	0.59	0.0748
15	17.32	352	Pentacosane	−0.95	0
16	17.62		UI	0.57	0.084*
17	18.53	366	Hexacosane	−0.94	0.0001
18	19.7	380	Heptacosane	−0.96	0
29	19.95		UI	0.65	0.0433
20	20		UI	−0.54	0.1056*
21	20.34	426	Tetracosan-1-ol	−0.77	0.0099
22	20.85	394	Octacosane	−0.83	0.003
23	20.91	410	Squalene	0.71	0.0213
24	21		UI	0.89	0.0006
25	21.47	440	Pentacosan-1-ol	−0.94	0
26	21.95	408	Nonacosane	−0.79	0.0069
27	22.52	454	Hexacosan-1-ol	−0.90	0.0004
28	23.04	422	Triacontane	−0.83	0.0029
29	23.21		UI	0.78	0.0072
30	23.57	468	Heptacosan-1-ol	−0.94	0
31	24.12	436	Hentriacontane	−0.93	0.0001
32	24.48	458	Cholesterol	0.87	0.001
33	24.72	482	Octacosan-1-ol	−0.92	0.0001
34	25		UI	0.64	0.0485
35	25.34	191/410	29-Methyl-(17α,21 α)-hopane	−0.90	0.0004
36	25.35	450	Dotriacontane	−0.91	0.0003
37	26.8	464	Tritriacontane	−0.87	0.0011
38	27.57	510	Triacontan-1-ol	−0.55	0.1010*
39	31.62	538	Dotriacontan-1-ol	−0.81	0.0045
40	45.27	502	Heptatriacontadien-2-ol	0.75	0.0133
41	47.1		UI	0.90	0.0004
42	47.4		UI	0.86	0.0013
43	53.04	594	Nonatriacontatrien-3-ol	0.81	0.0044

* Compounds not significantly related to SST change.

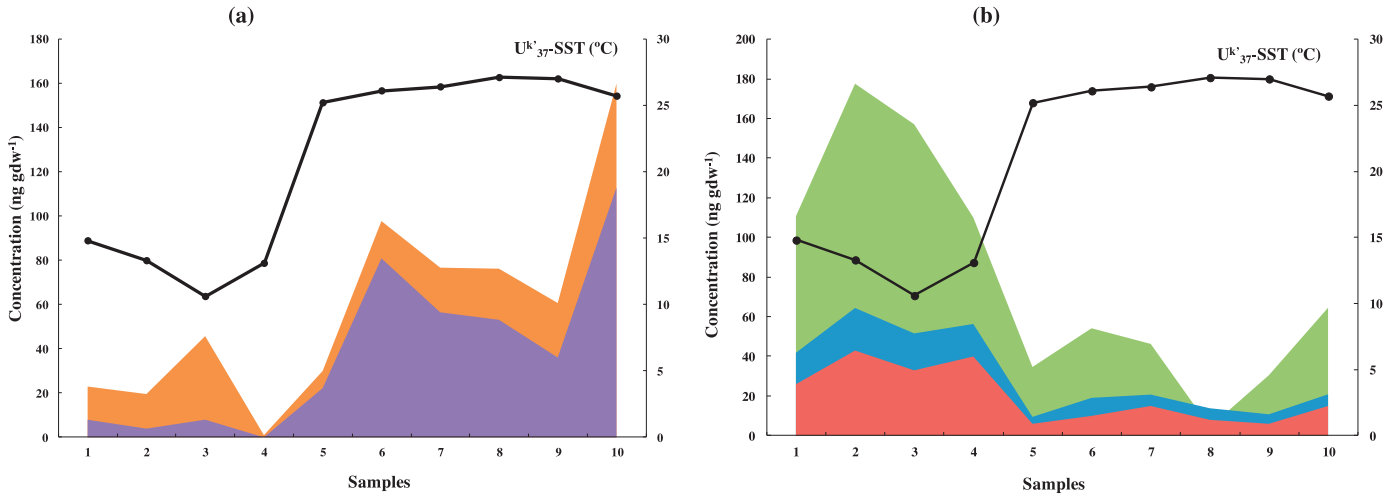


Fig. 5. Concentration profiles of (a) heptatriacontadien-2-ol (orange area), nonatriacontatrien-3-ol (purple area) and (b) pentacosane (red area), heptacosane (blue area), hentriacontane (green area) of the 10 sea sediment samples compared with the sea surface temperature (SST) which is shown by the black line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3.2. Sediment acidic fraction PLS results

In contrast to previous results with sediment neutral fraction, acidic fraction of the sediment samples required a significantly larger number of LVs to explain 90% of the **y** (SST) variance. Consequently, the PLSR model for SST prediction of acidic fraction samples were more complex than those built from neutral fraction samples, which needed only one LV to explain 93% of the **y** (SST) variance (Section 3.3.1). This is a consequence of a more intricate relationship between source precursor compounds found in the acid fraction and SST changes. Since the synthesis of acidic compounds, such as fatty acids, in the pelagic marine environment is controlled by local environment, SST values should also be one of the key factors affecting abundance patterns of fatty acids. However, it should also be considered that terrestrial vascular plants are another important source of fatty acids in aquatic environments [59] and that their abundance decreases more rapidly down core than neutral compounds, as a result of higher removal rate [60,61]. These compounds are common in marine and continental organisms and a relationship with their precursors is difficult. In addition, the rapid decrease in total fatty acid abundance synthesized by aquatic and terrestrial organisms prior to or during accumulation in the sediments appears to be a consequence of diagenetic effects, including benthic ingestion, chemical and physical availability, temperature and the degree of oxidation of the sediments [53,62]. It is likely that easier diagenetic alteration of the more labile fatty acids prevents preservation of any possible original correlation with SST.

4. Conclusions

Application of PCA, PLSR and VIP method to chromatographic profiles obtained by GC–MS analysis of IODP-U1318 marine sediments afforded the identification and determination of organic compounds which were more sensitive to SST changes over the last 11.7 Ma. In particular, the neutral fraction concentrations of long chain *n*-alkanes and long chain *n*-alkan-1-ols compounds associated to terrestrial inputs correlated inversely with SST. In contrast, the neutral fraction concentration of short chain *n*-alkan-1-ols, alkenols, cholesterol and squalene compounds that are typical of marine algae and zooplankton, correlated positively with SST. Conversely, compounds in the sedimentary acidic fraction showed more complex pattern correlations with SST difficult to interpret, probably due to the wider diversity of their sources, which contribute to their wide spread presence in marine systems and to their diagenetic alteration.

Acknowledgements

Special thanks go to the program CONSOLIDER-INGENIO 2010 (GRACCIE, CSD2007-00067) and to Ministerio de Economía y Competitividad, Spain, for a grant to the research project CTQ2012-38616-C02-01. B. M. thanks the CSIC-Ramón y Cajal post-doctoral program RYC-2013-14073.

References

[1] A. Hannachi, A Primer for EOF Analysis of Climate Data, Department of Meteorology, University of Reading, UK, 2004, pp. 1–33.
[2] A. Hannachi, I.T. Jolliffe, D.B. Stephenson, N. Trendafilov, In search of simple structures in climate: simplifying EOFs, Int. J. Climatol. 26 (2006) 7–28.
[3] J.F. McManus, G.C. Bond, W.S. Broecker, S. Johnsen, L. Labeyrie, S. Higgins, High-resolution climate records from the north Atlantic during the last interglacial, Nature 371 (1994) 326–329.
[4] N.J. Shackleton, M. Hall, E. Vicent, Phase relationship between millennial-scale events 64,000–24,000 years ago, Paleoceanography 15 (2000) 565–569.
[5] H. Elderfield, G. Ganssen, Past temperature and $\delta^{18}O$ of surface ocean waters inferred from foraminiferal Mg/Ca ratios, Nature 405 (2000) 442–445.
[6] T.I. Eglinton, G. Eglinton, Molecular proxies for paleoclimatology, Earth Planet Sci. Lett. 275 (2008) 1–16.
[7] M.N. Evans, S.E. Tolwinski-Ward, D.M. Thompson, K.J. Anchukaitis, Applications of proxy system modeling in high resolution paleoclimatology, Quat. Sci. Rev. 76 (2013) 16–28.
[8] E. Bard, Comparison of alkenone estimates with other paleotemperature proxies, Geochem. Geophys. Geosyst. 2 (2001) 1002.
[9] B. Martrat, J.O. Grimalt, N.J. Shackleton, L. de Abreu, M.A. Hutterli, T.F. Stocker, Four climate cycles of recurring deep and surface water destabilizations on the Iberian Margin, Science 317 (2007) 502–507.
[10] S.C. Brassell, G. Eglinton, I.T. Marlowe, U. Pflaumann, M. Sarnthein, Molecular stratigraphy: a new tool for climatic assessment, Nature 320 (1986) 129–133.
[11] F.G. Prahl, S.G. Wakeham, Calibration of unsaturation patterns in long-chain ketone compositions for palaeotemperature assessment, Nature 330 (1987) 367–369.
[12] F.G. Prahl, L.A. Muehlhausen, D.L. Zahnle, Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions, Geochim. Cosmochim. Acta 52 (1988) 2303–2310.
[13] J.F. López, J.O. Grimalt, Reassessment of the structural composition of the alkenone distributions in natural environments using an improved method for double bond location based on GC–MS analysis of cyclopropylimines, J. Am. Soc. Mass Spectrom. 17 (2006) 710–720.
[14] J. Imbrie, T.H. Van Andel, Vector analysis of heavy-mineral data, Geol. Soc. Am. Bull. 75 (1964) 1131–1156.
[15] D.B. Williams, W.A.S. Sarjeant, Organic-walled microfossils as depth and shoreline indicators, Mar. Geol. 5 (1967) 389–412.
[16] S.C. Brassell, R.G. Brereton, G. Eglinton, J.O. Grimalt, G. Liebezeit, I.T. Marlowe, U. Pflaumann, M. Sarnthein, Palaeoclimatic signals recognized by chemometric treatment of molecular stratigraphic data, Org. Geochem. 10 (1986) 649–660.
[17] J.G. Poynter, P. Farrimond, S.C. Brassell, G. Eglinton, A molecular stratigraphic study of sediments from Holes 658A and 660A, ODP Leg 108, Proc. ODP, Sci Results 108 (1989) 387–394.
[18] M.H. Conte, G. Eglinton, Alkenone and alkenoate distributions within the euphotic zone of the eastern North Atlantic: correlation with production temperature, Deep-Sea Res. 40 (Pt. I) (1993) 1935–1961.
[19] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[20] K.H. Esbensen, P. Geladi, 2.13—Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 211–226.
[21] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 1966, pp. 391–420.
[22] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[23] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[24] T.G. Ferdelman, A. Kano, T. Williams, J.P. Henriet, Expedition 307 scientists, site U1318, Proceedings of the Integrated Ocean Drilling Program, 307, Washington, D.C., 2006 (Integrated Ocean Drilling Program Management International, Inc.).
[25] B. De Mol, P. Van Rensbergen, S. Pillen, K. Van Herreweghe, D. Van Rooij, A. McDonnell, V. Huvenne, M. Ivanov, R. Swennen, J.P. Henriet, Large deep-water coral banks in the Porcupine Basin, southwest of Ireland, Mar. Geol. 188 (2002) 193–231.
[26] J. Raddatz, A. Rüggeberg, S. Margreth, W.C. Dullo, Paleoenvironmental reconstruction of Challenger Mound initiation in the Porcupine Seabight, NE Atlantic, Mar. Geol. 282 (2011) 79–90.
[27] A. Kano, T.G. Ferdelman, T. Williams, J.P. Henriet, T. Ishikawa, N. Kawagoe, C. Takashima, Y. Kakizaki, K. Abe, S. Saburo, E.L. Browning, L. Xianghui, Integrated Ocean Drilling Program Expedition 307 Scientists: age constraints on the origin and growth history of a deep-water coral mound in the northeast Atlantic drilled during Integrated Ocean Drilling Program Expedition 307, Geology 35 (2007) 1051–1054.
[28] J. Villanueva, J.O. Grimalt, Pitfalls in the chromatographic determination of the alkenone $U_{37}^{K'}$ index for paleotemperature estimation, J. Chromatogr. A 723 (1996) 285–291.
[29] J. Villanueva, C. Pelejero, J.O. Grimalt, Clean-up procedures for the unbiased estimation of C_{37} alkenone sea surface temperatures and terrigenous *n*-alkane inputs in paleoceanography, J. Chromatogr. A 757 (1997) 145–151.
[30] P.J. Müller, G. Kirst, G. Ruhland, I. von Storch, A. Rosell-Melé, Calibration of the alkenone paleotemperature index $U_{37}^{K'}$ based on core-tops from the eastern South Atlantic and the global ocean (60(N-60(S), Geochim. Cosmochim. Acta 62 (1998) 1757–1772.
[31] K.H. Liland, Multivariate methods in metabolomics—from pre-processing to dimension reduction and statistical analysis, TrAC-Trend. Anal. Chem. 30 (2011) 827–841.
[32] P. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636.
[33] P. Eilers, H. Boelens, Baseline correction with asymmetric least squares smoothing. http://www.science.uva.nl/~~hboelens/publications/draftpub/Eilers_2005.pdf, 2005.

[34] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, J. Chromatogr. A 805 (1988) 17–35.
[35] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, J. Chromatogr. A 996 (2003) 141–155.
[36] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A 1096 (2005) 101–110.
[37] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241.
[38] T. Skov, F. van den Berg, G. Tomasi, R. Bro, Automated alignment of chromatographic data, J. Chemom. 20 (2006) 484–497.
[39] M.E. Mann, R.S. Bradley, M.K. Hughes, Global-scale temperature patterns and climate forcing over the past six centuries, Nature 392 (1998) 779–787.
[40] S. Wold, E. Johansson, M. Cocchi, PLS—partial least-squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.
[41] S. Wold, PLS for multivariate linear modeling, in: H. van de Waterbeemd (Ed.), QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry, Verlag-Chemie, Weinheim, Germany, 1994, pp. 195–218.
[42] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.
[43] R.B. Gagosian, G.E. Nigrelli, Limnol. Oceanogr. 24 (1979) 838–849.
[44] K. Grice, W.C.M. Klein Breteler, S. Schouten, V. Grossi, J.W. de Leeuw, J.S. Sinninghe Damasté, Effects of zooplankton herbivory on biomarker proxy records, Paleoceanography 13 (1998) 686–693.
[45] M. Blumer, R.R.L. Guillard, T. Chase, Hydrocarbons of marine phytoplankton, Mar. Biol. 8 (1971) 183–189.
[46] M.I. Venkatesan, I.R. Kaplan, Distribution and transport of hydrocarbons in surface sediments of the Alaskan outer continental shelf, Geochim. Cosmochim. Acta 46 (1982) 2135–2149.
[47] A. Gogou, E.G. Stephanou, Marine organic geochemistry of the Eastern Mediterranean: 2. Polar biomarkers in Cretan Sea surficial sediments, Mar. Chem. 85 (2004) 1–25.
[48] J.F. Rontani, D. Marchand, J.K. Volkman, $NaBH_4$ reduction of alkenones to the corresponding alkenols: a useful tool fro their caracteritzation in natural samples, Org. Geochem. 32 (2001) 1329–1341.
[49] M.H. Conte, J.K. Volkman, G. Eglinton, Lipid biomarkers of the Haptophyta, in: J.C. Green, B.S.C. Leadbeater (Eds.), The Haptophyte Algae Systematics Association, Special volume 51, Clarendon, Oxford, England, 1994, pp. 351–377.
[50] G. Eglinton, R.J. Hamilton, Leaf epicuticular waxes, Science 156 (1967) 1322–1335.
[51] J.R. Ehleringer, T.E. Cerling, B.R. Helliker, C_4 photosynthesis atmospheric CO_2, and climate, Oecologia 112 (1997) 285–299.
[52] K. Koch, H.J. Ensikat, The hydrophobic coatings of plant surfaces: epicuticular wax crystals and their morphologies, crystallinity and molecular self-assembly, Micron 39 (2008) 759–772.
[53] E.A. Canuel, C.S. Martens, Reactivity of recently deposited organic matter: Degradation of lipid compounds near the sediment-water interface, Geochim. Cosmochim. Acta 60 (1996) 1793–1806.
[54] B.R.T. Simoneit, Organic matter in eolian dusts over the Atlantic Ocean, Mar. Chem. 5 (1977) 443–464.
[55] C. López-Martínez, J.O. Grimalt, B. Hoogakker, J. Gruetzner, M.J. Vautravers, I.N. McCave, Abrupt wind regime changes in the North Atlantic Ocean during the past 30,000–60,000 years, Paleoceanography 21 (2006) PA4215.
[56] F. Lamy, R. Gersonde, G. Winckler, O. Esper, A. Jaeschke, G. Kuhn, J. Ullermann, A. Martinez-Garcia, F. Lambert, R. Kilian, Increased dust deposition in the Pacific Southern Ocean during glacial periods, Science 343 (2014) 403–407.
[57] J.K. Volkman, R.B. Johns, F.T. Gillan, G.J. Perry, H.J. Bavor Jr., Microbial lipids of an intertidal sediment—I. Fatty acids and hydrocarbons, Geochim. Cosmochim. Acta 44 (1984) 1133–1143.
[58] M. Rohmer, P. Bouvier, G. Ourisson, Molecular evolution of biomembranes: structural equivalents and phylogenetic precursors of sterols, Proc. Natl. Acad. Sci. U. S. A. 76 (1979) 847–851.
[59] J. Dalsgaard, M. St. John, G. Kattner, D. Müller-Navarra, W. Hagen, Fatty acid trophic markers in the pelagic marine environment, Advances in Marine Biology, vol. 46, Academic Press, 2003, pp. 225–340.
[60] L.L. Belicka, R.W. Macdonald, M.B. Yunker, H.R. Harvey, The role of depositional regime on carbon transport and preservation in Arctic Ocean sediments, Mar. Chem. 86 (2004) 65–88.
[61] M.B. Yunker, L.L. Belicka, H.R. Harvey, R.W. Macdonald, Tracing the inputs and fate of marine and terrigenous organic matter in Arctic Ocean sediments: a multivariate analysis of lipid biomarkers, Deep-Sea Res. 52 (Pt. II) (2005) 3478–3508.
[62] H.R. Harvey, R.D. Fallon, J.S. Patton, The effect of organic matter and oxygen on the degradation of bacterial membrane lipids in marine sediments, Geochim. Cosmochim. Acta 50 (1986) 795–804.

5.1.1 DISCUSSIÓ DELS RESULATS OBTINGUTS

Els cromatogrames TIC obtinguts en l'anàlisi de les mostres de sediments marins van ser agrupats en dues matrius de dades, una matriu per la fracció neutra i una altra matriu per a la fracció àcida. Les dimensions d'aquestes era de 10 files (mostres) i 2940 columnes (temps de retenció). Les dues matrius es van analitzar primer per PCA i després per PLSR conjuntament amb el vector corresponent de temperatures.

En relació als resultats del PCA, per a la matriu de dades de la fracció neutra es va escollir un model de dos components principals amb una variància explicada del 87.79%, dels quals el primer explicava el 77.45% de la variància. El primer component permetia separar les mostres segons si corresponien al grup de temperatures altes (>25 ºC) o baixes (<15 ºC) (vegeu figura 2a, Article 4). El segon component (10.34% de la variància) no estava relacionat amb el canvi de temperatures de la superfície marina (*Sea Surface Temperature*, SST) i podria correspondre a un altre tipus de variació natural no investigada en aquest estudi. L'abundància dels compostos orgànics podria estar influenciada, a part de la SST, per altres factors abiòtics que influencien la productivitat marina com són la llum, els nutrients, el contingut d'oxigen o bé la salinitat de l'aigua. Aquests factors juguen un paper molt important en el creixement i abundància de les poblacions marines. En el cas de l'anàlisi de la matriu de la fracció àcida, la selecció de dos components explicava una variància total del 81.74%. No obstant, en el cas de la fracció àcida, la representació gràfica dels *scores* obtinguts per aquests dos components no va permetre esbrinar de forma clara la relació de les mostres amb els valors de SST (vegeu figura 2b, Article 4).

Les matrius de dades de les dues fraccions (neutres i àcids), **X** (10 x 2940), es van analitzar amb el mètode PLSR, per separat, en relació a les SST associades a cada mostra (**y**). L'aplicació del PLSR va permetre una millor interpretació de la informació continguda en les dades de les dues fraccions (neutres i àcids) dels compostos analitzats. El model PLSR de la matriu de dades de la fracció neutra va resultar en un sol component que explicava un 64.59% de la variància de **X** (vegeu figura 3, Article 4) i un 93.01% de la variància de **y** (SST). Posteriorment, es van seleccionar aquelles variables (compostos orgànics) que tenien un valor VIP major a u. Tal i com posteriorment s'argumenta al capítol 6 d'aquesta Tesi i a l'Article 5, en aquest cas es va preferir el mètode de selecció de variables VIP que el mètode SR.

A la figura 4 de l'Article 4 es representen els compostos que finalment es van seleccionar de la fracció neutra en relació als canvis de SST. Els compostos orgànics les concentracions dels quals mostraven correlacions positives amb la SST es marquen en vermell, i es marquen en blau els compostos orgànics les concentracions dels quals mostraven correlacions negatives amb la SST. Tal i com es pot observar en els resultats de la taula 2 de l'Article 4, una gran part dels compostos orgànics de la fracció neutra tenien un coeficient de correlació amb la SST major a 0.7, amb una significació $p < 0.05$. Molts d'aquests compostos orgànics es van identificar, però d'altres que també eren interpretables com a marcadors del canvi de la SST no van poder ser finalment identificats. Alguns d'aquests compostos no identificats presentaven un coeficient de correlació amb la SST major a 0.8, per tant, en estudis successius seria interessant aprofundir en la seva identificació, ja que podrien permetre una millora en la interpretació dels resultats.

La identificació dels compostos orgànics de la fracció neutra va permetre elaborar la hipòtesi sobre la relació que hi havia entre els canvis de concentració dels compostos i la variació de SST. L'esqualè, per exemple, estava correlacionat positivament amb la SST i l'elevada concentració d'aquest lípid reflecteix l'existència de grans quantitats de fitoplàncton i zooplàncton al mar (Blumer et al., 1971; Venkatesan i Kaplan, 1982). La concentració dels alcan-1-ols de cadena curta també va incrementar a altes SST, doncs la presència d'aquests compostos també depèn de l'abundància de plàncton i/o bactèries (Gogou i Stephanou, 2004). A la figura 5a de l'Article 4 es mostra com les concentracions de dos alquenols (heptatriacontadien-2-ol, nonatriacontatrien-3-ol) incrementaven amb les SST més altes. Aquests alquenols són produïts per la reducció bacteriana no selectiva de les alquenones en sediments anòxics o a partir de la biosíntesi de les algues que produeixen alquenones (Conte et al., 1994; Rontani et al., 2001). Conseqüentment, les altes concentracions d'aquests compostos reflecteixen l'augment de la productivitat biosintètica marina quan les SST eren més altes.

Els alcans de cadena llarga, des del tetracosane fins el tritriacontane, presentaven concentracions més altes a les mostres de sediments associades SST baixes (vegeu figura 5b, Article 4). Els alcans de cadena llarga amb un nombre de carbonis imparell són els principals lípids de les ceres epicuticulars de les plantes vasculars i són compostos comuns en la pols eòlica (Eglinton i Hamilton, 1967). Aquestes ceres es preserven molt bé en els sediments marins al llarg del temps (Canuel i Martens, 1996). Els components de les ceres epicuticulars són arrossegats de la superfície de les fulles per l'efecte de la pluja o del vent i són transportats al mar pel vent o bé per

moviments fluvials o del gel (Simoneit, 1977). D'acord amb aquesta hipòtesi, estudis anteriors com els de (López i Grimalt, 2006) i (Lamy et al., 2014) van trobar més abundància d'aquests compostos orgànics a l'Atlàntic Nord i a l'Oceà Pacífic durant els períodes glacials. Per altra banda, l'abundància d'alcan-1-ols de cadena llarga a les mostres de sediment estava inversament correlacionada amb la SST. Els compostos alcan-1-ols de cadena llarga són constituents ubics de les ceres de les plantes vasculars (Eglinton i Hamilton, 1967), juntament amb els alcans de cadena llarga.

