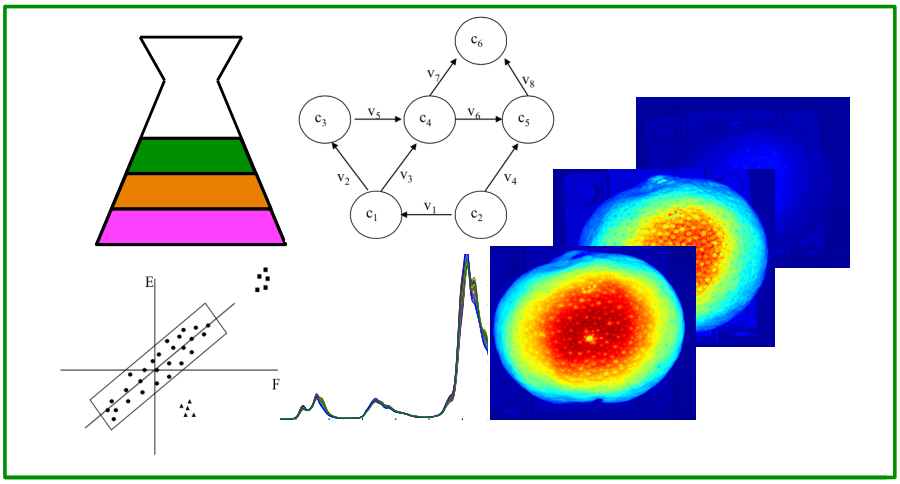
Proceedings of the VI Chemometrics Workshop for Young Researchers

October 1-2, 2015

Universitat Politècnica de València

Valencia, Spain



Editors

José Manuel Prats-Montalbán

Alberto Ferrer

Raffaele Vitale

Abel Folch-Fortuny

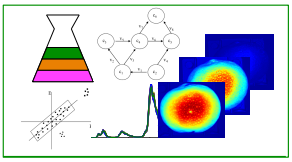
EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Congresos UPV Collection

The contents of this publication have been evaluated by the Scientific Committee which it relates and procedure set out http://mseg.webs.upv.es/vi_cwyr/index.html

www.lalibreria.upv.es / Ref.: 6227_01_01_01

ISBN: 978-84-9048-345-9 (print version)



Comittees

Scientific committee

- **Lutgarde Buydens**, Professor. Radboud University
- **Alberto Ferrer**, Professor. Universitat Politècnica de València (UPV)
- **Anna de Juan**, Associate Professor. Universitat de Barcelona
- **Onno de Noord,** Shell Global Solutions
- **Age Smilde**, Professor. Biosystems Data Analysis Group. Swammerdam Institute for Life Sciences. University of Amsterdam

Organizing committee

- **José Manuel Prats-Montalbán**, Associate Professor. UPV
- **Alberto Ferrer**, Full Professor. UPV
- **Raffaele Vitale,** PhD student. UPV
- **Abel Folch-Fortuny,** PhD student. UPV
- **Daniel Palací López,** PhD student. UPV
- **Eric Aguado Sarrió,** PhD student. UPV

Workshop secretariat

José Manuel Prats-Montalbán

Multivariate Statistical Engineering Group

Department of Applied Statistics, Operations Research and Quality

Cno. de Vera s/n, Edificio 7A, Valencia, Spain.

jopramon@eio.upv.es

Tel: +34 963877007 ext. 74949

Fax: +34 963877499

Contents

Thursday, October 1st, 2015

08:30 - 09:00 Registration

09:00 - 09.15 **Opening Session: The relevance of Chemometrics for research and industry** (Anna de Juan and Alberto Ferrer)

Chemometrics for Multivariate Statistical Process Control

09:15 – 09:50 ***Chemometrics in process industry*** (Onno de Noord)

10.00 – 11:00 **Oral presentations**

11:00 – 11:30 Coffee Break

11.30 – 12.50 **Oral presentations**

13:00-15:00 Lunch

Chemometrics for Omics data

15:00 – 15:35 ***Challenges in analyzing metabolomics data*** (A. Smilde)

15.40 – 17.00 **Oral presentations**

17.00 – 17.30 Coffee Break

17.30 – 18.50 **Oral presentations**

19.00 – 21.00 Free time. Spanish Chemometrics Network meeting

21.00 Workshop Gala Dinner

Friday, October 2nd, 2015

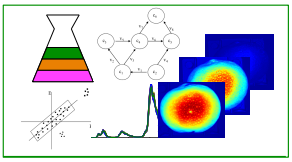
Chemometrics for Medicine, Analytical chemistry, Food Technology and Imaging

10.00 – 10.35	***Chemometrics for a Better Diagnosis and Understanding of Multiple Sclerosis*** (Lutgarde Buydens)
10.45 – 11.25	**Oral presentations**
11.25 - 11.55	Coffee Break
11.55 – 13:15	**Oral presentations**
13.15 – 13.30	Closing Session
13.30 – 15.00	Lunch

List of participants

ÁLVAREZ SEGURA, TAMARA
AGUADO SARRIÓ, ERIC
BENABOU, SANAE
BOQUÉ, RICARD
BORRÀS, EVA
BUYDENS, LUTGARDE
BURGOS SIMÓN, CLARA
DALMAU SOLÀ, NÚRIA
DE JUAN, ANNA
DE NOORD, ONNO
ELIA, ANDREA
FERRER RIQUELME, ALBERTO JOSE
FOLCH FORTUNY, ABEL
HERVÁS, DAVID
HURTADO SÁNCHEZ, MARÍA DEL CARMEN
JAUMOT, JOAQUIM
KULIGOWSKI, JULIA
LÓPEZ UREÑA, SERGIO
MARTINEZ, EKAITZ
MONAGO MARAÑA, OLGA
MUÑOZ DE LA PEÑA CASTRILLO, ARSENIO
ORTIZ VILLANUEVA, ELENA
ORTIZ, MARÍA CRUZ
PALACÍ LÓPEZ, DANIEL GONZALO
PÉREZ, ROCÍO
PRATS MONTALBÁN, JOSÉ MANUEL
QUINTAS SORIANO, GUILLERMO
SÁNCHEZ ILLANA, ÁNGEL
SANJUAN HERRAEZ, JUAN DANIEL
SARABIA, LUIS ANTONIO
SMILDE, AGE
VITALE, RAFFAELE

INVITED-TALKS



Chemometrics in Process Industry

Onno E. de Noord
Shell Global Solutions International BV, Amsterdam, The Netherlands

Abstract

Many aspects of chemometrics, according to a broad definition of the discipline, can be encountered in process industry. Most directly related to manufacturing processes are applications of multivariate technologies in Process Analytical Chemistry (PAC) and Process Chemometrics. In Process Chemometrics measurements from base layer instrumentation, such as temperature, pressure and flow sensors are used to extract information that can be used for Advanced Process Monitoring (APM), process diagnostics and trouble shooting. In PAC we use higher level chemical data, often coming from process spectrometers such as NIR. Multivariate methods are used to extract information on product qualities, process streams and feedstock characteristics. Typically the demands for industrial applications are different from applications in (R&D) laboratories, in particular in terms of robustness.

Challenges in analyzing metabolomics data

Age K. Smilde[1,2,3,4]

1. *Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Amsterdam, The Netherlands.*
2. *Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands.*
3. *Department of Food Science, Faculty of Sciences, University of Copenhagen, Denmark.*
4. *COPSAC, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.*

Abstract

Metabolomics relies heavily on advanced instrumental techniques such as GC-MS, NMR and LC-MS. It can be used to probe metabolism in cellular organisms, to analyze metabolites in body-fluid samples, plant extracts and food to name a few examples. The purpose of these measurements is dictated by the biological question underlying the study. In analyzing metabolomics data several crucial steps have to be undertaken; one of those is processing the data in such a way that the biological question is answered. Due to the large amount of data and the high demands posed on the quality of the results of the study, data processing is a very important step. To this end techniques as PCA and PLS can be used but these are not sufficient: also dedicated tools such as ASCA, association networks and data fusion approaches are needed. Along with some challenges and pitfalls, these tools will be discussed and illustrated with real-life examples.

Chemometrics for a Better Diagnosis and Understanding of Multiple Sclerosis

Lutgarde Buydens

Radboud University Nijmegen, Institute for Molecules and Materials, The Netherlands

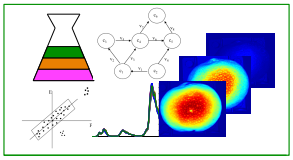
Abstract

While Multiple sclerosis is a major disabling disease of the Central nervous System (CNS) in young adults, little is known on the real cause of this disease; even diagnosis in an early stage is a non-solved issue.

Cerebrospinal Fluid (CSF) is the bio fluid, which is in closest interaction with the Central Nervous System (CNS). It is therefore the bio fluid that best mirrors the biochemical status and processes in brain and CNS. Biochemical changes are therefore most likely to be found by means of a comprehensive analysis of the CSF Other bio fluids such as plasma may also contain crucial information.

Comprehensive analysis by a large variety of analytical technologies, yield however complex data for which chemometric data analysis and data mining have become crucial tools. Since no analytical platform on its own yields a comprehensive image of the biochemical status, data fusion has become widespread in the last decade. Many methods have been proposed, most of them restrict to a linear fusion strategy However, it is not realistic to assume that all biological or (bio)chemical data display this simple linear behavior. In that case linear methods are bound to fail. In this lecture alternative approaches will be presented. One is based on the hierarchical fusion of mid-level fusion models. Non-linear kernel fusion model allow to cope specifically with nonlinearities (1). We use our pseudo-sample approach (2,3) to reveal the contribution of the individual variables. In the lecture we will present results of fusion of CSF and plasma analysis data for a better diagnosis and search for biomarkers for Multiple Sclerosis.

ORAL COMMUNICATIONS



Forensic characterization of similar and dissimilar sets of textile fiber extracts by three-way Excitation-Emission Matrix Fluorescence Spectroscopy in combination with second-order PARAFAC and MCR-ALS

A. Muñoz de la Peña[a,b], Matthew Rex[b], Hector C. Goicoechea[c], A. D. Campiglia[b,d]

[a]Department of Analytical Chemistry and IACYS, University of Extremadura, Badajoz, 06006, Spain, arsenio@unex.es, [b]Department of Chemistry, University of Central Florida, 4111 Libra Drive, P. O. Box 25000, Orlando, Florida 32816-2366, United States, mattew.rex@ucf.edu, [c]Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional de Litoral, Santa Fe S3000ZAA, Argentina, hgoico@fbcb.unl.edu.ar, [d]National Center for Forensic Science, University of Central Florida, 12354 Research Parkway, Suita 225, Orlando, Florida 32826, United States, andres.campiglia@ucf.edu
.

Abstract

Trace textile fiber evidence is found at numerous crime scenes and plays an important role in linking a suspect to the respective scene. In this communication, investigations into the fluorescence of the fiber dyes, and the fibers themselves, as well as methodology for discriminating between fibers using room temperature fluorescence (RTF) are reported. Three-way Excitation Emission Matrix (EEM) data was found to give the greatest amount of spectral information and provide the highest level of discrimination between non similar and similar fibers with the aid of second order PARAFAC and MCR-ALS chemometric analysis.

