Revista de la Societat Catalana de Química, núm. 10 (2011), p. 35-47
Filial de l'Institut d'Estudis Catalans
DOI: 10.2436/20.2003.01.27 ISSN: 2013-9853 (ed. electrònica) ISSN: 1576-8961 (ed. impresa)
http://revistes.iec.cat/index.php/RSCQ

Resolució multivariant en química: a la recerca de la bella simplicitat de la mesura

Multivariate resolution in chemistry: seeking the beautiful simplicity of the measurement

Anna de Juan,[1,*] Joaquim Jaumot,[1] Raimundo Gargallo[1] i Romà Tauler[2]

[1] *Universitat de Barcelona. Facultat de Química. Departament de Química Analítica. Grup de Quimiometria*

[2] *IDAEA-CSIC, Barcelona*

Al Dr. Casassas, agraint el mestratge científic i els bons consells

Resum. En quimiometria, existeix una família de mètodes, coneguts sota la denominació comuna de *resolució multivariant*, que permeten descriure fenòmens molt complexos i diversos a partir de la combinació de poques contribucions bàsiques fàcilment interpretables. La versatilitat i simplicitat dels models de resolució multivariant han fet que s'hagin aplicat amb èxit en el modelatge de processos d'origen químic i bioquímic i en la interpretació de dades *-òmiques* i ambientals.

Paraules clau: Quimiometria, resolució multivariant, MCR-ALS.

Abstract. There is a family of chemometric methods, called multivariate resolution methods, which allows the description of very diverse and complex phenomena through the combination of a small number of basic and easily interpretable contributions. The versatility and simplicity of the multivariate resolution models explains their successful application in modeling chemical and biochemical processes and in the interpretation of* –omic *and environmental data.

Keywords: Chemometrics, multivariate resolution, MCR-ALS.

Introducció

En l'art, els infinits matisos melòdics, de llums i de colors que percebem en una obra acabada són el resultat del mestratge de l'artista a l'hora de combinar un nombre molt petit de notes musicals o de pigments. La bellesa que deriva d'aquesta aparent complexitat no és res més que el resultat de les mescles sorgides de la inspiració de l'artista i d'uns elements bàsics extremament senzills.

En la ciència, com en l'art, la mesura obtinguda sembla cada cop més difícil d'interpretar, però la complexitat que suggereix l'observació directa és sovint aparent i també pot ser descrita per uns models bàsics altament simples i, per què no dir-ho, bells, ja que ens condueixen al coneixement científic de manera clara i precisa.

La quimiometria sorgeix com a disciplina que posa les eines estadístiques i matemàtiques al servei de la interpretació de la mesura química.[1,2] No és, doncs, una branca teòrica aliena al coneixement químic, sinó que pren el problema químic com a punt de referència i desenvolupa i adapta mètodes d'anàlisi de dades potents i rigorosos a les característiques específiques de la naturalesa del problema en estudi i del tipus de mesura que s'utilitza per a la seva investigació. Independentment del tipus de recerca que es dugui a terme, en quimiometria sempre regna el principi de parsimònia, l'esperit del qual fou introduït pel filòsof Guillem d'Occam,[3] que estableix que el millor mètode o model és sempre l'alternativa més simple que permet explicar el fenomen objecte d'estudi.[4]

Els mètodes de resolució multivariant són un dels millors paradigmes del principi de parsimònia. Són matemàticament molt simples, donen models petits de mida, fan ús del comportament natural de la mesura química i proporcionen resultats directes amb sentit químic.[5] Per fer-ne una ràpida descripció, tan sols cal dir que proporcionen models lineals i additius que descriuen l'evolució/variació d'un sistema (químic, biològic o ambiental) com la suma sospesada d'un nombre petit de contribucions bàsiques o components. La manera en què es produeix la transició des de la informació de les dades originals fins al model de contribucions bàsiques i la variació de la definició de *component* per adaptar els mètodes de resolució a l'estudi de problemàtiques extremament diverses són els temes que seran desenvolupats a les seccions següents.

Correspondència: Anna de Juan. Universitat de Barcelona. Facultat de Química.
Departament de Química Analítica. Grup de Quimiometria
C. de Martí i Franquès, 1-11. 08028 Barcelona
Tel.: +34 934 039 778. Fax: +34 934 021 233
A. e.: *anna.dejuan@ub.edu*

Fonaments dels mètodes de resolució multivariant. La recuperació del model de la mesura

La millor manera d'explicar el funcionament dels mètodes de resolució multivariant és triar un exemple químic que s'hi adapti perfectament, com ara les dades espectroscòpiques. La mesura espectroscòpica segueix la llei de Lambert-Beer, que és formalment idèntica al model bàsic dels mètodes de resolució multivariant. Tal com ja s'ha indicat anteriorment, aquests mètodes s'orienten a descriure l'evolució d'un sistema químic seguit mitjançant una mesura multivariant. Tal sistema podria ser un procés seguit espectroscòpicament, tot considerant el terme *procés* en un sentit ampli, que inclouria des d'una reacció química fins a una elució cromatogràfica o qualsevol canvi fisicoquímic que es reflectís mitjançant una variació del senyal espectroscòpic adquirit.

La informació procedent d'un procés seguit espectroscòpicament es pot organitzar en una matriu (taula) de dades, **D**, en la qual les files són els espectres recollits a cada estadi del procés (valor de pH, temps d'elució, etc.) (figura 1). Si pensem en la naturalesa de la mesura espectroscòpica, la taula de mesures originals pot ser descrita com la suma del senyal proporcionat pels components purs del procés (espècies químiques, compostos eluïts, etc.) (figura 1*a*). El senyal pur d'un component pot ser expressat a partir de la forma del seu espectre pur (s_i^T), sospesada en cada estadi del procés pel valor de concentració o d'abundància que li correspon (c_i). En termes matemàtics, això vol dir que la informació de la taula completa del senyal d'un component pur pot ser expressada mitjançant una díada de vectors ($c_i s_i^T$) (figura 1*b*). El model global de la mesura del procés es pot expressar d'una manera més compacta agrupant la