Els compostos orgànics fòssils de la fracció neutra de les mostres de sediments investigades tenen orígens tant marí com terrestre. La concentració dels lípids d'origen marí va augmentar amb l'augment de la SST, indicant una alta productivitat marina i una major producció de biomassa. En canvi, els compostos orgànics d'origen terrestre presentaven una major acumulació en els sediments marins en presència de baixes temperatures de l'aigua superficial del mar. Els processos climàtics a baixes temperatures van propiciar el transport de majors fluxos de materials terrígens al lloc on es va extraure la mostra del sediment IODP U1318C. Aquest fet és consistent amb observacions anteriors fetes a l'Atlàntic nord durant diversos períodes glacials (López i Grimalt, 2006).

L'aplicació del mètode PLSR a la matriu de dades de la fracció àcida de compostos orgànics fòssils va necessitar un major nombre de variables latents per arribar a explicar un 90% de la variància de **y** (SST). Aquests resultats són conseqüència d'una relació complexa entre la variació de la concentració dels compostos orgànics de la fracció àcida, el seu origen i la variació de la temperatura superficial del mar (SST). Les plantes vasculars també són una font important dels àcids grassos dels organismes presents en ambients aquàtics (Dalsgaard et al., 2003). L'abundància dels àcids grassos en els sediments decreix més ràpidament que la dels compostos orgànics de la fracció neutra, com a conseqüència de la seva major velocitat d'eliminació per processos de degradació (Belicka et al., 2004; Yunker et al., 2005). Aquesta ràpida degradació dels compostos àcids als sediments és resultat dels diferents efectes digenètics, els quals inclouen la ingestió bentònica, l'afectació física i química, la temperatura i el grau d'oxidació dels sediments (Canuel i Martens, 1996; Harvey et al., 1986). L'alteració diagènetica dels àcids grassos més làbils impedeix la preservació de qualsevol possible correlació original amb la SST.

Com a resum d'aquest capítol es constata que existeix una forta correlació entre els perfils cromatogràfics de la fracció neutra de les mostres de sediments marins investigats i la variació de la temperatura superficial de l'aigua de mar (SST) associada a aquests sediments. Tot i que ja existeixen molts estudis científics que correlacionen individualment diversos marcadors químics (climàtics) i la SST, en el present capítol i treball corresponent (Article 4) es relacionen de forma simultània tots els compostos orgànics obtinguts per GC-MS de mostres de sediments marins amb els canvis de temperatura deguts a canvis climàtics en el passat. En aquest cas s'ha fet servir una aproximació similar a la que s'utilitza en els estudis de metabolòmica, descrita en capítols anteriors. Moltes dels avenços actuals de l'aplicació de les eines quimiomètriques en estudis metabolòmics (vegeu secció 2.3 de la introducció de la Tesi, capítol 2) són, per tant, també extrapolables als estudis paleoclimàtics com els que s'han presentat en aquest capítol.

Capítol 6

Comparació de mètodes per a la selecció de variables

6.1 INTRODUCCIÓ

Un problema comú en moltes àrees de coneixement químic i biològic, i en particular en estudis òmics no dirigits (*untarget*) (Capítols 4 i 5) és que la majoria de variables són generalment irrellevants per a un problema determinat de discriminació o classificació. En el cas de la metabolòmica es generen grans conjunts de dades multivariants obtinguts mitjançant l'anàlisi per LC-MS, i l'objectiu principal és determinar a partir dels senyals mesurats (pics cromatogràfics) quins són els possibles marcadors (metabòlits) que responen als factors específics investigats. Una manera de fer la selecció de variables és fer totes les combinacions de variables possibles i seleccionar aquelles que són millors. Això sembla factible, però a la pràctica és difícil degut al gran nombre de variables involucrades i pot representar molt esforç i temps de càlcul. Per això, s'han desenvolupat diferents mètodes que seleccionen el grup de variables que és òptim per a una finalitat específica (Andersen i Bro, 2010). L'aplicabilitat d'aquests mètodes depèn de la naturalesa de les dades i del propòsit de l'estudi.

El desenvolupament de mètodes adequats de selecció de variables és un dels aspectes més importants en quimiometria. La reducció de la dimensió d'una taula de dades multivariant esdevé essencial quan aquesta és complexa i està composta per centenars o milers de dades originals. La principal idea de la selecció de variables és eliminar aquelles que no contribueixen a l'estructura latent de les dades o, dit d'altre manera, seleccionar únicament aquelles variables que contribueixen a l'estructura latent de les dades i ajudar a predir o explicar el fenomen investigat. Per tant, aquesta selecció pot facilitar la predicció i/o interpretació del problema químic i/o biològic investigat. El mètode de regressió per mínims quadrats parcials (*Partial Least Squares*, PLS) ha guanyat popularitat en estudis metabolòmics degut a la seva capacitat de relacionar un gran nombre de variables correlacionades amb una resposta determinada a través d'un model de regressió o classificació multivariant. Existeixen diferents mètodes per a seleccionar les variables més importants en problemes de regressió i classificació. Entre aquests, cal esmentar el mètode que avalua la importància de la variables en la projecció (*Variable Importance in Projection*, VIP) (Chong i Jun, 2005; Wold et al., 1993) , el mètode que calcula el quocient de selectivitat (*Selectivity Ratio*, SR) (Rajalahti et al., 2009a; Rajalahti et al., 2009b) , el procediment basat en mínims quadrats parcials per intervals (*interval Partial Least Squares*, *i*PLS)(Norgaard et al., 2000) i els algorismes genètics (*Genetic Algorithm*, GA)(Leardi, 2000; Leardi et al., 1992). De tota

manera, i tal i com s'ha explicat a la introducció d'aquesta memòria i a l'Article 5 els mètodes VIP i SR són els que actualment s'estan utilitzant més freqüentment en estudis d'òmica a partir de mètodes quimiomètrics.

En aquest capítol es compara l'aplicació dels mètodes de selecció de variables VIP i SR a partir de la regressió per mínims quadrats parcials (PLS), els quals s'han descrit a la secció 2.5.5 de la introducció de la Tesi (capítol 2). Aquests dos mètodes es comparen mitjançant la seva aplicació a tres conjunts de dades de diferent naturalesa: 1) paràmetres fisicoquímics de qualitat i sensorials sobre un mateix grup de mostres d'aigua (conjunt de dades de qualitat sensorial de l'aigua) (Platikanov et al., 2013); 2) perfils GC-MS de compostos orgànics fòssils acumulats als sediments marins relacionats amb la temperatura superficial del mar (*Sea Surface Temperature*, SST) (conjunt de dades climàtiques, vegeu capítol 5); i 3) expressions transcriptòmiques de femelles de *Daphnia magna* exposades a dosis subletals d'inhibidors selectius de recaptació de serotonina (*Selective Serotonin Reuptake Inhibidors*, SSRI) relacionades amb el seu nombre de cries (conjunt de dades de transcriptòmica) (Campos et al., 2013). Per a la discussió sobre quin era el mètode de selecció de variables més adequat es van calcular els coeficients de correlació (r) i els nivells de significació (valor p) per a cada una de les variables seleccionades amb la finalitat d'interpretar els fenòmens subjacents per a cada conjunt de dades.

En l'estudi del conjunt de dades sobre la qualitat sensorial de l'aigua es van determinar els paràmetres fisicoquímics (**X**) més influents associats amb la puntuació global sensorial (del gust) de l'aigua (**y**), per part d'un panell de professionals especialitzats per dur a terme assaigs de degustació. Aquest estudi sensorial es va realitzar sobre aigües minerals embotellades i aigües de la xarxa pública de distribució, les quals tenien continguts químics i orígens diversos. Aquest conjunt de dades era especialment interessant perquè representava un conjunt de dades (matriu de dades) amb un nombre limitat de variables 13 paràmetres fisicoquímics (matriu **X**) i una puntuació sensorial mitjana (vector **y**) sobre 25 mostres d'aigua diferents (Platikanov et al., 2013).

En el cas de l'estudi de dades climàtiques és van investigar les variacions dels perfils cromatogràfics GC-MS de 10 mostres de sediments marins en relació els canvis de la temperatura superficial del mar (*Sea Surface Temperature*, SST) associats a les mateixes mostres de sediments (**y**). Es van estudiar els resultats de l'aplicació dels mètodes de selecció VIP i SR en els cromatogrames TIC GC-MS (10 mostres i 2940 temps de retenció, matriu **X1**), i també en les

concentracions de 77 compostos prèviament quantificats a partir dels cromatogrames TIC (10 mostres i 77 compostos, matriu **X2**).

En l'estudi de les dades de transcriptòmica es van analitzar 1207 gens de 12 mostres de *Daphnia Magna* a partir de xips d'ADN (àcid desoxiribonucleic) (*microarrays*), matriu **X**, i la seva relació amb el seu nombre de cries (vector **y**). La finalitat d'aquests estudi era la d'investigar quins són els mecanismes moleculars que expliquen els efectes hormètics observats en les respostes a l'estrès de la transcripció en organismes de *Daphnia magna* adults i juvenils exposats a dosis subletals de SSRI. L'anàlisi de xips d'ADN és un exemple típic on s'obtenen matrius de dades de grans dimensions. Aquests conjunts de dades contenen des de milers a desenes de milers de perfils d'expressió gènica (variables), però generalment es disposa nomes d'un nombre limitat (desenes) de mostres. Per aquest motiu aquest grup de dades ha estat també escollit per a la comparació dels mètodes de selecció de variables, VIP i SR, en presentar aspectes diferents als dels altres dos grups de dades també investigats en aquest capítol.

Research article

Journal of CHEMOMETRICS

Received: 11 March 2015, Revised: 17 June 2015, Accepted: 19 June 2015, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.2736

Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation

Mireia Farrés[a], Stefan Platikanov[a], Stefan Tsakovski[b] and Romà Tauler[a]*

This study compares the application of two variable selection methods in partial least squares regression (PLSR), the variable importance in projection (VIP) method and the selectivity ratio (SR) method. For this purpose, three different data sets were analysed: (a) physiochemical water quality parameters related to sensorial data, (b) gas chromatography–mass spectrometry (GC-MS) chemical (organic compound) profiles from fossil sea sediment samples related to sea surface temperature (SST) changes, and (c) exposed genes of *Daphnia magna* female samples related to their total offspring production. Correlation coefficients (*r*), levels of significance (*p*-value) and interpretation of the underlying experimental phenomena allowed the discussion about the best approach for variable selection in each case. The comparison of the two variable selection methods in the first water quality data set showed that the SR method is more accurate for sensorial prediction. For the climate data set, when raw total ion current (TIC) GC-MS chromatograms were considered, variables selected using the VIP method were easier to interpret compared with those selected by the SR method. However, when only some chromatographic peak areas (concentrations) were considered, the SR method was more efficient for prediction, and the VIP method selected the most relevant variables for the interpretation of SST changes. Finally, for the transcriptomic data set, the SR method was found again to be more reliable for prediction purposes. Copyright © 2015 John Wiley & Sons, Ltd.

Additional supporting information can be found in the online version of this article at the publisher's website.

Keywords: variable importance in projection; selectivity ratio; variable selection; partial least squares

1. INTRODUCTION

The analysis of multivariate and megavariate data has become increasingly important in diverse scientific fields like in gene expression microarray data analysis; gas and liquid chromatography-mass spectrometry (GC- and LC-MS); Fourier transform infrared spectroscopy; Raman, nuclear magnetic resonance, and MS spectroscopies; and hyperspectral imaging, among other. In all these cases, variable selection techniques are a critical step to obtain a good prediction performance and to explain the underlying phenomena. For predicting one or several parameters from a multivariate data set, multivariate linear calibration models based on latent variables like principal component regression [1] and partial least squares regression (PLSR) [2–4] methods are used. These methods can process very large data sets even when the number of variables is much larger than the number of samples [5]. In many cases, most of these variables are little relevant to the investigated problem as they represent variation not related to the response to be modelled and their number can be drastically reduced with minor loss of information. Variable selection methods help selecting a small set of very relevant predictor variables which are correlated to a particular response variable. Variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance the model interpretability with parsimonious representation.

There are many approaches that have been proposed as variable selection methods; a large number of them have been extensively described in previous works [5,6]. One of the most popular variable selection methods at present is the variable importance in partial least squares (PLS) projection method, which was proposed in 1993 by Wold *et al*. [7] as 'variable influence on projection' (VIP) which is also known as 'variable importance in projection' scores or VIP scores by [8]. VIP scores are useful in understanding **X** space predictor variables that best explain **y** variance. VIP method selects those **X** variables that contribute most to the underlying variation in the **X** variables. This includes the variation not related to **y**, but describing interferences, that is so-called orthogonal variation. Target projection (TP) with selectivity ratio (SR) is another popular method, which was also proposed as a tool for variable selection in multivariate data

* *Correspondence to: R. Tauler, Department of Environmental Chemistry (IDAEA-CSIC), Jordi Girona 18, 08034, Barcelona, Spain*
E-mail: Roma.Tauler@idaea.csic.es

a *M. Farrés, S. Platikanov, R. Tauler*
Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18, 08034, Barcleona, Spain

b *S. Tsakovski*
Department of Analytical Chemistry, Faculty of Chemistry, Sofia University, James Bourchier Blvd, 1164, Sofia, Bulgaria

analysis [9]. Variable SRs are obtained by calculating the ratio of explained to residual variance of the **X** variables on the **y** target-projected component. SR is a method for **X** variable ranking in relation to explanation of **y** variance, especially useful for prediction.

VIP and SR are two of the most frequently used methods in chemometrics for variable selection. VIP scores selection method has been extensively used in different fields and thus for a variety of data types [10–16]. Recently, Galindo-Prieto and coauthors [17] proposed a new VIP approach for orthogonal projections to PLS latent structures to enhance model interpretability. Although the number of applications in scientific works of the SR method is lower than the VIP scores method, SR method is significantly increasing its use at present [9,18–21]. VIP and SR methods have been compared in some scientific works together with other variable selection methods [9,22,23]. Rajalahti and coauthors [9] compared SR and VIP on mass spectral profiles and stated that VIP approach was not working for biomarker selection, because it proposed too many false biomarkers. Likewise, Tran and coauthors [23] compared SR and VIP methods on NIR spectra data sets, and it was found that SR was more reliable for data sets with noisy variables when compared with the VIP, despite that SR was too conservative. More recently, [24] investigated PLS discriminant analysis (PLS-DA) combined with VIP and SR on GC coupled with flame ionization detection data. In this case, the effect of variable selection was monitored using a bootstrap procedure. It was concluded that SR presented best predictive abilities than VIP, and that the VIP method was biased to select those variables that were present at the highest concentrations and presented large absolute sizes (they have large variances in the PLS-DA model).

In this work, VIP and SR variable selection methods are compared for the analysis of three different data sets. In the first data set (water quality data set), physiochemical water quality parameters were related to sensorial data of the same water samples [25]. In the second data set (climate data set), GC-MS chemical (organic compound) profiles from fossil sea sediment samples were correlated with sea surface temperature (SST) changes [16]. And, in the third data set (transcriptomic data set), genes from *Daphnia magna* female samples exposed to sublethal doses of the selective serotonin reuptake inhibitors (SSRIs: fluoxetine and fluvoxamine) were correlated to their corresponding total offspring production [26].

A discussion about the reasons why the two variable selection methods (VIP and SR) differ when applied to the same three data sets is presented. Also, the possible advantages and disadvantages of applying these two methods are examined.

2. THEORY

2.1. Partial least squares regression (PLSR)

In this work PLSR [3,4,27] analysis was applied to the three data sets (water quality data, climate data, and transcriptome data) to investigate the more influent variables in the corresponding model and interpret them. PLSR is a multivariate linear regression method used to find correlation models between predictor variables (**X** data matrix) and response variables (usually arranged in a **y** vector) measured on the same set of samples. PLSR provides information about the correlation structures of the variables and about their structural similarities or dissimilarities.

2.2. Variable importance in projection (VIP)

The VIP selection method was first published by Wold and coauthors [7]. VIP scores summarise the influence of individual **X** variables on the PLS model. VIP scores are calculated as the weighted sum of squares of the PLS weights, **w***, which take into account the amount of explained **y** variance in each extracted latent variable (dimension). VIP scores give a measure useful to select what are the variables which contribute the most to the **y** variance explanation. For a given model and data set there will always be only one VIP scores-vector, summarizing all components and **y** variables.

The VIP score for the j^{th} variable is given as

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{f=1}^{F} \mathrm{w}_{jf}^2 \cdot \mathrm{SSY}_f \cdot J}{\mathrm{SSY}_{\mathrm{total}} \cdot F}} \tag{1}$$

where w_{jf} is the weight value for j variable and f component and SSY_f is the sum of squares of explained variance for the f^{th} component and J number of **X** variables. SSY_{total} is the total sum of squares explained of the dependent variable, and F is the total number of components. The $\mathbf{w}_{jf}^2$ gives the importance of the j^{th} variable in each f^{th} component, and VIP_j is a measure of the global contribution of j variable in the complete PLS model. In case of one-dimensional **Y** space, **y**, holds

$$\mathrm{SSY}_f = \mathbf{b}_f^2 \mathbf{t}_f' \mathbf{t}_f \qquad \mathrm{SSY}_{\mathrm{total}} = \mathbf{b}^2 \mathbf{T}' \mathbf{T} \tag{2}$$

where **T** is the **X** scores matrix and **b** is the PLS inner relation vector of coefficients.

Since the average of the squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion for variable selection [28]. This is not a statistically justified limit, and it can be shown that it is very sensitive to the presence of non-relevant information pertaining to **X** [23].

2.3. Selectivity ratio (SR)

The SR [9,21] method is a visualization tool for searching what are the important variables of a multivariate data set in the prediction of a particular property. The ratio between the explained and the residual (unexplained) variance for each variable in the TP vector defines the SR for the variable in question. This TP utilises both the predictive ability (regression vector) and the explanatory ability (spectral variance/covariance matrix) for the calculation of the SR. Given the PLS regression vector, $\mathbf{b}_{PLS}$, Target projection is performed via the projection of the rows of **X** onto the normalised regression coefficients vector $\mathbf{b}_{PLS}$ in Eqn (3). In this equation t_{TP} is proportional to the predicted values, $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{PLS}$. The loadings, $\mathbf{p}_{TP}$, are obtained by projecting the columns of **X** onto the score vectors, t_{TP}, which again is proportional to $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{PLS}$ in Eqns (3) and (4).

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{b}_{PLS} / \|\mathbf{b}_{PLS}\| \tag{3}$$

$$\mathbf{p}_{TP} = \mathbf{X}'\mathbf{t}_{TP} / (\mathbf{t}_{TP}'\mathbf{t}_{TP}) \tag{4}$$

The ratio of the explained variance ($SS_{i,explained}$) and of the residual variance for each variable ($SS_{i,residual}$) in the sum of squares

in Eqns (5) and (6), respectively, is used then to determine the variable importance, SR, in Eqn (7)

$$SS_{i,\text{explained}} = \|\mathbf{t}_{TP}\mathbf{p}_{TPi}{}'\|^2 \quad (5)$$

$$SS_{i,\text{residual}} = \|\mathbf{e}_{TPi}\|^2 \quad (6)$$

$$SR_i = SS_{i,\text{explained}}/SS_{i,\text{residual}} \quad (7)$$

Rajalahti and coauthors [21] proposed an *F*-test to define as a boundary between variable regions with high discriminating ability and less interesting regions. In order to determinate which variable has a high discriminatory ability and to reject the null hypothesis (explained and residual variances are the same), the calculated *F* value (F_{calc}), which is equal to SR_i from Eqn (7), has to exceed the critical value for the *F* distribution, F_{crit}.

$$F_{calc} = SR_i > F_{crit} = F(\alpha, N\text{-}2, N\text{-}3) \quad (8)$$

where *N* is the sample size and α the significance level. The number of degrees of freedom for the numerator (explained variance) in Eqn (8) is equal to sample size *N* minus two degrees of freedom, one because of the calculation of the variable's mean and one because of the introduction of the target component (*N*-2). For the denominator (residual variance) one extra degree of freedom is lost when the explained variance is subtracted from the original variance of the variable. Thus, the remaining degrees of freedom of the denominator are (*N*-3) [21]. In this work, the *F*-test (95%) criterion has been chosen to select the marker candidate.

3. DATA SETS DESCRIPTION

3.1. Water quality data set

This data set (refer to Table S1) consisted of 13 physicochemical parameters (**X** predictor variables) and of one overall score for the water taste and flavour evaluation (**y** predicted variable) of 25 bottled mineral and tap water samples covering a wide range of mineralization and chemical composition from different sources [25].

Water samples were analysed using standard methods. Water blends and dilutions were allowed for 48 h of equilibration before analysis. Sodium, potassium, calcium and magnesium and silica concentration levels were analysed by inductively coupled plasma–optical emission spectrometry. Conductivity at 20 °C, pH and bicarbonate levels were determined by robotic titrosampler. Chloride, nitrate and sulphate concentrations were analysed by ionic chromatography. Thermal desorption spectroscopy (dry residue at 180 °C) levels were measured by gravimetry. Free residual chlorine was analysed by N,N-diethyl-p-phenylenediamine (DPD) colorimetric method. Sensory analysis were carried out by a panel of selected people [29], which were trained according to a flavour profile analysis method previously developed in Devesa *et al.* [30]. Results of these analyses are given in Table S1, which summarises the collected mineral and tap water samples, their mineral composition described as physicochemical parameters and their overall flavour scores. Further information about samples and analysis procedure is given in Platikanov *et al.* [25].

3.2. Climate data set

This climate data set (refer to Figure S1) consisted of 10 total ion current (TIC) GC-MS profiles (2940 retention times) of the neutral fractions extracted from 10 fossil stratified sediment marine samples (**X** predictor variables, **X1**) and of their corresponding alkenone-based reconstructed SST (**y** predicted variable). Sediment marine records were taken from site IODP-U1318C, in the continental margin southwest of UK islands during the Integrated Ocean Drilling Program (IODP) Expedition 307 at Challenger Mound in Porcupine Seabight, 420.9 m below seafloor [31]. The procedures and equipment for extracting, isolating and quantifying the fossil compounds have been described by Villanueva and Grimalt and Villanueva *et al.* [32,33]. Purified extract samples were diluted in toluene and analysed using GC-MS. Mass spectra were acquired in electron ionization mode scanning from *m/z* 42 to 700. Previous to their analysis, the chromatographic profiles of the 10 samples were baseline corrected using asymmetric least squares (AsLS) method [34,35], peak aligned using correlation optimised warping (COW) algorithm [36,37] and mean-centred. Further details of experimental methodology and of data pre-treatment are given in Farrés *et al.* [16]. Annual mean SST values were reconstructed using a global core-top calibration method and the alkenone unsaturation index $U^{k'}_{37}$ [38] and mean-centred prior to PLSR analysis.

Additionally, a second **X** matrix of predictor variables, **X2** (refer to Table S2), was obtained using 77 fossil compounds recognised in the original chromatogram and identified by comparison of their corresponding measured spectra to library data and synthetic standards [16]. Their chromatographic areas were then integrated to estimate the relative concentration of each of the identified organic compounds. This new **X2** matrix was then correlated to the **y** vector having the SST values associated to each of the 10 analysed samples.

3.3. Transcriptome data set

Transcriptome data set consists of the microarray analysis transcriptional stress responses (1207 exposed gene fragments) of 12 *D. magna* female samples (**X** matrix of predictor variables) exposed to sublethal doses of the SSRIs: fluoxetine and fluvoxamine, and their corresponding total offspring production (**y** predicted variables). Tests were performed to determine the effects of the chemicals of interest on transcriptomic responses and reproduction rates on adult stages. Microarrays analyses were performed on isolated RNA from the samples. Microarray results were validated with real-time quantitative polymerase chain reaction . Further information about the experiments is found in Campos *et al.* [26].