Keywords: *Forensic science, textile fiber extracts, similar and dissimilar fibers forensic comparison, PARAFAC, MCR-ALS.*

Resumen

En numerosos escenarios de crimenes se encuentran evidencias basadas en trazas de fibras de tejidos, que juegan un papel importante para relacionar un sospechoso al correspondiente escenario. En esta comunicación se presentan investigaciones acerca de la fluorescencia de los colorantes de las fibras, y de las fibras mismas, así como una metodología para discriminar entre fibras utilizando fluorescencia a temperatura ambiente (RTF). Se encontró que las Matrices de Excitacion Emission (EEM), datos de tres vías, proporcionaron la mayor cantidad de informacion spectral y el mayor nivel

de discriminacion entre fibras similares y distintas, en combinación con los métodos quimiométricos de segundo orden PARAFAC y MCR-ALS.

Palabras clave: *Ciencia Forense, Extractos de fibras de tejidos, comparación forense entre fibras similares y distintas, PARAFAC, MCR-ALS.*

Introduction

Fibers are key trace evidence often found at a crime scenario. Analytical techniques that can either discriminate between similar fibers or match a known to a questioned fiber are highly valuable in forensic science. Cloths based on fibers usually contain additives such as dyes to impart color to a textile fiber.

When fibers cannot be discriminated by non-destructive tests, a common approach is to solvent extract the questioned and the known fiber for further dye analysis. Established techniques for the analysis of fiber extracts include ultraviolet and visible absorption spectrometry, thin-layer chromatography and high-performance liquid chromatography (HPLC). (Blackledge 2007). Although the discriminating power of these techniques is well suited for those cases where the optical and/or chromatographic behaviors of dyes from a questioned and a known source are different, their selectivity falls short to differentiate between two fibers that have been dyed with highly similar dyes.This work focuses on the total fluorescence emission of fiber extracts. To the extent of our literature search, little efforts have been made to investigate the full potential of luminescence techniques for the problem at hand. Fluorescence microscopy for forensic fiber analysis has been reported (Chao 2007, Abu-Rous et al. 2007), but measurements were made with band-pass filters that take little advantage on spectral information. Recently, single fiber identification with nondestructive first-order fluorescence emisison principal component cluster analysis has been reported by our group (Appalaneni et al. 2014). The subject is of current interest, as is reported in a recent review on forensic comparison of synthetic fibers (Farah et al. 2015).

Our approach takes RTF spectroscopy to a higher level of selectivity. In addition to the contribution of the textile dye to the fluorescence spectrum of the fiber extract, we investigate the contribution of intrinsic fluorescence impurities – i.e. impurities imbedded into the fibers during the fabrication of the garments - as a reproducible source for fiber comparison. The accurate comparison of visually indistinguishable EEMs is best accomplished with the aid of chemometric analysis.The accurate comparison of EEMs requires the algorithm to determine the number of fluorescence components that contribute to the data set of excitation and emission spectra and the emission and excitation profiles corresponding to each component. Among the algorithms that exist to compare almost identical EEMs, we chose second order parallel factor analysis (PARAFAC) and multivariate curve resolution alternating least squares (MCR-ALS) (Anderson and Bro 2003, Tauler 1995).

Experimental

Fibers were individually pulled from cloths using tweezers. Each fiber was cut into a strand of appropriate length (4cm, 2cm or 5mm) using scissors or razor blades. Tweezers, scissors and razor blades were previously cleaned with methanol and visually examined under ultraviolet light (254 nm) to prevent the presence of fluorescence contamination. Each 4cm or 2cm strand was cut into pieces of approximately 5mm in length. 5mm strands were used as such. Fibers were solvent extracted following the procedure recommended by the Federal Bureau of Investigations (FBI) (*Forensic Science Communication* 1999). All pieces from one fiber were placed in a 6x50mm glass culture tube. 200 μL of extracting solvent were added to each tube. The tubes were sealed by melting with a propane torch. Sealed tubes were placed in an oven at 100° C for one hour. Tubes were removed from the oven, scored and broken open. The solvent was removed with a micro-pipette and placed in a plastic vial for storage.

Excitation and fluorescence spectra were recorded using a commercial spectrofluorometer (FluoroMax-P from Horiba Jobin-Yvon) equipped with a continuous 100 W pulsed xenon lamp with broadband illumination from 200 to 2000 nm. Measurements were made by pouring un-degassed liquid solutions into micro-quartz cuvettes (1cm path length x 2mm width) that held a maximum volume of 400 μL. EEM from fiber extracts were collected at 5 nm excitation steps and 1 nm emission steps, from longer to shorter wavelengths to reduce the risk of potential photo-degradation due to extensive sample excitation.

Chemometric Analysis

All chemometric calculations were done using MATLAB 7.0. Routines for PARAFAC and MCR-ALS were available in the Internet thanks to Bro (http://www.models.kvl.dk/source/) and Tauler (http://www.ub.edu/mcr/welcome.html), respectively. A useful MATLAB graphical interface was used for easy data manipulation and graphics presentation (Olivieri 2014). This interface provided a simple means of loading the data matrices into the MATLAB working space before running PARAFAC and MCR-ALS.

Results and Discussion

Fibers and EEMs Collection for PARAFAC analysis.

The investigated fibers were separated in three different sets. Set #1 included ten nylon fibers pre-dyed with Acid Red 151 and collected from different areas of the same piece of cloth. Set # 2 included ten nylon fibers pre-dyed with Acid Red 151 but collected from a different piece of cloth than the fibers in set #1. Set # 3 included ten cotton fibers pre-dyed with Direct Blue 1 and collected from the same piece of cloth. Each fiber from each set was individually extracted with ethanol and one EEM per fiber extract was recorded.

The spectral de-convolution of EEMs via PARAFAC provided the best fit for a five fluorescence component mixture in the three cases. Figure 1 depicts the statistical grouping of the 30 EEM recorded from the individual extracts of the 30 fibers. The intercepts of the three PC values, which correspond to the maximum intensities of the first three spectral components of each EEM, place each EEM within a statistical (elliptical) domain. The statistical domain is defined using the bivariate method to achieve a confidence level of 95%. Although PARAFAC groups EEMs from set #1 and #2 within the same statistical group, fiber extracts from set #1 show EEMs with similar PC values while EEM from set #2 show PC values scattered over the entire elliptical domain. Under this prospective, PARAFAC provides some sort of discrimination between fiber extracts from set #1 and #2.

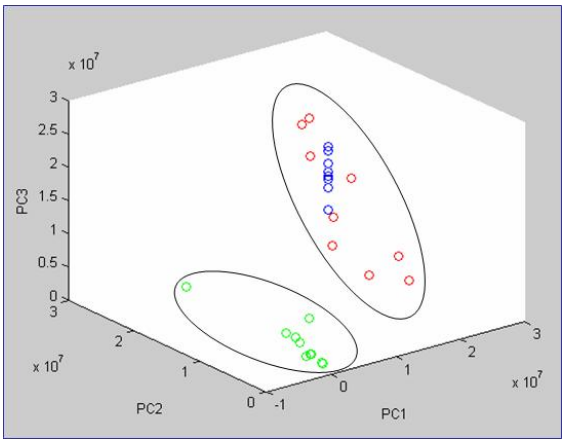


Figure 1. Statistical Grouping of 30 EEM from the 30 Fibers in Sets #1, #2 and #3. Set#1 is represented by blue circles, Set #2 by red and Set #3 by green circles.

Fibers and EEMs Collection for MCR-ALS Analysis

All MCR-ALS comparisons were made among EEMs recorded from ethanol extracts of each of visually indistinguishable pre-dyed nylon fibers. Acid Red 151 was extracted from two different pieces of cloths, ten fibers per cloth. Acid Yellow 17 and Acid Yellow 23 fibers were collected from one piece of cloth. Figure 2 correlates the emission spectra of the five fluorescence components in each type of the Acid Yellow 17 and Acid Yellow 23 fiber extracts. The five correlations were made comparing the spectral intensities of the corresponding components in each type of extract at each excitation and fluorescence wavelength. From the calculated values of the correlation coefficients, it becomes readily apparent that only three of the five components exhibit similar spectral profiles. Close comparison of the five pairs of excitation and fluorescence spectra support correlation coefficients close to unity for three predicted components. Based on the prediction that two components only exist in one type of fiber extract, MCR-ALS is able to discriminate among

these two types of visually indistinguishable fibers. Similarly, MCR-ALS was able to discriminate among Acid Red 151 extracts of fibers collected from two different cloths.

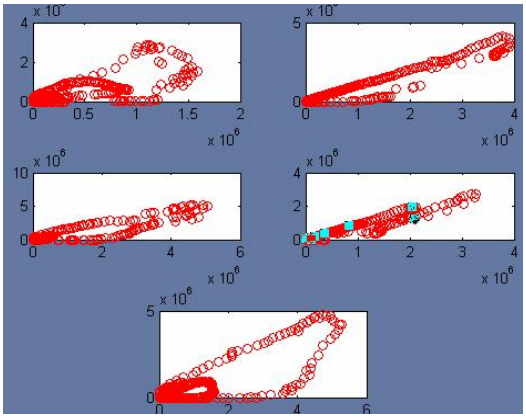


Figure 2. Correlation of Five emission profiles extracted from EEM of extracts taken from nylon fibers dyed with Acid Yellow 17 and 23. Five correlation coefficients are as follows top left-0.7564; top right-0.9696; middle left-0.9480; middle right-0.9677, bottom-0.8300

Acknowledgments

The authors are grateful to US National Institute of Justice (Grant # 2011-DN-BX-K553), UNL and CONICET, and Ministerio de Economía y Competitividad of Spain (Project CTQ2014-52309-P) and Gobierno de Extremadura (GR15-Research Group FQM003), both co-financed by European FEDER funds, for financially supporting this work.

References

[1] Abu-Rous, M., Schuster, K.C., Adlassnig, W., Lichtscheidl, I. Melliand Internat. 2007, 13, 382.

[2] Anderson, C.M., Bro, R., Journal of Chemometrics, 2003, 17, 200.

[3] Appalaneni, K., Heider, E., Moore, A.F.T., Campiglia, A.D., Anal. Chem., 2014, 86, 6774-6780.

[4] Blackledge, R.D. (ed) Forensic Analysis on the Cutting Edge. Wiley & Sons, Hoboken, NJ 2007

[5] Cho, L.L., Forensic Science Journal 2007, 6, 55.

[6] Farah, S., Kunduru, K.R., Tsach, T., Bentolila, A., Domb, A.J., Polymers Advances Technologies, DOI: 10.1002/pat.3540

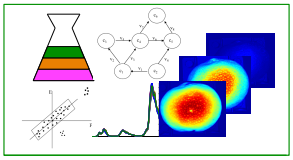
[7] http://www.models.kvl.dk/source/

[8] Forensic Science Communication., 1999, Vol 1, No. 1, http://www.fbi.gov/hq/lab/fsc/backissu/april1999/houckapb.htm

[9] Olivieri, A.C., Escandar, G.M., Practical Three-way Calibration, Elsevier, 2014.