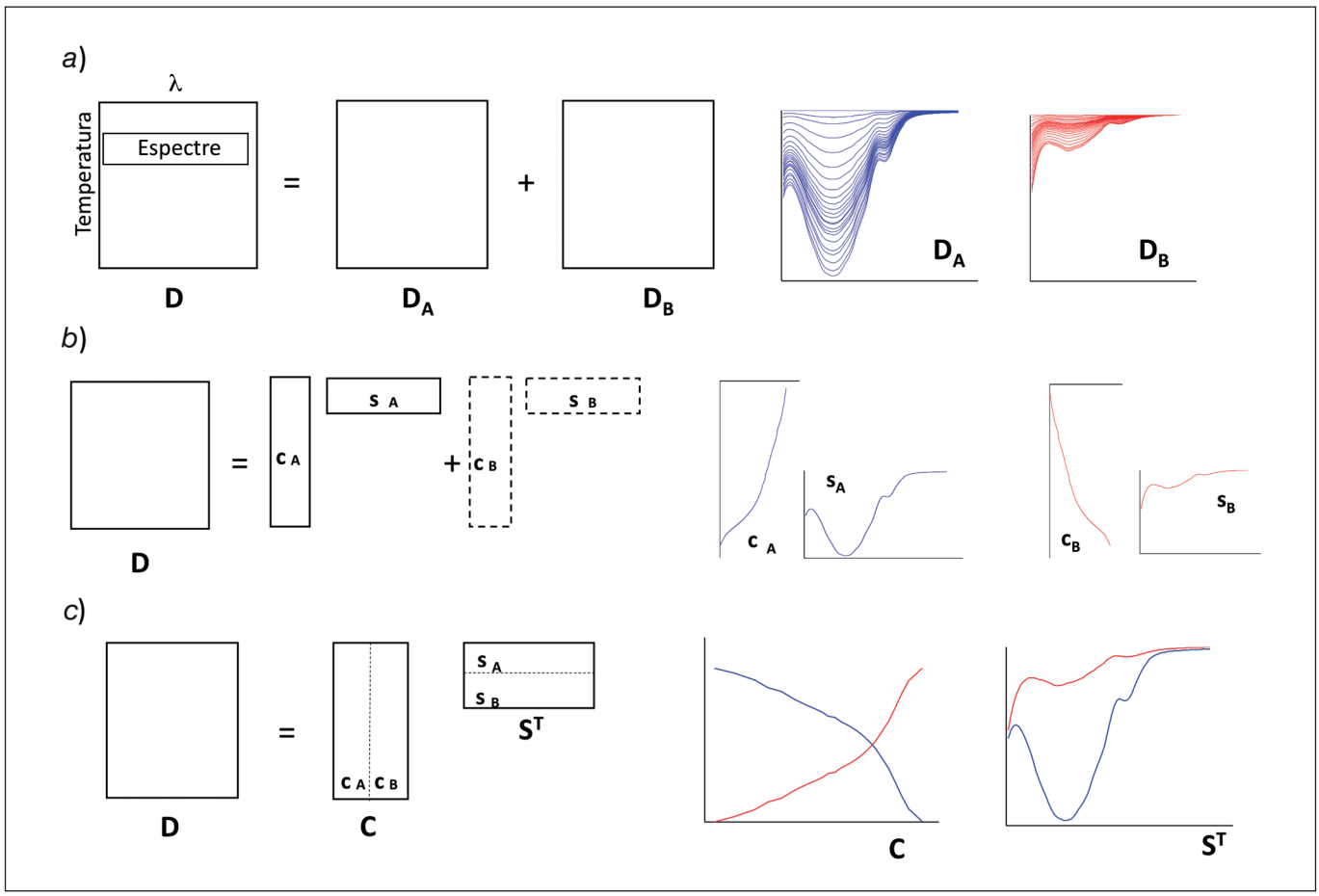


FIGURA 1. Taula de dades d'un procés seguit espectroscòpicament expressada com a: *a*) suma de les taules de senyals purs dels components del procés; *b*) suma de les díades $\mathbf{c}_i\mathbf{s}_i^T$ de cada component del procés, i *c*) model bilineal $\mathbf{CS}^T$.

suma de les díades de tots els components purs com a producte de dues matrius ($\mathbf{S}^T$ i **C**) que contenen els espectres purs dels components del procés i els seus perfils de concentració associats, respectivament. L'expressió final $\mathbf{D} = \mathbf{CS}^T$ és la forma matricial de la llei de Lambert-Beer i també el model bilineal bàsic dels mètodes de resolució multivariant (en dades experimentals, cal afegir una matriu **E** a aquesta equació, que reflecteix l'error experimental present a les dades).[5–7] Cal notar, en aquest punt, dos aspectes particularment rellevants. Primer: que la informació de partida de la matriu **D**, que pot constar de centenars o milers de files i columnes que contenen un senyal en el qual totes les contribucions del procés es troben barrejades, és perfectament descrita per un model de dimensions petites, ja que $\mathbf{S}^T$ i **C** tenen tantes files i columnes, respectivament, com components purs participen en el procés. Segon: que els perfils d'aquestes matrius (espectres purs i perfils de concentració) atorguen una informació química sobre el procés directament interpretable.

Si bé la mesura espectroscòpica té un model subjacent extremament senzill i interpretable, calen eines per poder obtenirlo a partir de la mesura original. Aquestes eines són els mètodes de resolució multivariant, que treballen amb la informació de partida mancada de selectivitat, **D**, per proporcionar el model bilineal i interpretable que la descriu, $\mathbf{CS}^T$.

Existeixen molts algorismes de resolució multivariant, però ens centrarem en la descripció d'un dels més simples i potents: el mètode de resolució de corbes per mínims quadrats alternats (*multivariate curve resolution-alternating least squares*, MCR-ALS).[8–11] Aquest mètode iteratiu consta de les etapes següents:

1) Determinació del nombre de components per a la descripció de la variació de la matriu de dades originals. Es pot saber *a priori* o pot ser obtingut amb altres mètodes quimiomètrics, com ara l'anàlisi de components principals (PCA).[12]
2) Generació d'una estimació inicial de la matriu **C** o $\mathbf{S}^T$. Es pot fer a partir del coneixement químic del sistema o amb mètodes quimiomètrics, com ara l'anàlisi de factors emergents (EFA)[13] o els algorismes de selecció de variables pures, com ara SIMPLISMA.[14]
3) Càlcul de la matriu **C** i de la matriu $\mathbf{S}^T$ a partir de la matriu **D** i de la matriu complementària del model bilineal, procedent de l'estimació inicial o del càlcul de la iteració anterior, per mínims quadrats sota restriccions.
4) Comprovació de la correcta reproducció de la matriu inicial **D** a partir del model bilineal calculat, $\mathbf{CS}^T$, segons uns criteris de convergència. Si no s'assoleix el criteri proposat, es retorna a l'etapa 3.

De fet, el pas 3 és l'etapa essencial de l'algorisme, i ho és en tant que la informació química que es posseeix sobre el sistema i sobre les propietats dels perfils de concentració i de resposta pot ser introduïda en l'optimització iterativa sota la forma de restriccions. Una restricció és qualsevol propietat química o matemàtica que compleixi sistemàticament els perfils del model bilineal.[8,9,15] L'aplicació de les restriccions als perfils de les matrius **C** i $\mathbf{S}^T$ els atorga sentit químic. Com a restriccions més habituals, es poden esmentar les següents:

– No-negativitat: força els perfils a tenir elements iguals o majors que zero. Això és aplicable a valors de concentració i a molts senyals espectroscòpics, com ara l'espectroscòpia d'absorció UV-visible o la fluorescència.
– Unimodalitat: permet la presència d'un únic màxim per perfil. Útil en perfils de concentració o senyals en forma de pic (per exemple, pics cromatogràfics o alguns senyals electroquímics) i en perfils de concentració de processos en els quals els components apareixen i desapareixen de manera seqüencial.
– Sistema tancat: és l'expressió del balanç de massa en sistemes en reacció.
– Selectivitat: força l'absència de tots els components menys el selectiu en algunes finestres de perfils de concentració o de respostes.

Hi ha altres restriccions que fan referència a l'estructura matemàtica de les dades o a la incorporació d'un model fisicoquímic explícit per modelar els perfils de concentració, però no es descriuran, atesa la seva complexitat i el fet que es poden trobar a les referències adjuntes.[9,16–18] Algunes característiques importants de les restriccions són que la seva aplicació és opcional i flexible (**C** i $\mathbf{S}^T$ es poden restringir de manera diferent, de la mateixa manera que cadascun dels seus perfils) i que se'n pot controlar la tolerància en el compliment estricte de les condicions que les defineixen. La versatilitat en l'aplicació de les restriccions fa que l'algorisme MCR-ALS s'adapti de manera específica a l'estudi d'un gran nombre de problemes químics i que en respecti la diversitat.

La resolució multivariant en l'anàlisi de dades espectroscòpiques. L'aplicació natural del mètode

Vista l'explicació de la secció anterior, no és estrany que la interpretació de processos seguits espectroscòpicament hagi estat històricament l'àrea natural d'aplicació dels mètodes de resolució multivariant. En aquest cas, el model de la mesura instrumental i el model bilineal són idèntics i hi ha una associació directa entre el concepte *component del model* (contribució) i el concepte *component químic* (entès com a 'compost, espècie química, forma fisicoquímica o estereoquímica que du associat un espectre pur clarament definit').

Malgrat la claredat entre l'associació dels models espectroscòpics i bilineals de resolució, la interpretació de processos no és un problema senzill, particularment quan estan associats a sistemes químics o biològics de gran complexitat, com ara les macromolècules de DNA o les proteïnes.[19–22] Per tal de comprendre un procés biològic o un procés complex en general, sovint no és suficient la realització d'un únic experiment, sinó que cal utilitzar experiments que es realitzin en condicions experimentals diferents i també cal seguir l'evolució del procés amb espectroscòpies diverses, que permetin detectar esdeveniments químics que es produeixen a diferents nivells estructurals o moleculars. La comprensió total del procés s'assoleix quan totes aquestes dades s'interpreten de manera conjunta.

Si bé els fonaments dels mètodes de resolució multivariant s'han explicat prenent com a referència l'anàlisi d'una única taula de dades, el model bilineal bàsic utilitzat en resolució s'estén a estructures de dades formades per l'acoblament de diverses matrius. Són el que s'anomena *estructures multiconjunt* (*multiset*). Aquestes estructures es poden configurar de maneres molt diverses (figura 2). Així, es poden acoblar matrius de dades en la direcció de les files (*row-wise augmented matrix*), que provinguin del seguiment del mateix procés amb tècniques espectroscòpiques diferents, o bé en la direcció de les columnes (*column-wise augmented matrix*), que siguin el