4. RESULTS AND DISCUSSION

4.1. Water quality data set results

The application of PLSR to physicochemical and sensory data allowed discrimination among panellist evaluators according to their preferences for different water types.

PLSR was applied first to the water quality data set which include sensorial data values (as **y** variable). **X** (water physicochemical parameters) and **y** (water taste scores) variables were autoscaled. Using leave-one-out cross-validation, PLSR modelling resulted in a two-latent variables model which accounted for 68.43% of the **X** data variance and around 89% of the **y** variance.

Table I shows the pair-wise correlation coefficients (*r*) between water quality physicochemical parameters (column variables of **X** matrix) and water taste scores (**y**-vector), with their level of significance (*p*-value). The selection obtained using VIP and SR approaches is provided in Figure 1 together with the list of variables correlated to the predicted response **y**.

Figure 1, as given in Tran and coauthors [23], shows the comparison of the previously shown correlation coefficients in absolute value, *r* (Table I), with the variables selected using VIP and SR methods. The variables along the *x*-axis are arranged according to their correlation coefficients as it is shown in the map legend at the right side (whiter colour means higher correlation). VIP and SR selected variables are represented in white and non-selected variables are in black.

Table I. Correlation coefficients (*r*) between water quality physicochemical parameters and water taste scores with their significance level (*p*-values) (water quality data set)

Variable	R	*p*-value
Conductivity	−0.8587	0[a]
TDS	−0.8197	0[a]
Cl^-	−0.8403	0[a]
SO_4^{2-}	−0.6105	0.0012
$NO3^-$	−0.798	0[a]
$HCO3^-$	−0.3882	0.0552
Ca^{2-}	−0.6143	0.0011
Mg^{2+}	−0.6433	0.0005
Na^+	−0.7622	0[a]
K^+	−0.6923	0.0001
pH	0.0945	0.6532
Si	−0.0611	0.7716
Cl_2	−0.5156	0.0083

[a]Relevant variables in both selectivity methods (VIP and SR).

Figure 1 shows that almost the same number of variables were selected by the VIP scores and SR methods, considering a threshold greater than one for VIP scores and 2.04 for SR (F-test, 95%). When the correlation coefficient of variables selected by VIP and SR methods were compared, variables with absolute value of *r* higher than 0.76 (Table I) were the same in both variable selection methods (Figure 1). Variable number 10 (Table I), potassium, with an *r* value of −0.69 was selected by the VIP method but not by the SR method. Originally, potassium was in such low concentrations in all water samples that was no perceivable in taste by panellists, only sodium (Table I) was in such a concentration that had effects on taste [25]. In the VIP calculation, because of the **X** variable autoscaling, the weighted sum of squares of the PLSR potassium weights resulted to be almost equal to sodium weights, and VIP scores were similar. The selection of potassium variable by the VIP method should be considered doubtful because of their low correlation coefficient value with **y** compared with other selected variables, but potassium had spurious correlation with sodium, this is the reason why it is considered in the PLS model. For this example, the SR method resulted to be more accurate because it did not select potassium as an important variable.

4.2. Climate data set results

The application of PLSR allowed the identification of organic compounds whose concentrations in sea sediment stratified samples were changing more with SST.

PLSR was initially applied to TIC GC-MS chromatograms of the neutral fractions extracted from 10 fossil sediment marine samples (**X1** variables, with dimensions of 10×2940) and to the reconstructed SST variables, SST values (**y** variable, with dimensions 10×1). Using leave-one-out cross-validation, first PLSR latent variable (LV1) accounted for 64.59% of **X1** data variance

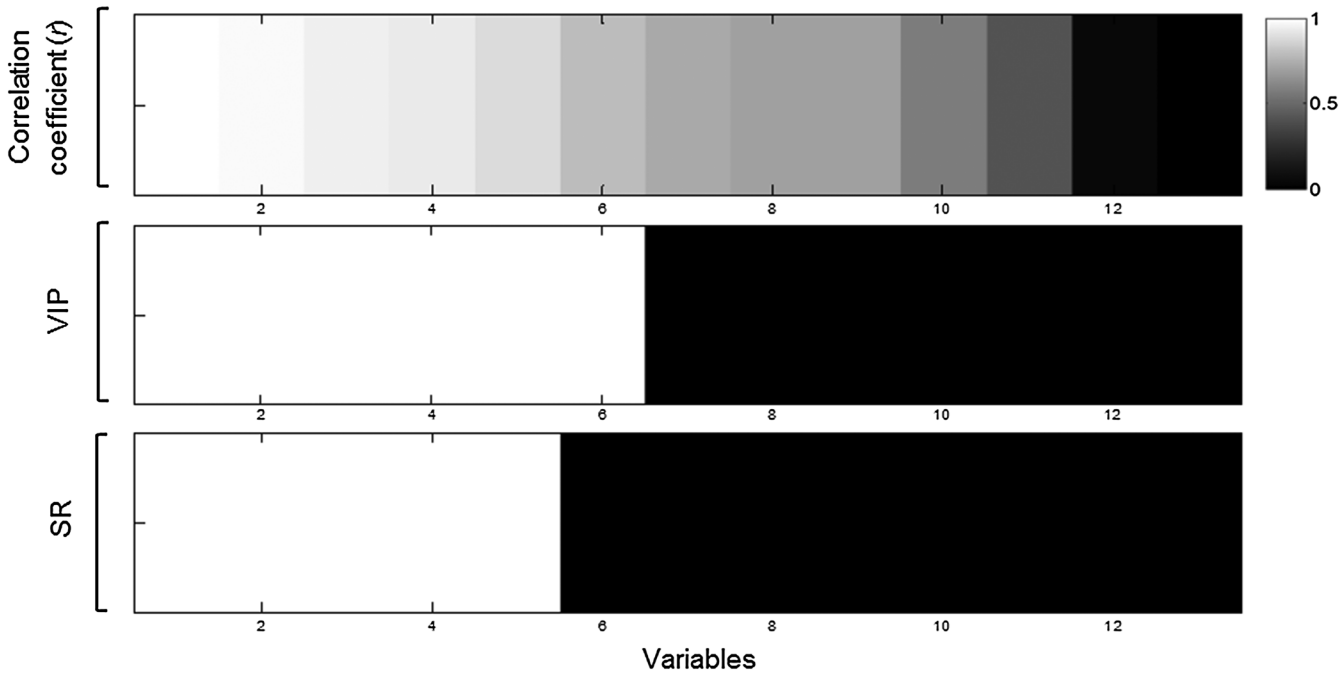


Figure 1. Correlation coefficients (*r*), in absolute value, between water quality physicochemical parameters and water taste scores compared with the variables selected using VIP and SR methods (water quality data set). Selected variables are represented in white and non-selected variables are in black. Explanation of this figure is given in the text.

and as much as 93.01% of the dependent variable **y** (SST) variance. PLSR results confirmed therefore, the presence of a strong correlation pattern between the 10 TIC chromatographic elution profiles of the neutral fraction and the SST changes in sediment samples. Likewise, Matikainen *et al*. [39] performed repeated double cross validation (RDCV) as an alternative method to estimate the number of PLS components.

From the application of PLSR to the preprocessed TIC chromatograms (**X1** matrix), the VIP method selected 50 variables with VIP scores greater than one (refer to Table S3), and the SR method selected 58 variables with SR scores greater than 3.73 (*F*-test, 95%) (refer to Table S4). In Figure 2, the selection obtained using VIP and a SR method is provided, together with the list of variables that should have a relation to the predicted response **y**. All SR variables had absolute *r* values higher than 0.75 whereas there were a subset of the VIP variables which had absolute *r* values lower than 0.6. The SR method gives a higher sensitivity and the VIP method a lower specificity. However, there is a subset of variables selected by the VIP method (Figure 2) with high correlation coefficients with **y** but not selected by the SR method.

When the SR variable selection method was applied to **X1**, SR scores resulted higher when the corresponding correlation coefficients with the **y** variable (refer to Figure 2) were also higher, close to one, regardless the shape of the chromatographic peak. In some cases, significant SR values were obtained at regions of the chromatographic profile where no peak was present. In Figure 3 two examples are given, where TIC chromatogram regions at retention times 11.63 and 12.10 min were chemically meaningless, but SR values were rather high. In a previous work [40] this issue was solved using a threshold value. However, the selection of a threshold value in our case was not easy because it led to the modification of the original chromatographic profiles and it gave biased results. In contrast, in Figure 3, all VIP scores are placed at retention times where, indeed, there are chromatographic peaks.

PLSR analysis was also applied to the autoscaled concentrations of the identified organic compounds in the GC-MS analysis (**X2** matrix of predictor variables, of dimensions of 10×77) of sea sediment samples. Using a leave-one-out cross-validation strategy, PLSR LV1 accounted for 42.70% of the **X2** data variance and for almost 93% of the **y** data variance. In this case, from the 77 variables tested of **X2**, there were 42 with VIP scores greater than one. And, 16 variables were significant according to the *F*-test (95%) for the SR method (refer to Figure 4). All these 16 SR variables were in the same group as the 42 VIP variables. In order to improve the results, significant multivariate correlation (sMC) method was applied for variable selection [23]. Results were similar to those obtained for VIP and they were omitted for brevity.

Correlation coefficients of the 16 variables selected by both, VIP and SR, methods were greater than 0.8 ($p < 0.005$). There were three variables with absolute values of *r* near to one ($p < 0.005$), which were rejected by SR and not by VIP (variables 4, 15 and 73 of concentration **X2** matrix, in Table S5). Two of these variables, identified as nonadecane and

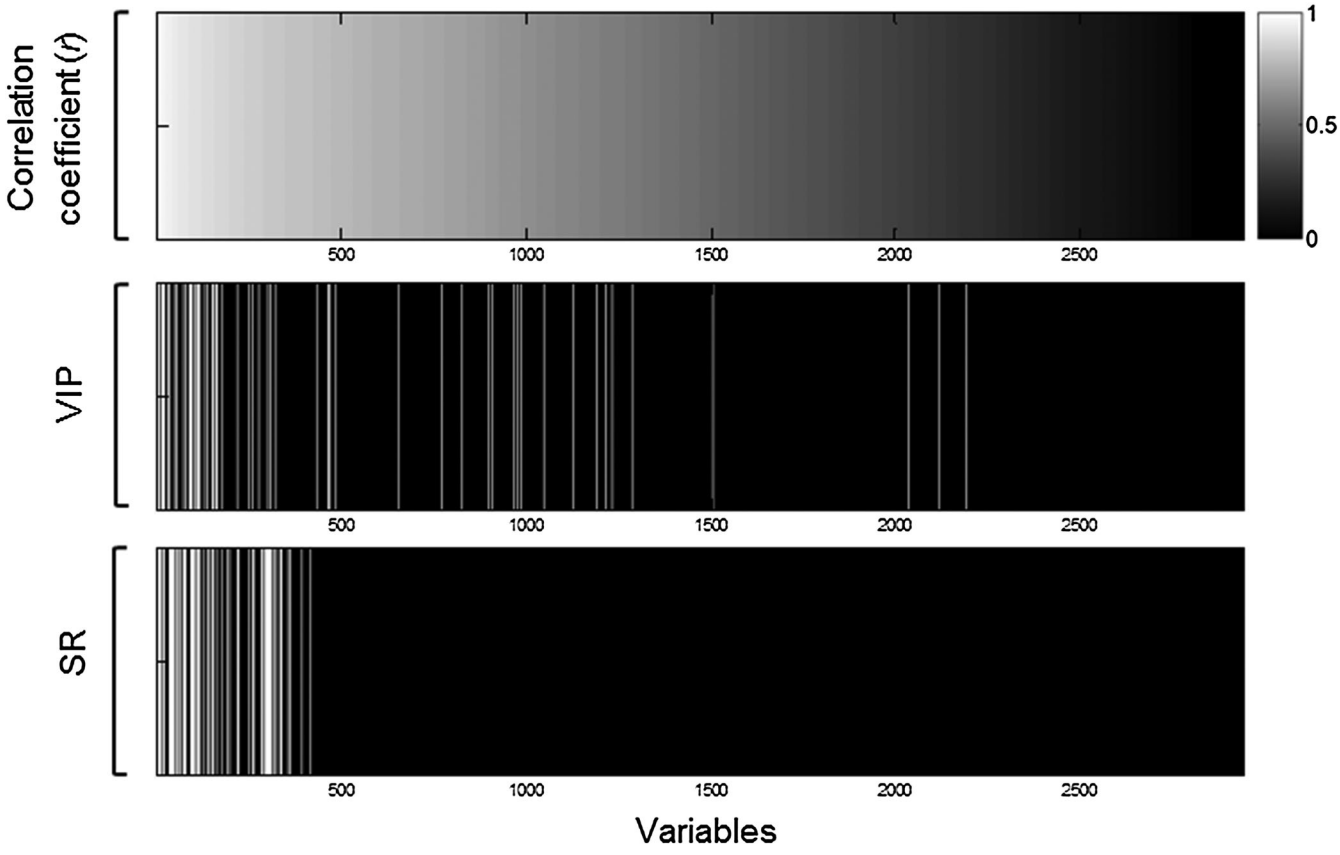


Figure 2. Correlation coefficients (*r*), in absolute value, between TIC chromatograms and sea surface temperature compared with the variables selected using VIP and SR methods (climate data set). Selected variables are represented in white and non-selected variables are in black. Explanation of this figure is given in the text.

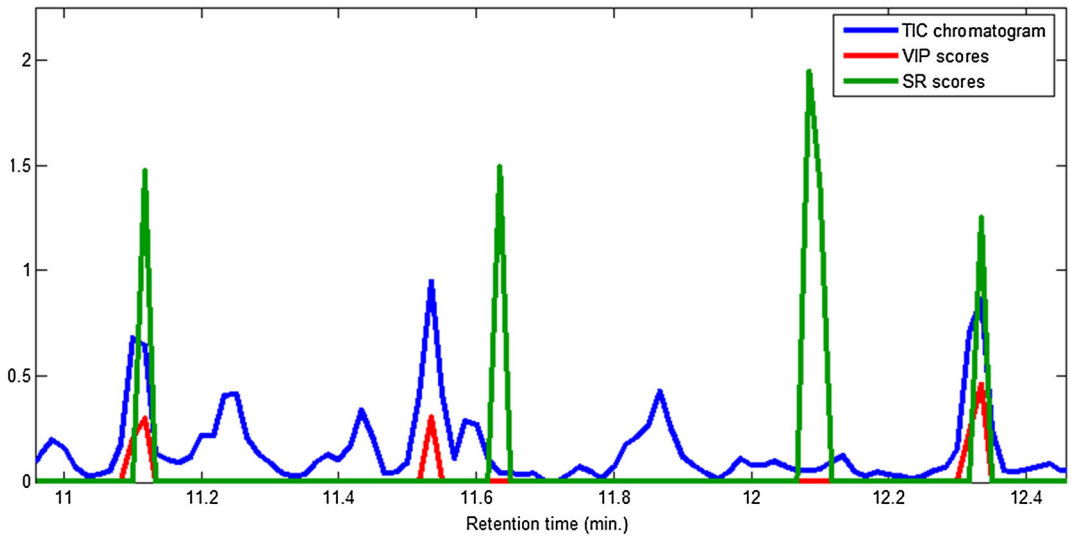


Figure 3. Comparison of a small segment of a representative TIC chromatogram (blue line) with VIP scores (red line) and SR scores (green line), over their threshold values (see in the preceding texts). Values were normalised to a common scale to allow their simultaneous representation.

tetracosane, were relevant for the interpretation of **y** variance (SST changes) [16]. All other variables selected by VIP and not by SR methods presented an absolute *r* value between 0.62 and 0.72 ($p < 0.005$).

When variables selected using VIP method from the GC-MS TIC chromatograms, **X1** matrix, (50 in total) were compared with those selected from the concentrations of the 77 identified organic compounds, **X2** matrix, (42 in total) 24 common variables were encountered to be relevant in both analysis. All 24 variables had an absolute *r* value higher than 0.75 ($p < 0.05$) for the TIC chromatograms, and higher than 0.62 ($p < 0.05$) for the concentrations. When comparing the variables selected by SR scores from TIC chromatograms (58 in total) with those selected from peak areas (16 in total), 11 variables proved to be common in both analyses with absolute *r* values higher than 0.90 ($p < 0.0005$).

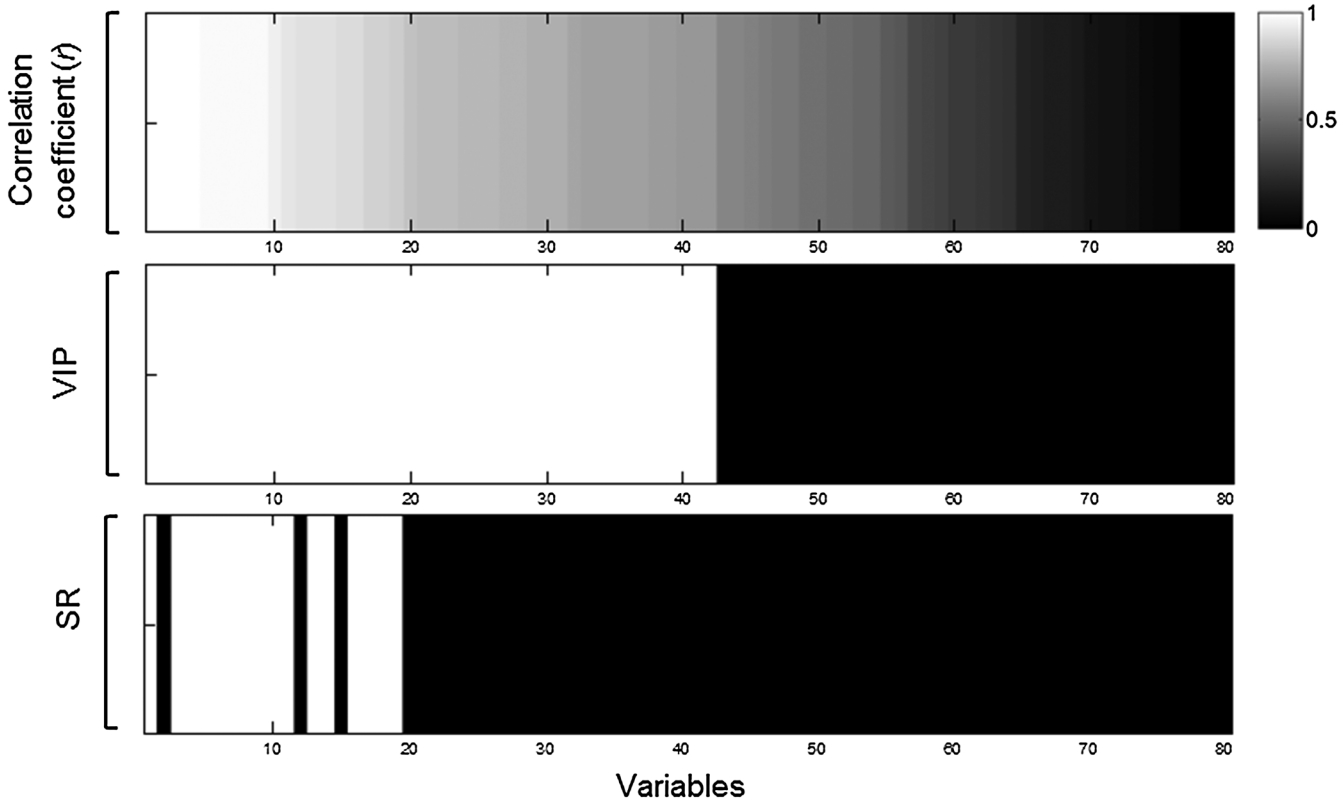


Figure 4. Correlation coefficients (*r*), in absolute value, between identified chromatographic peak areas and sea surface temperature compared with the variables selected using VIP and SR methods (climate data set). Selected variables are represented in white and non-selected variables are in black. Explanation of this figure is given in the text.

To summarise, 11 variables were selected to be important by VIP and SR methods using both the TIC chromatograms data set (**X1**) and the concentrations data set (**X2**) (refer to Table II). Table II shows that short chain *n*-alkanes (eicosane, heneicosane, docosane and tricosane) and long chain *n*-alkanes (from pentacosane to dotriacontane) were in high abundances at lowest SST values.

Short chain *n*-alkanes originate from a variety of sources, such as marine algae and bacteria [41,42]. Long chain *n*-alkanes have a terrestrial source because they are major lipid molecules in epicuticular waxes of vascular plants, and they are common components of eolian winds dust. As discussed in Farrés *et al.* [16], climatic processes at lower temperatures transported higher fluxes of terrigenous material (terrestrial compounds) at IODP U1318C site.

The analysed TIC chromatograms contained many variables (retention times) and many of them had no chemical meaning (not related with chemical compounds). Therefore when they were analysed by PLSR, clear differences between SR and VIP variable selection methods were encountered. Since the SR method is based on a variance ratio evaluation (refer to in the preceding texts) and each retention time is an individual variable, many of the selected variables by SR were not chemically interpretable (Figure 3). Although correlation coefficients of selected variables by SR were also high, they could not be considered to be reliable markers of the sought changes in SST values.

VIP scores obtained in the PLSR analysis of the TIC chromatograms data set (**X1**) with one single component (LV1) were easier to interpret and gave a profile with chromatographic shape. In contrast to variables selected by the SR method (Figure 2), VIP-selected variables had absolute *r* values lower than 0.6 (Figure 2 and Table S3), which means that some of them were less correlated with SST changes. However, the number of relevant variables selected by the SR method was lower for the concentrations data set (**X2**) than for the TIC chromatograms data set (**X1**). This was because in this case, when compound concentrations (peak areas) were used, baseline, background and noisy contributions were discarded and the possibility of false positives was drastically diminished.

Table II. Correlation coefficients (*r*) and *p*-values of identified chemical compounds finally selected by VIP and SR methods from TIC chromatographic and peak area data sets (climate data set)

Compound	Correlation coefficient	*p*-value	Retention time (min)	M^+ (*m/z*)
Eicosane	−0.95	0	11.1	282
Heneicosane	−0.93	0.0001	12.3	296
Docosane	−0.93	0.0001	13.6	310
Tricosane	−0.96	0	14.8	324
Pentacosane	−0.95	0	17.3	352
Hexacosane	−0.94	0.0001	18.5	366
Heptacosane	−0.96	0	19.7	380
Nonacosane	−0.79	0.0069	22	408
Hentriacontane	−0.93	0.0001	24.1	436
Octacosan-1-ol	−0.92	0.0001	24.7	482
Dotriacontane	−0.91	0.0003	25.3	450

Retention times and *m/z* values of the identified chemical compounds are also given.

4.3. Transcriptomic data set results

PLSR was used to investigate the correlations between transcriptome data and total offspring production. And, genes contributing more to the prediction of the reproduction responses were selected and compared using the proposed variable selection methods.

PLSR analysis of the transcriptomic data set consisted of control and SSRIs treated *D.magna* samples. Using leave-one-out cross-validation PLSR modelling resulted in a two-latent variables model that accounted for 57.8% of **X** data variance and explained around 80% of the total **y** variance.

VIP scores greater than 1 and SR scores higher than 3.14 (*F*-test, 95%) were selected and compared. Only 6 variables resulted significant by the SR scores method, with 5 of them coincident with those selected by the VIP scores method (271 in total). Owing to the fact that variable selection methods aim at selecting a small set of very relevant variables and that with a VIP threshold greater than one gave too many variables, VIP threshold value was incremented to 2 and 154 variables resulted important. In this group, the same 5 coincident variables (with VIP and SR method) were again included. Finally, when VIP threshold was incremented up to 3, a total number of 84 variables resulted to be important, but now only 3 of them coincided with those variables selected by SR scores (Figure 3).