[10] Tauler, R., Chemom. Intell. Lab. Syst., 1995, 30, 133.

[11] Tauler, R., De Juan, A., "Multivariate Curve Resolution Homepage", http://www.ub.edu/mcr/welcome.html



From powder properties to in-vitro performance of dry powder inhalers: a multivariate approach using Partial Least Square Regression

Andrea Elia[a,b], Marina Cocchi[a], Ciro Cottini[a], Daniela Riolo[a], Claudio Cafiero[a], Roberto Bosi[a], Emilio Lutero[a]

[a]Department of chemical and Geological Sciences, Via Campi 103, 41126 Modena, Italy[b], Chiesi Farmaceutici S.p.A., Dept. of Chemistry, Manufacturing & Control Parma, Italy. e-mail: A.Elia@chiesi.com

Abstract

A multivariate apprach was used in order to study the correlations among the physical properties of bulk powders and the in-vitro performance of dry powder inhalers (DPI). PLS models were obtained for the prediction of the DPI performance using data from bulk powder characterization. Variable importance in projection (VIP) was used in order to assess the most influential rheological variables to estimate the performance.

Keywords: *Dry powder inhaler; PLS regression; VIP; rheological tests; DUSA; NGI; performance; correlation..*

Introduction

This study aims at investigating the correlations among the physical properties of bulk powders and the in-vitro performance of dry powder inhalers (DPI) (Atkins 2005), in order to generate models for predicting the DPI performance using data from bulk powder characterization. Samples of bulk powder, belonging to scale-up process batches having different formulations, process parameter and bulk size, were characterized by rheological, density and particle size tests. In vitro performance was evaluated by a dosage unit sampling apparatus (DUSA) and a next generation impactor (NGI). Correlation between rheological, technological and performance properties were established using partial least square regression (PLS) (Brereton 1990, Wold et al. 1993). The Variable importance in projection (VIP) parameters were used to rank the most influent rheological variables for modelling the

device performance (Favilla et al. 2013, Mehmood et al. 2012). Particle size, density and rate of flowability are significant for predicting the delivered dose of the API and the total quantity of powder related to each dose. Rehological variables, describing the degree of cohesiveness and the flow properties of powder, resulted correlated to the total amount of the active ingredient for different formulations. PLS-2 models were then tested on new samples. DUSA variables resulted better predicted compared to NGI variables. The prediction error for external test set data was respectively 2.1% for the quantity of total powder and 1.9% for the quantity of active ingredient delivered at each dose.

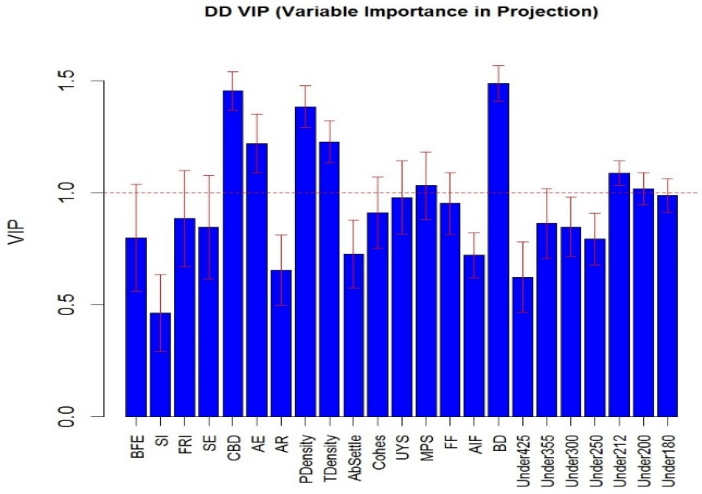


Figure 1. Variable Importance in Projection for Delivered Dose in PLS-2 Model calculated for DUSA variables.

References

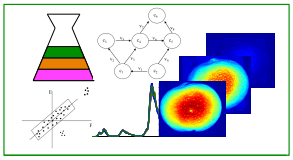
[1] Atkins, Paul J. Dry powder inhalers: an overview. Respiratory care 50.10 (2005): 1304-1312.

[2] Brereton, R. G. Chemometrics: applications of mathematics and statistics to laboratory systems. 1990. E. Horwood.

[3] Favilla, S., Durante, C., Vigni, M. L., & Cocchi, M. Assessing feature relevance in NPLS models by VIP. Chemometrics and Intelligent Laboratory Systems, (2013): 129, 76-86.

[4] Mehmood, Tahir, et al. A review of variable selection methods in partial least squares regression.. Chemometrics and Intelligent Laboratory Systems 118 (2012): 62-69.

[5] Wold, S., E. Johansson, and M. Cocchi. PLS—partial least squares projections to latent structures. 3D QSAR in drug design 1 (1993): 523-550.



Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in wine samples.

M.C. Hurtado-Sánchez[a], M. Cabrera Bañegil[b], T. Galeano Díaz[a], I. Durán Merás[a]

[a]Department of Analytical Chemistry and IACYs, University of Extremadura, 06006 Badajoz, Spain,
[b]Centro de Investigaciones Científicas y Tecnológicas de Extremadura, 06187 Badajoz, Spain

Abstract

The potential of front-face fluorescence spectroscopy combined with chemometric methods was investigated for the quantification of the main polyphenols presents in wine samples. Second-order mualtivariate algorithm was choiced with this aim, by employing excitation-emission matrices as analytical signal. Both PARAFAC and U-PLS/RBL algorithms were assessed and the last one was finally selected as optimum for the quantification of catequin, epicatequin, vanillic acid, caffeic acid, gallic acid and resveratrol in red wine samples. U-PLS/RBL provided the best results and was selected as the optimum algorithm.

Keywords: *Polyphenols, Wine, Front-face fluorescence spectroscopy, Excitation-emission matrix, Second-order multivariate algorithms.*

Resumen

Se ha evaluado el potencial de la técnica de fron-face acoplada a métodos quimiométricos para la cuantificación de los principales polifenoles presentes en muestras de vino. Para este fin se seleccionaron algoritmos de calibración multivariante de segundo orden, mediante el empleo de matrices de excitación-emisión como señal analítica. Tanto PARAFAC como U-PLS/RBL fueron los algoritmos evaluados para la cuantificación de catequina, epicatequina, ácido vanílico, ácido cafeico, ácido gálico y el resveratrol en muestras de vino tinto. U-PLS/RBL proporcionó los mejores resultados y se fue seleccionado como óptimo.

Palabras clave: *Polifenoles, Vino, Fluorescencia front-face, Matrices de excitación-emisión, Algoritmos de calibración multivariante de segundo orden.*

Introduction

Wine is a widely consumed beverage in Europe, especially in countries like Spain, Italy, and France. It is a complex solution containing different components with strong antimicrobial properties, such as a low pH (from 3 to 4), relatively high ethanol concentrations (10–15%) and some antimicrobial components (Gañan *et al*., 2009). In this sample, considerable interest has been focused on the bioactive phenolic compounds, notably anthocyanins, flavanols, flavonols and resveratrol, since they possess many biological activities, such as antioxidant, cardioprotective, anticancer, anti-inflammation, antiaging and antimicrobial properties (Vallverdú-Queralt *et al*., 2015). Moreover, these compounds determine important sensorial characteristics, such as color, mouth-feel, astringency and bitterness. They are the main components responsible for the differences between red and white wines, especially for the color, taste, and mouth-feel sensations of red wines (Ivanova-Petropulos *et al*., 2015). The phenolic composition of the wine depends on the raw material and the type of vinification followed, which affects physical phenomena (diffusion from the solid parts, extraction of wood compounds, etc.), and chemical and biochemical phenomena (oxidation, degradation, condensation, etc.(Mulero *et al*, 2015).

Many methods have already been developed to characterize and quantifie phenolic compounds in wine by employing a wide variety of analytical techniques, being HPLC coupled to fluorimetric or mass detectors the technique most employed with this aim. However, these methods are very time consuming and are not particularly advantageous for the quantification of a large number of samples. In this sense, characterization and quantification of compounds in wines and other kinds of samples based on analytical methods combined with chemometric treatment of data provides excellent robustness and efficiency. Physicochemical parameters, concentrations of wine components and instrumental signals can be used as multivariate data. Wine features, including origin, variety and winemaking practices, can be evaluated from this source of information (Saurina, 2010).

Otherwise, front-face fluorescence allows measurement of the fluorescence of powdered, turbid, and concentrated samples, as well as complex food matrices such as meat, fish and dairy products. In front-face fluorescence spectroscopy, the surface of a sample is simply illuminated by excitation light, and the emitted fluorescence from the same surface is measured, which minimises reflected light, scattered radiation and depolarisation phenomena. In this way, this technique allow the analysis of complex samples without pretreatment. Moreover, it is non-destructive, rapid, easy to use and not expensive.

Dufour *et al*. (2006) employed rapid fluorescence measurements applied directly on wines for monitoring the variety, the typicality and the vintage of a collection of French and German wines. This study showed that front-face fluorescence spectroscopy combined with chemometrics offers a promising approach for the authentication of wines. Airado-

Rodríguez *et al*. (2011) also showed the potential of the autofluorescence of wine, through the measurements of excitation-emission matrices (EEMs) of untreated wine samples by front-face fluorescence, combined with the three-way method PARAFAC for the purpose of discrimination of wine according to the appellation of origin.

Bearing in mind this information, front-face fluorescence spectroscopy combined with second-order multivariate algorithms method seems to be a very attractive methodology for the determination of phenolic compounds in wine. In this work, both PARAFAC and U-PLS/RBL second-order algorithms has been assessed for the quantification of catequin, epicatequin, vanillic acid, caffeic acid, gallic acid and resveratrol in red wine samples following this methodology.

Experimental

To perform the chemometrics analysis, firstly, a calibration set was built with a full factorial design in the concentration ranges between 2.2-16.5, 1.72-9.72, 0.95-4.79, 1.92-9.6, 11.31-64.08 and 0.20-5.81μg mL^{-1} for catequin, epicatequin, vanillic acid, caffeic acid, gallic acid and resveratrol, respectively. The corresponding volumes of the standard solutions of each analyte were transferred into 10.00 mL volumetric flasks, containing 3 mL of tartrate buffer, pH 3.6, and 1.5 mL of ethanol and ultrapure water was added to the mark. For the analysis of catequin, epicatequin, vanillic acid, caffeic acid, gallic acid, excitation-emission matrix of each solution were obtaianed in mode front-face in the ranges 240-290 nm (each 5 nm, excitation) and 290-450 nm (each 0.5 nm, emission. Another excitation-emission matrix of each solution were recorded in the wide spectral excitation range from 300 to 350 nm (each 5 nm) and emission range from 330 to 400 nm (each 0.5 nm), for the analysis of resveratrol. The instrument was set up as follows: wavelength scanning speed, 300 nm/min, monochromators band pass excitation/emission (nm/nm), 2.5/5 and detector voltage, 700 V.