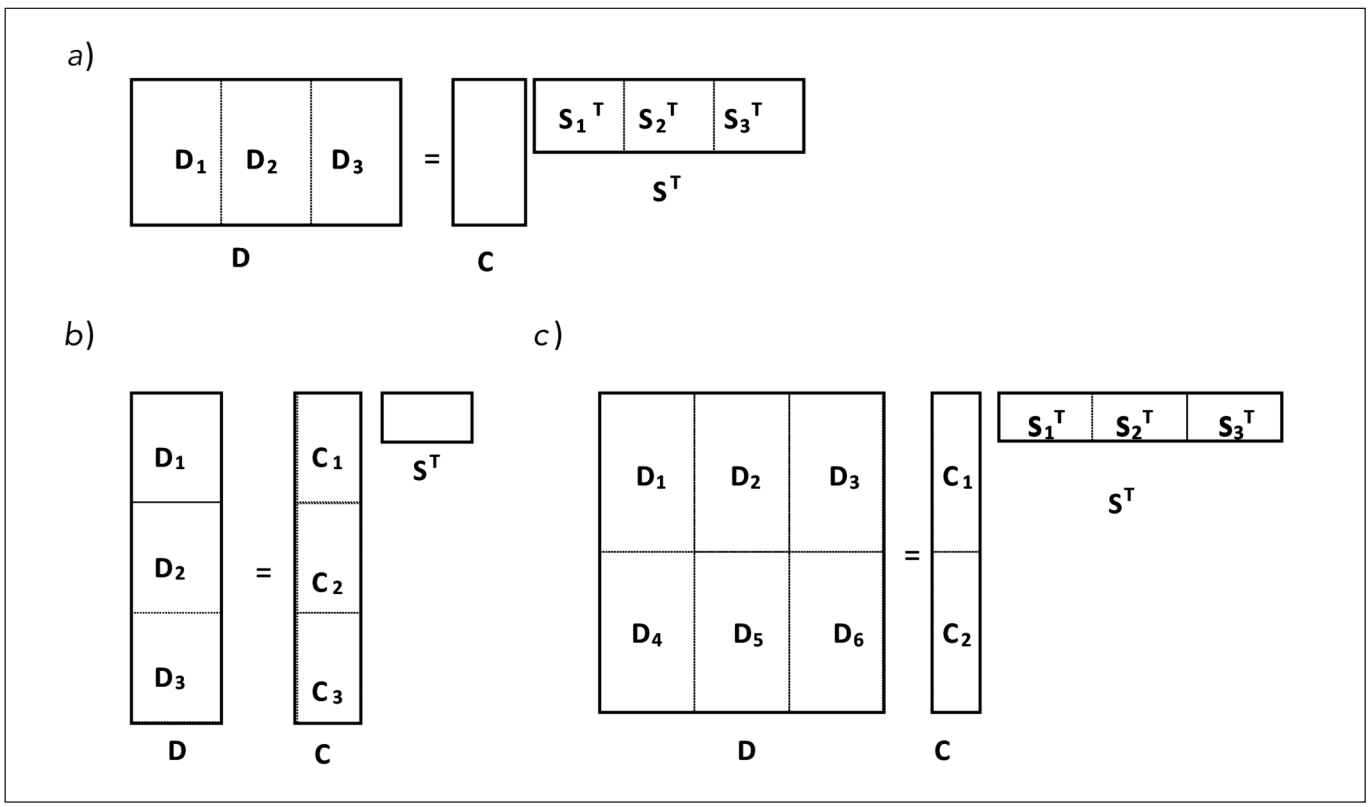


FIGURA 2. Tipus d'estructures de dades multiconjunt i models bilineals associats: *a*) matriu augmentada en la direcció de les files (*row-wise augmented data matrix*); *b*) matriu augmentada en la direcció de les columnes (*column-wise augmented data matrix*), i *c*) matriu augmentada en la direcció de les files i de les columnes (*row-wise and column-wise augmented data matrix*).

resultat de realitzar experiments replicats o en condicions diferents sobre el mateix sistema utilitzant la mateixa tècnica espectroscòpica. També es pot realitzar el doble acoblament en la direcció de les files i de les columnes (*row-wise and column-wise augmented matrix*). L'ús d'aquests acoblaments proporciona resultats extremament robustos i permet la interpretació global i exhaustiva de l'evolució del procés i de les característiques estructurals dels components que hi participen.[9,10]

A tall d'exemple de l'aplicació de la resolució multivariant a estructures multiconjunt procedents del seguiment d'un procés biològic, presentem el cas de l'estudi de les transicions de la proteïna hemoglobina induïdes per la variació del pH.[23] En aquest cas, s'han utilitzat cinc tècniques espectroscòpiques per tal de veure quines són les transformacions que pateix la proteïna en els diferents nivells estructurals i en la coordinació del seu grup *hemo-*. Les tècniques han estat el dicroisme circular (DC) a la regió de l'UV llunyà, sensible a les variacions en l'estructura secundària de la proteïna (formació d'hèlix, fulles beta, etc.); el DC a la regió de l'UV proper i la fluorescència, sensible a canvis en l'estructura terciària (formació de l'estructura globular); el DC a la regió Soret, que percep canvis associats a la coordinació del grup *hemo-*, i l'espectroscòpia d'absorció UV-visible, que aprecia els canvis anteriors de manera global. S'ha recollit per a cadascuna de les tècniques una sèrie d'espectres mesurats en cadascun dels valors de pH del procés (figura 3*a*).

Les cinc matrius de dades procedents d'aquestes tècniques es poden agrupar per donar una estructura multiconjunt multitècnica, com la presentada a la figura 2*a*. És important esmentar que, en aplicar restriccions a estructures multiconjunt, també es poden fer diferències entre les condicions

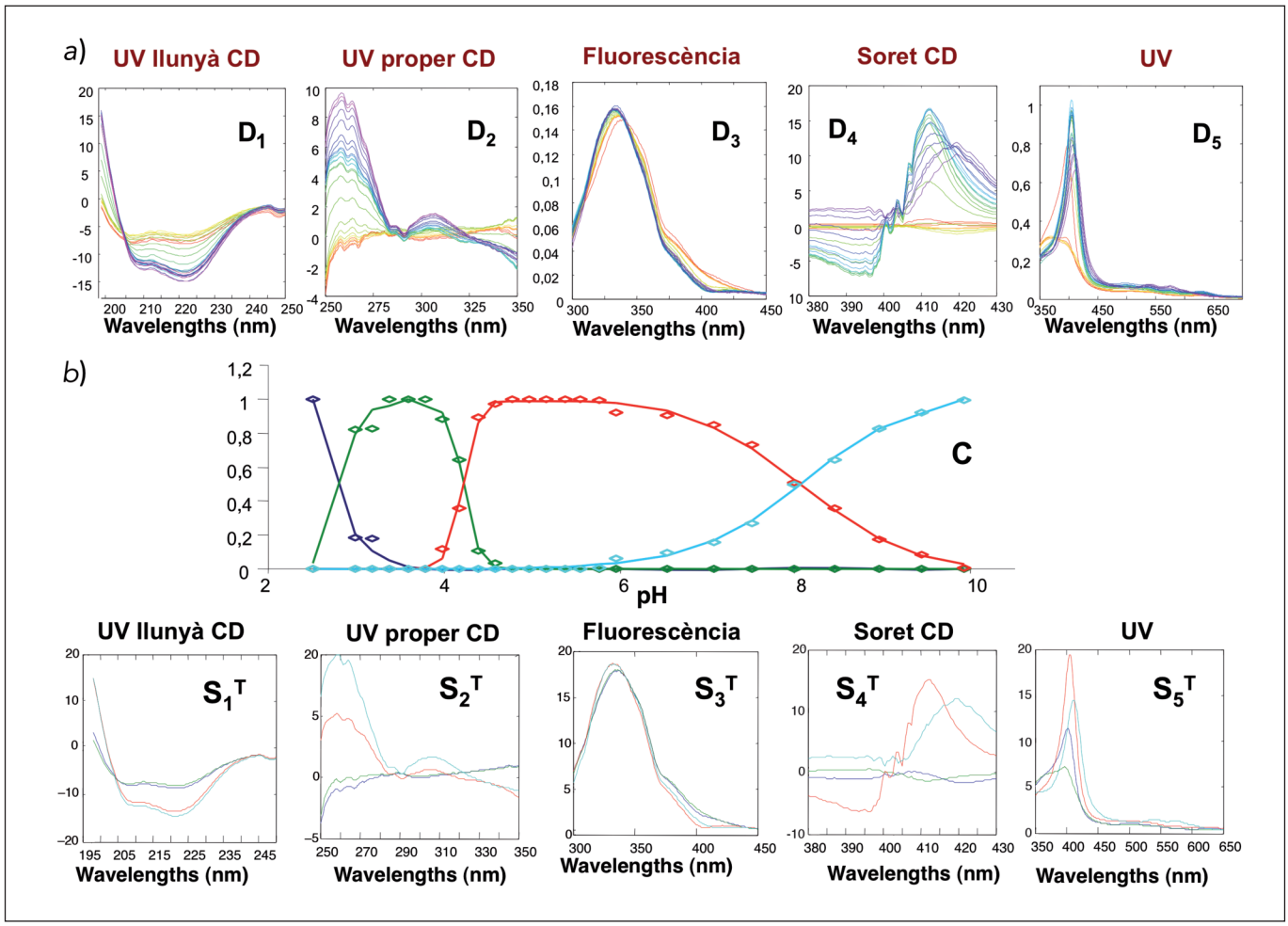


FIGURA 3. Transicions conformacionals de l'hemoglobina amb el pH: *a*) dades espectroscòpiques recollides durant el procés, i *b*) perfils de concentració i espectres resolts de les conformacions proteiques involucrades en el procés.

aplicades a les diferents submatrius (en aquest cas, a les submatrius $\mathbf{S}_i$ que formen la matriu augmentada $\mathbf{S}^T$).[9,10] Així, en aquest cas particular, mentre que els espectres purs de les submatrius de fluorescència i absorció molecular seran sotmesos a la no-negativitat, les submatrius de dicroisme circular no seran sotmeses a aquesta restricció, ja que els valors d'el·lipticitat negatius són possibles en aquest tipus de tècnica. Pel que fa a l'única matriu de concentracions, s'han aplicat restriccions de no-negativitat, unimodalitat i sistema tancat per modelar les transicions del sistema. Cal insistir en la gran robustesa de la descripció de l'evolució del procés proporcionada per l'única matriu $\mathbf{C}$, ja que els perfils obtinguts són vàlids per descriure la variació de totes les tècniques espectroscòpiques emprades i no pateixen del soroll experimental o de variacions poc definides o inexistents en alguna de les tècniques instrumentals utilitzades.