Correlation coefficients (*r*) between the variables selected by SR (*F*-test 95%) and VIP scores (threshold of 2 and 3) were calculated (refer to Table S6). Fifty-six variables with VIP scores higher than 2 presented absolutes correlation coefficients of 0.6 or higher ($p < 0.05$), and 42 variables with VIP scores higher than 3 presented absolutes correlation coefficients of 0.6 or higher ($p < 0.05$). It has to be pointed out that some of the variables selected by the VIP scores method even of a threshold of 3, as shown in Figure 5, were little correlated with the independent variable. VIP method finds variables that are important not only because of their possible correlation with the **y** variable but also because they describe significantly the **X** variance [5]. In contrast, the six variables selected by the SR scores method presented always a large absolute correlation coefficient value, higher than 0.64 ($p < 0.05$) (Figure 5). Therefore, in terms of number of false positives, the SR method should be considered a better variable selection method.

In this transcriptomic data example, there were a large number of variables which were selected as relevant using the VIP scores method which were not selected as relevant using the SR method, even though their correlation coefficients with the **y** variables were also high (refer to Table S6). As stated previously by other authors [23], as a consequence of the reduced number of degrees of freedom in the *F*-test used in the SR method a very small number of variables are finally selected, which is sometimes a very conservative decision. A large number of false positives are therefore excluded by the SR variable selection method, yet this can imply the cost of excluding also some relevant variables. Therefore, the number of false negatives may be the main disadvantage of the SR method, as it has been shown for this transcriptomic data set and also previously for the climatic data set (refer to section 3.2), where some highly correlated variables with the **y** vector were discarded by the SR variable selection method.

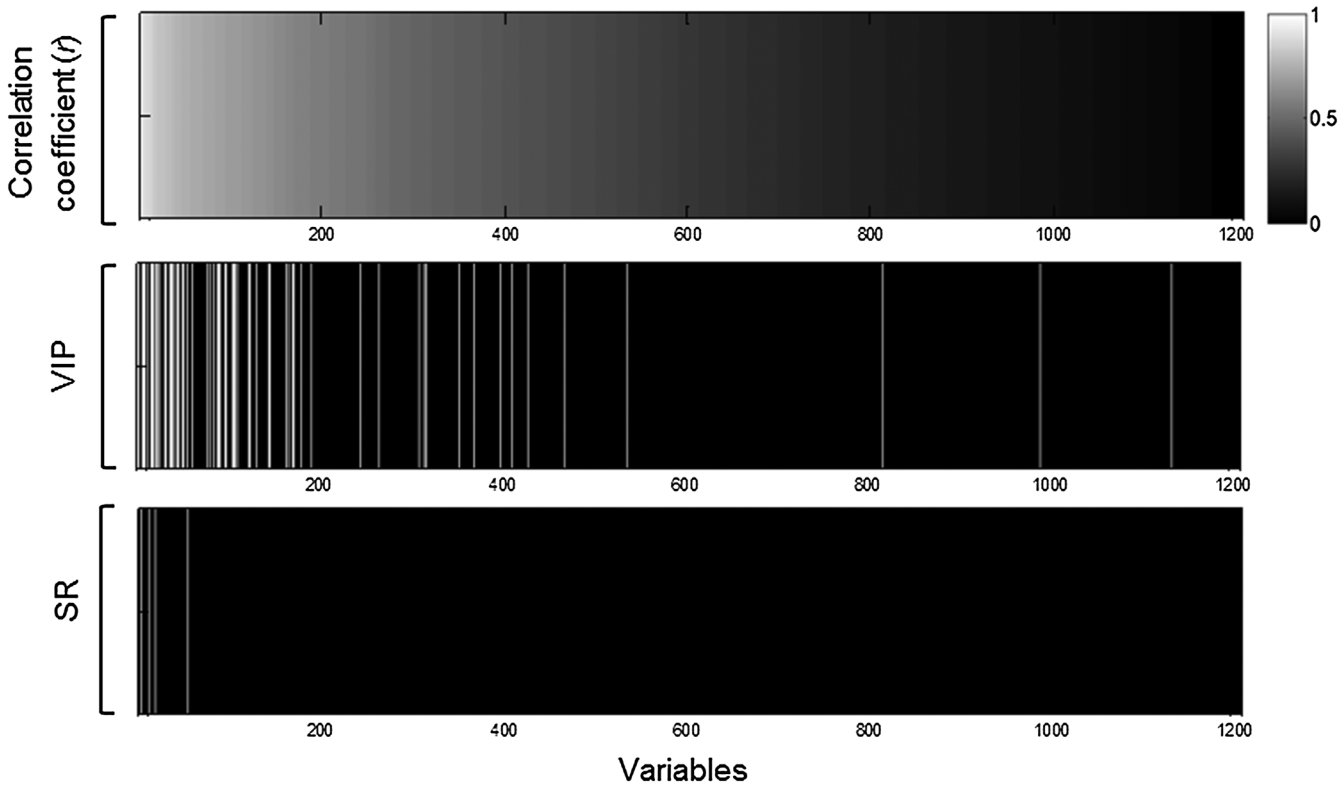


Figure 5. Correlation coefficients (*r*), in absolute value, between exposed genes and total offspring production compared with the variables selected using VIP and SR methods (transcriptomic data set). Selected variables are represented in white and non-selected variables are in black. Explanation of this figure is given in the text.

Two new PLSR analyses were performed using the same data set of this section 3.3. But, in this case, the initial total number of 1207 variables was reduced selecting only those relevant from the previous developed PLSR model. In the first analysis, the VIP selected (threshold of 3) variables were used (**X** matrix of predictor variables of dimensions of 1207×84) and in the second analysis, the SR selected variables were used (**X** matrix of predictor variables of dimensions of 1207×6). PLSR modelling using VIP selected variables resulted in two-latent variables model that accounted for 47.69% of **X** variance and for 93.35% of **y** variance. The second PLSR model using SR selected variables with two latent variables 2 LV, accounted for 89% of **X** variance and for almost 81% of **y** variance. If the aim of the variable selection method is to extract a low number of possible biomarkers of the investigated treatment for prediction (total offspring production of females treated with SSRIs), SR method should be considered more accurate. As clearly shown previously, half of the variables selected by VIP already explained a 93% of **y** variance. Whereas in the case of the SR method, with only 6 variables selected by SR, already a total of 81% of the **y** variance was explained. In terms of transcriptomic data interpretation, 6 genes may not be enough for testing the hypothesis of the study. In the present study, this hypothesis was whether low concentrations of SSRIs affected offspring production and/or juvenile developmental rates though different mechanisms of action [26]. The number of genes involved in *D. magna* metabolic pathways is most probably hundreds, and their identification is a complex task; therefore, selection (by SR method) of only 6 genes is insufficient/scarce for a global interpretation of the changes in the metabolic pathways. In this case the SR threshold could be lowered as it was performed in previous work [26], where those genes with SR values higher than the mean were finally considered to be the most relevant for explaining the offspring.

5. CONCLUSIONS

In this paper, VIP and the SR variable selection methods have been compared for three different data sets. In general, the VIP method selects a higher number of variables than the SR method. In this work, variables selected by the VIP method were sometimes false positive candidates, while those selected by the SR method gave false negative candidates. VIP scores variable selection method was more reliable than the SR method for raw large chromatographic data sets such as the TIC GC-MS data in the climate data set. In contrast, for other types of preprocessed or transformed data sets, like the physicochemical variables data set and the concentrations from the climate data set, both methods detected efficiently the most relevant variables. However, in climate data set SR did not consider some important variables for **y** variance interpretation. In the transcriptomic data set, variables selected by the SR method predicted most the **y** variance, whereas, only a low number of variables selected by VIPs were contributing to the description of the **y** variance. Final decision about the best approach should be performed according to the optimal description of the experimental data and the aim of the variable selection.

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.

REFERENCES

1. Næs T, Martens H. Principal component regression in NIR analysis: viewpoints, background details and selection of components. *J. Chemometr.* 1988; **2**(2): 155–167.
2. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**(0): 1–17.
3. Wold H. *Estimation of Principal Components and Related Models by Iterative Least squares*. In *Multivariate Analysis*, Krishnaiah PR (ed.). Academic Press: New York, 1966; 391–420.
4. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* 2001; **58**(2): 109–130.
5. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J. Chemometr.* 2010; **24**(11–12): 728–737.
6. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemometr. Intell. Lab. Syst.* 2012; **118**(0): 62–69.
7. Wold S, Johansson A, Cochi M (eds). *PLS-partial least squares projections to latent structures*. ESCOM Science Publishers: Leiden, 1993; 523–550.
8. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. Multi- and megavariate data analysis, Part 1, basic principles and applications: Umetrics AB, 2006.
9. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometr. Intell. Lab. Syst.* 2009; **95**(1): 35–48.
10. Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.* 2003; **112**(5–6): 581–592.
11. Modlich O, Prisack H-B, Munnes M, Audretsch W, Bojar H. Predictors of primary breast cancers responsiveness to preoperative Epirubicin/Cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. *J. Transl. Med.* 2005; **3**(1): 32.
12. Serranti S, Cesare D, Marini F, Bonifazi G. Classification of oat and groat kernels using NIR hyperspectral imaging. *Talanta* 2013; **103** (0): 276–284.
13. Eriksson L, Hermens JM, Johansson E, Verhaar HM, Wold S. Multivariate analysis of aquatic toxicity data with PLS. *Aquat. Sci.* 1995; **57**(3): 217–241.
14. Yang J, Zhao X, Liu X, Wang C, Gao P, Wang J, Li L, Gu J, Yang S, Xu G. High performance liquid chromatography−mass spectrometry for metabonomics: potential biomarkers for acute deterioration of liver function in chronic hepatitis B. *J. Proteome Res.* 2006; **5**(3): 554–561.
15. Eriksson L, Gottfries J, Johansson E Wold S. Time-resolved QSAR: an approach to PLS modelling of three-way biological data. *Chemometr. Intell. Lab. Syst.* 2004; **73**(1): 73–84.
16. Farrés M, Martrat B, Mol BD, Grimalt JO, Tauler R. Extraction of climatic signals from fossil organic compounds in marine sediments up to 11.7 Ma old (IODP-U1318). *Anal. Chim. Acta* 2015; **879**(0): 1–9.
17. Galindo-Prieto B, Eriksson L Trygg J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemometr.* 2014; **28**: 623–632.
18. Gomez-Carracedo MP, Ferre J, Andrade JM, Fernandez-Varela R, Boque R. Objective chemical fingerprinting of oil spills by partial least-squares discriminant analysis. *Anal. Bioanal. Chem.* 2012; **403** (7): 2027–2037.
19. Rajalahti T, Kroksveen AC, Arneberg R, Berven FS, Vedeler CA, Myhr K-M, Kvalheim OM. A multivariate approach to reveal biomarker signatures for disease classification: application to mass spectral profiles of cerebrospinal fluid from patients with multiple sclerosis. *J. Proteome Res.* 2010; **9**(7): 3608–3620.
20. Karimi S, Hemmateenejad B. Identification of discriminatory variables in proteomics data analysis by clustering of variables. *Anal. Chim. Acta* 2013; **767**(0): 35–43.
21. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal. Chem.* 2009; **81**(7): 2581–2590.
22. Andries JPM, Heyden YV, Buydens LMC. Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: comparison of properties for ranking. *Anal. Chim. Acta* 2013; **760**(0): 34–45.
23. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemometr. Intell. Lab. Syst.* 2014; **138**(0): 153–160.
24. Krakowska B, Stanimirova I, Orzel J, Daszykowski M, Grabowski I, Zaleszczyk G, Sznajder M. Detection of discoloration in diesel fuel based on gas chromatographic fingerprints. *Anal. Bioanal. Chem.* 2014; **407**: 1–12.
25. Platikanov S, Garcia V, Fonseca I, Rullán E, Devesa R, Tauler R. Influence of minerals on the taste of bottled and tap water: a chemometric approach. *Water Res.* 2013; **47**(2): 693–704.
26. Campos B, Garcia-Reyero N, Rivetti C, Escalon L, Habib T, Tauler R, Tsakovski S, Piña B, Barata C. Identification of metabolic pathways in Daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors and 4-nonylphenol using transcriptomic and phenotypic responses. *Environ. Sci. Technol.* 2013; **47**(16): 9434–9443.
27. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
28. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst.* 2005; **78**(1–2): 103–112.
29. Devesa R, Fabrellas C, Cardeñoso R, Matia L, Ventura F, Salvatella N. The panel of Aigües de Barcelona: 15 years of history. *Water Sci. Technol.* 2004; **49**: 145–151.
30. Devesa R, Cardeñoso R, Matía L. Contribution of the FPA tasting panel to decision making about drinking water treatment facilities. *Water Sci. Technol.* 2007; **55**: 127–135.
31. Ferdelman TG, Kano A, Williams T, Henriet J-P, Scientists E. Expedition 307 Scientists. Site U1318. Proc IODP: Washington, DC (Integrated Ocean Drilling Program Management International, Inc., 2006.
32. Villanueva J, Grimalt JO. Pitfalls in the chromatographic determination of the alkenone U37k index for paleotemperature estimation. *J. Chromatogr. A* 1996; **723**(2): 285–291.
33. Villanueva J, Pelejero C, Grimalt JO. Clean-up procedures for the unbiased estimation of C-37 alkenone sea surface temperatures and terrigenous n-alkane inputs in paleoceanography. *J. Chromatogr. A* 1997; **757**(1–2): 145–151.
34. Eilers PHC. A Perfect Smoother. *Anal. Chem.* 2003; **75**(14): 3631–3636.
35. Eilers P, Boelens H. Baseline correction with asymmetric least squares smoothing, 2005.
36. Nielsen N-PV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998; **805**(1–2): 17–35.
37. Tomasi G, Berg FVD, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometr.* 2004; **18**(5): 231–241.
38. Müller PJ, Kirst G, Ruhland G, von Storch I, Rosell-Melé A. Calibration of the alkenone paleotemperature index U37K' based on core-tops from the eastern South Atlantic and the global ocean (60°N–60°S). *Geochim. Cosmochim. Acta* 1998; **62**(10): 1757–1772.
39. Matikainen M, Rajalahti T, Peltoniemi M, Parvinen P, Juppo A. Determinants of new product launch success in the pharmaceutical industry. *J. Pharmaceut. Innovat.* 2015; **10**(2): 175–189.
40. Kvalheim OM, Chan H-Y, Benzie IFF, Szeto Y-T, Tzang AH-C, Mok DK-W, Chau F-T. Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. *Chemometr. Intell. Lab. Syst.* 2011; **107**(1): 98–105.
41. Blumer M, Guillard RRL, Chase T. Hydrocarbons of marine phytoplankton. *Mar. Biol.* 1971; **8**(3): 183–189.
42. Volkman JK, Johns RB, Gillan FT, Perry GJ, Bavor HJ, Jr. Microbial lipids of an intertidal sediment—I. Fatty acids and hydrocarbons. *Geochim. Cosmochim. Acta* 1980; **44**(8): 1133–1143.

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article at the publisher's website.

SUPPLEMENTARY MATERIAL

Table 1. Physicochemical water parameters of water quality data set samples and the overall score for the water taste and flavour evaluation (Mean liking) of 25 bottled mineral and tap water samples.

Samples	Conductivity (µS/cm)	TDS (mg/L)	Cl^- (mg/L)	SO_4^{2-} (mg/L)	NO_3^- (mg/L)	HCO3− (mg/L)	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	Na^+ (mg/L)	K^+ (mg/L)	pH	Si (mg/L)	Cl_2 (mg/L)	Mean liking
LM·min1	38	26	0.7	1.3	3.5	18	5	1	1.3	0.3	6.7	8.9	0d	6.1
$NaHCO_3$·min_1	278	193	17	11.6	0.4	145.4	12.4	0.8	51.5	1.5	7.3	16	0	6.4
$NaHCO_3$·min_2	539	347	27.4	21.3	0.7	285.1	22.8	1.6	87.7	4	7.9	32.2	0	5.5
$Ca(HCO_3)_2$·min_1	201	121	4.2	12.2	0.9	124.2	21.3	12.6	2.7	0.5	8.1	5	0	6.9
$Ca(HCO_3)_2$·min_2	375	262	7.8	21.9	1.9	285	56.9	25.5	5.3	1.1	8	7.5	0	6.8
$Ca(HCO_3)_2$/$CaSO_4$·min_1	388	287	5.5	109	1.1	115.9	70.4	16.1	1.1	0.8	7.9	4.8	0	6.5
$Ca(HCO_3)_2$/$CaSO_4$·min_2	993	844	7.9	328.9	4.3	399	203.8	43.1	5	1.9	7.5	9.7	0	6.4
$CaSO_4$·min_1	220	189	2.7	86.8	0.2	32.5	35.1	6	0.5	0.3	7.6	0.5	0	6.7
$CaSO_4$·min_2	786	662	3.7	385.8	1.1	122.9	162.9	27.6	2.3	1.1	7.3	3	0	5.5
NaCl·min_1	274	165	74.3	1.2	3.5	25.9	6.3	1	121.6	0.4	6.9	10.9	0	6.1
NaCl·min_2	635	389	193.9	1.5	3.9	20.8	6.2	0.9	31.3	0.4	7	9.6	0	4.9
NaCl/$NaHCO_3$·min_1	289	195	49.6	6.5	1.8	85.6	9.6	1	52.8	0.9	7.2	13	0	6.1
NaCl/$NaHCO_3$·min_2	591	373	112.2	12.1	2.5	164.4	12.2	1.3	116	1.7	7.2	20.8	0	5.5
LM·tap_1	50	45	6.2	4.9	0.8	22.9	5.6	0.8	4.7	0.7	7.2	7.3	0.54	5.3
$Ca(HCO_3)_2$·tap_1	208	128	0.1	3.8	1.7	142.1	42.7	0.2	6.2	1.4	7.5	0.2	0.59	6.1
$Ca(HCO_3)_2$·tap_2	500	346	20.3	17.9	8.4	307	116.7	6.3	8.6	1	7.6	12.2	0.79	5.3
$Ca(HCO_3)_2$/$CaSO_4$·tap_1	988	801	25	311	21	337.5	189.4	52.1	10.9	2.1	7.5	10.4	0.56	5.2
$Ca(HCO_3)_2$/$CaSO_4$·tap_2	1075	893	32.8	405	14.3	283.9	200	44.8	20.6	2.4	7.8	8.3	0.45	4.5

$NaHCO_3$/NaCl/$Ca(HCO_3)_2$ ·tap_1	660	429	107.9	74.8	6.3	158.4	64.7	13	58.7	11.3	7.3	3	0.64	4.8
$NaHCO_3$/$Ca(HCO_3)_2$·tap_1	282	183	17.4	47	4.6	116	38.3	7.8	9.7	2.5	7.6	3.4	0.66	5.9
$NaHCO_3$/$Ca(HCO_3)_2$·tap_2	418	281	30.2	54.3	8.2	164.7	59.7	12.3	18.2	3.1	7.8	4.4	0.62	5.5
HM·tap1 (high mineralization)c	1392	895	288	149	12	241.6	102.4	29	138.9	32.2	7.9	4.6	0.28	3.6
HM·tap_2	1459	1009	302.5	183.6	10.8	171.3	110.9	33.3	165.1	31.5	7.7	4.7	0.52	3.6
HM·tap_3	1700	1346	139.8	401	27.1	362.8	272.9	80.7	140	10.5	7.4	10.4	0.33	3.7
HM·tap_4	2617	1983	506.5	608.4	31.4	334.2	290.1	85.6	286.8	10.1	7.2	12.4	0.63	2.6

Figure 1. Plot of 10 total ion current (TIC) chromatograms (**X1**, 2940 retention times, 55 min.) of the neutral fraction extracted from 10 fossil stratified sediment marine samples.

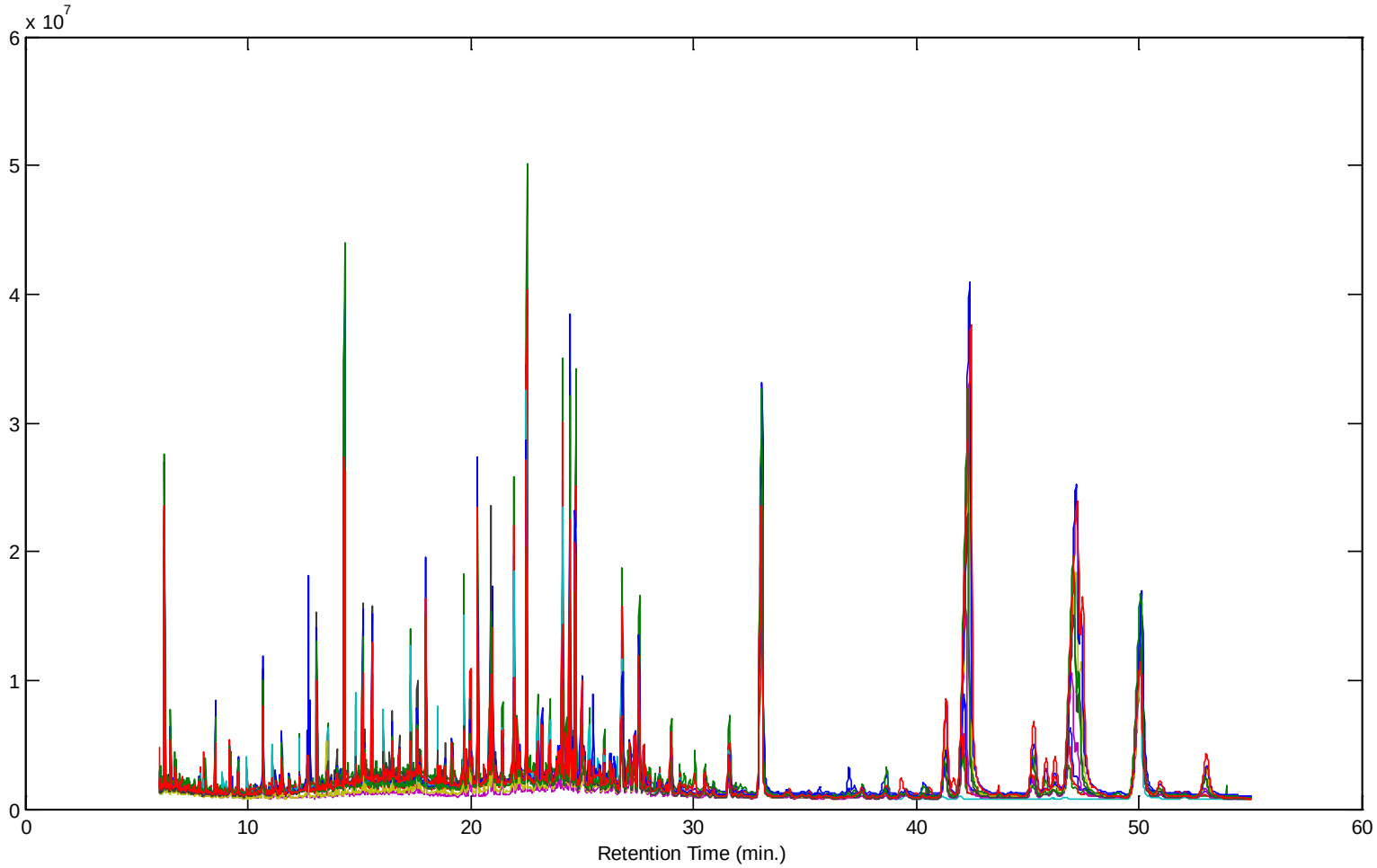


Table 2. Areas of the 77 identified chromatographic peaks from the TIC GC-MS chromatograms (**X2**) and their retention time (RT).