A set of 6 validation samples was prepared and processed in a similar way, having analyte concentrations different from the calibration ones and selected at random from the corresponding calibration ranges. All calculations were done using MatLab R2008a, using the MVC2 routine, an integrated MatLab toolbox for second-order calibration developed by A. C. Olivieri *et al*. (2009).

Once method was validated with standards, the analysis of wine samples was conducted. For that, the corresponding excitation-emission matrix of each analyzed red wine was recorded without any pretreatment.

With the objective of reaching a better knowledge of the concentration of the target analytes in analyzed wines samples, each wine was also chromatographied and the eluate was

fluorimetrically monitored at the $\lambda_{exc}/\lambda_{em}$ (nm/nm) wavelengths: 270/313, 270/360, 270/433. The chromatographic separation was carried out using a ZORBAX Eclipse XDB C-18 (4.6 x 50 mm, 1.8 μm) column An appropriate optimized gradient program was chosen to carry out the elution of samples, with the objective of getting a good separation and a good profile of the fluorescent compounds in wine at each specific $\lambda_{exc}/\lambda_{em}$ pair: as mobile phase methanol-formic acid-water (10:2:88, v:v) as solvent A and methanol-formic acid-water (90:2:8, v:v) as solvent B with the following gradient program: 0 min, 100% A; 0-15 min, 85% A; 15-25, min 70% A; 25-34 min, 30% A.

Results and discussion

Both PARAFAC and U-PLS/RBL algorithms have been evaluated regarding the quantitation of the analytes in wine samples. A first phase in the data processing with different second-order algorithms is the estimation of the number of responsive components. In the case of PARAFAC, the core consistency diagnostic test (CORCONDIA) was employed for this aim, while with U-PLS the usual procedure is the well-known leave-one-sample cross-validation procedure, according to Haaland and Thomas' criterion. Among the excitation and emission ranges, an optimal region was selected for both calibration and prediction purposes in order to improve the results obtained and to remove Rayleigh scattering.

In the analysis of validation samples, PARAFAC provided good results for the quantification of vanillic acid and resveratrol, while U-PLS showed good recoveries for all the analytes. For that reason, this algorithm was selected for the analysis of the red wine samples. Wine samples were not submitted to any prior treatment. In this case, RBL was also required, with two unexpected components in most cases. Until now, good results just for vanillic acid, caffeic acid and resveratrol have been obtained. The concentration values obtained in the analysis of these analytes in some red wine samples by the U-PLS/RBL developed method and by chromatographic method are summarized in Table 1.

Table 1. Comparison between the results obtained by U-PLS/RBL and HPLC-DAD (μg mL^{-1})

	Vanillic acid		Caffeic acid		Resveratrol	
	U-PLS/ RBL	HPLC-FLD	U-PLS/ RBL	HPLC-FLD	U-PLS/ RBL	HPLC-FLD
Wine 1	1.05	0.89	7.77	8.03	0.58	0.67
Wine 2	1.66	1.69	11.26	10.44	0.49	0.53
Wine 3	1.74	2.06	12.87	9.3	1.74	1.91
Wine 4	1.64	1.41	11.94	13.98	0.21	0.21
Wine 5	1.12	1.13	9.39	9.05	0.95	0.95
Wine 6	1.35	1.19	12.77	12.14	2.27	1.74

Conclusions

Some of the principal phenolic compounds of wine such as gallic, caffeic, *p*-coumaric and vanillic acids, catechin, epicatechin, and resveratrol have been studied using front-face fluorescence spectroscopy in combination with chemometric techniques. U-PLS/RBL provided good results for the quantification of caffeic and vanillic acids and resveratrol in red wine samples. The method represents a new example of the power of coupling the partial least-squares algorithm with residual bilinearization for the resolution of very complex systems. Moreover, it can be classified as a green chemistry-precedure because allows good selectivity and sensitivity in avoiding the use of toxic organic solvents. In addition, the method is fast, and not need any previous pretreatment of the samples.

References

[1] Airado-Rodríguez D., Durán Merás I. Galeano Díaz T. Petter Wold, J. (2011). *Front-face fluorescence spectroscopy: A new tool for control in the wine industry*. J. Food Compos. Anal. 24, 257-264.

[2] Dufour E. Letort A., Laguet A., Lebecque A., Serra J.N. (2006*). Investigation of variety, typicality and vintage of French and German wines using front-face fluorescence spectroscopy*. Anal. Chim. Acta 563, 292-299.

[3] Gañan M., Martínez-Rodríguez A. J., Carrascosa A. V. (2009). *Antimicrobial activity of phenolic compounds of wine against Campylobacter jejuni.* Food Control 20, 739-742.

[4] Ivanova-Petropulos V., Hermosín-Gutiérrez I., Boros B., Stefova M., Stafilov T., Vojnoski B., Dörnyei Á., Kilár F. (2015). *Phenolic compounds and antioxidant activity of Macedonian red wines*. J. Food Compos. Anal. 41, 1-14.

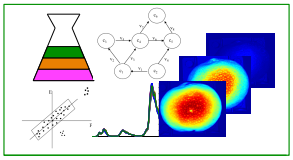
[5] Mulero J., Martínez G., Oliva J., Cermeño S., Cayuela J. M., Zafrilla P., Martínez-Cachá A., Barba A. (2015). Phenolic compounds and antioxidant activity of red wine made from grapes treated with different fungicides.

[6] Olivieri A.C., Wu H.L., Yu R.Q., (2009). *MVC2: a MATLAB graphical interface toolbox for second-order multivariate calibration.* Chemom. Intell. Lab. Syst. 96, 246-251.

[7] Saurina J. (2010). *Characterization of wines using compositional profiles and chemometrics*. TrAC Trends Anal. Chem. 29, 234-245.

[8] Vallverdú-Queralt A., Boix N., Piqué E., Gómez-Catalan J., Medina-Remon A., Sasot G., Mercader-Martí M., Llobet J. M., Lamuela-Raventos R. M. (2015). *Identification of phenolic compounds in red wine extract samples and zebrafish embryos by HPLC-ESI-LTQ-Orbitrap-MS.* Food Chem. 181, 146-151.

The authors are grateful to the and Ministerio de Economía y Competitividad of Spain (Proyecto CTQ2014-52309-P) and the Gobierno de Extremadura (Ayuda a Grupos 2015-Research Group FQM003), both co-financed by the European FEDER funds, for financial support for financial support.



Test de migración de aminas aromáticas primarias desde utensilios de poliamida a un simulante alimentario utilizando fluorescencia molecular de excitación-emisión y PARAFAC

Silvia Sanllorente[a], María Cruz Ortiz[a] y Luis A. Sarabia[b]

[a]Dpto. de Química y [b]Dpto de Matemáticas y Computación de la Universidad de Burgos, Plaza Misael Bañuelos s/n, 09001 Burgos, e-mail: silsan@ubu.es, mcortiz@ubu.es, lsarabia@ubu.es

Abstract

A procedure based on PARAFAC decomposition and the standard addition method applied to EEM data was proposed. The unequivocal identification and quantification of three primary aromatic amines was possible despite the high overlapping signals after a strategy applied to recover the trilinearity. Also the migration test cinetic has been modelled.

Keywords: *PARAFAC, Excitation-Emission Matrix, Primary aromatic amines, migration testing, Reglamento UE 10/2011.*

Resumen

Se propone un procedimiento basado en una descomposición PARAFAC y una adición estándar aplicada a datos de EEM. Después de una estrategia para recuperar la trilinealidad, ha sido posible la identificación inequívoca y la cuantificación de tres aminas primarias aromáticas a pesar del elevado solapamiento de las señales. También se modela la cinética de migración.

Palabras clave: *PARAFAC, matrices de Excitación-Emisión, Aminas aromáticas primarias, test de migración, Reglamento UE 10/2011.*

Introducción

En este trabajo se ha realizado un estudio de migración de aminas primarias aromáticas (PAAs) desde utensilios de cocina de poliamida (nylon) a simulante alimentario (solución acuosa al 3% de ácido acético [Reglamento UE 10/2011]). Las aminas analizadas han sido la anilina (A), 2,4-diaminotolueno (2,4-TDA) y 4,4'-diaminodifenilmetano (4,4'-DMA) [EUR 24815 EN 2011]. Estas dos últimas están englobadas en el grupo 2B (posibles carcinógenos). Los utensilios de nylon, como cazos, espumaderas y cucharas, que de forma

habitual se utilizan para cocinar y freír contienen PAAs que pueden migrar desde estos artículos a los alimentos. La UE ha establecido un límite legal sobre el nivel permitido en la migración de 10 μg kg^{-1}.

La espectroscopía de fluorescencia molecular presenta un elevado potencial debido a su alta sensibilidad y facilidad de uso. Sin embargo, las señales registradas procedentes de los analitos de interés pueden estar solapadas e incluso amortiguadas por la presencia de otras moléculas o iones. Las matrices de excitación emisión (EEM) son especialmente adecuadas para ser analizadas con técnicas de tres-vías, con el fin de obtener la separación de las señales de los fluoróforos incluso en presencia de efecto amortiguador (Rubio L, et al, 2013). Además, cuando se utilizan métodos como PARAFAC con datos trilineales la descomposición factorial es única, de modo que los factores obtenidos matemáticamente se corresponden con los fluoróforos presentes en la muestra. La ventaja de "segundo orden" permite la identificación inequívoca y la cuantificación en presencia de analitos no calibrados.

En este trabajo se muestra un procedimiento para recuperar la trilinealidad y utilizar la propiedad de segundo orden para determinar la cantidad de aminas e identificarlas usando PARAFAC y una adición estándar. En este caso los fluoroforos de la matriz, que permanecen constantes al realizar la adición estándar, provocan un grave fallo de trilinealidad. El procedimiento quimiométrico utilizado para recuperar la trinealidad consiste en restar los factores relacionados con los fluoróforos de la matriz del tensor de datos original. De este modo se pueden identificar sin ambigüedad y cuantificar las aminas en las muestras de los test de migración. Este procedimiento ya ha sido utilizado en Rubio et al (2013) en la determinación de pesticidas en matrices complejas.

1.Experimental

Los espectros de excitación-emisión fluorescente de cada muestra se registraron a temperatura ambiente en un espectrofluorímetro PerkinElmer LS 50B rango 295-394 nm, cada 1 nm para la emisión y 220-275 nm ,cada 5 nm en excitación; velocidad de 1500 nm min^{-1}.

Los patrones de calibrado utilizados en la adición estándar son muestras que contienen las tres PAAs en concentraciones que se reflejan en la Tabla 1, todos ellos se prepararon en metanol y fueron añadidos a un extracto obtenido con el test de migración al que había sido sometida la cuchara (2 horas en contacto con ácido acético 3% v/v a 100 ºC). Este simulante es el indicado en la normativa para alimentos ácidos. Este procedimiento simula la matriz de la muestra en la que luego se van a medir las aminas. El tensor de tamaño 21×100×12, en el que 21 son el número de muestras, 100 las intensidades de emisión y 12 las de excitación se le denominará en lo que sigue T1. Tensor que se utiliza entre otras finalidades para detectar los factores que están en la matriz. Por otro lado, las muestras para la cinética de migración se preparan sumergiendo otra cuchara (del mismo lote) en el simulante a 100 ºC y tomando una muestra cada 15 minutos, renovando el simulante en

cada muestreo. Estos 6 extractos se evaporan a sequedad y se reconstituyen en metanol. Todas estas muestras constituyen el tensor 2 (T2) de tamaño 6×100×12 que ha se ha utilizado para realizar la curva de la cinética de migración. La Fig. 1A, muestra las curvas de nivel de una muestra de extracto de la cuchara después de la migración. Las Figs. 1B-1D son las curvas de nivel del mismo extracto al que se han adicionado las tres mezclas marcadas con asterisco en la Tabla 1. Se observa un elevado grado de solapamiento.