La figura 3*b* mostra els resultats de la resolució associada a aquest procés. S'hi observen quatre contribucions, la naturalesa de les quals podrà ser elucidada a partir de la informació que fa referència als valors de pH als quals les transicions entre components tenen lloc i a les característiques espectrals que s'observen en els seus espectres purs associats. Així, la transició a pH 4 suggereix la desnaturalització de la proteïna (és a dir, una pèrdua de les estructures secundària, terciària i del grup *hemo*-). Aquesta hipòtesi es corrobora en estudiar l'espectre pur de les dues conformacions involucrades en la transició, que passen de tenir una gran estructura en totes les regions de DC en la forma nativa de la proteïna (abundant a valors de pH entre 4 i 8) a mostrar senyals molt menys intensos o inexistents, típics de la conformació desnaturalitzada (més abundant a valors de pH inferiors a 4). També s'identifica clarament la transició que té lloc a un pH al voltant de 8, que reflecteix el pas de l'hemoglobina a l'oxihemoglobina. En aquest cas, l'espectre pur associat a l'oxihemoglobina presenta clars signes de la identitat d'aquesta estructura, com ara un desplaçament cap al roig a la banda de la regió Soret i l'aparició de petites bandes entre la zona de 500 i 600 nm en l'espectroscòpia d'absorció molecular.[24] A més, els espectres resolts també indiquen que aquest canvi en la coordinació no afecta l'estructura secundària de la proteïna (espectres purs idèntics a la zona de l'UV llunyà de CD), però sí que té un efecte en l'estructura globular terciària de la proteïna. L'altra transició a valors de pH més àcids s'efectua entre formes desnaturalitzades de la proteïna.

És important el fet de notar que tota aquesta informació s'hauria obtingut de manera parcial i menys ben definida si s'haguessin estudiat aquestes tècniques separadament. Així, doncs, certs esdeveniments, com ara la coordinació del grup *hemo*-, no afecten tots els nivells estructurals de la proteïna i, per tant, no serien detectats amb algunes de les tècniques. I tampoc no seria possible la detecció i el modelatge d'espècies que no presentessin senyal en certes regions (com ara la conformació desnaturalitzada a la regió Soret). Per tant, sempre que sigui possible, cal treure partit de la potència de l'anàlisi d'estructures multiconjunt en la interpretació de processos, ja que sempre proporcionarà una visió global igual o millor que la proporcionada per l'anàlisi individual de les taules de dades i, en qualsevol cas, sempre serà més fiable i robusta.

En l'exemple presentat, la complexitat prové del procés químic estudiat. Hi ha, però, altres casos de dades espectroscòpiques en els quals l'aparent dificultat procedeix de la naturalesa intrínseca de la mesura instrumental: aquest és el cas de les imatges hiperespectrals. La imatge hiperespectral d'una mostra química està formada per milers d'espectres recollits en diferents punts de la superfície de la mostra i anomenats *píxels*.[25] L'interès de les imatges hiperespectrals rau en l'obtenció simultània d'informació estructural i espacial sobre els components de la mostra. La informació estructural es dedueix de la forma global dels espectres o de la interpretació de les seves bandes espectrals, mentre que la informació de distribució espacial, que és la característica diferencial d'una imatge, prové del fet que cadascun dels espectres recollits s'associa a una posició espacial concreta de la mostra. Per tant, la variació espectral entre els píxels de la nostra imatge no és res més que el reflex de la variabilitat de la distribució espacial dels components a la mostra analitzada.[25,26]

Les imatges hiperespectrals poden ser visualitzades d'una manera molt clara com un cub de dades, en el qual dues de les dimensions s'associen a les coordenades espacials dels píxels, mentre que la tercera és la dimensió espectral (figura 4*a*). Si aquest cub es desplega, s'obté una taula de dades espectroscòpiques, en la qual els espectres dels píxels estan l'un sota l'altre i s'hi poden aplicar els mètodes de resolució multivariant, tal com hem vist a l'exemple anterior del procés. Els resultats de la resolució multivariant proporcionen de manera directa els espectres purs dels constituents de la imatge (matriu $\mathbf{S}^T$) i, per tal de recuperar la informació relativa a la seva distribució espacial, tan sols cal replegar cadascun dels perfils

de concentració (matriu **C**) segons l'estructura espacial original de la imatge (figura 4*a*).[27] A tall d'exemple, es mostren els resultats de la resolució d'una imatge hiperespectral Raman procedent d'un càlcul urinari (figura 4*b*).[25] En ella s'observen els mapes de distribució dels tres components del càlcul, identificats com a *whewellita*, *whedellita* i *dahlita*, gràcies a la bona correspondència entre la forma dels espectres resolts (en negre) i els espectres de referència procedents d'una biblioteca d'espectres Raman associada a compostos freqüents en càlculs urinaris (en vermell).[28]

L'exemple presentat mostra que les mesures espectroscòpiques d'estructura aparentment complexa, com ara les imatges hiperespectrals i d'altres que es visualitzen de forma tridimensional, com ara les procedents de cromatografies bidimensionals amb detecció espectroscòpica[29] o de dades espectroscòpiques 2D,[30] també poden ser analitzades amb mètodes de resolució multivariant amb uns òptims resultats.

La resolució multivariant en dades ambientals i *–òmiques.* L'adaptació del concepte de *component*

L'adequació dels mètodes de resolució multivariant a les mesures espectroscòpiques no presenta cap dubte, ja que ambdós comparteixen el mateix model bilineal subjacent. Ara bé, és encara lícita o útil la resolució multivariant per a l'estudi d'altres taules de dades en les quals el model bàsic és desconegut o inexistent? Cal acollir-se de nou al principi de parsi-