Variable	RT (min.)	1	2	3	4	5	6	7	8	9	10
1	6.27	45	50	51	45	44	55	53	49	52	58
2	6.55	8	7	8	8	8	9	9	8	10	10
3	9.18	10	9	14	3	4	10	10	10	9	14
4	9.95	6	7	8	10	1	1	3	1	1	3
5	10.68	28	26	21	17	10	20	37	27	22	25
6	11.12	8	10	10	12	2	2	5	3	2	5
7	11.87	4	5	4	3	2	3	8	6	5	6
8	12.33	11	15	13	17	2	4	7	4	3	5
9	12.72	53	12	12	6	7	9	15	17	10	12
10	12.80	18	1	1	1	0	1	1	1	1	1
11	13.10	27	34	23	19	14	31	60	40	34	38
12	13.58	13	17	14	18	3	5	10	4	5	8
13	14.85	16	24	19	26	4	4	9	4	4	5
14	15.58	17	31	22	17	14	27	61	45	37	47
15	16.08	12	18	15	21	2	4	12	4	4	6
16	16.80	5	8	6	4	4	6	15	9	8	11
17	17.32	26	43	33	40	6	10	15	8	6	15
18	18.00	41	58	50	35	23	49	56	54	42	62
19	18.53	16	25	21	23	3	6	12	5	4	7
20	19.18	8	13	10	8	5	9	16	11	9	14
21	19.70	42	64	52	56	9	19	21	14	11	21
22	20.32	61	104	76	64	38	66	76	82	61	99
23	20.85	17	24	22	22	20	20	14	5	3	7
24	20.92	12	11	6	6	20	20	103	39	45	33
25	21.43	14	31	21	19	7	11	16	14	11	18
26	21.95	64	104	89	75	19	33	37	26	20	37
27	22.10	11	16	7	11	6	10	11	9	7	21
28	22.52	132	234	179	143	53	82	77	97	69	135
29	23.02	15	27	22	21	7	15	13	10	7	14
30	23.52	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.4	0.2	0.5
31	23.57	19	26	24	22	8	15	11	10	10	19
32	24.12	111	177	157	110	35	54	46	4	30	64
33	24.30	12	19	7	12	4	15	7	12	5	18
34	24.47	48	67	46	36	43	94	198	143	125	108
35	24.59	8	13	12	4	6	9	13	11	8	11
36	24.72	122	180	132	116	50	81	65	88	63	111
37	25.34	1	2	1	2	0	0	0	1	0	1
38	25.35	15	33	21	26	5	9	11	7	5	13
39	26.02	12	22	19	15	7	11	8	10	8	14
40	26.30	8	15	10	6	5	11	10	11	8	13
41	26.47	7	4	5	4	4	7	9	12	7	8
42	26.80	65	117	101	93	31	48	40	67	29	55
43	27.57	62	97	73	70	36	67	56	68	50	79

44	27.80	12	16	20	6	6	18	15	18	13	26
45	28.49	9	9	13	12	8	9	10	6	8	10
46	28.84	3	5	5	2	2	3	3	3	2	4
47	29.02	14	43	31	20	10	21	18	24	17	43
48	29.40	8	15	11	12	4	8	7	7	5	9
49	30.54	10	19	17	13	4	9	9	9	5	13
50	31.62	33	61	45	35	17	31	27	24	19	36
51	34.34	3	4	4	2	2	5	4	4	3	9
52	35.24	3	3	5	1	1	2	1	2	1	2
53	35.67	5	2	2	0	1	2	1	1	1	1
54	36.02	2	2	2	2	1	2	1	1	1	3
55	36.70	3	3	1	1	1	1	1	1	1	1
56	37.00	22	4	8	12	2	7	3	5	3	1
57	37.24	7	7	4	1	1	3	2	0	1	0
58	37.62	5	14	9	7	4	9	10	6	5	14
59	37.95	2	1	2	0	1	1	1	0	0	0
60	38.69	19	33	16	5	5	14	11	8	7	6
61	41.35	88	55	120	3	42	104	66	58	46	162
62	42.15	101	52	78	3	299	990	724	869	664	1323
63	42.69	5	8	3	1	0	0	0	0	0	0
64	43.15	1	4	2	1	0	1	1	0	0	0
65	44.27	3	2	4	1	1	3	2	1	1	4
66	44.69	1	2	4	0	0	2	2	2	2	2
67	45.27	23	20	46	1	30	98	77	76	61	160
68	45.82	35	28	67	1	11	11	4	5	2	6
69	46.24	28	23	69	2	12	23	13	12	7	32
70	46.92	114	56	97	5	324	898	659	758	545	1152
71	47.62	3	7	2	0	0	0	0	0	0	0
72	48.30	2	4	1	0	0	0	0	0	0	0
73	49.14	6	5	3	0	1	4	4	5	5	2
74	51.00	2	1	3	0	5	18	13	12	8	31
75	51.79	4	4	11	0	1	1	0	0	0	0
76	52.17	6	8	11	3	4	4	4	2	4	2
77	53.02	8	4	8	0	22	81	56	53	36	113

Table 3 Correlation coefficients (*r*) and level of significance (*p*-value) between variables selected by VIP method from TIC GC-MS chromatograms and sea surface temperature (**X1** climate data set) and their retention time (RT).

RT (min.)	**r**	***p*-value**
6.25	0.7264	0.0173
8.53	0.6837	0.0293
9.95	-0.9626	0
11.12	-0.9459	0
11.53	0.6633	0.0366

12.33	-0.9306	0.0001
12.70	-0.2495	0.4869
13.08	0.6309	0.0505
13.58	-0.9276	0.0001
14.33	0.2608	0.4667
14.82	-0.9583	0
15.17	0.6938	0.0261
15.23	0.6875	0.028
15.57	0.6582	0.0385
16.08	-0.9187	0.0002
16.47	0.7346	0.0155
17.32	-0.953	0
17.62	0.5721	0.084
18.53	-0.9381	0.0001
19.70	-0.9592	0
19.95	0.6546	0.04
20.00	-0.6468	0.0433
20.33	-0.3477	0.3248
20.85	-0.6347	0.0487
20.92	0.573	0.0834
21.00	0.8291	0.003
21.42	0.7103	0.0213
21.47	-0.8894	0.0006
21.95	-0.9421	0
22.52	-0.787	0.0069
23.04	-0.902	0.0004
23.22	0.8305	0.0029
23.55	-0.9017	0.0004
24.12	-0.9316	0.0001
24.49	0.7948	0.006
24.69	-0.9238	0.0001
24.75	0.4841	0.1563
25.00	0.6351	0.0485
25.35	-0.9105	0.0003
26.80	-0.8681	0.0011
27.57	-0.4591	0.1819
31.64	-0.8103	0.0045
33.14	0.4844	0.156
45.25	0.7457	0.0133
47.10	0.8936	0.0005
47.40	0.8668	0.0012
50.02	0.564	0.0895
53.02	0.8123	0.0043

Table 4. Correlation coefficients (*r*) and level of significance (*p*-value) between variables selected by SR method from TIC GC-MS chromatograms and sea surface temperature (**X1** climate data set) and their retention time (RT).

RT (min.)	R	*p*-value
7.73	0.94	0.0001
7.92	-0.96	0.0000
8.60	0.87	0.0010
8.75	0.85	0.0017
8.88	-0.98	0.0000
9.03	-0.96	0.0000
9.95	-0.96	0.0000
11.12	-0.95	0.0000
11.63	-0.94	0.0000
12.08	-0.98	0.0000
12.33	-0.93	0.0001
13.30	-0.93	0.0001
13.58	-0.93	0.0001
14.82	-0.96	0.0000
16.08	-0.92	0.0002
16.13	-0.91	0.0003
16.85	-0.91	0.0003
17.32	-0.95	0.0000
18.53	-0.94	0.0001
19.28	-0.98	0.0000
19.33	-0.96	0.0000
19.70	-0.96	0.0000
20.43	-0.95	0.0000
21.00	0.83	0.0030
21.35	0.87	0.0012
21.52	-0.95	0.0000
21.57	-0.88	0.0008
21.90	-0.92	0.0002
21.95	-0.94	0.0000
22.82	-0.99	0.0000
23.09	-0.87	0.0011
23.22	0.83	0.0029
23.27	0.84	0.0024
23.34	0.89	0.0006
24.12	-0.93	0.0001
24.69	-0.92	0.0001
24.87	-0.90	0.0004
25.35	-0.91	0.0003
28.20	-0.92	0.0002
28.75	0.84	0.0025

39.09	0.90	0.0003
39.14	0.90	0.0004
43.05	0.79	0.0071
44.89	-0.93	0.0001
44.94	-0.91	0.0002
45.12	0.80	0.0056
45.19	0.80	0.0054
45.52	0.85	0.0019
47.10	0.89	0.0005
47.40	0.87	0.0012
50.27	0.89	0.0006
50.60	0.78	0.0078
53.02	0.81	0.0043
53.27	0.82	0.0035
53.35	0.82	0.0039
53.75	0.85	0.0021
53.80	0.84	0.0022
53.89	0.84	0.0023

Table 5. Correlation coefficients (*r*) and level of significance (*p*-value) between identified chromatographic peak areas and sea surface temperature (**X2** climate data set) and their retention time (RT).

Variable	**RT (min.)**	**R**	***p*-value**
1	6.27	0.42	0.2293
2	6.55	0.66	0.0394
3	9.18	0.04	0.9167
4	9.95	-0.94	0.0001
5	10.68	0.11	0.7712
6	11.12	-0.93	0.0001
7	11.87	0.33	0.3575
8	12.33	-0.92	0.0001
9	12.72	-0.24	0.5029
10	12.80	-0.32	0.3684
11	13.10	0.46	0.1807
12	13.58	-0.89	0.0006
13	14.85	-0.93	0.0001
14	15.58	0.58	0.0819
15	16.08	-0.85	0.0018
16	16.80	0.47	0.1683
17	17.32	-0.93	0.0001
18	18.00	0.08	0.8326
19	18.53	-0.92	0.0001

20	19.18	0.15	0.6812
21	19.70	-0.95	0.0000
22	20.32	-0.15	0.6726
23	20.85	-0.70	0.0247
24	20.92	0.63	0.0499
25	21.43	-0.66	0.0393
26	21.95	-0.93	0.0001
27	22.10	-0.04	0.9062
28	22.52*	-0.81	0.0049
29	23.02*	-0.83	0.0031
30	23.52	0.28	0.4293
31	23.57*	-0.86	0.0014
32	24.12	-0.92	0.0001
33	24.30	-0.19	0.6049
34	24.47	0.71	0.0228
35	24.59	0.05	0.9005
36	24.72	-0.79	0.0062
37	25.34	-0.74	0.0150
38	25.35	-0.85	0.0019
39	26.02	-0.80	0.0056
40	26.30	-0.01	0.9725
41	26.47	0.62	0.0546
42	26.8*	-0.85	0.0019
43	27.57	-0.50	0.1440
44	27.80	0.17	0.6386
45	28.49	-0.66	0.0371
46	28.84	-0.47	0.1673
47	29.02	-0.26	0.4724
48	29.40	-0.76	0.0104
49	30.54	-0.74	0.0136
50	31.62	-0.73	0.0172
51	34.34	0.29	0.4183
52	35.24	-0.65	0.0403
53	35.67	-0.35	0.3237
54	36.02	-0.41	0.2432
55	36.70	-0.52	0.1227
56	37.00	-0.58	0.0815
57	37.24	-0.66	0.0384
58	37.62	-0.15	0.6880
59	37.95	-0.52	0.1245
60	38.69	-0.55	0.1026
61	39.59	0.00	0.9943
62	40.37	-0.49	0.1513
63	42.69	-0.75	0.0132

64	43.15	-0.70	0.0250
65	44.27	-0.27	0.4515
66	44.69	-0.10	0.7792
67	45.27	0.64	0.0442
68	45.82	-0.71	0.0213
69	46.24	-0.47	0.1660
70	46.92	0.83	0.0027
71	47.62	-0.64	0.0456
72	48.30	-0.63	0.0492
73	49.14	0.09	0.7942
74	51.00	0.67	0.0338
75	51.79	-0.72	0.0179
76	52.17	-0.73	0.0164
77	53.02	0.74	0.0144

Table 6. Correlation coefficients (*r*), in absolute value, in descending order and level of significance (*p*-values) between exposed genes selected (by VIP and SR methods) according to a threshold and total offspring production (transcriptomic data set).

VIP						**SR**	
treshold 1		treshold 2		threshold 3		Threshold 3.14 F-test (95%)	
R	*p*-value	R	*p*-value	R	*p*-value	R	*p*-value
0.8752	0.0002	0.8752	0.0002	0.8752	0.0002	0.8752	0.0002
0.8277	0.0009	0.8277	0.0009	0.8277	0.0009	0.788	0.0023
0.8011	0.0017	0.8011	0.0017	0.8011	0.0017	0.7872	0.0024
0.7995	0.0018	0.7995	0.0018	0.7995	0.0018	0.7548	0.0045
0.788	0.0023	0.788	0.0023	0.7665	0.0036	0.7174	0.0086
0.7872	0.0024	0.7872	0.0024	0.7646	0.0038	0.643	0.0241
0.7792	0.0028	0.7665	0.0036	0.7577	0.0043		
0.7665	0.0036	0.7646	0.0038	0.7548	0.0045		
0.7646	0.0038	0.7577	0.0043	0.7525	0.0047		
0.7577	0.0043	0.7548	0.0045	0.7211	0.0081		
0.7576	0.0043	0.7525	0.0047	0.7193	0.0084		
0.7548	0.0045	0.7497	0.005	0.7174	0.0086		
0.7525	0.0047	0.7278	0.0073	0.7169	0.0087		
0.7497	0.005	0.7211	0.0081	0.71	0.0097		
0.7278	0.0073	0.7193	0.0084	0.7087	0.0099		
0.724	0.0078	0.7174	0.0086	0.7057	0.0103		
0.7211	0.0081	0.7169	0.0087	0.6999	0.0113		
0.7193	0.0084	0.71	0.0097	0.6987	0.0115		
0.7174	0.0086	0.7087	0.0099	0.6893	0.0132		

0.7169	0.0087	0.7057	0.0103	0.6882	0.0134
0.7109	0.0095	0.7017	0.011	0.6835	0.0143
0.71	0.0097	0.6999	0.0113	0.6802	0.0149
0.7087	0.0099	0.6987	0.0115	0.6766	0.0157
0.7057	0.0103	0.6921	0.0126	0.6749	0.0161
0.7017	0.011	0.6893	0.0132	0.6682	0.0175
0.6999	0.0113	0.6882	0.0134	0.663	0.0188
0.6987	0.0115	0.6835	0.0143	0.6627	0.0189
0.6982	0.0116	0.6815	0.0147	0.6575	0.0201
0.6932	0.0124	0.6802	0.0149	0.6563	0.0205
0.6921	0.0126	0.6766	0.0157	0.6497	0.0222
0.6899	0.013	0.6749	0.0161	0.6464	0.0231
0.6893	0.0132	0.6682	0.0175	0.6425	0.0243
0.6882	0.0134	0.663	0.0188	0.637	0.0259
0.6876	0.0135	0.6627	0.0189	0.626	0.0294
0.6835	0.0143	0.6575	0.0201	0.6229	0.0305
0.6815	0.0147	0.6563	0.0205	0.6191	0.0318
0.6802	0.0149	0.6517	0.0217	0.6175	0.0324
0.6766	0.0157	0.6497	0.0222	0.613	0.0341
0.6749	0.0161	0.6464	0.0231	0.606	0.0367
0.6682	0.0175	0.6425	0.0243	0.6055	0.0369
0.6646	0.0184	0.6405	0.0248	0.6028	0.038
0.663	0.0188	0.6402	0.0249	0.6012	0.0387
0.6627	0.0189	0.637	0.0259	0.5994	0.0394
0.6608	0.0193	0.6305	0.028	0.5956	0.041
0.6575	0.0201	0.627	0.0291	0.5954	0.0411
0.6563	0.0205	0.626	0.0294	0.5928	0.0422
0.6517	0.0217	0.6229	0.0305	0.5905	0.0432
0.6497	0.0222	0.6191	0.0318	0.5896	0.0436
0.6464	0.0231	0.6175	0.0324	0.5891	0.0439
0.6451	0.0235	0.614	0.0337	0.5887	0.044
0.6443	0.0237	0.613	0.0341	0.5885	0.0441
0.6425	0.0243	0.606	0.0367	0.5771	0.0495
0.6405	0.0248	0.6055	0.0369	0.5748	0.0506
0.6402	0.0249	0.6028	0.038	0.5737	0.0511
0.6388	0.0253	0.6012	0.0387	0.571	0.0525
0.6386	0.0254	0.6011	0.0387	0.5486	0.0647
0.637	0.0259	0.5994	0.0394	0.5472	0.0656
0.6342	0.0268	0.5956	0.041	0.5235	0.0807
0.6328	0.0272	0.5954	0.0411	0.5201	0.083
0.6305	0.028	0.5928	0.0422	0.5189	0.0839
0.627	0.0291	0.5905	0.0432	0.5182	0.0844
0.626	0.0294	0.5896	0.0436	0.5178	0.0847
0.6244	0.03	0.5891	0.0439	0.5123	0.0886

0.6237	0.0302	0.5887	0.044	0.5013	0.0969
0.6229	0.0305	0.5885	0.0441	0.4914	0.1047
0.6191	0.0318	0.5799	0.0481	0.4788	0.1153
0.6175	0.0324	0.5771	0.0495	0.4682	0.1247
0.6157	0.033	0.5748	0.0506	0.4475	0.1446
0.614	0.0337	0.5737	0.0511	0.4332	0.1595
0.613	0.0341	0.5729	0.0515	0.4141	0.1809
0.6113	0.0347	0.5723	0.0518	0.411	0.1844
0.606	0.0367	0.571	0.0525	0.4107	0.1848
0.6055	0.0369	0.5691	0.0534	0.3926	0.2068
0.6028	0.038	0.5671	0.0545	0.3845	0.2172
0.6025	0.0381	0.5578	0.0595	0.3673	0.2402
0.6012	0.0387	0.5559	0.0606	0.3597	0.2508
0.6012	0.0387	0.5486	0.0647	0.3496	0.2652
0.6011	0.0387	0.5472	0.0656	0.3294	0.2958
0.5994	0.0394	0.5459	0.0664	0.2936	0.3544
0.5956	0.041	0.5452	0.0668	0.1767	0.5828
0.5954	0.0411	0.5396	0.0702	0.1503	0.6411
0.5928	0.0422	0.5394	0.0703	0.1243	0.7002
0.5928	0.0422	0.5352	0.0729	0.0877	0.7863
0.5926	0.0423	0.5235	0.0807	0.0287	0.9294
0.5905	0.0432	0.5201	0.083		
0.5896	0.0436	0.5189	0.0839		
0.5891	0.0439	0.5182	0.0844		
0.5887	0.044	0.5178	0.0847		
0.5885	0.0441	0.5123	0.0886		
0.5874	0.0446	0.5108	0.0897		
0.5861	0.0452	0.5065	0.0929		
0.582	0.0471	0.5013	0.0969		
0.5799	0.0481	0.496	0.101		
0.5787	0.0487	0.4935	0.103		
0.5771	0.0495	0.4914	0.1047		
0.5748	0.0506	0.491	0.105		
0.5737	0.0511	0.4869	0.1084		
0.5729	0.0515	0.4859	0.1092		
0.5723	0.0518	0.484	0.1109		
0.5711	0.0524	0.4813	0.1132		
0.571	0.0525	0.48	0.1143		
0.5691	0.0534	0.4788	0.1153		
0.5671	0.0545	0.4787	0.1154		
0.5667	0.0547	0.4749	0.1187		
0.5626	0.0569	0.4707	0.1225		
0.5578	0.0595	0.47	0.1231		
0.5571	0.0599	0.4682	0.1247		

0.5559	0.0606	0.4661	0.1266
0.5554	0.0608	0.46	0.1324
0.5486	0.0647	0.4498	0.1424
0.5472	0.0656	0.4475	0.1446
0.5459	0.0664	0.4372	0.1552
0.5452	0.0668	0.4344	0.1582
0.5396	0.0702	0.4336	0.1591
0.5394	0.0703	0.4332	0.1595
0.5352	0.0729	0.4279	0.1653
0.5348	0.0732	0.4275	0.1657
0.5341	0.0737	0.42	0.1741
0.5339	0.0738	0.4186	0.1757
0.5329	0.0744	0.4152	0.1796
0.5237	0.0805	0.4141	0.1809
0.5235	0.0807	0.4118	0.1834
0.5201	0.083	0.411	0.1844
0.5189	0.0839	0.4107	0.1848
0.5182	0.0844	0.3997	0.198
0.5178	0.0847	0.3926	0.2068
0.5155	0.0863	0.3896	0.2106
0.5152	0.0865	0.3845	0.2172
0.5142	0.0872	0.3792	0.2241
0.5123	0.0886	0.3673	0.2402
0.5123	0.0886	0.3663	0.2416
0.5108	0.0897	0.3597	0.2508
0.5065	0.0929	0.3497	0.2651
0.5013	0.0969	0.3496	0.2652
0.5009	0.0972	0.3496	0.2653
0.4982	0.0993	0.3492	0.2659
0.4963	0.1008	0.3294	0.2958
0.496	0.101	0.3219	0.3075
0.4935	0.103	0.3197	0.311
0.4914	0.1047	0.2936	0.3544
0.491	0.105	0.2927	0.3559
0.4892	0.1065	0.2573	0.4196
0.4869	0.1084	0.2371	0.4581
0.4859	0.1092	0.2264	0.4792
0.4847	0.1103	0.1991	0.5349
0.484	0.1109	0.1989	0.5354
0.4813	0.1132	0.1767	0.5828
0.48	0.1143	0.1739	0.5889
0.4798	0.1144	0.1503	0.6411
0.4788	0.1153	0.1251	0.6984
0.4787	0.1154	0.1243	0.7002

0.4749	0.1187	0.0877	0.7863
0.4748	0.1189	0.0287	0.9294
0.4707	0.1225	0.0017	0.9958
0.4701	0.123		
0.47	0.1231		
0.4682	0.1247		
0.4661	0.1266		
0.4623	0.1302		
0.46	0.1324		
0.4562	0.136		
0.4498	0.1424		
0.4475	0.1446		
0.4411	0.1512		
0.4377	0.1547		
0.4375	0.155		
0.4372	0.1552		
0.4344	0.1582		
0.4336	0.1591		
0.4332	0.1595		
0.4322	0.1606		
0.4304	0.1625		
0.4295	0.1635		
0.4288	0.1642		
0.4279	0.1653		
0.4275	0.1657		
0.4255	0.1679		
0.42	0.1741		
0.4186	0.1757		
0.4159	0.1787		
0.4152	0.1796		
0.4141	0.1809		
0.4141	0.1808		
0.4118	0.1834		
0.411	0.1844		
0.4107	0.1848		
0.4094	0.1863		
0.4086	0.1872		
0.3997	0.198		
0.3988	0.1991		
0.3966	0.2018		
0.3965	0.202		
0.3936	0.2055		
0.3926	0.2068		
0.3896	0.2106		

0.3845	0.2172
0.3828	0.2194
0.3792	0.2241
0.3773	0.2266
0.3738	0.2313
0.3724	0.2332
0.3687	0.2382
0.3673	0.2402
0.3663	0.2416
0.3646	0.2439
0.3609	0.2491
0.3604	0.2499
0.3597	0.2508
0.3577	0.2536
0.3497	0.2651
0.3496	0.2652
0.3496	0.2653
0.3492	0.2659
0.3477	0.2681
0.3461	0.2705
0.3443	0.2731
0.3307	0.2938
0.3306	0.2938
0.3294	0.2958
0.3245	0.3034
0.324	0.3043
0.3219	0.3075
0.3205	0.3097
0.3197	0.311
0.3184	0.3131
0.3137	0.3207
0.3064	0.3327
0.2962	0.35
0.2937	0.3541
0.2936	0.3544
0.2927	0.3559
0.2915	0.3579
0.2894	0.3617
0.2746	0.3877
0.2729	0.3907
0.2698	0.3964
0.2676	0.4003
0.2573	0.4196
0.2413	0.4499

0.2382	0.456
0.2371	0.4581
0.2331	0.466
0.2264	0.4792
0.1991	0.5349
0.1989	0.5354
0.18	0.5757
0.1767	0.5828
0.1739	0.5889
0.1503	0.6411
0.147	0.6484
0.1439	0.6555
0.1363	0.6726
0.1359	0.6735
0.1316	0.6836
0.1252	0.6982
0.1251	0.6984
0.1243	0.7002
0.119	0.7126
0.1084	0.7374
0.1071	0.7405
0.1012	0.7544
0.0991	0.7594
0.0955	0.7678
0.0881	0.7855
0.0877	0.7863
0.0876	0.7867
0.0823	0.7994
0.0605	0.8517
0.0287	0.9294
0.0077	0.981
0.0017	0.9958

6.1.1 DISCUSSIÓ DELS RESULTATS OBTINGUTS

1) L'anàlisi amb validació creuada (*leave-one-out*) del conjunt de dades de qualitat de l'aigua autoescalades va resultar en un model de dues variables latents que explicaven un 68.43% de la variància de la matriu **X** i un 89% de la variància de **y**. A la figura 1 de l'Article 5 es pot comprovar que els dos mètodes de selecció, VIP i SR, van identificar gairebé les mateixes variables tenint en compte els valors llindar de cada un dels mètodes de selecció, més gran que u per al mètode VIP i de 2.04 per al mètode SR (test *F*, 95%). Només la variable corresponent al potassi va ser seleccionada pel mètode VIP i no pel SR. La selecció del potassi com a variable rellevant és qüestionable, ja que el valor absolut del seu coeficient de correlació amb la variable sensorial del gust de l'aigua (**y**) és baix si es compara amb el que tenen les altres variables (vegeu taula 1, Article 5). El valor relativament alt del seu VIP indicaria l'existència en aquest cas d'una correlació espúria d'aquesta variable sensorial amb el potassi.