Tabla 1. Diseño de los patrones de calibrado de mezclas de las tres aminas.

Núm.	$C_{anilina}$ (ppb)	C_{DMA} (ppb)	C_{TDA} (ppb)	Núm.	$C_{anilina}$ (ppb)	C_{DMA} (ppb)	C_{TDA} (ppb)
1	2.5	5	25	9	10	2.5	100
2	20	10	25	10	20	5	100
3	10	20	25	11	2.5	10	100
4	5	2.5	25	12	5	20	100
5(*)	20	2.5	50	13(*)	2.5	2.5	200
6	5	5	50	14	10	5	200
7	10	10	50	15	5	10	200
8(*)	2.5	20	50	16	20	20	200

El software utilizado ha sido: PLS_Toolbox 6.0.1 para la versión de Matlab 7.12.0.635 para la realización de los modelos PARAFAC y STATGRAPHICS Centurion XVII.

2.Resultados y conclusiones

El modelo PARAFAC construido incialmente con el tensor T1 no es trilineal, tiene un índice CORCONDIA del 13%. La Fig, 2 muestra los loadings en los tres perfiles y puede observarse en la Fig 2A. un perfil constante correspondiente al fluoróforo contenido en la matriz. Se recupera la trilinealidad restando este factor del tensor origial. La descomposición PARAFAC resultante es trilineal, necesita 3 factores, y tiene un índice CORCONDIA igual a 98%. Siendo idénticos los otros dos perfiles (excitación y emisión) a los obtenidos en la descomposición anterior. La identificación de las tres aminas se hace calculando el coeficiente de correlación entre los perfiles espectrales obtenidos de PARAFAC y los espectros originales de cada una de las muestras puras. En la cuchara sólo se detecta DMA.

La ecuación $C_{DMA} = a + \left(131.4 - a\right)e^{-b(t-15)}$ ajustada para la cinética del DMA, muestra un decaimiento exponencial (Fig. 3). El modelo explica el 97.7% de la concentración de DMA.

En conclusión, se propone un procedimiento rápido y sencillo que permite, con señales procedentes de fluorescencia de excitación-emisión junto con una descomposición PARAFAC, cuantificar e identificar inequívocamente tres aminas aromáticas primarias. Además se ha realizado un test de migración y se ha obtenido la ecuación de su cinética para el DMA, única amina encontrada en la cuchara de poliamida analizada.

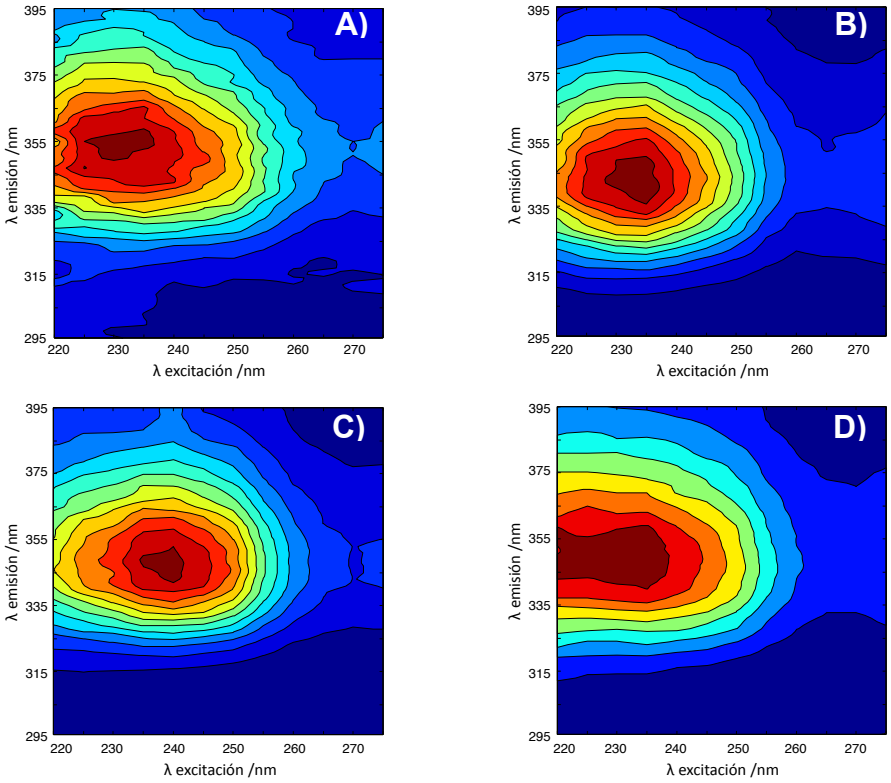


Figura 1. A) muestra de cuchara λ_{exc}=230, λ_{em}=353 (máximo intensidad =60.2; B) muestra de cuchara con la mezcla 5*; λ_{exc}=235, λ_{em}=342 (máximo=252.7); C) muestra de cuchara con mezcla 8*; λ_{exc}=240, λ_{em}=348 (máximo=244.1); D) muestra de cuchara con la mezcla 13*; λ_{exc}=235, λ_{em}=349 (máximo=174.3).

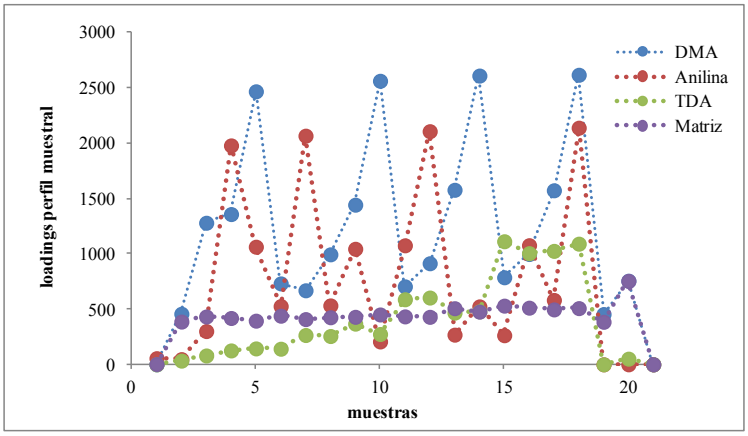


Figura 2. Loading del modelo PARAFAC (T1) A) perfil muestral: muestras 3-18 (Tabla 1), muestras 1 y 21 blancos, muestras 2 y 19-20 muestras test de migración.

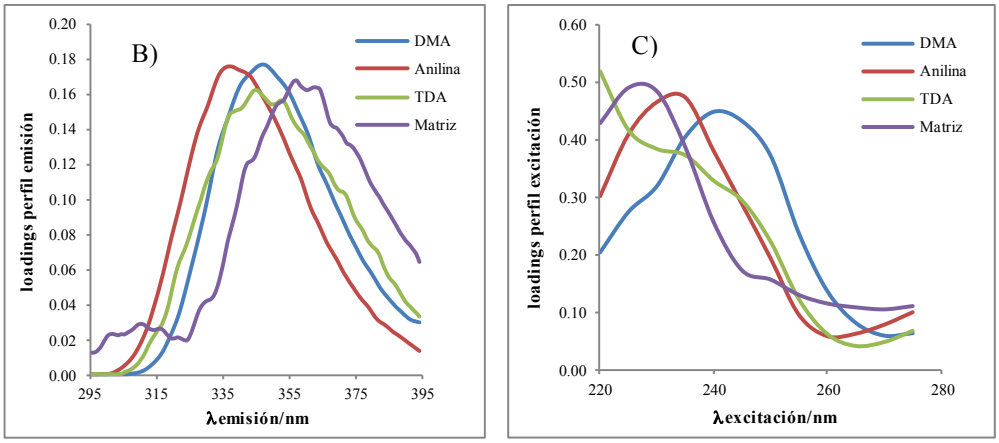


Figura 2 cont.. Loading del modelo PARAFAC (T1) B) Perfil de emisión, C) Perfil de excitación.

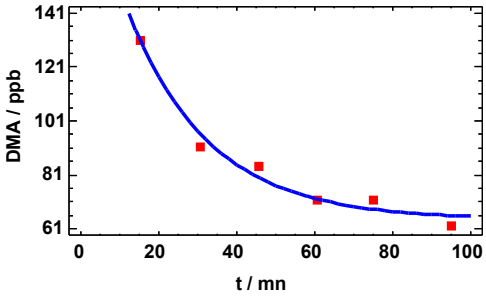


Figura 3.. Datos experimentales de la cinetica de migración del DMA y modelo ajustada.

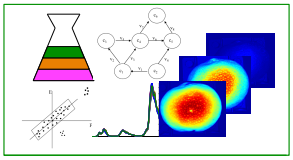
Agradecimientos: Los autores agradecen la financiación del Ministerio de Economía y Competitividad (CTQ2014-53157-R)

Referencias

[1] EUR 24815 EN 2011. Guidelines on testing articles in contact with foodstuffs for the specific migration of primary aromatic amines (PAAs) from polyamide kitchen utensils (with specific support of Regulation on imports from China and Hong Kong)

[2] Regulation UE 10/2011 of the European Parliament and the Council of 14 January 2011 on plastic materials and articles intended to come into contact with food.

[3] Rubio L., Ortiz M.C., Sarabia L.A., (2014), Identification and quantification of carbamate pesticides in dried lime tree flowers by means of excitation-emission molecular fluorescence and Parallel Factor Analysis when quenching effect exists, Analytica Chimica, 820 (2014) 9-22.



Second order algorithms to differentiate and quantify minor components in paprika samples

Olga Monago Maraña[1a], Isabel Durán Merás[2a], Teresa Galeano Díaz[3a] and Arsenio Muñoz de la Peña[4a]

[a] Department of Analytical Chemistry and IACYS, University of Extremadura, Badajoz, 06006, Spain. 1: olgamonago@unex.es; 2: iduran@unex.es; 3: tgaleano@unex.es; 4: arsenio@unex.es

Abstract

In this communication several studies made in paprika samples are presented. One of them was a PARAFAC analysis to cluster paprika samples according their origin. Two groups were obtained, one of them with the samples belonging to the PDO *"Pimentón de La Vera*" and another group with those samples not belonging to this PDO. Another study performed was the quantification of quercetin and kaempferol by using fluorescence coupled to second order algorithms (U-PLS/RBL and N-PLS/RBL). Good results were obtained with synthetic samples and in the case of real samples these were better for quercetin, which is present in higher concentration in paprika samples .