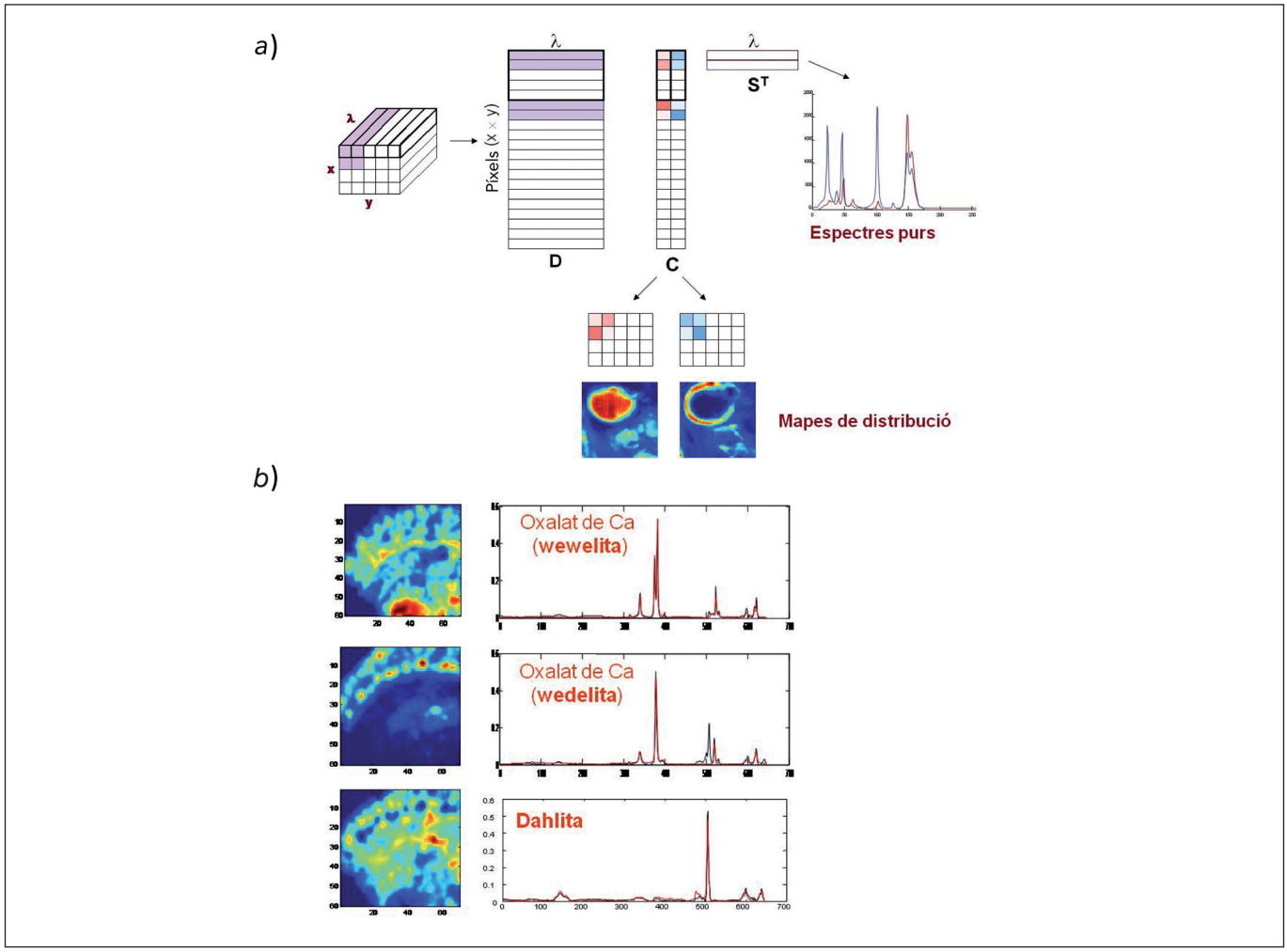
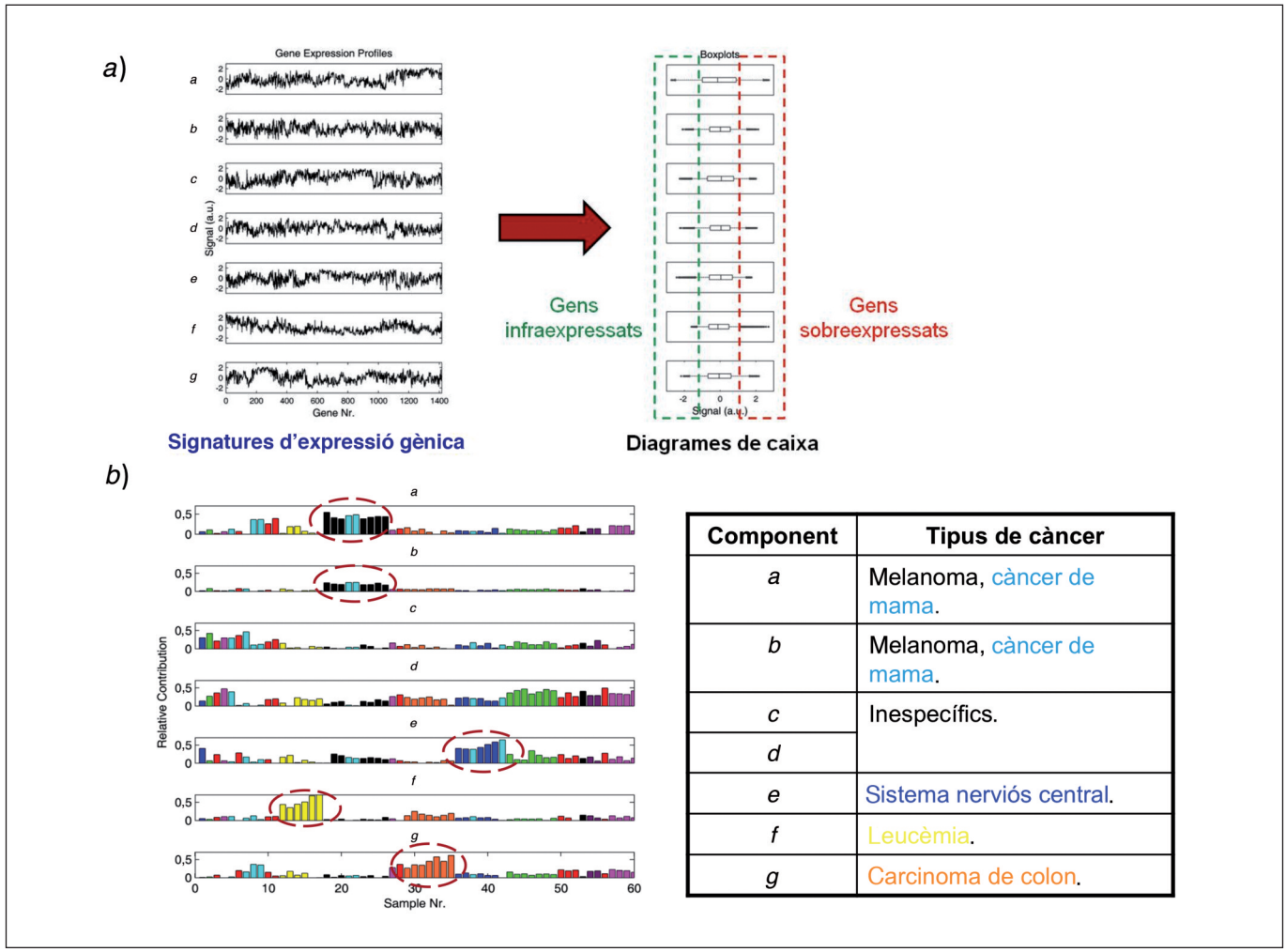


FIGURA 4. *a*) Resolució multivariant aplicada a l'anàlisi d'imatges hiperespectrals, i *b*) mapes de distribució i espectres resolts obtinguts a partir de la imatge hiperespectral Raman d'un càlcul urinari.

mònia i respondre que, a falta d'un model explícit de comportament, el fet de provar de trobar pautes o relacions entre mostres o variables a partir del model més simple, és a dir, el bilineal, sembla la primera opció que cal assajar. En fer aquesta tria, caldrà afrontar nous aspectes, com ara la reformulació del concepte *component*, del significat científic de les contribucions bàsiques del model bilineal. En aquest sentit, les dades *-òmiques* i les dades ambientals s'han revelat com a camps en els quals la resolució multivariant proporciona resultats sorprenentment útils i interpretables.

Un dels exemples més emblemàtics i pioners de dades *-òmiques* són les mesures procedents de micromatrius de DNA (*DNA microarrays*).[31] Aquestes micromatrius són plaques que contenen centenars o milers de pouets, cadascun dels quals s'utilitza per estudiar l'expressió d'un gen determinat en hibridar-se a una seqüència d'una línia cel·lular d'interès, que pot correspondre a una patologia o a una condició vital concreta. Es pot visualitzar una micromatriu de DNA com una taula d'experiments, en la qual les files designen les línies cel·lulars objecte d'estudi i les columnes, els gens dels quals es vol estudiar la resposta en aquelles línies cel·lulars. Sense entrar en detall en la tecnologia de les micromatrius de DNA, només cal dir que el comportament d'un gen davant d'una certa línia cel·lular pot ser de tres tipus: normal, la qual cosa vol dir que interacciona de la mateixa manera amb la línia cel·lular en estudi que amb una línia cel·lular de control, en la qual la patologia o condició en estudi és absent; sobreexpressat, quan el gen interacciona de manera clarament preferent amb la línia cel·lular d'interès, o infraexpressat, quan la interacció preferent es dóna amb la línia cel·lular de control. Els gens d'interès sempre seran aquells que es trobin sobreex-



Component	Tipus de càncer
a	Melanoma, càncer de mama.
b	Melanoma, càncer de mama.
c	Inespecífics.
d	
e	Sistema nerviós central.
f	Leucèmia.
g	Carcinoma de colon.

FIGURA 5. Resultats de la resolució multivariant procedents de l'anàlisi d'una micromatriu de DNA: *a*) signatures gèniques resoltes i diagrames de caixa relacionats, i *b*) perfils mostrals relatius a les signatures gèniques bàsiques. Els tipus diferents de línies cel·lulars tumorals es diferencien pel color.

pressats o infraexpressats davant d'una certa línia cel·lular, ja que són els que presenten un comportament clarament diferenciat de la resta en presència de la patologia o condició d'interès.