2) Respecte el grup de dades climàtiques, en primer lloc es va aplicar el procediment PLSR a als cromatogrames TIC de la fracció neutre dels compostos orgànics (matriu **X1** de dimensions 10 x 2940), i en segon lloc es va aplicar només a les concentracions de 77 compostos orgànics prèviament identificats i quantificats (matriu **X2** de dimensions 10 x 77). En els dos casos es va investigar la seva correspondència amb les diferents SST associades a cada una de les deu mostres analitzades (vector **y**).

En el primer cas (matriu **X1**) amb una sola variable latent n'hi va haver prou per explicar un 65.51% de la variància de **X1** i un 93.01% de la variància de **y** (SST). En aquest cas, els VIP amb un valor llindar de u van permetre seleccionar 50 variables i el SR, amb un valor llindar de 3.73 (test *F*, 95%), va destacar 58 variables (vegeu figura 2, Article 5). Totes les variables determinades pel mètode SR tenien valors absoluts del coeficients de correlació (*r*) en relació a les SST majors a 0.75, mentre que algunes de variables seleccionades pel mètode VIP tenien valors absoluts de *r* menors a 0.6. En aquest cas es pot afirmar que el mètode SR dóna una major selectivitat i el mètode VIP una menor especificitat. No obstant això, quan el mètode SR es va aplicar a la matriu **X1** formada per perfils cromatogràfics TIC (*Total Ion Current*), algunes de les regions seleccionades (temps de retenció) no tenien forma de pic cromatogràfic reconeixible (vegeu figura 3, Article 5). Conseqüentment, moltes de les variables seleccionades pel mètode SR no es van

poder considerar com a rellevants ja que mancaven de significat químic. En canvi, a la mateixa figura 3 de l'Article 5 es pot observar com les variables seleccionades pel mètode VIP sí que es corresponien de forma adequada amb els pics cromatogràfics obtinguts ja que el gràfic resultant dels VIP tenia mimetisme amb els perfils TIC GC-MS originals.

En el segon cas (matriu **X2**) una variable latent va explicar un 42.7% de la variància de **X2** i gairebé el 93% de la variància de **y**. En aquest cas, 42 variables van ser seleccionades pel mètode VIP i 16 pel mètode SR (vegeu figura 4, Article 5). Les 16 variables seleccionades pel mètode SR formaven part d'un subconjunt determinat de les variables destacades pels VIP. Els valors absoluts dels coeficients de correlació en relació a les SST d'aquestes 16 variables eren majors a 0.8 ($p < 0.05$). El fet que el mètode SR en aquest cas seleccionés un nombre de variables menor que el mètode VIP es deu a que quan es van seleccionar concentracions de compostos (àrees dels pics cromatogràfics) autoescalades, la línia de base, el soroll de fons i altres contribucions en els cromatogrames TIC s'eliminen i la possibilitat de falsos positius es va reduir dràsticament. Hi va haver tres variables que van ser seleccionades pel mètode VIP i rebutjades pel SR, amb un valor absolut de *r* molt pròxim a 1. Dues d'aquestes variables no destacades pel mètode SR van ser identificades com a nonadecane i tetracosane, les quals representaven compostos rellevants per a la interpretació dels canvis en la temperatura superficial del mar (SST) (vegeu capítol 5). Tal i com altres autors (Tran et al., 2014) ja van posar de manifest anteriorment, com a conseqüència de l'ús d'un nombre reduït de graus de llibertat al test *F*, el mètode SR selecciona un nombre molt baix de variables, i això de vegades pot suposar una decisió massa conservadora, i conseqüentment el nombre de falsos negatius pot ser el principal desavantatge del mètode SR.

3) El resultat de l'aplicació del PLSR de la matriu de dades de xips d'ADN en relació a les dosis d'estrès (SSRI) va donar un model de dues variables latents que explicaven un 57.8% de la variància de **X** i un 80% de la variància de **y**. En aquest cas, després de l'aplicació del mètode de selecció de variables VIP amb un valor llindar de u, el nombre de gens seleccionat com a importants va ser molt elevat (271 en total) comparat amb el nombre de gens seleccionat pel mètode SR (test *F*, 95%) que va ser de només 6. L'objectiu de la selecció de variables és triar un grup reduït de variables rellevants, per tant, el valor llindar dels VIP es va incrementar fins a 3. En aquest cas es van seleccionar un total de 84 variables, només 3 de les quals coincidien amb els gens seleccionats pel SR (vegeu figura 5, Article 5). De les 84 variables seleccionades pels VIP (valor llindar de 3), només 42 presentaven un coeficient de correlació absolut major de 0.6 ($p <$

0.05). En el mètode VIP, no només són importants les variables que tenen una bona correlació amb la variable **y**, sinó també aquelles que descriuen de manera significativa la variància de **X** (Andersen i Bro, 2010). En aquest cas s'il·lustra que en termes del nombre de falsos positius, el mètode SR és més adequat per a la selecció de variables en dades de xips d'ADN. Tot i això, tal i com s'ha esmentat en els resultats de les dades climàtiques (matriu **X2**) i en el treball de Tran i coautors (Tran et al., 2014), el nombre falsos negatius representen un desavantatge del mètode SR.

En aquest mateix estudi es van analitzar dues noves matrius, una amb les variables seleccionades pels VIP amb un llindar de 3 (matriu de dimensions 84 x 1207) i una altra amb les variables seleccionades pel mètode SR (matriu de dimensions 6 x 1207) i es van generar dos nous models PLSR en relació a les dosis de SSRI (**y**). El model resultant de les variables dels VIP va constar de dues variables latents que explicaven el 47.69% de la variància de **X** i el 93.35% de la variància de **y**. El model PLSR de les variables seleccionades pel mètode SR també va resultar en dues variables latents que explicaven el 89% de la variància de **X** i gairebé un 81% de la variància de **y**. Tenint en compte aquests resultats i el fet els mètodes de selecció de variables pretenen extraure un nombre reduït dels millors marcadors que expliquin els canvis en la variable dependent, el mètode SR es va considerar més adequat. El nombre de gens implicats en les vies metabòliques de l'organisme *D.magna* és probablement de centenars i la seva identificació és una tasca complexa. Per tant, la selecció de només 6 gens era massa petita per a la finalitat buscada. En aquest cas, es suggeria rebaixar el llindar dels valors de SR tal i com ja s'havia fet en treballs anteriors (Campos et al., 2013), on es van seleccionar finalment com a rellevants aquells gens que tenien un valor SR superior a la mitjana de tots ells.

A partir dels resultats obtinguts en aquest capítol i dels resultats de l'estudi dels perfils cromatogràfics de llevat exposats a Cu(II) (Article 3), es clou que el nivell de tall generalment proposat per als VIP, més gran que u (Chong i Jun, 2005), no és un valor adequat per tots els casos, ja que en molts casos particulars és necessari augmentar els valors d'aquest llindar per sobre de u o inclús per sobre de dos. Per exemple, en el cas de les dades de transcriptòmica (Article 5) va ser necessari elevar el valor llindar fins a 3 per tenir finalment un grup adequat de variables rellevants.

En tots els casos estudiats, el mètode SR va seleccionar les variables amb un valor absolut alt del coeficient de correlació amb la variable **y**, tal i com es pot observar a les Figures 1, 4 i 5 de l'Article 5.

Com a compendi dels resultats obtinguts en els diferents exemples d'aplicació dels mètodes de selecció de variables VIP i SR, es pot apuntar que la sensibilitat (fracció de positius reals que s'identifiquen correctament) del mètode SR és major que la del mètode VIP, tal i com es conclou també en el treball de Krakowska i coautors (Krakowska et al., 2014) descrit a la introducció de l'Article 5. Per altra banda, el mètode VIP pot detectar moltes variables veritablement rellevants en totes les situacions considerades, però també pot seleccionar incorrectament moltes variables irrellevants, per tant, l'especificitat (proporció de negatius que s'identifiquen correctament) d'aquest mètode tendeix a ser baixa tot i tenir una bona sensibilitat.

Capítol 7

Conclusions

L'anàlisi dirigida (*target*) i no dirigida (*untarget*) de dades metabolòmiques obtingudes per cromatografia líquida o de gasos amb detecció per espectrometria de masses, conjuntament amb l'aplicació de diferents mètodes quimiomètrics, ha permès l'extracció d'informació química i bioquímica dels diversos sistemes experimentals estudiats en aquesta Tesi. En aquestes anàlisis s'ha aprofundit en la selecció de compostos marcadors dels efectes produïts per diferents condicions ambientals o factors estressants aplicats a diferents tipus de mostres naturals.

Les conclusions específiques referents als estudis realitzats són:

Estudi dels productes al·leloquímics del blat:

1) S'han demostrat els avantatges de l'aplicació dels mètodes ANOVA-PCA i ASCA en l'anàlisi dirigida dels metabòlits secundaris (al·leloquímics) obtinguts de la planta del blat per a l'avaluació de diferents factors externs (estadi de creixement, tipus de cultiu, varietat i tipus de mostra) a partir de dades TIC LC-MS.

2) Els factors externs que han tingut efectes significatius en el cultiu de la planta del blat són: estadi de creixement, tipus de mostra i interacció entre l'estadi de creixement i tipus de mostra.

3) La concentració dels metabòlits al·leloquímics DIMBOA-Glc, DIMBOA, HMBOA i MBOA ha canviat de forma sistemàtica durant el creixement de la planta del blat en relació als factors significatius. La primera etapa de creixement de la planta del blat ha presentat molta variabilitat natural amb independència dels altres factors considerats. En les següents etapes de creixement, els factors estudiats han tingut una major influència sobre les concentracions dels metabòlits al·leloquímics investigats. En les etapes més tardanes de creixement, les mostres obtingudes dels brots i de les arrels de la planta del blat han estat manifestament diferenciades entre si en relació als metabòlits investigats.

Estudi de dades metabolòmiques del llevat:

1) S'ha desenvolupat un mètode d'anàlisi dels metabòlits del llevat a partir de la seva separació per cromatografia líquida mitjançant columnes d'interacció hidrofílica (HILIC).

2) S'ha proposat una estratègia per al tractament quimiomètric i interpretació dels resultats d'experiments metabolòmics no dirigits. Aquesta estratègia ha permès avaluar l'efecte de l'aplicació de condicions estressants sobre el cultiu del llevat (canvi de temperatura de cultiu i concentracions creixents de Cu(II)).

3) S'han observat modificacions en el metaboloma del llevat quan aquest s'ha cultivat a 42ºC (condicions estressants). Les concentracions de 43 metabòlits de les mostres estressades han canviat significativament en relació a les mostres control. Alguns dels metabòlits relacionats amb el creixement cel·lular del llevat han estat afectats pel canvi de temperatura de cultiu. El llevat ha seguit un patró consistent d'adaptació metabòlica al canvi de temperatura de cultiu.

4) S'han observat canvis en el metaboloma del llevat quan aquest s'ha cultivat a concentracions (subletals) creixents de Cu(II). L'exposició del llevat a Cu(II) ha causat l'increment de les espècies reactives de l'oxigen (ROS) al medi, les quals han provocat la reducció del metabolisme del llevat i alteracions el seu ADN. Les cèl·lules del llevat han contrarestat l'estrès oxidatiu mitjançant la protecció dels seus components cel·lulars i l'activació de mecanismes de reparació de l'ADN.

Estudi de dades paleoclimàtiques:

1) S'ha establert la correlació existent entre els compostos orgànics acumulats als sediments marins (IODP-U1318) durant l'època del Miocè i la temperatura superficial del mar (SST) a partir del mètode VIP-PLSR aplicat als perfils cromatogràfics TIC de les mostres de sediment.

2) Les concentracions del grup de lípids marins van augmentar a altes SST, fet que indica una alta productivitat marina i una major producció de biomassa. Les concentracions del grup de

compostos d'origen terrestre indiquen majors aportacions a temperatures més baixes. El tractament quimiomèric de les dades paleoclimàtiques ha permès la diferenciació entre les aportacions dels compostos del grup de lípids d'origen marí i terrestre.

3) Els compostos orgànics de la fracció àcida han mostrat patrons complexos en relació als canvis de SST, difícils d'interpretar. La gran diversitat de fonts d'àcids grassos i l'alteració dels àcids grassos més làbils pels efectes diagenètics han dificultat l'establiment de qualsevol correlació original amb la SST.

Conclusions de tipus quimiomètric:

1) El mètode ASCA ha resultat ser més precís i fiable en la interpretació dels resultats en comparació amb el mètode d'ANOVA-PCA. El mètode ASCA ha millorat l'aplicació de PCA en termes d'interpretabilitat i de selecció de les característiques i dels aspectes rellevants subjacents en les dades metabolòmiques. El test de permutació del mètode ASCA ha permès determinar els factors externs que han tingut efectes significatius en el cultiu de la planta del blat, i la seva possible interacció.

2) El procediment de compressió de dades LC-MS basat en la selecció de les regions d'interès, ROI, ha resultat més adequada que la compressió mitjançant el procediment d'agrupació en contenidors d'igual mida (*binning*). El procediment ROI ha permès reduir la dimensió de la matriu de dades de forma considerable sense perdre resolució espectral. El procediment ROI ha permès aplicar el mètode MCR-ALS als cromatogrames LC-MS complerts i no només a finestres de temps d'aquests.

3) El mètode de resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS) ha permès l'estimació dels perfils de concentració i dels espectres purs dels metabòlits en l'anàlisi no dirigida dels perfils LC-MS del llevat (*Saccharomyces cerevisiae*) sotmès a condicions ambientals estressants (temperatura i Cu(II)). L'estratègia proposada ha simplificat considerablement la interpretació de les dades metabolòmiques LC-MS i ha permès la proposta de nous biomarcadors.

4) S'ha demostrat que el mètode de selecció de variables SR té una alta sensibilitat i especificitat en les dades fisicoquímiques de qualitat de l'aigua, en les concentracions dels compostos orgànics de les dades climàtiques i en els microxips d'ADN (*microarrays*) de les dades de transcriptòmica. El nombre de falsos negatius ha estat el principal inconvenient en l'aplicació del mètode SR. El mètode de selecció de variables VIP ha permès detectar variables rellevants en conjunts de dades de diferent naturalesa (paràmetres fisicoquímics, cromatogrames TIC i *microarrays*). S'ha demostrat que el mètode VIP té bona sensibilitat i presenta baixa especificitat. S'ha demostrat que el mètode VIP és més rigorós en contraposició al mètode SR per a la selecció de variables en els cromatogrames TIC.

Referències

(National Research Council (US), 2007)
National Research Council (US) Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology (2007). Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment. Washington(DC): National Academies Press.

(Allen et al., 2003)
Allen, J., Davey, H.M., Broadhurst, D., Heald, J.K., Rowland, J.J., Oliver, S.G., and Kell, D.B. (2003). High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nature Biotechnology *21*, 692-696.

(Andersen et al., 2010)
Andersen, C.M., and Bro, R. (2010). Variable selection in regression—a tutorial. Journal of Chemometrics *24*, 728-737.

(Agradoña et al. 1980)
Argandoña, V.H., Luza, J.G., Niemeyer, H.M., and Corcuera, L.J. (1980). Role of hydroxamic acids in the resistance of cereals to aphids. Phytochemistry *19*, 1665-1668.

(Ariño, 2011)
Ariño, J. (2011). Els llevats com a organisme model de recerca en biologia, in Organismes model en biologia *62*, 45-59 (Corominas M., Valls M., Ed.) Treballs de la Societat Catalana de Biologia.

(Avery et al., 1996)
Avery, S.V., Howlett, N.G., and Radice, S. (1996). Copper toxicity towards Saccharomyces cerevisiae: dependence on plasma membrane fatty acid composition. Applied and Environmental Microbiology *62*, 3960-3966.

(Bajad et al., 2006)
Bajad, S.U., Lu, W., Kimball, E.H., Yuan, J., Peterson, C., and Rabinowitz, J.D. (2006). Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. Journal of Chromatography A *1125*, 76-88.

(Barker i Rayens, 2003)
Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics *17*, 166-173.

(Bedair i Sumner, 2008)
Bedair, M., and Sumner, L.W. (2008). Current and emerging mass-spectrometry technologies for metabolomics. TrAC Trends in Analytical Chemistry *27*, 238-250.

(Belicka et al., 2004)
Belicka, L.L., Macdonald, R.W., Yunker, M.B., and Harvey, H.R. (2004). The role of depositional regime on carbon transport and preservation in Arctic Ocean sediments. Marine Chemistry *86*, 65-88.

(Beltran et al., 2012)
Beltran, A., Suarez, M., Rodríguez, M.A., Vinaixa, M., Samino, S., Arola, L., Correig, X., and Yanes, O. (2012). Assessment of Compatibility between Extraction Methods for NMR- and LC/MS-Based Metabolomics. Analytical Chemistry *84*, 5838-5844.

(Benaroudj et al., 2001)
Benaroudj, N., Lee, D.H., and Goldberg, A.L. (2001). Trehalose Accumulation during Cellular Stress Protects Cells and Cellular Proteins from Damage by Oxygen Radicals. Journal of Biological Chemistry *276*, 24261-24267.

(Bennett et al., 2000)
Bennett, D.A., and Waters, M.D. (2000). Applying biomarker research. Environmental Health Perspectives *108*, 907-910.

(Bentaleb et al., 1999)
Bentaleb, I., Grimalt, J.O., Vidussi, F., Marty, J.C., Martin, V., Denis, M., Hatté, C., and Fontugne, M. (1999). The C37 alkenone record of seawater temperature during seasonal thermocline stratification. Marine Chemistry *64*, 301-313.

(Bilello, 2005)
Bilello, J.A. (2005). The agony and ecstasy of "OMIC" technologies in drug development. Current Molecular Medicine *5*, 39-52.

(Blum et al., 1992)
Blum, U., Gerig, T.M., Worsham, A.D., Holappa, L.D., and King, L.D. (1992). Allelopathic activity in wheat-conventional and wheat-no-till soils: Development of soil extract bioassays. Journal of Chemical Ecology *18*, 2191-2221.

(Blumer et al., 1971)
Blumer, M., Guillard, R.R.L., and Chase, T. (1971). Hydrocarbons of marine phytoplankton. Marine Biology *8*, 183-189.

(Boer et al., 2010)
Boer, V.M., Crutchfield, C.A., Bradley, P.H., Botstein, D., and Rabinowitz, J.D. (2010). Growth-limiting Intracellular Metabolites in Yeast Growing under Diverse Nutrient Limitations. Molecular Biology of the Cell *21*, 198-211.

(Boudah et al., 2013)
Boudah, S., Paris, A., and Junot, C. (2013). Chapter Four - Liquid Chromatography Coupled to Mass Spectrometry-Based Metabolomics and the Concept of Biomarker, in Advances in Botanical Research, 159-218 (Dominique R., Ed.) Academic Press.

(Bouhifd et al., 2013)
Bouhifd, M., Hartung, T., Hogberg, H.T., Kleensang, A., and Zhao, L. (2013). Review: Toxicometabolomics. Journal of Applied Toxicology *33*, 1365-1383.

(Brandolini et al., 2002)
Brandolini, V., Tedeschi, P., Capece, A., Maietti, A., Mazzotta, D., Salzano, G., Paparella, A., and Romano, P. (2002). Saccharomyces cerevisiae wine strains differing in copper resistance exhibit different capability to reduce copper content in wine. World Journal of Microbiology and Biotechnology *18*, 499-503.

(Brassell et al., 1986a)
Brassell, S.C., Brereton, R.G., Eglinton, G., Grimalt, J., Liebezeit, G., Marlowe, I.T., Pflaumann, U., and Sarnthein, M. (1986a). Palaeoclimatic signals recognized by chemometric treatment of molecular stratigraphic data. Organic Geochemistry *10*, 649-660.

(Brassell et al., 1986b)
Brassell, S.C., Eglinton, G., Marlowe, I.T., Pflaumann, U., and Sarnthein, M. (1986b). Molecular stratigraphy: a new tool for climatic assessment. Nature *320*, 129-133.

(Bremner, 1998)
Bremner, I. (1998). Manifestations of copper excess. The American Journal of Clinical Nutrition *67*, 1069S-1073S.

(Bro and Jong, 1997)
Bro, R., and De Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. Journal of Chemometrics *11*, 393-401.

(Bro et al., 2008)
Bro, R., Kjeldahl, K., Smilde, A.K., and Kiers, H.A.L. (2008). Cross-validation of component models: A critical look at current methods. Analytical and Bioanalytical Chemistry *390*, 1241-1251.

(Cambier et al., 1999)
Cambier, V., Hance, T., and de Hoffmann, E. (1999). Non-injured maize contains several 1,4-benzoxazin-3-one related compounds but only as glucoconjugates. Phytochemical Analysis *10*, 119-126.

(Campos et al., 2013)
Campos, B., Garcia-Reyero, N., Rivetti, C., Escalon, L., Habib, T., Tauler, R., Tsakovski, S., Piña, B., and Barata, C. (2013). Identification of Metabolic Pathways in Daphnia magna Explaining Hormetic Effects of Selective Serotonin

Reuptake Inhibitors and 4-Nonylphenol Using Transcriptomic and Phenotypic Responses. Environmental Science & Technology *47*, 9434-9443.

(Canelas et al., 2009)
Canelas, A.B., ten Pierick, A., Ras, C., Seifar, R.M., van Dam, J.C., van Gulik, W.M., and Heijnen, J.J. (2009). Quantitative Evaluation of Intracellular Metabolite Extraction Techniques for Yeast Metabolomics. Analytical Chemistry *81*, 7379-7389.

(Canuel et al., 1996)
Canuel, E.A., and Martens, C.S. (1996). Reactivity of recently deposited organic matter: Degradation of lipid compounds near the sediment-water interface. Geochimica et Cosmochimica Acta *60*, 1793-1806.

(Castrillo et al., 2003)
Castrillo, J.I., Hayes, A., Mohammed, S., Gaskell, S.J., and Oliver, S.G. (2003). An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. Phytochemistry *62*, 929-937.

(Castrillo i Oliver, 2006)
Castrillo, J.I., and Oliver, S.G. (2006). Metabolomics and Systems Biology in *Saccharomyces cerevisiae*, in Fungal Genomics, 3-18 (Brown, A., Ed.) Springer Berlin Heidelberg.

(Castrillo et al., 2007)
Castrillo, J.I., Zeef, L.A., Hoyle, D.C., Zhang, N., Hayes, A., Gardner, D.C., Cornell, M.J., Petty, J., Hakes, L., Wardleworth, L., et al. (2007). Growth control of the eukaryote cell: a systems biology study in yeast. Journal of Biology *6*, 4.

(Chong i Jun, 2005)
Chong, I.-G., and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems *78*, 103-112.

(Collins et al., 2015)
Collins, K.M., Oregioni, A., Robertson, L.E., Kelly, G., and Ramos, A. (2015). Protein–RNA specificity by high-throughput principal component analysis of NMR spectra. Nucleic Acids Research.