Keywords: *flavonoids, paprika, fluorescence, parallel factor analysis, unfolded-partial least-squares with residual bilinearization, multidimensional-partial least-squares with residual bilinearization*

Resumen

En esta comunicación se presentan varios estudios realizados en muestras de pimentón. Uno de ellos consistió en un análisis por PARAFAC para agrupar las muestras de acuerdo a su origen. Se obtuvo un agrupamiento de las muestras según perteneciesen o no a la DOP "Pimentón de La Vera". Otro estudio realizado en pimentón fue la cuantificación de quercetina y kaempferol llevada a cabo mediante fluorescencia y algoritmos de segundo orden (U-PLS/RBL y N-PLS/RBL), obteniéndose mejores resultados para la quercetina, ya que se encontraba en mayor concentración en estas muestras.

.***Palabras clave:*** *flavonoides, pimentón, fluorescencia, análisis paralelo de factores, mínimos cuadrados parciales desdoblados con bilinearización residual, mínimos cuadrados parciales multidimensionales con bilinezarización residual.*

Introduction

In Spain, La Vera (Extremadura) is one of the main geographical areas where paprika is cultivated and produced. This product is a red powder obtained by grinding the dried pepper pods of some varieties of *Capsicum annum L.* This natural food product is commonly used as spice and colorant in cookery and to provide redness to meat products and commercial sacues (Palacios-Morillo et al., 2014). La Vera paprika is recognized under Protected Designation of Origin (PDO) by the European Union and it is obtained from peppers which are dried by means of a characteristic drying system, peppers are smoked-dried (oak or holm oak wood fire), and the rest of peppers produced in other Spanish areas or in other countries are sun dried or hot air dried (Bartolomé et al., 2011). It is a slow process, lasting ten to fifteen days and it confers on the paprika its three fundamental characteristics: aroma, flavor, and color, providing the necessary heat for the perfect dehydration of the fruits. For this reason, in order to avoid fraudulent mixtures, it is important to have tools to differentiate products according to their belonging or not to the PDO.

This product is particularly rich in organic microcomponents with antioxidants properties (carotenes, tocopherols, capsaicinoids, flavonoid glycosides (flavonoids bound to various sugars)…), whose content in paprika depend on the variety of the peppers used to obtain the powder or the system employed. These compounds present absorbent and fluorescence properties which could be used to determinate them.

Recently, spectroscopic and separative techniques together with multivariate and multiway chemometric tools have been commonly used for reducing the time of analysis and providing more information (Borràs et al., 2015). In this sense, analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments have been used (Reinholds et al., 2015) and in order to discriminate foods according to some characteristic properties (Airado-Rodríguez et al., 2009; Palacios-Morillo et al., 2014; Di Bella et al., 2015). Between the most used chemometric techniques Prncipal Component Analysis (PCA) and Parallel Factor Analysis (PARAFAC) for first and second order data, respectively, are found. In addition, second-order algorithms present an advantage, which is the improved ability to get accurate concentration estimates of analytes of interest, even in the presence of uncalibrated interfering components (Escandar et al., 2014, Muñoz de la Peña et al., 2015). However, in the case of a complex matrix, such as the paprika, few studies are found using spectroscopic techniques coupled to chemometric tools to classificate samples according their characterictic properties or quantify some of their main minor components.

In this work, two parts with different aims were developed. One of these parts was exploring the possibilities of the fluorescence properties of some of the main minor components present in paprika, trying to differentiate paprika samples according to they belong or not to

the Protected Design of Origin (PDO) *"Pimentón de La Vera"*. Another part of this work was intended to quantify a mixture of flavonoids in paprika samples using spectrofluorimetry coupled to second order algorithms (PARAFAC, U-PLS/RBL and N-PLS/RBL).

Experimental procedures and results

a) Classification of paprika samples according their origin by using fluorescence coupled to PARAFAC.

The fluorescence components were extracted from 0.1 g of paprika samples with 20 mL of ethanol in an ultrasound bath for 10 min. The EEM matrices were registered. Excitation wavelengths were increased from 200 to 295 nm at 5 nm steps; for each excitation wavelength, the emission spectrum was obtained in the range 300 – 400 nm at 1 nm steps.

With the aim of evaluating capabilities of EEMs to distinguish between samples of different origin, a PARAFAC model was contructed using the EEMs of a set of 48 samples of "Pimentón de La Vera" paprika samples and 19 of paprika samples fron other producers. A pretreatment of data set to remove the Rayleigh signals in all the EEMs used for PARAFAC analysis was performed (Bahram et al., 2006).

Great differences in fluorescence signal were observed between the extracts of "*La Vera"* paprika samples and the extracts of the other types of paprika. The appropriate number of components for constructing the PARAFAC model was chosen and the 3D structures of the three PARAFAC components were obtained. The loading corresponding to the first component did not match with the standard compunds mentioned in the introduction (carotenes, tocopherols, flavonoids or capsaicinoids) and the loading corresponding to the second component was very similar to the EEM of the α-tocopherol.

The score values corresponding to each PARAFAC component were plotted against each other in order to visualize possible systematic information contained in fluorescence data, with respect to the variable origin of the sample. It can be observed two clusters of the paprika samples according to their origin.

The results showed that this methodology could be improved and used as a rapid tool to ensure the authenticity of *"La Vera"* paprika samples.

b) Quantification of flavonoid compounds in paprika samples by using U-PLS/RBL, N-PLS/RBL and PARAFAC.

Firstly, flavonoids were extracted from paprika samples using a previously optimized extraction procedure (0.5 g of sample and 20 mL of MeOH as extractant solvent for 30 min in the ultrasound bath). The analytes were retained in a C18 cartridge and eluted with 2.5 mL MeOH/H_2O (85/25 v/v), after a previous cleaning step, in order to concentrate samples.

Secondly, an acid hydrolysis was performed and excitation-emission fluorescence matrices (EEMs) were obtained (excitation wavelengths from 400 to 470 nm at 5 - nm steps and emission wavelengths from 480 - 600 nm at 2 nm) using an optimized procedure based on the use of a basic medium to maximize the fluorescence signals of the analytes.

In this wavelengths region, where flavonoids exhibited fluorescence, a PARAFAC analysis was performed similarly the previous case. However, in this case the samples were not differentiated as well as in the previously described. After that, a calibration set was constructed from the EEMs of standard samples and U-PLS/RBL, N-PLS/RBL and PARAFAC second-order calibration models were optimized and utilized for the quantification of the main fluorescente flavonoids of paprika, quercetin and kaempferol, in a group of samples belonging to the PDO "*Pimentón de La Vera*", and another group of other different samples. PARAFAC did not allow to differentiate both analytes, however, U-PLS/RBL and N-PLS/RBL allowed differentiate and quantify them separately. The results were compared with those obtained by a previously developed HPLC method and they were better in the case of quercetin because of the higher concentrations of this flavonoid in the analyzed paprika samples.

Conclusions

In the first part of this work fluorescence obtained in the UV region coupled to PARAFAC was utilized for the differentiation of paprika samples according to whether they belonged to the PDO "Pimentón de La Vera" or not. A clustering of PDO samples was appreciated.

In the second part flavonoid compounds (quercetin and kaempferol) were quantify by using U-PLS/RBL and N-PLS/RBL. The results were correlated with whose obtained by liquid chromatography. Better results were obtained in the case of quercetin.

Acknowledgements

The authors are grateful to the Ministerio de Economía y Competitividad of Spain (Project CTQ2014-52309-P) and the Gobierno de Extremadura (GR15-Research Group FQM003), both co-financed by the European FEDER funds, for financially supporting this work. Olga Monago Maraña is grateful to the Ministerio de Educación, Cultura y Deporte of Spain for a FPU grant (Resolución de 18 de noviembre de 2013, de la Secretaría de Estado de Educación, Formación Profesional y Universidades, BOE n◦ 279, de 21/11/13, reference number FPU13/02249).

References

[1] Airado-Rodríguez, D., Galeano-Díaz, T., Durán-Merás, I. (2009). Usefulness of fluorescence excitation-emission matrices in combination with PARAFAC, as fingerprint of red wines. Journal of Agricultural and Food Chemistry, 57, 1711-1720.

[2] Bahram, M., Bro, R., Stedmon, C., Afkhami, A. (2006). Handling of Rayleigh and Raman scatter for PARAFAC modelling of fluorescence data using interpolation. Journal of chemometrics, 20, 99-105.

[3] Bartolomé, T., Coleto, J.M., Velázquez, R. 2011. <<Pimentón de La Vera>>: un caso paradigmático de denominación de origen protegida. In M.R. Lucas, M. Saraiva & A. Rosa (Eds.), A qualidade. Numa perspectiva multi e interdisciplinar, Vol 2 (1st ed) (pp. 117 - 125). Lisboa, Portugal: Ediçoes Sílabo, Lda.

[4] Borrás, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L. 2015. Data fusión methodologies for food and beverage authentication and quality asesment – A review. Analytical Chimica Acta. In Press. http://dx.doi.org/10.1016/j.aca.2015.04.042

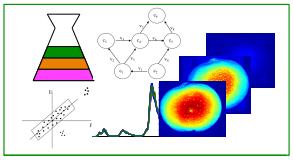
[5] Di Bella G., Lo Turco V., Giorgia Potorti A., Bua G.D. 2015. Geographical discrimination of Italian honey by multi-element analysis with a chemometric approach. Journal of Food Composition and Analysis. In Press. http://dx.doi.org/10.1016/j.jfca.2015.05.003

[6] Escandar, G.M., Goicoechea, H.C., Muñoz de la Peña, A., Olivieri A.C. (2014). Second- and higher-order data generation and calibration: A tutorial. Analytical Chimica Acta, 806, 8 - 26.

[7] Muñoz de la Peña, A., Olivieri, A.C., Escandar, G.M., Goicoechea, H.C. (2015). Fundamentals and analytical applications of multi-way calibration. Elsevier Editorial (Chapter 7 and 8)

[8] Palacios-Morillo, A., Marcos Jurado, J., Alcázar, A. & De Pablos, F. (2014). Geographical characterization of Spanish PDO paprika by multivariate analysis of multielemental content. Talanta, 128, 15-22.

[9] Reinholds I., Bartkevics V., Silvis I.C.J., van Ruth S.M. 2015. Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: A review Journal of Food Composition and Analysis. In Press. http://dx.doi.org/10.1016/j.jfca.2015.05.004



Prediction of olive oil sensory descriptors using instrumental data fusion and partial least squares (PLS) regression

Eva Borràs, Joan Ferré, Ricard Boqué, Montserrat Mestres, Laura Aceña, Olga Busto
Dpto. Química Analítica y Química Orgánica, Universitat Rovira i Virgili, Tarragona. eva.borras@urv.cat, joan.ferre@urv.cat, ricard.boque@urv.cat

Abstract

Three instrumental techniques, headspace-mass spectrometry (HS-MS), mid-infrared spectroscopy (MIR) and UV-visible spectrophotometry (UV-vis) have been combined to quantify virgin olive oil sensory descriptors. The reference sensory values were provided by an official taste panel. Different data fusion strategies were studied to improve the predictions. Best PLS regression models were obtained for musty and fruity attributes. For all the attributes data fusion strategies shown an improvement of the predictions compared to individual techniques.