Des del punt de vista de la mesura, els resultats numèrics d'un experiment de micromatrius de DNA mantenen l'estructura de la taula d'experiments. Cada experiment proporciona un únic valor numèric: un valor d'expressió gènica, que és gran i positiu, si el gen està sobreexpressat; gran i negatiu, quan està infraexpressat, i proper a zero, quan presenta un comportament normal. Per tant, la taula de dades (matriu) procedent d'una micromatriu de DNA té tantes files com línies cel·lulars i tantes columnes com gens assajats. Cada fila d'aquesta taula conté el perfil d'expressió gènica d'una línia cel·lular particular, és a dir, la resposta multivariant formada per l'expressió gènica de tots els gens assajats.

Arribats a aquest punt, cal pensar quina mena d'informació trobaríem en aplicar mètodes de resolució multivariant a aquest tipus de mesures, què significarien les contribucions bàsiques obtingudes i quin sentit biològic tindrien les matrius del model bilineal**.** La hipòtesi a considerar és que, de la mateixa manera que la combinació d'un nombre petit d'espectres purs pot descriure tota la variabilitat de formes espectrals que es mesuren durant un procés, potser pot existir un nombre limitat de signatures gèniques bàsiques que, combinades adequadament, puguin reproduir el perfil d'expressió gènica de qualsevol línia cel·lular de la micromatriu en estudi. Aquestes signatures bàsiques mostrarien, com a informació d'interès, grups de gens que sistemàticament apareixen sobreexpressats i infraexpressats de manera simultània. En aquest context, per fer més entenedor el model bilineal, podríem utilitzar la notació $\mathbf{D} = \mathbf{S}\mathbf{G}^T$, on $\mathbf{G}^T$ seria la matriu que conté els perfils de les signatures gèniques bàsiques de la micromatriu i **S** estaria formada pels perfils mostrals (*sample profiles*) relacionats. Cada perfil mostral inclouria l'abundància (importància) d'una determinada signatura gènica bàsica per reproduir l'expressió gènica de les diferents línies cel·lulars. En observar els perfils mostrals, es podria veure si certes patologies estan lligades de manera dominant a una signatura gènica en particular o si la combinació de signatures per a la seva descripció és més inespecífica.[32,33]

Per a una millor comprensió de la teoria, es descriu a continuació l'exemple de l'estudi d'una micromatriu de DNA formada per seixanta línies de cèl·lules canceroses associades a diferents tipus de tumors, sobre les quals s'han assajat mil quatrecents setze gens.[32] Es tracta del conjunt de dades NCI60, posat a lliure disposició científica pel grup de Ross i els seus col·laboradors.[34] En utilitzar el mètode de resolució, les úniques restriccions aplicades han estat la no-negativitat per definir els perfils mostrals (les contribucions de les signatures gèniques bàsiques han de ser positives) i una condició de normalització per a les signatures gèniques bàsiques obtingudes, que s'adapta a la naturalesa de la mesura instrumental utilitzada.

Les dades van poder ser descrites amb set contribucions, set signatures gèniques. Tal com es veu a la figura 5*a*, les signatures gèniques no donen una informació directa interpretable. A partir dels seus valors, es construeixen diagrames de caixa, però són els valors extrems superiors i inferiors, que assenyalen els gens sobreexpressats o infraexpressats, respectivament, els que són objecte d'estudi posterior. La naturalesa d'aquests gens es compara amb bases de dades d'informació gènica ontològica per confirmar associacions gèniques de les que s'hagi pogut tenir constància amb anterioritat, o bé serveixen de punt de partida per a la descoberta de noves associacions gèniques. Per a més detalls sobre la informació gènica de les contribucions trobades pels mètodes de resolució en aquest exemple, adrecem el lector a la referència.[32] Són de més fàcil interpretació els perfils mostrals resolts, presentats a la figura 5*b*, on les línies cel·lulars dels diferents tipus de tumor es marquen amb un codi de color. En aquest cas, es veu clarament que certes signatures gèniques bàsiques s'associen d'una manera molt dominant a certs tipus de tumor (com ara la signatura *g*, al carcinoma de colon; la *f*, a la leucèmia; la *e*, a tumors del sistema nerviós central, o les *a* i *b*, al melanoma i al càncer de mama). D'altres són més inespecífiques i es podrien associar a signatures gèniques pròpies de qualsevol procés cancerós o de qualsevol estat cel·lular.

En el context de les dades *–òmiques* i, en concret, de les genòmiques, cal insistir en la utilitat dels mètodes de resolució com una eina exploratòria de relacions gen–gen i gen–línia cel·lular molt potent, per a la qual no cal establir hipòtesis prèvies de tipus biològic. Els resultats obtinguts poden ser validats amb altres fonts d'informació independents o poden proporcionar noves vies d'investigació no considerades.

L'estudi, el control i la vigilància ambientals són també un camp en el qual es generen quantitats ingents de dades que cal tractar i interpretar adequadament i per a les quals calen

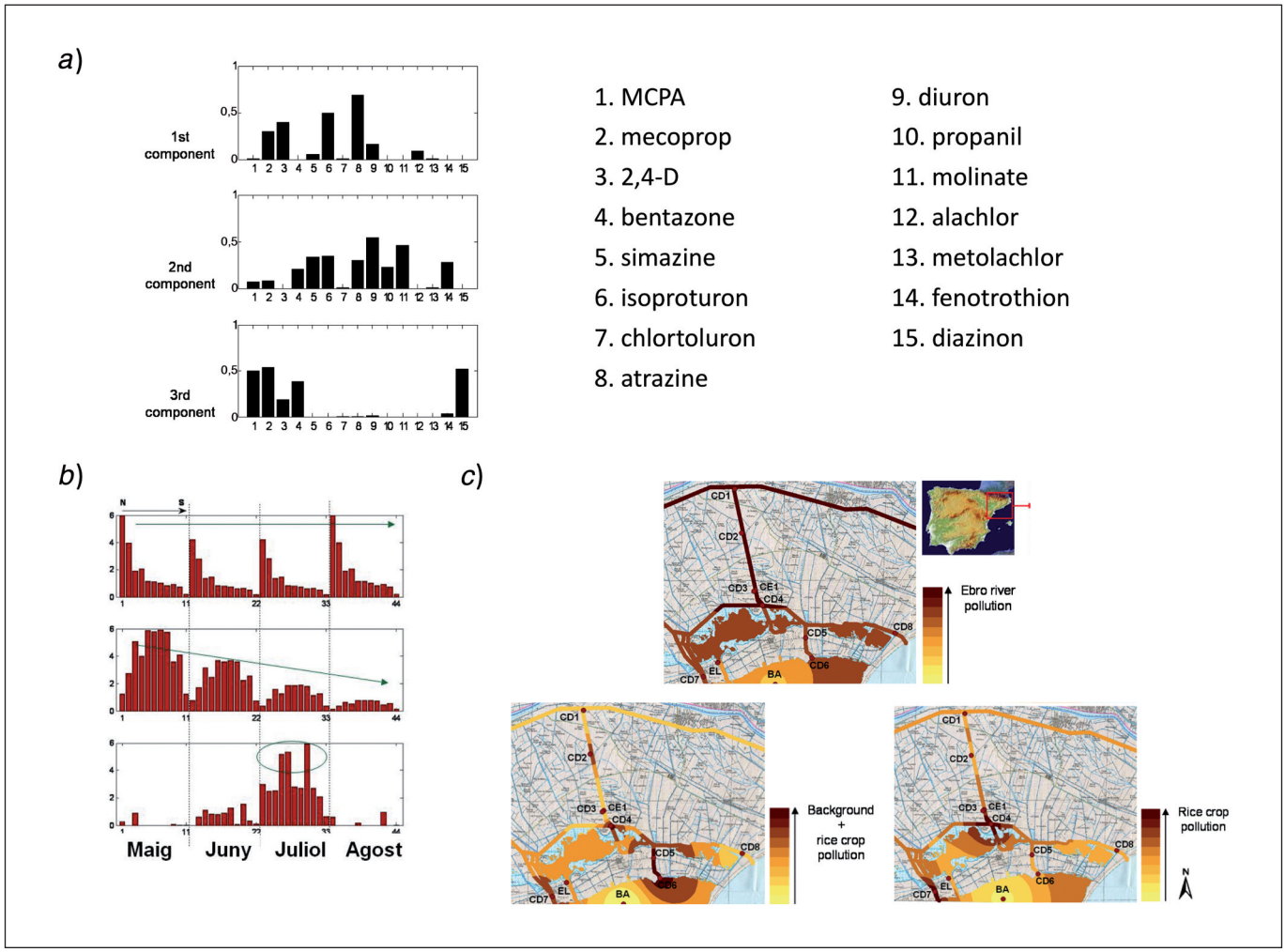


FIGURA 6. Resultats de la resolució multivariant aplicada a l'estudi de la contaminació del delta de l'Ebre: *a*) perfils composicionals; *b*) perfils geograficotemporals de les fonts de contaminació, i *c*) representació de les pautes de contaminació geogràfiques amb la incorporació dels resultats de la resolució multivariant en un sistema d'informació geogràfica (GIS).