(Conte et al., 1993)
Conte, M.H., and Eglinton, G. (1993). Alkenone and alkenoate distributions within the euphotic zone of the eastern North Atlantic: correlation with production temperature. Deep Sea Research Part I: Oceanographic Research Papers *40*, 1935-1961.

(Conte et al., 1995)
Conte, M.H., Thompson, A., Eglinton, G., and Green, J.C. (1995). Lipid biomarker diversity in the coccolithophorid *Emiliana huxleyi* (prymnesiophyceae) and the related spcecies Gephyrocapsa oceanica. Journal of Phycology *31*, 272-282.

(Conte et al., 1994)
Conte, M.H., Volkman, J.K., and Eglinton, G. (1994). Lipid biomarkers of the Haptophyta, in The Haptophyte Algae, 351-377 (Green J.C. and Leadbeater B.S.C, eds.) Clarendon Press.

(Copaja et al., 1999)
Copaja, S.V., Nicol, D., and Wratten, S.D. (1999). Accumulation of hydroxamic acids during wheat germination. Phytochemistry *50*, 17-24.

(Creek et al., 2014)
Creek, D., Dunn, W., Fiehn, O., Griffin, J., Hall, R., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E., Sumner, L., Trengove, R., Wolfender J.L. (2014). Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics *10*, 350-353.

(Crowley i North, 1988)
Crowley, T.J., and North, G.R. (1988). Abrupt Climate Change and Extinction Events in Earth History. Science *240*, 996-1002.

(Dalsgaard et al., 2003)
Dalsgaard, J., St. John, M., Kattner, G., Müller-Navarra, D., and Hagen, W. (2003). Fatty acid trophic markers in the pelagic marine environment, in Advances in Marine Biology, 225-340 Academic Press.

(de Figueiredo et al., 2011)
de Figueiredo, L.F., Gossmann, T.I., Ziegler, M., and Schuster, S. (2011). Pathway analysis of NAD+ metabolism. Biochemical Journal *439*, 341-348.

(de Jesus Pereira et al., 2003)
de Jesus Pereira, E., Panek, A.D., and Eleutherio, E.C.A. (2003). Protection against oxidation during dehydration of yeast. Cell Stress & Chaperones *8*, 120-124.

(de Juan et al., 2014)
de Juan, A., Jaumot, J., and Tauler, R. (2014). Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Analytical Methods *6*, 4964-4976.

(de Juan i Tauler, 2003)
de Juan, A., and Tauler, R. (2003). Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. Analytica Chimica Acta *500*, 195-210.

(de Juan i Tauler, 2006)
de Juan, A., and Tauler, R. (2006). Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. Critical Reviews in Analytical Chemistry *36*, 163 - 176.

(De Vos et al., 2007)
De Vos, R.C.H., Moco, S., Lommen, A., Keurentjes, J.J.B., Bino, R.J., and Hall, R.D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nature Protocols *2*, 778-791.

(Draper i Smith, 1998)
Draper, N.R., and Smith, H. (1998). Fitting a Straight Line by Least Squares, in Applied Regression Analysis, 15-46 John Wiley & Sons, Inc.

(Ducruix et al., 2008)
Ducruix, C., Vailhen, D., Werner, E., Fievet, J.B., Bourguignon, J., Tabet, J.-C., Ezan, E., and Junot, C. (2008). Metabolomic investigation of the response of the model plant Arabidopsis thaliana to cadmium exposure: Evaluation of data pretreatment methods for further statistical analyses. Chemometrics and Intelligent Laboratory Systems *91*, 67-77.

(Dunn et al., 2008)
Dunn, W.B., Broadhurst, D., Brown, M., Baker, P.N., Redman, C.W.G., Kenny, L.C., and Kell, D.B. (2008). Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. Journal of Chromatography B *871*, 288-298.

(Dunn et al., 2011)
Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R., and Griffin, J.L. (2011). Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chemical Society reviews *40*, 387-426.

(Eglinton i Hamilton, 1967)
Eglinton, G., and Hamilton, R.J. (1967). Leaf Epicuticular Waxes. Science *156*, 1322-1335.

(Eilers i Boelens, 2005)
Eilers, P., and Boelens, H. (2005). Baseline Correction with Asymmetric Least Squares Smoothing.

(Eilers, 2003a)
Eilers, P.H.C. (2003a). Parametric Time Warping. Analytical Chemistry *76*, 404-411.

(Eilers, 2003b)
Eilers, P.H.C. (2003b). A Perfect Smoother. Analytical Chemistry *75*, 3631-3636.

(Einhellig, 1996)
Einhellig, F.A. (1996). Interactions Involving Allelopathy in Cropping Systems. Agron. J. *88*, 886-893.
Ellis, D.I., and Goodacre, R. (2006). Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst *131*, 875-885.

(Esbensen i Geladi, 2009)
Esbensen, K.H., and Geladi, P. (2009). 2.13 - Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice, in Comprehensive Chemometrics, 211-226 (Brown S.D., Tauler, R., Walczak, B., Ed.) Oxford: Elsevier.

(Estruch, 2000)
Estruch, F. (2000). Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. FEMS Microbiology Reviews *24*, 469-486.

(Fernández-Varela et al., 2015)
Fernández-Varela, R., Tomasi, G., and Christensen, J.H. (2015). An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes. Journal of Chromatography A *1384*, 133-141.

(Fiehn et al., 2007)
Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L., Goodacre, R., Hardy, N., Taylor, C., et al. (2007). The metabolomics standards initiative (MSI). Metabolomics *3*, 175-178.

(Fomsgaard et al., 2004)
Fomsgaard, I.S., Mortensen, A.G., and Carlsen, S.C.K. (2004). Microbial transformation products of benzoxazolinone and benzoxazinone allelochemicals--a review. Chemosphere *54*, 1025-1038.

(Fuerst i Wagner, 1957)
Fuerst, R., and Wagner, R.P. (1957). An analysis of the free intracellular amino acids of certain strains of Neurospora. Archives of Biochemistry and Biophysics *70*, 311-326.

(Gaetke i Chow, 2003)
Gaetke, L.M., and Chow, C.K. (2003). Copper toxicity, oxidative stress, and antioxidant nutrients. Toxicology *189*, 147-163.

(Games et al., 1984)
Games, D.E., Alcock, N.J., van der Greef, J., Nyssen, L.M., Maarse, H., Ten, M.C., and de Brauw, N. (1984). Analysis of pepper and capsicum oleoresins by high-performance liquid chromatography—mass spectrometry and field desorption mass spectrometry. Journal of Chromatography A *294*, 269-279.

(Gates i Sweeley, 1978)
Gates, S.C., and Sweeley, C.C. (1978). Quantitative metabolic profiling based on gas chromatography. Clinical Chemistry *24*, 1663-1673.

(Gelpí, 2002)
Gelpí, E. (2002). Interfaces for coupled liquid-phase separation/mass spectrometry techniques. An update on recent developments. Journal of Mass Spectrometry *37*, 241-253.

(Gibbons et al., 2015)
Gibbons, H., O'Gorman, A., and Brennan, L. (2015). Metabolomics as a tool in nutritional research. Current Opinion in Lipidology *26*, 30-34.

(Goffeau et al., 1996)
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 Genes. Science *274*, 546-567.

(Gogou i Stephanou, 2004)
Gogou, A., and Stephanou, E.G. (2004). Marine organic geochemistry of the Eastern Mediterranean: 2. Polar biomarkers in Cretan Sea surficial sediments. Marine Chemistry *85*, 1-25.

(Golub et al., 1970)
Golub, G.H., Golub, C., and Reinsch (1970). Singular value decomposition and least squares solutions. Numerische Mathematik *14*, 403-420.

(Golub i Loan, 1996)
Golub, G.H., and Loan, C.F.V. (1996). Matrix Computations. The Johns Hopkins University Press (Baltimore and London).

(Gonzalez et al., 1997)
Gonzalez, B., François, J., and Renaud, M. (1997). A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. Yeast *13*, 1347-1355.

(Goodacre et al., 2004)
Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G., and Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology *22*, 245-252.

(Gorrochategui et al., 2015)
Gorrochategui, E., Jaumot, J., and Tauler, R. (2015). A protocol for LC-MS metabolomic data processing using chemometric tools. Protocol Exchange.

(Gough, 2002)
Gough, N.R. (2002). Science's signal transduction knowledge environment: the connections maps database. Ann N Y Acad Sci *971*, 585-587.

(Greef i Smilde, 2005)
Greef, J.v.d., and Smilde, A.K. (2005). Symbiosis of chemometrics and metabolomics: past, present, and future. Journal of Chemometrics *19*, 376-386.

(Griffiths et al., 2010)
Griffiths, W.J., Koal, T., Wang, Y., Kohl, M., Enot, D.P., and Deigner, H.-P. (2010). Targeted Metabolomics for Biomarker Discovery. Angewandte Chemie International Edition *49*, 5426-5445.

(Groušl et al., 2009)
Groušl, T., Ivanov, P., Frydlová, I., Vašicová, P., Janda, F., Vojtová, J., Malínská, K., Malcová, I., Nováková, L., Janošková, D., et al. (2009). Robust heat shock induces eIF2α-phosphorylation-independent assembly of stress granules containing eIF3 and 40S ribosomal subunits in budding yeast, Saccharomyces cerevisiae. Journal of Cell Science *122*, 2078-2088.

(Gullbert et al., 2004)
Gullberg, J., Jonsson, P., Nordström, A., Sjöström, M., and Moritz, T. (2004). Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. Analytical Biochemistry *331*, 283-295.

(Gurden et al., 2001)
Gurden, S.P., Westerhuis, J.A., Bro, R., and Smilde, A.K. (2001). A comparison of multiway regression and scaling methods. Chemometrics and Intelligent Laboratory Systems *59*, 121-136.

(Hancock, 1958)
Hancock, R. (1958). The intracellular amino acids of Staphylococcus aureus: Release and analysis. Biochimica et Biophysica Acta *28*, 402-412.

(Harper i Lynch, 1982)
Harper, S.H.T., and Lynch, J.M. (1982). The role of water-soluble components in phytotoxicity from decomposing straw. Plant Soil *65*, 11-17.

(Harrington et al., 2005)
Harrington, P.d.B., Vieira, N.E., Espinoza, J., Nien, J.K., Romero, R., and Yergey, A.L. (2005). Analysis of variance-principal component analysis: A soft tool for proteomic discovery. Analytica Chimica Acta *544*, 118-127.

(Harvey et al., 1986)
Harvey, H.R., Fallon, R.D., and Patton, J.S. (1986). The effect of organic matter and oxygen on the degradation of bacterial membrane lipids in marine sediments. Geochimica et Cosmochimica Acta *50*, 795-804.

(Hays et al., 1976)
Hays, J.D., Imbrie, J., and Shackleton, N.J. (1976). Variations in the Earth's Orbit: Pacemaker of the Ice Ages. Science *194*, 1121-1132.

(Hess et al., 1997)
Hess, M., Barralis, G., Bleiholder, H., Buhr, L., Eggers, T.H., Hack, H., and Stauss, R. (1997). Use of the extended BBCH scale—general for the descriptions of the growth stages of mono; and dicotyledonous weed species. Weed Research *37*, 433-441.

(Hirayama et al., 2009)
Hirayama, A., Kami, K., Sugimoto, M., Sugawara, M., Toki, N., Onozuka, H., Kinoshita, T., Saito, N., Ochiai, A., Tomita, M., et al. (2009). Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. Cancer Research *69*, 4918-4925.

(Hoefsloot et al., 2009)
Hoefsloot, H.C.J., Vis, D.J., Westerhuis, J.A., Smilde, A.K., and Jansen, J.J. (2009). 2.23 - Multiset Data Analysis: ANOVA Simultaneous Component Analysis and Related Methods, in Comprehensive Chemometrics, 453-472 (Brown S.D., Tauler, R., Walczak, B., Ed.) Oxford: Elsevier.

(Hohmann, 2002)
Hohmann, S. (2002). Osmotic Stress Signaling and Osmoadaptation in Yeasts. Microbiology and Molecular Biology Reviews *66*, 300-372.

(Horai et al., 2010)
Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry *45*, 703-714.

(Horning i Horning, 1971)
Horning, E.C., and Horning, M.G. (1971). Human Metabolic Profiles Obtained by GC and GC/MS. Journal of Chromatographic Science *9*, 129-140.

(Hottiger et al., 1987)
Hottiger, T., Boller, T., and Wiemken, A. (1987). Rapid changes of heat and desiccation tolerance correlated with changes of trehalose content in Saccharomyces cerevisiae cells subjected to temperature shifts. FEBS Letters *220*, 113-115.

(Iacobucci, 1995)
Iacobucci, D. (1995). Analysis of variance for unbalanced data, in AMA Winter Educators Conference: Marketing Theory and Practice, 337-343 (Stewart, D.W. and Vilcassim, N.J., Ed.).

(Ipcc, 2007)
Ipcc (2007). Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller, Ed.) Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

(Ipcc, 2014a)
Ipcc (2014a). Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White, Ed.) Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

(Ipcc, 2014b)
Ipcc (2014b). Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Barros, V.R., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White, Ed.). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

(Izawa et al., 1995)
Izawa, S., Inoue, Y., and Kimura, A. (1995). Oxidative stress response in yeast: effect of glutathione on adaptation to hydrogen peroxide stress in Saccharomyces cerevisiae. FEBS Letters *368*, 73-76.

(Jabran et al., 2015)
Jabran, K., Mahajan, G., Sardana, V., and Chauhan, B.S. (2015). Allelopathy for weed control in agricultural systems. Crop Protection *72*, 57-65.

(Janick et al., 1979)
Janick, J. (1979). Horticultural Science. W.H. Freeman (San Francisco).

(Jansen et al., 2008)
Jansen, J.J., Bro, R., Hoefsloot, H.C.J., Berg, F.W.J.v.d., Westerhuis, J.A., and Smilde, A.K. (2008). PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. Journal of Chemometrics *22*, 114-121.

(Jansen et al., 2005a)
Jansen, J.J., Hoefsloot, H.C.J., van der Greef, J., Timmerman, M.E., and Smilde, A.K. (2005a). Multilevel component analysis of time-resolved metabolic fingerprinting data. Analytica Chimica Acta *530*, 173-183.

(Jansen et al., 2005b)
Jansen, J.J., Hoefsloot, H.C.J., van der Greef, J., Timmerman, M.E., Westerhuis, J.A., and Smilde, A.K. (2005b). ASCA: analysis of multivariate data obtained from an experimental design. Journal of Chemometrics *19*, 469-481.

(Jaumot et al., 2005)
Jaumot, J., Gargallo, R., de Juan, A., and Tauler, R. (2005). A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. Chemometrics and Intelligent Laboratory Systems *76*, 101-110.

(Jewison et al., 2012)
Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A.C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., et al. (2012). YMDB: the Yeast Metabolome Database. Nucleic Acids Research *40*, D815-D820.

(Kaddurah-Daouk i Krishnan, 2008)
Kaddurah-Daouk, R., and Krishnan, K.R.R. (2008). Metabolomics: A Global Biochemical Approach to the Study of Central Nervous System Diseases. Neuropsychopharmacology *34*, 173-186.

(Kaneihsa et al., 2012)
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research *40*, D109-D114.

(Kato-Noguchi et al., 2010)
Kato-Noguchi, H., Macías, F.A., and Molinillo, J.M.G. (2010). Structure–activity relationship of benzoxazinones and related compounds with respect to the growth inhibition and α-amylase activity in cress seedlings. Journal of Plant Physiology *167*, 1221-1225.

(Kell i Goodacre, 2014)
Kell, D.B., and Goodacre, R. (2014). Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. Drug Discovery Today *19*, 171-182.

(Kerr i Churchill, 2007)
Kerr, M.K., and Churchill, G.A. (2007). Statistical design and the analysis of gene expression microarray data. Genetics Research *89*, 509-514.

(Kerr et al., 2000)
Kerr, M.K., Martin, M., and Churchill, G.A. (2000). Analysis of Variance for Gene Expression Microarray Data. Journal of Computational Biology *7*, 819-837.

(Keseler et al., 2013)
Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., et al. (2013). EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Research *41*, D605-D612.

(Kitano, 2002)
Kitano, H. (2002). Systems Biology: A Brief Overview. Science *295*, 1662-1664.

(Koch-Nolte et al., 2011)
Koch-Nolte, F., Fischer, S., Haag, F., and Ziegler, M. (2011). Compartmentation of NAD+-dependent signalling. FEBS Letters *585*, 1651-1656.

(Koek et al., 2006)
Koek, M.M., Muilwijk, B., van der Werf, M.J., and Hankemeier, T. (2006). Microbial Metabolomics with Gas Chromatography/Mass Spectrometry. Analytical Chemistry *78*, 1272-1281.

(Koning i Dam, 1992)
Koning, W.D., and Dam, K.V. (1992). A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. Analytical Biochemistry *204*, 118-123.

(Krakowska et al., 2014)
Krakowska, B., Stanimirova, I., Orzel, J., Daszykowski, M., Grabowski, I., Zaleszczyk, G., and Sznajder, M. (2014). Detection of discoloration in diesel fuel based on gas chromatographic fingerprints. Analytical and Bioanalytical Chemistry, 1-12.

(Kvalheim i Karstang, 1989)
Kvalheim, O.M., and Karstang, T.V. (1989). Interpretation of latent-variable regression models. Chemometrics and Intelligent Laboratory Systems *7*, 39-51.

(Lämmerhofer, 2010)
Lämmerhofer, M. (2010). HILIC and mixed-mode chromatography: The rising stars in separation science. Journal of Separation Science *33*, 679-680.

(Lamy et al., 2014)
Lamy, F., Gersonde, R., Winckler, G., Esper, O., Jaeschke, A., Kuhn, G., Ullermann, J., Martinez-Garcia, A., Lambert, F., and Kilian, R. (2014). Increased Dust Deposition in the Pacific Southern Ocean During Glacial Periods. Science *343*, 403-407.

(Laterra i Bazzalo, 1999)
Laterra, P., and Bazzalo, M.E. (1999). Seed-to-seed allelopathic effects between two invaders of burned Pampa grasslands. Weed Research *39*, 297-308.

(Lawson i Hanson, 1995)
Lawson, C., and Hanson, R. (1995). Solving Least Squares Problems. (Society for Industrial and Applied Mathematics).

(Leardi, 2000)
Leardi, R. (2000). Application of genetic algorithm–PLS for feature selection in spectral data sets. Journal of Chemometrics *14*, 643-655.

(Leardi et al., 1992)
Leardi, R., Boggia, R., and Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. Journal of Chemometrics *6*, 267-281.

(Linder, 2012)
Linder, M.C. (2012). The relationship of copper to DNA damage and damage prevention in humans. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *733*, 83-91.

(Lisec et al., 2006)
Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., and Fernie, A.R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. Nature Protocols *1*, 387-396.

(López i Grimalt, 2006)
López, J.F., and Grimalt, J.O. (2006). Reassessment of the Structural Composition of the Alkenone Distributions in Natural Environments Using an Improved Method for Double Bond Location Based on GC-MS Analysis of Cyclopropylimines. Journal of the American Society for Mass Spectrometry *17*, 710-720.

(Lu et al., 2008)
Lu, W., Bennett, B.D., and Rabinowitz, J.D. (2008). Analytical strategies for LC–MS-based targeted metabolomics. Journal of Chromatography B *871*, 236-242.

(Lundstedt et al., 1998)
Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, Å., Pettersen, J., and Bergman, R. (1998). Experimental design and optimization. Chemometrics and Intelligent Laboratory Systems *42*, 3-40.

(Ly-Verdu et al., 2015)
Ly-Verdu, S., Groger, T.M., Arteaga-Salas, J.M., Brandmaier, S., Kahle, M., Neschen, S., Harbe de Angelis, M., and Zimmermann, R. (2015). Combining metabolomic non-targeted GCxGC-ToF-MS analysis and chemometric ASCA-based study of variances to assess dietary influence on type 2 diabetes development in a mouse model. Anal Bioanal Chem *407*, 343-354.

(Mac Key, 2005)
Mac Key, J. (2005). Wheat: its concept, evolution and taxonomy, in Durum wheat breeding. Current approaches and future strategies, 3-61 (Royo, C. and Di Fonzo N., Ed.).

(Mamas et al., 2011)
Mamas, M., Dunn, W.B., Neyses, L., and Goodacre, R. (2011). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Archives of Toxicology *85*, 5-17.

(Mark et al., 2000)
Mark A. Warne, E.M.L., D. Osborn, J. M. Weeks, J. K. Nicholson (2000). An NMR-based metabonomic investigation of the toxic effects of 3-trifluoromethyl-aniline on the earthworm Eisenia veneta. Biomarkers *5*, 56-72.

(Marlowe et al., 1990)
Marlowe, I.T., Brassell, S.C., Eglinton, G., and Green, J.C. (1990). Long-chain alkenones and alkyl alkenoates and the fossil coccolith record of marine sediments. Chemical Geology *88*, 349-375.

(Marlowe et al., 1984)
Marlowe, I.T., Green, J.C., Neal, A.C., Brassell, S.C., Eglinton, G., and Course, P.A. (1984). Long chain (n-C37–C39) alkenones in the Prymnesiophyceae. Distribution of alkenones and other lipids and their taxonomic significance. British Phycological Journal *19*, 203-216.

(Mathiassen et al., 2006)
Mathiassen, S.K., Kudsk, P., and Mogensen, B.B. (2006). Herbicidal Effects of Soil-Incorporated Wheat. Journal of Agricultural and Food Chemistry *54*, 1058-1063.

(Mehmood et al., 2012)
Mehmood, T., Liland, K.H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems *118*, 62-69.

(Mensonides et al., 2013)
Mensonides, F.I.C., Hellingwerf, K.J., de Mattos, M.J.T., and Brul, S. (2013). Multiphasic adaptation of the transcriptome of Saccharomyces cerevisiae to heat stress. Food Research International *54*, 1103-1112.

(Miller, 2007)
Miller, M.G. (2007). Environmental Metabolomics: A SWOT Analysis (Strengths, Weaknesses, Opportunities, and Threats). Journal of Proteome Research *6*, 540-545.

(Mogensen et al., 2006)
Mogensen, B.B., Krongaard, T., Mathiassen, S.K., and Kudsk, P. (2006). Quantification of Benzoxazinone Derivatives in Wheat (Triticum aestivum) Varieties Grown under Contrasting Conditions in Denmark. Journal of Agricultural and Food Chemistry *54*, 1023-1030.

(Morano et al., 2012)
Morano, K.A., Grant, C.M., and Moye-Rowley, W.S. (2012). The Response to Heat Shock and Oxidative Stress in Saccharomyces cerevisiae. Genetics *190*, 1157-1195.

(Mortimer i Polsinelli, 1999)
Mortimer, R., and Polsinelli, M. (1999). On the origins of wine yeast. Research in Microbiology *150*, 199-204.

(Müller et al., 1998)
Müller, P.J., Kirst, G., Ruhland, G., von Storch, I., and Rosell-Melé, A. (1998). Calibration of the alkenone paleotemperature index U37K′ based on core-tops from the eastern South Atlantic and the global ocean (60°N-60°S). Geochimica et Cosmochimica Acta *62*, 1757-1772.

(Nair et al., 1990)
Nair, M., Whitenack, C., and Putnam, A. (1990). 2,2′-OXO-1, 1 ′-azobenzene A microbially transformed allelochemical from 2,3-Benzoxazolinone: I. Journal of Chemical Ecology *16*, 353-364.

(Nakagawa et al., 1995)
Nakagawa, E., Amano, T., Hirai, N., and Iwamura, H. (1995). Non-induced cyclic hydroxamic acids in wheat during juvenile stage of growth. Phytochemistry *38*, 1349-1354.

(Netzly et al., 1998)
Netzly, D., Riople, J., Eljeta, G., and Butler, L. (1988). Germination Stimulants of Witchweed (Striga asiatica) from Hydrophobic Root Exudate of Sorghum (Sorghum bicolor). Weed Science *36*, 441-446.