Keywords:.*Olive oil, sensory attributes, data fusion, PLS regression*

Introduction

Virgin olive oil is a highly appreciated vegetable oil with unique nutritional and organoleptic properties. Its sensory and chemical quality characteristics depend on olive variety, environmental factors, agronomic techniques and cultivation, production and storage conditions. The European Community (EC), the Codex Alimentarius and the International Olive Oil Council (IOOC) have accorded maximum values of specific parameters to guarantee olive oil quality.

To determine olive oil quality categories (extra-virgin, virgin or lampante) different physico-chemical and sensory parameters are evaluated. The only homologated method to assess olive oil sensory attributes is the evaluation by an official taste panel. However, subjectivity, human variability, lack of standards and low throughput per day are some inherent problems associated to this methodology.

Sensory attributes of olive oil are classified into ‘positive’ and ‘negative’. Positive attributes are mainly fruity, bitterness and pungency notes, as well as green grass, sweetness and astringency. The negative ones describe the defects of olive oil, and include fusty (along

with muddy sediment), musty-humidity, winey-vinegary, rancid and metallic. These sensory descriptors depend on the content of volatile and non-volatile minor components.

Alternative solutions to taste panels have been proposed, most of them using instrumental techniques, which offer advantages in terms of fastness, automation and precision. Volatile compounds can be analized by electronic noses (gas sensors or mass spectrometers) using different pre-concentration methods. Non-volatile compounds can be analyzed by electronic tongues (liquid sensors or vibrational spectroscopic techniques such as mid-infrared). Color, although not considered in the evaluation by the taste panel, may influence the quality of the olive oil and color measurements, i.e. by UV-vis spectrophotometry, can provide helpful information.

As olive oil sensory attributes are perceived as a mixture of gustative and olfactive sensations, the combination of data from different instrumental sources can provide complementary information and simplify the sensorial evaluation. Different data fusion approaches of the 'spectral fingerprints' obtained by different instrumental techniques can be applied to correlate to human sensory responses using multivariate pattern recognition techniques. In this study the main olive oil sensory attributes were quantified combining an electronic nose based on headspace mass spectrometry (HS-MS), an electronic tongue based on MIR spectroscopy and an electronic eye based on UV-vis spectrophotometry. Partial least-squares (PLS) was used to correlate sensory data provided by a human taste panel following the official method of the Olive Oil Council (COI/T20/Doc15).

Experimental part

Sensory analysis. Ten sensory attributes, six positive (fruity, bitter, pungent, green grass, sweet and astringent) and four defects (fusty, musty, winey and rancid) were evaluated by the panel for 343 olive oil samples from Catalonia during four harvests (2010-2014). Descriptors were scored in a scale between 0 and 10.

Instrumental analysis. The 343 samples were analyzed with three instrumental techniques: an MS based e-nose and MIR based e-tongue and UV-vis based e-eye. The e-nose consisted on collecting the sample headspace with a solid phase micro-extraction (SPME) fiber and transferring it to an HP5973N MS detector (avoiding chromatographic separation). The m/z range was 50-350 amu. The e-tongue was a FT-MIR Nexus (Thermo Nicolet) spectrometer using a ZnSe crystal ARK multi-bounce over the range 4000–600 cm^{-1} and at 4 cm^{-1} resolution. The e-eye was a UV-Visible Helios Gamma spectrophotometer (Thermo) acquiring within 300 – 1000 nm at 2 nm resolution.

Multivariate analysis. To remove the seasonal variation between samples, a preliminary orthogonalization of the HS-MS and MIR data was applied

To find the optimal prediction model for each attribute, different spectral regions were considered along with different pre-processing options. PLS regression models were built and leave-one-out cross-validated. The lowest root mean square error (RMSECV) was the criterion used to select the optimal number of factors. The final models' performance was confirmed by a test set validation. The average of ten different models was calculated using a random split into a training and test set, with 65% and 35% of the samples, respectively, in order to avoid test results depending on the particular split.

PLS regression models were built for the individual data blocks (MS, MIR and UV-vis), for two-block fused data (MS + MIR) and for three-block fused data (MS + MIR + UV-vis), using low- and mid-level data fusion strategies. In low-level fusion raw data from individual techniques were simply concatenated before model calculation and in mid-level fusion relevant features (independent scores from each individual PLS model) were extracted from the different data blocks and were concatenated into a single matrix.

Results and discussion

The best prediction models for each attribute, considering one-, two- and three-blocks, are summarized in Table 1, together with the detailed PLS results for a specific range of sensory intensities and a final relative root mean square error of prediction (rRMSEP). Best models were obtained for musty and fruity attributes, with R^2 higher than 0.6 and relative errors around 11%. Among the three instrumental techniques, in general, the best results were obtained with mass spectrometry, except for bitterness and fustiness. Good results with MS prove that volatile compounds may potentially contribute to aroma perception. In the case of bitterness the best one-block results were obtained by MIR, confirming the studies that have shown the relationship between this attribute and the polyphenol (non-volatile) content. However, for all the attributes, the best prediction results were obtained when applying data fusion, although in some cases the three-block data fusion only showed a slight improvement. Low-level data fusion was the best option to predict bitterness, pungency and astringency with only MS and MIR. Mid-level data fusion enhanced the prediction of the rest of the attributes, using two-blocks (MS + MIR) for mustiness and three-blocks (MS + MIR + UV-vis) for fruity, green grass, sweet, fusty, winey and rancid attributes.

Table 1. Test-validation PLS regression results (*) for all the attributes studied. Highlighted techniques are the best strategies selected for each attribute.

Attributes	Data Fusion	Technique	Range	R^2p		RMSEP		rRMSEP (%)
				mean	SD	mean	SD	
Positive attributes								
Fruity	One-block	MS	0 – 7	0.55	0.05	0.89	0.07	13.1
	Two-blocks[a]	Low-level		0.63	0.04	0.79	0.04	11.6
	Three-blocks[b]	Mid-level		0.62	0.09	0.77	0.04	11.3
Bitter	One-block	MIR	1 – 7	0.50	0.05	0.67	0.04	11.2
	Two-blocks[a]	Low-level		0.56	0.06	0.62	0.04	10.3
	Three-blocks[b]	Low-level		0.54	0.06	0.64	0.03	10.7
Pungent	One-block	MS	2 – 6.5	0.26	0.08	0.63	0.05	16.2
	Two-blocks[a]	Low-level		0.47	0.07	0.53	0.04	13.6
	Three-blocks[b]	Low-level		0.45	0.06	0.53	0.03	13.6
Green grass	One-block	MS	0 – 5	0.47	0.06	0.83	0.04	17.3
	Two-blocks[a]	Mid-level		0.54	0.07	0.77	0.06	16.0
	Three-blocks[b]	Mid-level		0.58	0.06	0.75	0.05	15.6
Sweet	One-block	MS	3.5 – 5.5	0.36	0.06	0.32	0.02	16.0
	Two-blocks[a]	Low-level		0.41	0.05	0.31	0.01	15.5
	Three-blocks[b]	Mid-level		0.44	0.07	0.30	0.00	15.0
Astringent	One-block	MS	0 – 4	0.40	0.07	0.78	0.04	19.0
	Two-blocks[a]	Low-level		0.56	0.05	0.66	0.03	16.1
	Three-blocks[b]	Low-level		0.53	0.03	0.68	0.03	16.6
Negative attributes								
Fusty	One-block	UV-vis	0 – 6.5	0.54	0.09	0.95	0.11	15.1
	Two-blocks[a]	Mid-level		0.54	0.10	0.92	0.10	14.6
	Three-blocks[b]	Mid-level		0.64	0.05	0.84	0.09	13.3
Musty	One-block	MS	0 – 7	0.64	0.06	0.93	0.09	13.5
	Two-blocks[a]	Mid-level		0.71	0.03	0.82	0.08	11.9
	Three-blocks[b]	Mid-level		0.71	0.06	0.82	0.07	11.9
Winey	One-block	MS	0 – 4	0.58	0.06	0.70	0.05	17.9
	Two-blocks[a]	Mid-level		0.59	0.05	0.69	0.04	17.7
	Three-blocks[b]	Mid-level		0.63	0.06	0.67	0.05	17.2
Rancid	One-block	MS	0 – 7	0.36	0.07	0.82	0.08	12.2
	Two-blocks[a]	Mid-level		0.46	0.09	0.79	0.11	11.8
	Three-blocks[b]	Mid-level		0.51	0.07	0.74	0.09	11.0

(*) Results presented as mean and SD (standard deviation) of the 10 models
R^2p: coefficient of determination of prediction; RMSEP: root mean square error of prediction (test-set); rRMSEP: relative RMSEP
MS: headspace-mass spectrometer; MIR: mid-infrared spectroscopy; UV-vis: ultraviolet-visible spectrophotometer
Two-blocks[a]: MS + MIR; Three-blocks[b]: MS + MIR + UV-vis

References

[1] Apetrei, C., Apetrei, I.M., Villanueva, S., de Saja, J.A., Gutierrez-Rosales, F., Rodríguez-Méndez, M.L., 2010. *Combination of an e-nose, an e-tongue and an e-eye for the characterisation of olive oils with different degree of bitterness*. Anal. Chim. Acta 663, 91–7.

[2] Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015. *Data fusion methodologies for* food *and beverage authentication and quality assessment – A review*. Anal. Chim. Acta. doi:10.1016/j.aca.2015.04.042

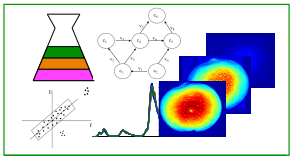
[3] Escuderos, M.E., García, M., Jiménez, A., Horrillo, M. del C., 2013. Edible and non-edible olive oils discrimination by the application of a sensory olfactory system based on tin dioxide sensors. Food Chem. 136, 1154–9.

[4] Ferré, J., Brown, S.D., 2001. Reduction of Model Complexity by Orthogonalization with Respect to Non-Relevant Spectral Changes. Appl. Spectrosc. 55, 708–714

[5] International Olive Council, 2013. COI/T.20/Doc. No 15/Rev. 6 - Sensory analysis of olive oil. Method for the organoleptic assessment of virgin olive oil. COI.

[6] Monteleone, E., Langstaff, S., 2014. Olive Oil Sensory Science, John Wiley. ed. John Wiley & Sons, Ltd, Chichester, UK.

[7] Sinelli, N., Cerretani, L., Egidio, V. Di, Bendini, A., Casiraghi, E., Di Egidio, V., Bendini, A., Casiraghi, E., 2010. Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity. Food Res. Int. 43, 369–375.



Visible-Near InfraRed Spectroscopy (Vis-NIRS) application for differentiation of fresh and frozen/thawed tuna fillets

Martínez E. [a*], Saitua E. [a], Rodríguez R. [a], Olabarrieta I. [a], Pérez I. [a] y Reis M.M. [b]

[a]AZTI-Tecnalia, Parque Tecnológico de Bizkaia, Astondo Bidea – Edif. 609, E-48160, Derio-Bizkaia. (* emartinez@azti.es, [b]Food Assurance and Meat Science Team, AgResearch, Ruakura Research Centre, 10 Bisley Road, Hamilton, New Zealand.