eines potents d'anàlisi de la informació.[35] És molt freqüent el fet de trobar-se amb mesures que es poden organitzar d'una manera bastant natural com a taules de dades de naturalesa diversa. Així, podem pensar en taules on les files designin punts geogràfics o de mostratge i les columnes, concentracions o continguts de compostos o de contaminants, o bé en taules on les direccions siguin de naturalesa temporal (mesos, anys, estacions, etc.)[36,37] o bé designin compartiments de l'ecosistema (aigua, sediments, plantes, éssers vius, etc.).[17,38] Així, es pot intuir que les combinacions d'informació poden ser molt diverses i, pel mateix motiu, que la presència d'estructures multiconjunt en aquest tipus d'estudi és també molt freqüent (per exemple, un agrupament de taules de dades geograficocomposicionals, cadascuna de les quals adquirida en una època de l'any diferent, o bé un conjunt de taules que fan referència als mateixos punts de mostratge, però que contenen informació diversa: concentració de contaminants, paràmetres fisicoquímics, etc.).[38,39]

També en aquest cas, com en el dels estudis de dades *-òmiques*, la reformulació del concepte *component* és important. En aquest cas, però, es pot resoldre d'una manera senzilla gràcies al suport de la definició de *component ambiental* proposada pels especialistes del camp. En efecte, els models de receptor que existeixen des de fa dècades en l'àrea del medi ambient postulen que les mesures globals de concentracions de contaminants o d'altres compostos que es troben en el medi ambient es poden descriure com el resultat de l'aportació de diferents fonts ambientals.[40,41] Aquestes fonts tenen una composició definida i se'ls poden associar pautes de variació geograficotemporals concretes. Així, doncs, en el context de les dades ambientals, un *component* és una font am-

biental i l'escenari mediambiental que representen una o més taules de dades pot ser descrit com la suma de les aportacions d'un nombre petit d'aquestes fonts. En una taula ambiental típica, on les files poden ser punts de mostratge i les columnes, concentracions de contaminants, el model bilineal de resolució $\mathbf{D} = \mathbf{CS}^T$ es reinterpreta dient que la matriu $\mathbf{S}^T$ conté els perfils composicionals de les fonts mediambientals (és a dir, la representació de la proporció relativa dels diferents compostos que la componen) i que la matriu **C** conté els perfils geogràfics relacionats, que marquen l'aportació de cada font en funció del punt geogràfic objecte d'estudi. La matriu **C** pot marcar pautes geogràfiques de variació d'una font, però també pautes temporals o pautes de variació entre compartiments d'un ecosistema; tot dependrà de la definició de les files de la taula de dades que s'estudiï.

Com a exemple que il·lustra les particularitats de les dades ambientals i el resultat proporcionat pels mètodes de resolució, es presenta un estudi sobre la contaminació del delta de l'Ebre.[39] Per tal d'entendre aquest fenomen mediambiental i les característiques composicionals, geogràfiques i temporals de les fonts de contaminació d'aquesta àrea, es van efectuar mesures de la concentració de quinze pesticides en onze punts geogràfics del delta, que incloïen zones més properes al llit fluvial, canals de drenatge i d'irrigació dels arrossars i zones de la desembocadura més properes al mar. La determinació de la concentració dels pesticides es va dur a terme un cop al mes, durant els mesos de maig a agost de 2006. Totes aquestes observacions van donar lloc a quatre taules de dades geograficocomposicionals, cadascuna d'elles relativa a un dels mesos en els quals es van realitzar la presa de mostra i les determinacions analítiques corresponents. Aquestes taules de dades van ser organitzades com una estructura multiconjunt, del tipus presentat a la figura 2*b*, en la qual la direcció comuna de les columnes la formen els quinze pesticides determinats i les quatre taules de dades dels diferents mesos, amb els onze punts geogràfics mostrejats, es disposen l'una sota l'altra. En la resolució d'aquesta estructura multiconjunt es va aplicar la restricció de no-negativitat a tots els perfils (les aportacions de les fonts de contaminació mai no poden ser negatives) i es va forçar que la pauta de variació geogràfica d'algunes fonts fos idèntica als quatre mesos assajats (restricció de trilinealitat parcial).[9,39]

En aquest cas, el fenomen de contaminació del delta va poder ser descrit mitjançant tres fonts de contaminació, de les quals es van obtenir els perfils de composició, geogràfics i temporals. La figura 6*a* mostra els perfils composicionals de les fonts. La primera s'interpreta com una contribució de fons del riu, ja que conté en major proporció pesticides que són d'ús generalitzat en agricultura; la segona s'associa d'una manera molt directa al conreu de l'arròs, ja que els pesticides dominants són específics per a aquest tipus de cultiu, i la tercera es creu més lligada a certes pràctiques de reg dels arrossars. La figura 6*b* presenta els perfils geograficotemporals de cadascuna d'aquestes fonts; les línies verticals puntejades a cada gràfic separen la informació recollida al llarg dels diferents mesos. Atès que les dues primeres fonts (contribució de fons del riu i conreu de l'arròs) són sempre presents en els mesos estudiats, es van forçar els seus perfils a tenir la mateixa forma en cadascun dels mesos, la qual cosa va donar resultats satisfactoris i interpretables. Per al cas de la contribució de fons del riu, s'aprecia una major abundància en els punts més alts del curs del riu (els punts geogràfics, a les taules de dades, estan ordenats de manera decreixent, segons la distància a la desembocadura), mentre que en el cas dels arrossars, els punts que mostren més contribució de la font són els ubicats a les zones de cultiu. En la tercera font de contaminació, la informació geogràfica varia entre mesos, però el mes de juliol, en el qual és més present, els punts de més abundància coincideixen amb canals d'irrigació o de drenatge. L'evolució temporal es reflecteix en la intensitat relativa de l'aportació de les fonts al llarg dels diferents mesos. Hi ha una gran estabilitat en la font de contribució de fons del riu, ja que sempre és present, mentre que la intensitat de la font relativa al conreu de l'arròs baixa en el decurs dels mesos, ja que l'administració dels pesticides es realitza pel mes de maig i, amb el temps, els pesticides es degraden o es dilueixen. La tercera font és clarament més present en el mes de juliol, el mes més sec, en el qual cal realitzar certes pràctiques de retroirrigació per assegurar la humitat suficient als conreus.

La informació presentada és coherent i de gran interès, però, tot i el petit nombre de punts de mostratge analitzats en aquest exemple, es pot intuir la problemàtica d'interpretar les pautes de variació geogràfica de les fonts mediambientals en taules de dades en les quals els punts de mostratge són molt més nombrosos i en les quals l'ús d'un perfil lineal no permet incorporar la situació geogràfica relativa d'uns punts respecte dels altres. Per això, en aquest tipus d'aplicacions, la interpretació de les pautes de variació geogràfica de les fonts millora extraordinàriament fent ús dels sistemes d'informació geogràfica (GIS).[39,42] En el GIS, els valors d'abundància del perfil geogràfic resolt de cada font són utilitzats com a informació

de partida per configurar un mapa d'abundància de la font, en el qual els valors de les coordenades geogràfiques no mostrejades seran estimats per interpolació o *krigatge* dels valors trobats mitjançant els mètodes de resolució per als punts geogràfics reals estudiats. La figura 6c mostra els mapes trobats per GIS a l'exemple presentat. La informació descrita al paràgraf anterior es visualitza i s'interpreta d'una manera molt més clara sobre el seu entorn geogràfic real.

Conclusions

S'ha mostrat la utilitat dels mètodes de resolució multivariant en contextos força diversos. La flexibilitat i la facilitat d'aplicació d'aquests mètodes, que fa que s'adaptin a les especificitats dels problemes estudiats, i la simplicitat del model obtingut, que atorga una interpretabilitat directa i clara als resultats obtinguts, són els punts clau que justifiquen l'ús cada vegada més estès d'aquestes eines. Ben establerts per als camps d'aplicació més coneguts, l'explotació dels mètodes de resolució no està exhaurida i cal esperar trobar nous camps d'aplicació, noves maneres d'incorporar coneixement científic per a la millora dels perfils obtinguts i nous usos dels perfils bàsics resolts, que comprimeixen la informació de les dades brutes d'una manera eficaç i interpretable com a punt de partida en estudis addicionals. No faltarà creativitat ni una forta voluntat multidisciplinària per explorar totes aquestes noves i atractives possibilitats en el futur.

Agraïments

Els autors volen agrair el suport econòmic del Govern espanyol (Projecte CTQ2009-11572) i el reconeixement com a grup de recerca consolidat (2009 SGR 45) a la Generalitat de Catalunya.