(Nevedomskaya et al., 2010)
Nevedomskaya, E., Ramautar, R., Derks, R., Westbroek, I., Zondag, G., van der Pluijm, I., Deelder, A.M., and Mayboroda, O.A. (2010). CE-MS for Metabolic Profiling of Volume-Limited Urine Samples: Application to Accelerated Aging TTD Mice. Journal of Proteome Research *9*, 4869-4874.

(Nicholson et al., 1983)
Nicholson, J.K., Buckingham, M.J., and Sadler, P.J. (1983). High resolution 1H n.m.r. studies of vertebrate blood and plasma. Biochemical Journal *211*, 605-615.

(Nicholson i Lindon, 2008)
Nicholson, J.K., and Lindon, J.C. (2008). Systems biology: Metabonomics. Nature *455*, 1054-1056.

(Nicholson et al., 1999)
Nicholson, J.K., Lindon, J.C., and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica: The fate and safety evaluation of foreign compounds in biological systems *29*, 1181 - 1189.

(Nicholson i Wilson, 1989)
Nicholson, J.K., and Wilson, I.D. (1989). High resolution proton magnetic resonance spectroscopy of biological fluids. Progress in Nuclear Magnetic Resonance Spectroscopy *21*, 449-501.

(Nicholson i Wilson, 2003)
Nicholson, J.K., and Wilson, I.D. (2003). Understanding 'Global' Systems Biology: Metabonomics and the Continuum of Metabolism. Nature Reviews Drug Discovery *2*, 668-676.

(Nielsen et al., 1998)
Nielsen, N.-P.V., Carstensen, J.M., and Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. Journal of Chromatography A *805*, 17-35.

(Niemeyer, 1988)
Niemeyer, H.M. (1988). Hydroxamic acids (4-hydroxy-1,4-benzoxazin-3-ones), defence chemicals in the gramineae. Phytochemistry *27*, 3349-3358.

(Norgaard et al., 2000)
Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., and Engelsen, S.B. (2000). Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Applied Spectroscopy *54*, 413-419.

(Odell et al., 2000)
Odell, M., Sriskanda, V., Shuman, S., and Nikolov, D.B. (2000). Crystal Structure of Eukaryotic DNA Ligase–Adenylate Illuminates the Mechanism of Nick Sensing and Strand Joining. Molecular Cell *6*, 1183-1193.

(Ogata et al., 1999)
Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *27*, 29-34.

(Orešič, 2009)
Orešič, M. (2009). Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. Nutrition, Metabolism and Cardiovascular Diseases *19*, 816-824.

(Paillard, 2001)
Paillard, D. (2001). Glacial cycles: Toward a new paradigm. Reviews of Geophysics *39*, 325-346.

(Patti et al., 2012)
Patti, G.J., Yanes, O., and Siuzdak, G. (2012). Metabolomics: the apogee of the omic triology. Nature reviews. Molecular cell biology *13*, 263-269.

(Pérez i Ormenoñuñez, 1991)
Pérez, F. and Ormenoñuñez, J. (1991). Difference in hydroxamic acid content in roots and root exudates of wheat (Triticum aestivum L.) and rye (Secale cereale L.): Possible role in allelopathy. Journal of Chemical Ecology *17*, 1037-1043.

(Perez, 1990)
Perez, F.J. (1990). Allelopathic effect of hydroxamic acids from cereals on Avena sativa and A. Fatua. Phytochemistry *29*, 773-776.

(Pérez et al., 2009)
Pérez, I.S.n., Culzoni, M.a.J., Siano, G.G., García, M.a.D.G., Goicoechea, H.c.C., and Galera, M.a.M.n. (2009). Detection of Unintended Stress Effects Based on a Metabonomic Study in Tomato Fruits after Treatment with Carbofuran Pesticide. Capabilities of MCR-ALS Applied to LC-MS Three-Way Data Arrays. Analytical Chemistry *81*, 8335-8346.

(Petti et al., 2011)
Petti, A.A., Crutchfield, C.A., Rabinowitz, J.D., and Botstein, D. (2011). Survival of starving yeast is correlated with oxidative stress response and nonrespiratory mitochondrial function. Proceedings of the National Academy of Sciences *108*, E1089–E1098.

(Pierce et al., 2005)
Pierce, K.M., Hope, J.L., Johnson, K.J., Wright, B.W., and Synovec, R.E. (2005). 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 *1096*, 101-110.

(Platikanov, et al. 2013)
Platikanov, S., Garcia, V., Fonseca, I., Rullán, E., Devesa, R., and Tauler, R. (2013). Influence of minerals on the taste of bottled and tap water: A chemometric approach. Water Research *47*, 693-704.

(Poynter et al., 1989)
Poynter, J., Farrimond, P., Brassell, S., and Eglinton, G. (1989). Molecular stratigraphic study of sediments from Holes 658A and 660A, Leg 108, in Proceedings of the Ocean Drilling Program, Scientific Results, 387-394 US Government Printing Office Washington DC.

(Prahl et al., 1989)
Prahl, F.G., de Lange, G.J., Lyle, M., and Sparrow, M.A. (1989). Post-depositional stability of long-chain alkenones under contrasting redox conditions. Nature *341*, 434-437.

(Prahl et al., 1988)
Prahl, F.G., Muehlhausen, L.A., and Zahnle, D.L. (1988). Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. Geochimica et Cosmochimica Acta *52*, 2303-2310.

(Prahl i Wakeham, 1987)
Prahl, F.G., and Wakeham, S.G. (1987). Calibration of unsaturation patterns in long-chain ketone compositions for palaeotemperature assessment. Nature *330*, 367-369.

(Prasad Maharjan i Ferenci, 2003)
Prasad Maharjan, R., and Ferenci, T. (2003). Global metabolite analysis: the influence of extraction methodology on metabolome profiles of Escherichia coli. Analytical Biochemistry *313*, 145-154.

(Pretorius, 2000)
Pretorius, I.S. (2000). Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. Yeast *16*, 675-729.

(Querol et al., 1994)
Querol, A., Barrio, E., and Ramón, D. (1994). Population dynamics of natural Saccharomyces strains during wine fermentation. International Journal of Food Microbiology *21*, 315-323.

(Rabinowitz i Kimball, 2007)
Rabinowitz, J.D., and Kimball, E. (2007). Acidic Acetonitrile for Cellular Metabolome Extraction from Escherichia coli. Analytical Chemistry *79*, 6167-6173.

(Rajalahti et al., 2009a)
Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.-M., Ulvik, R.J., and Kvalheim, O.M. (2009a). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometrics and Intelligent Laboratory Systems *95*, 35-48.

(Rajalahti et al., 2009b)
Rajalahti, T., Arneberg, R., Kroksveen, A.C., Berle, M., Myhr, K.-M., and Kvalheim, O.M. (2009b). Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. Analytical Chemistry *81*, 2581-2590.

(Richards, 1934)
Richards, O.W. (1934). The analysis of growth as illustrated by yeast, in Cold Spring Harbor Symposia on Quantitative Biology, 157-166 Cold Spring Harbor Laboratory Press.

(Richard i Haynes, 1932)
Richards, O.W., and Haynes, F.W. (1932). Oxygen consumption and carbon dioxide production during the growth of yeast. Plant Physiology *7*, 139-144.

(Robertson, 2005)
Robertson, D.G. (2005). Metabonomics in Toxicology: A Review. Toxicological Sciences *85*, 809-822.

(Robertson i Frevert, 2013)
Robertson, D.G., and Frevert, U. (2013). Metabolomics in drug discovery and development. Clinical pharmacology and therapeutics *94*, 559-561.

(Rontani et al., 2001)
Rontani, J.-F., Marchand, D., and Volkman, J.K. (2001). NaBH4 reduction of alkenones to the corresponding alkenols: a useful tool for their characterisation in natural samples. Organic Geochemistry *32*, 1329-1341.

(Rosell-Melé, 2003)
Rosell-Melé, A. (2003). Biomarkers as proxies of climate change, in Global change in the Holocene (A.W. Mackay, A.W, Battarbee, R.W., Briks, J.B. and Oldfield, F., Ed.)

(Salek et al., 2013)
Salek, R.M., Steinbeck, C., Viant, M.R., Goodacre, R., and Dunn, W.B. (2013). The role of reporting standards for metabolite annotation and identification in metabolomic studies. GigaScience *2*, 13-13.

(Savorani et al., 2013)
Savorani, F., Rasmussen, M.A., Mikkelsen, M.S., and Engelsen, S.B. (2013). A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. Food Research International *54*, 1131-1145.

(Savorani et al., 2010)
Savorani, F., Tomasi, G., and Engelsen, S.B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. Journal of Magnetic Resonance *202*, 190-202.

(Shouten et al., 2002)
Schouten, S., Hopmans, E.C., Schefuß, E., and Sinninghe Damsté, J.S. (2002). Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? Earth and Planetary Science Letters *204*, 265-274.

(Schulz et al., 2013)
Schulz, M., Marocco, A., Tabaglio, V., Macias, F., and Molinillo, J.G. (2013). Benzoxazinoids in Rye Allelopathy - From Discovery to Application in Sustainable Weed Control and Organic Farming. Journal of Chemical Ecology *39*, 154-174.

(Sène et al., 2000)
Sène, M., Doré, T., and Pellissier, F. (2000). Effect of Phenolic Acids in Soil under and Between Rows of a Prior Sorghum (Sorghum bicolor) Crop on Germination, Emergence, and Seedling Growth of Peanut (Arachis hypogea). Journal of Chemical Ecology *26*, 625-637.

(Shao, 1993)
Shao, J. (1993). Linear Model Selection by Cross-validation. Journal of the American Statistical Association *88*, 486-494.

(Shaw i Mitchell-Olds, 1993)
Shaw, R.G., and Mitchell-Olds, T. (1993). Anova for Unbalanced Data: An Overview. Ecology *74*, 1638-1645.

(Sherman, 2002)
Sherman, F. (2002). Getting started with yeast, in Methods in Enzymology, 3-41 (G. Christine, and R.F. Gerald, Ed.) Academic Press.

(Sicker et al., 2000)
Sicker, D., Frey, M., Schulz, M., and Gierl, A. (2000). Role of natural benzoxazinones in the survival strategy of plants, in International Review of Cytology, 319-346 (Kwang, W.J., Ed.) Academic Press.

(Simoneit, 1977)
Simoneit, B.R.T. (1977). Organic matter in eolian dusts over the Atlantic Ocean. Marine Chemistry *5*, 443-464.

(Skov et al., 2006)
Skov, T., Berg, F.v.d., Tomasi, G., and Bro, R. (2006). Automated alignment of chromatographic data. Journal of Chemometrics *20*, 484-497.

(Smilde et al., 2004)
Smilde, A.K., Bro, R., and Geladi, P. (2004). Multi-way analysis. Applications in the chemical sciences, Chichester: Wiley.

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

(Smilde et al., 2012)
Smilde, A.K., Timmerman, M.E., Hendriks, M.M.W.B., Jansen, J.J., and Hoefsloot, H.C.J. (2012). Generic framework for high-dimensional fixed-effects ANOVA. Briefings in Bioinformatics *13*, 524-535.

(Smith et al., 2009)
Smith, L.M., Maher, A.D., Want, E.J., Elliott, P., Stamler, J., Hawkes, G.E., Holmes, E., Lindon, J.C., and Nicholson, J.K. (2009). Large-Scale Human Metabolic Phenotyping and Molecular Epidemiological Studies via 1H NMR Spectroscopy of Urine: Investigation of Borate Preservation. Analytical Chemistry *81*, 4847-4856.

(Smolinska et al., 2012)
Smolinska, A., Blanchet, L., Buydens, L.M.C., and Wijmenga, S.S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. Analytica Chimica Acta *750*, 82-97.

(Soga et al., 2003)
Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., and Nishioka, T. (2003). Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry. Journal of Proteome Research *2*, 488-494.

(Søltoft et al., 2008)
Søltoft, M., Jørgensen, L.N., Svensmark, B., and Fomsgaard, I.S. (2008). Benzoxazinoid concentrations show correlation with Fusarium Head Blight resistance in Danish wheat varieties. Biochemical Systematics and Ecology *36*, 245-259.

(Sriskanda et al., 2001)
Sriskanda, V., Moyer, R.W., and Shuman, S. (2001). NAD+-dependent DNA Ligase Encoded by a Eukaryotic Virus. Journal of Biological Chemistry *276*, 36100-36109.

(Stanimirova et al., 2011)
Stanimirova, I., Michalik, K., Drzazga, Z., Trzeciak, H., Wentzell, P.D., and Walczak, B. (2011). Interpretation of analysis of variance models using principal component analysis to assess the effect of a maternal anticancer treatment on the mineralization of rat bones. Analytica Chimica Acta *689*, 1-7.

(Strassburg et al., 2010)
Strassburg, K., Walther, D., Takahashi, H., Kanaya, S., and Kopka, J. (2010). Dynamic Transcriptional and Metabolic Responses in Yeast Adapting to Temperature Stress. OMICS: A Journal of Integrative Biology *14*, 249-259.

(Sumner et al., 2007)
Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.M., Fiehn, O., Goodacre, R., Griffin, J.L., et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics: Official journal of the Metabolomic Society *3*, 211-221.

(Tang et al., 2014)
Tang, D.-Q., Zou, L., Yin, X.-X., and Ong, C.N. (2014). HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS. Mass Spectrometry Reviews

(Tas i van der Greef, 1994)
Tas, A.C., and van der Greef, J. (1994). Mass spectrometric profiling and pattern recognition. Mass Spectrometry Reviews *13*, 155-181.

(Tauler, 1995)
Tauler, R. (1995). Multivariate curve resolution applied to second order data. Chemometrics and Intelligent Laboratory Systems *30*, 133-146.

(Tauler i Barceló, 1993)
Tauler, R., and Barceló, D. (1993). Multivariate curve resolution applied to liquid chromatography—diode array detection. TrAC Trends in Analytical Chemistry *12*, 319-327.

(Tauler et al., 1993)
Tauler, R., Kowalski, B., and Fleming, S. (1993). Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. Analytical Chemistry *65*, 2040-2047.

(Tauler et al., 1995)
Tauler, R., Smilde, A., and Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. Journal of Chemometrics *9*, 31-58.

(Timbrell, 1998)
Timbrell, J.A. (1998). Biomarkers in toxicology. Toxicology *129*, 1-12.

(Tolstikov i Fiehn, 2002)
Tolstikov, V.V., and Fiehn, O. (2002). Analysis of Highly Polar Compounds of Plant Origin: Combination of Hydrophilic Interaction Chromatography and Electrospray Ion Trap Mass Spectrometry. Analytical Biochemistry *301*, 298-307.

(Tolstikov et al., 2003)
Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N., and Fiehn, O. (2003). Monolithic Silica-Based Capillary Reversed-Phase Liquid Chromatography/Electrospray Mass Spectrometry for Plant Metabolomics. Analytical Chemistry *75*, 6737-6740.

(Torres et al., 1996)
Torres, A., Oliva, R.M., Castellano, D., and Cross, P. (1996). Introduction. A Science for the Future., in First World Congress on Allelopathy (Cadiz, Spain).

(Tran et al., 2014)
Tran, T.N., Afanador, N.L., Buydens, L.M.C., and Blanchet, L. (2014). Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). Chemometrics and Intelligent Laboratory Systems *138*, 153-160.

(Troyer, 2011)
Troyer, J.R. (2001). In the beginning: The multiple discovery of the first hormone herbicides. Weed Science *49*, 290-297.

(Trygg et al., 2006)
Trygg, J., Gullberg, J., Johansson, A.I., Jonsson, P., and Moritz, T. (2006). Chemometrics in Metabolomics — An Introduction. In Plant Metabolomics. K. Saito, R. Dixon, and L. Willmitzer, eds. (Springer Berlin Heidelberg), pp. 117-128.

(Urano et al., 2010)
Urano, K., Kurihara, Y., Seki, M., and Shinozaki, K. (2010). 'Omics' analyses of regulatory networks in plant abiotic stress responses. Current Opinion in Plant Biology *13*, 132-138.

(van der Greef et al., 1983)
van der Greef, J., Tas, A.C., Bouwman, J., Ten Noever de Brauw, M.C., and Schreurs, W.H.P. (1983). Evaluation of field-desorption and fast atom-bombardment mass spectrometric profiles by pattern recognition techniques. Analytica Chimica Acta *150*, 45-52.

(Venkatesan i Kaplan, 1982)
Venkatesan, M.I., and Kaplan, I.R. (1982). Distribution and transport of hydrocarbons in surface sediments of the alaskan outer continental shelf. Geochimica et Cosmochimica Acta *46*, 2135-2149.

(Viant i Sommer, 2013)
Viant, M., and Sommer, U. (2013). Mass spectrometry based environmental metabolomics: a primer and review. Metabolomics *9*, 144-158.

(Viant, 2008)
Viant, M.R. (2008). Recent developments in environmental metabolomics. Molecular Biosystems *4*, 980-986.

(Viant, 2009)
Viant, M.R. (2009). Applications of metabolomics to the environmental sciences. Metabolomics *5*, 1-2.

(Villagrasa et al., 2006a)
Villagrasa, M., Guillamón, M., Eljarrat, E., and Barceló, D. (2006a). Determination of Benzoxazinone Derivatives in Plants by Combining Pressurized Liquid Extraction−Solid-Phase Extraction Followed by Liquid Chromatography−Electrospray Mass Spectrometry. Journal of Agricultural and Food Chemistry *54*, 1001-1008.

(Villagrassa et al., 2007)
Villagrasa, M., Guillamón, M., Eljarrat, E., and Barceló, D. (2007). Matrix effect in liquid chromatography-electrospray ionization mass spectrometry analysis of benzoxazinoid derivatives in plant material. Journal of Chromatography A *1157*, 108-114.

(Villagrassa et al., 2006b)
Villagrasa, M., Guillamón, M., Labandeira, A., Taberner, A., Eljarrat, E., and Barceló, D. (2006b). Benzoxazinoid Allelochemicals in Wheat: Distribution among Foliage, Roots, and Seeds. Journal of Agricultural and Food Chemistry *54*, 1009-1015.

(Villagrasa et al., 2008)
Villagrasa, M., Guillamón, M., Navarro, A., Eljarrat, E., and Barceló, D. (2008). Development of a pressurized liquid extraction-solid-phase extraction followed by liquid chromatography-electrospray ionization tandem mass spectrometry method for the quantitative determination of benzoxazolinones and their degradation products in agricultural soil. Journal of Chromatography A *1179*, 190-197.

(Villanueva i Grimalt, 1996)
Villanueva, J., and Grimalt, J.O. (1996). Pitfalls in the chromatographic determination of the alkenone U37k index for paleotemperature estimation. Journal of Chromatography A *723*, 285-291.

(Villanueva et al., 1997)
Villanueva, J., Pelejero, C., and Grimalt, J.O. (1997). Clean-up procedures for the unbiased estimation of C-37 alkenone sea surface temperatures and terrigenous n-alkane inputs in paleoceanography. Journal of Chromatography A *757*, 145-151.

(Villas-Bôas et al., 2005)
Villas-Bôas, S.G., Højer-Pedersen, J., Åkesson, M., Smedsgaard, J., and Nielsen, J. (2005). Global metabolite analysis of yeast: evaluation of sample preparation methods. Yeast *22*, 1155-1169.

(Vis et al., 2007)
Vis, D., Westerhuis, J., Smilde, A., and van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. BMC Bioinformatics *8*, 322.

(Volkman et al., 1980)
Volkman, J.K., Johns, R.B., Gillan, F.T., Perry, G.J., and Bavor Jr, H.J. (1980). Microbial lipids of an intertidal sediment—I. Fatty acids and hydrocarbons. Geochimica et Cosmochimica Acta *44*, 1133-1143.

(Wahlroos i Virtanen, 1958)
Wahlroos, Ö., and Virtanen, A.I. (1958). On the Antifungal Effect of Benzoxazolinone and 6-Methoxybenzoxazolinone, Respectively, on Fusarium nivale. Acta Chemica Scandinavica *12*, 124-128.

(Weljie et al., 2006)
Weljie, A., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. (2006). Targeted Profiling:Â Quantitative Analysis of1H NMR Metabolomics Data. Analytical chemistry *78*, 4430-4442.

(Westerhoff i Palsson, 2004)
Westerhoff, H.V., and Palsson, B.O. (2004). The evolution of molecular biology into systems biology. Nat Biotech *22*, 1249-1252.

(Willis, 1993)
Willis, R.J. (1993). Terminology and trends in allelopathy. Allelopathy Journal *1*, 6 - 28.

(Winder et al., 2011)
Winder, C.L., Dunn, W.B., and Goodacre, R. (2011). TARDIS-based microbial metabolomics: time and relative differences in systems. Trends in microbiology *19*, 315-322.

(Winding et al., 2005)
Windig, W., Gallagher, N.B., Shaver, J.M., and Wise, B.M. (2005). A new approach for interactive self-modeling mixture analysis. Chemometrics and Intelligent Laboratory Systems *77*, 85-96.

(Winding i Stephernson, 1992)
Windig, W., and Stephenson, D.A. (1992). Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach. Analytical Chemistry *64*, 2735-2742.

(Wishart, 2008)
Wishart, D.S. (2008). Quantitative metabolomics using NMR. TrAC Trends in Analytical Chemistry *27*, 228-237.

(Wold, 1966)
Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least squares., in Multivariate Analysis, 391-420 Academic Press.

(Wold, 1995)
Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems *30*, 109-115.

(Wold et al., 1987)
Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems *2*, 37-52.

(Wold et al. 1993)
Wold, S., Johansson, A., and Cochi, M., eds. (1993). PLS-partial least squares projections to latent structures, in 3D QSAR in Drug Design. Theory, Methods and Applications 523-550 (Kubinyi, H., Ed.) ESCOM Science Publishers (Leiden).

(Wold et al. 2001)
Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems *58*, 109-130.

(Woods, 1960)
Woods, F. (1960). Biological antagonisms due to phytotoxic root exudates. Botanical Review *26*, 546-569.

(Work, 1949)
Work, E. (1949). Chromatographic investigations of amino acids from micro-organisms: I. The amino acids of Corynebacterium diphtheria. Biochimica et Biophysica Acta *3*, 400-411.

(Wu et al., 2000a)
Wu, H., Haig, T., Pratley, J., Lemerle, D., and An, M. (2000a). Allelochemicals in Wheat (Triticum Aestivum L.): Variation of Phenolic Acids in Root Tissues. Journal of Agricultural and Food Chemistry *48*, 5321-5325.

(Wu et al., 2000b)
Wu, H., Haig, T., Pratley, J., Lemerle, D., and An, M. (2000b). Distribution and Exudation of Allelochemicals in Wheat Triticum aestivum. Journal of Chemical Ecology *26*, 2141-2154.

(Wu et al., 2001)
Wu, H., Haig, T., Pratley, J., Lemerle, D., and An, M. (2001). Allelochemicals in Wheat (Triticum aestivum L.): Production and Exudation of 2,4-Dihydroxy-7-Methoxy-1,4-Benzoxazin-3-One. Journal of Chemical Ecology *27*, 1691-1700.

(Yunker et al., 2005)
Yunker, M.B., Belicka, L.L., Harvey, H.R., and Macdonald, R.W. (2005). Tracing the inputs and fate of marine and terrigenous organic matter in Arctic Ocean sediments: A multivariate analysis of lipid biomarkers. Deep Sea Research Part II: Topical Studies in Oceanography *52*, 3478-3508.

(Zadoks et al., 1974)
Zadoks, J.C., Chang, T.T., and Konzak, C.F. (1974). A decimal code for the growth stages of cereals. Weed Research *14*, 415-421.

(Zeaiter i Rutledge, 2009)
Zeaiter, M., and Rutledge, D. (2009). 3.04 - Preprocessing Methods, in Comprehensive Chemometrics, 121-231 (Brown S.D., Tauler, R., Walczak, B., Ed.) Oxford: Elsevier.

(Zwanenburg et al., 2011)
Zwanenburg, G., Hoefsloot, H.C.J., Westerhuis, J.A., Jansen, J.J., and Smilde, A.K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. Journal of Chemometrics *25*, 561-567.