Referències

[1] Massart, D. L.; Vandeginste, B. G. M.; Lewi, P. J.; Smeyers-Verbeke, J.; Buydens, L. M. C.; Jong, S. de. *Handbook of Chemometrics and Qualimetrics*. Elsevier, **1998**.
[2] *Comprehensive Chemometrics*. Brown, S.; Tauler, T.; Walczak, B. (ed.). Elsevier, **2009**.
[3] Hoffmann, R.; Minkin, V. I.; Carpenter, B. K. *Ockham's Razor and Chemistry, HYLE—Inter. J. Philosophy of Chemistry* **1997**, *3*, 3.
[4] Seasholtz, M. B.; Kowalski, B. R. *Anal. Chim. Acta* **1993**, *277*, 165.
[5] Rutan, S. C.; Juan, A. de; Tauler, R. a *Comprehensive Chemometrics*. Brown, S.; Tauler, R.; Walczak, B. (ed.). Elsevier, vol. 2, **2009**, 249.
[6] Juan, A. de; Tauler, R. *Anal. Chim. Acta* **2003**, *500*, 195.
[7] Juan, A. de; Tauler, R. *Crit. Rev. Anal. Chem.* **2006**, *36*, 163.
[8] Juan, A. de; Rutan, S. C.; Tauler, R. a *Comprehensive Chemometrics*. Brown, S.; Tauler, R.; Walczak, B. (ed.). Elsevier, vol. 2, **2009**, 325.
[9] Tauler, R.; Maeder, M.; Juan, A. de a *Comprehensive Chemometrics*. Brown, S.; Tauler, R.; Walczak, B. (ed.). Elsevier, vol. 2, **2009**, 473.
[10] Tauler. R. *Chemom. Intell. Lab. Sys.* **1995**, *30*, 133.
[11] Jaumot, J.; Gargallo, R.; Juan, A. de; Tauler, R. *Chemom. Intell. Lab. Sys.* **2005**, *76*, 101.
[12] Golub, G. H.; Reinsch, C. *Numer. Math.* **1970**, *14*, 403.
[13] Maeder, M. *Anal. Chem.* **1987**, *59*, 527.
[14] Windig, W.; Guilment, J. *Anal Chem.* **1991**, *63*, 1425.
[15] Tauler, R.; Smilde, A. K.; Kowalski, B. R. *J. Chemometr.* **1995**, *9*, 31.
[16] Tauler, R.; Marqués, I.; Casassas, E. *J. Chemometr.* **2001**, *15*, 55.
[17] Peré-Trepat, E.; Ginebreda, A.; Tauler, R. *Chemom. Intell. Lab. Sys.* **2007**, *88*, 69.
[18] Juan, A. de; Maeder, M.; Martínez, M.; Tauler, R. *Chemom. Intell. Lab. Sys.* **2000**, *54*, 123.
[19] Jaumot, J.; Eritja, R.; Tauler, R.; Gargallo, R. *Nucleic Acids Res.* **2006**, *34*, 206.
[20] Bucek, P.; Jaumot, J.; Aviñó, A.; Eritja, R.; Gargallo, R. *Chemistry*: *A European Journal* **2009**, *15*, 12663.
[21] Navea, S.; Juan, A. de; Tauler, R. *Anal. Chem.* **2002**, *64*, 6031.
[22] Navea, S.; Juan, A. de; Tauler, R. *Anal. Chem.* **2003**, *75*, 5592.
[23] Muñoz, G.; Juan, A de. *Anal. Chim. Acta*. **2007**, *595*, 198.
[24] Berova, N.; Nakanishi, K.; Woody, R. *Circular dichroism: Principles and Applications*. Wiley, **2000**.
[25] *Infrared and Raman spectroscopic imaging*. Salzer, R.; Siesler, H. W. (ed.). Wiley-VCH, **2009**.
[26] Grahn, H. F.; Geladi, P. *Techniques and applications of hyperspectral image analysis*. Wiley, **2007**.
[27] Juan, A. de; Tauler, R.; Dyson, R.; Marcolli, C.; Rault, M.; *Maeder M. TrAC-Trends in Anal. Chem.* **2004**, *23*, 70.
[28] Dao, N. Q.; Daudon, M. *Infrared and Raman spectra of calculi*. Elsevier.
[29] Stoll, D. R.; Li, X.; Wang, X.; Carr, P. W.; Porter, S. E. G.; Rutan, S. C. *J. Chromatogr. A* **2007**, *1168*, 3.
[30] Jaumot, J.; Marchan, V.; Gargallo, R.; Grandas, A.; Tauler, R. *Anal. Chem.* **2004**, *76*, 7094.
[31] Causton, H. C.; Quackenbush, J.; Brazma, A. *Microarray: Gene Expression Data Analysis*. Oxford, **2003**.

[32] Jaumot, J.; Tauler, R.; Gargallo, R. *Anal. Biochem.* **2006**, *358*, 76.
[33] Jaumot, J.; Piña, B.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2010**. DOI: 10.1016/j.chemolab.2010.04.004.
[34] Ross, D. T.; Scherf, U.; Eisen, M. B.; Perou, C. M.; Rees, C.; Spellman, P.; Iyer, V.; Jeffrey, S. S.; Rijn, M. van de; Waltham, M.; Pergamenschikov, A.; Lee, J. C. E.; Lashkari, D.; Shalon, D.; Myers, T. G.; Weinstein, J. N.; Botstein, D.; Brown, P. O. *Nat. Genet.* **2000**, *24*, 227.
[35] Tauler, R. «Interpretation of environmental data using chemometrics», a *Sample Handling and Trace Analysis of Pollutants: Techniques, Applications and Quality Assurance*. Barceló, D. (ed.). Elsevier, **2000**, cap. 16, p. 689.
[36] Alier, M.; Felipe-Sotelo, M.; Hernández, I.; Tauler, R. *Anal. Chim. Acta* **2009**, *642*, 77.
[37] Felipe-Sotelo, M.; Gustems, L.; Hernández, I.; Terrado, M.; Tauler, R. *Atmospher. Environ.* **2006**, *40*, 7421.
[38] Terrado, M.; Barceló, D.; Tauler, R. *Anal. Chim. Acta* **2010**, *657*, 19.
[39] Terrado, M.; Barceló, D.; Tauler, R. *Environ. Sci. Tech.* **2009**, *43*, 5321.
[40] Hopke. P. K. *Receptor Modeling in Environmental Chemistry*. Wiley & Sons, **1985**.
[41] Hopke, P. K. *Journal of Chemometrics* **2003**, *17*, 265.
[42] Terrado, M.; Barceló, D.; Tauler, R. *Talanta* **2006**, *70*, 691.


A. de Juan


J. Jaumot


R. Gargallo


R. Tauler

Anna de Juan és professora titular del Departament de Química Analítica de la Universitat de Barcelona. El seu àmbit de recerca és la quimiometria i el desenvolupament i aplicacions dels mètodes de resolució multivariant. Ha realitzat unes setanta publicacions en llibres i revistes internacionals i ha presentat més de cent comunicacions a congressos. El 2004, va rebre l'Elsevier Chemometrics Award, i el 2009, el Kowalski Award. En l'actualitat, pertany als equips editorials assessors de les revistes *Chemometrics and Intelligent Laboratory Systems* i *Analytica Chimica Acta.*

Joaquim Jaumot (nascut el 1978) és actualment professor lector al Departament de Química Analítica de la Universitat de Barcelona. La seva recerca està centrada en el desenvolupament d'eines quimiomètriques de resolució multivariant i la seva aplicació a l'estudi de problemes bioanalítics. És autor d'una vintena d'articles en revistes científiques internacionals i ha codirigit diversos treballs de màster.

Raimundo Gargallo (nascut el 1968) ha estat investigador Ramón y Cajal i és actualment professor agregat al Departament de Química Analítica de la Universitat de Barcelona. La seva recerca està centrada en l'estudi d'estructures complexes d'àcids nucleics, com ara les estructures G-quàdruples i els motius-i (*i-motif*). És autor d'una cinquantena d'articles en revistes científiques internacionals i ha codirigit diferents tesis doctorals i treballs de màster.

Romà Tauler (Barcelona, 1955) és professor d'investigació de l'Institut de Diagnòstic Ambiental i Estudis de l'Aigua (IDÆA, CSIC). És l'editor en cap de la revista *Journal of Chemometrics and International Laboratory Systems* i de l'obra de referència en quatre volums *Major Reference Work: Comprehensive Chemometrics*, *Chemical and Biochemical Data Analysis*. Ha rebut diferents premis relacionats amb la quimiometria, com el 2009 Award for Achievements in Chemometrics i el 2009 Kowalski Award. Ha publicat uns dos-cents trenta treballs de recerca indexats a la *Web of Knowledge*, amb més de cinc mil set-centes citacions (novembre 2011). Actualment, és el president de la Societat Catalana de Química.