Abstract

The fillet of fresh tuna is an expensive product sold on local and international market. When fished at locations distant to market whole fresh tuna is frozen at temperature bellow -60 ºC to extent its shelf-life and it is sold as a frozen or frozen/thawed product. However, sometimes a fraudulent practice is found in the market when fillets or loins from thawed tuna are sold as fresh at a higher price. When freezing and thawing operations are carried out at proper conditions it is difficult to differentiate between fresh and frozen/thawed fillets. In this work, we investigate the ability of Visible-Near InfraRed Spectroscopy (Vis-NIRS) to detect whether a sample of tuna is fresh or whether it has been frozen/thawed. Fresh fillets obtained locally were subdivided in samples, which were scanned by Vis-NIRS and subsequently frozen. After four days the samples were thawed at 4°C for 24 hours and re-scanned by Vis-NIRS, i.e. each sample was scanned before and after freezing/thawing. Vis-NIR spectra (415 nm to 925 nm) were collected in two surfaces of the samples (original and bloomed surface). Two multivariate methods were compared to evaluate the Vis-NIRS as tool for differentiation between fresh and frozen/thawed samples, i.e.: Partial Least Square Discriminant Analysis (PLS-DA); and Multi-Level Partial Least Square Discriminant Analysis (MLPLS-DA). Repeated double cross-validation was applied to compare the performance of the two approaches. MLPLS-DA showed higher success rate (96%, sensitivity=96%, specificity = 96%) compared to PLS-DA (81%, sensitivity=91%, specificity = 71%). Regression coefficients resulting from MLPLS-DA showed two spectral ranges of importance, one in the visible range and the other one in the near infrared spectral range. Overall results suggested that Vis-NIRS was able to detect the difference between fresh and frozen/thawed tuna samples.

Keywords: *MLPLS-DA, PLS-DA, Vis-NIRS, NIR, fresh tuna, frozen tuna, thawed tuna.*

Introduction

This work investigates the ability of Visible-Near InfraRed Spectroscopy (Vis-NIRS) to detect the effect of freezing/thawing applied to fillets of tuna (*Thunnus thuynnus*), contributing to the development of a non-invasive method for detection of the freshness of tuna fillets. As first step, in this investigation we compared Vis-NIR spectra collected from tuna samples before and after being frozen/thawed. The samples were obtained by cutting off 15 fillets (790g ±215g) from different tuna fish, in pieces (67g±30g). Thus, the obtained samples or pieces varied in size and fat content.. By using this procedure we expected to produce a variation on the effect of freezing across samples, which would allow investigating Vis-NIRS in a more challenging situation. However, this could also mean that the effect of freezing could be affected by sample to sample variation (e.g. sample size). To deal with this type of problem it has been proposed the use of Multi-Level Partial Least Square-Discriminant Analysis (MLPLS-DA), which allows separating the effect of treatment from variation among samples. Thus we applied MLPLS-DA and Partial Least Square-Discriminant Analysis (PLS-DA) to evaluate ability of Vis-NIRS to detect whether a sample of tuna is fresh or frozen/thawed.

Vis-NIR data were collected in two sub-sets. In the first one (sub-set 1, n=12 fillets) each fillet was taken from 4 °C to a room at 16 °C, and cut into 9 pieces, which were covered with plastic film (the fresh cut surface called 'T' was left upwards), and left for one hour in the room until they were scanned at 12.6±1.8 ºC. For three additional fillets (sub-set 2, n=3 fillets), similar procedure was used with a slight modification: after the fillet was cut into 9 pieces, they were left at room temperature (16 °C) covered with plastic film for one hour and then, a 5 mm slice of the transversal area was cut off for each of the nine pieces, which were left for another hour covered with plastic film at 16 ºC, and they were scanned at 17.4±0.5ºC. The modification on the procedure for sub-set 2 was applied to generate a set of samples with a higher temperature for the fresh samples on Vis-NIRS scanning. Vis-NIR spectra were collected on the original external surface (called 'S') and on the transversal surface ('T'), generated by cutting, for each of the nine pieces. After collecting the spectra, the samples were wrapped and frozen at -80 °C. After four days three samples of each fillet were transferred to 4 °C and left to thaw for twenty four hours. Then, the samples were transferred to a room at 16 °C and left to equilibrate for one hour, where they were unwrapped and weighted, and a slice of five millimetres was cut off from the same traversal section which was scanned by Vis-NIRS when fresh. The samples were then left for 1 hour with this fresh surface upwards and covered with plastic film. Afterwards, Vis-NIR spectra

were collected on surfaces 'S' (original external surface of the samples) and 'T' (fresh/bloomed surface) at 17.5±1.4 and 17.8±1.0 ºC for subset 1 and subset 2 respectively.

The NIR equipment used to collect the data is composed by a spectrometer (AvaSpec 2048, Avantes, Netherlands), a light source (AvaLight-HAL,Avantes, Netherlands) and a fibre optic probe (FCR-7UVIR400-2-2.5x100,Avantes, Netherland), which is composed of 7 optical fibres of 200µm core (6 illumination-fibres and one read fibre). The reflectance spectra acquired is integrated in AvaSoft 8 software (Avantes, Netherlands).

Double cross validation (DCV) was applied to select the parameters of the classification models (*Filzmoser et al. 2009, Westerhuis et al. 2010, E. Szymaníska et al. 2012*). In double cross validation procedure the data set is first split in 'ndcv' sets (ndcv=3): one is used as test data set and the others are combined and used as calibration data set. The calibration data is used to fit a model. To fit this model it is necessary to identify the number of latent variables, which is done using cross validation procedure. In this case, the calibration data set is subdivided in 'ncv' sub-sets (ncv=4). The number of latent variables is increased from 1 to a maximum number 'nlvmax' (nlvmax=15). For each number of latent variables (1 to nlvmax) ncv models are fitted, by leaving each of ncv sub-sets out of the model fitting, and using the model fitted without that set to predict the data from the sub-set left out. Each sub-set is left out once, and at the end the predictions of the sub-sets which were left out are combined and the performance of the model is evaluated using the number of misclassified samples. The number of latent variables corresponding to the best performance (the minimum number of misclassified samples) is chosen. Then, a model is fitted with all samples from the calibration data set using the chosen number of latent variables and it is applied to the data from the samples of the test data set. This process is performed until all the ndcv sets have been used as test data set once. Double cross validation is repeated twenty times and the average of the twenty predictions for each sample is presented. In this study the procedure was repeated three times, for each time only one sample per fillet was used. This procedure was applied for MLPLS-DA and PLS-DA in similar subsets.

MLPLS-DA and PLS-DA were carried out using toolbox from Biosystems Data Analysis Group from the University of Amsterdam (*MLPLSDA*) using Matlab R2013a (Version 8.1.0.604, The MathWorks, Inc.). Data visualization was carried out in R v 3.2.0 (*R Core Team, 2015*) with package 'lattice'.

The results of prediction for MLPLS-DA and PLS-DA are presented in Figure 1. In this case, predictions for fresh ('Frsh') samples are expected to be lower than zero and samples that had been frozen/thawed ('Frzn/Thwd') higher than zero. Figure 1 shows that both models are able to detect the difference between fresh and frozen/thawed samples. However, PLS-DA seems to be affected by the 'initial' state of the sample, which is

suggested by the presence of offsets in the predictions of fresh samples. This is corrected by MLPLS-DA which considers the same sample as control and treatment. There were two sub sets of the samples (sub-set 1 and 2), which present different scanning temperatures for the fresh samples (12.6±1.8°C/17.5±1.4°C and 17.4±0.5°C /17.8±1.0°C). There is no indication that the temperature at time of scanning affected the predictions of MLPLS-DA as shown on Figure 1, where predictions for sub-set 1 (lower scanning temperature at fresh) are shown on the bottom row and for sub-set 2 (higher scanning temperature at fresh) are shown on the top row.

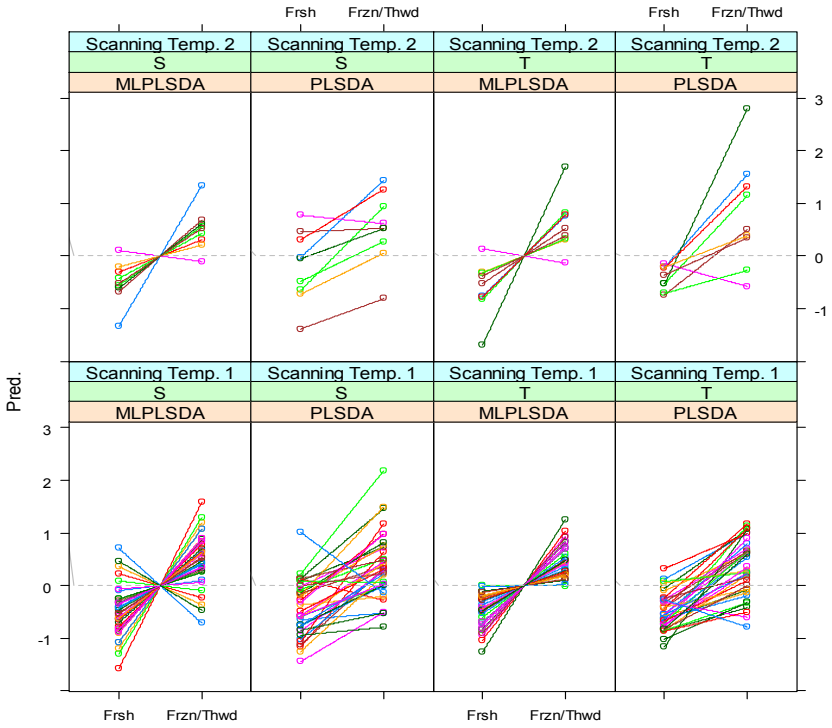


Figure 1 – Predictions of MLPLS-DA and PLS-DA fitted to differentiate between fresh (Frsh) and frozen/thawed (Frzn/Thwd) tuna fillets. Predictions for fresh are expected to be lower than zero and for frozen/thawed higher than zero. Each line corresponds to one sample connecting prediction when fresh with prediction after being frozen/thawed. 'S' and 'T' represent the surface where Vis-NIR spectra were collected.Table 1 – Performance based on the number of misclassified samples for predictions of models MLPLS-DA and PLS-DA fitted to differentiate between fresh and frozen/thawed tuna fillets.

Table 1. Sensitivity/Specificity: percentage of positives/negatives correctly identified. NER: Non-error rate.

	Surface	NER	Sensitivity(fresh)/ Specificity(frozen/thawed)	Sensitivity(frozen/thawed)/ Specificity(fresh)
MLPLS-DA	S	87%	87%	87%
	T	96%	96%	96%
PLS-DA	S	78%	78%	78%
	T	81%	91%	71%

Overall, the highest rate of correct classification is observed in spectra collected in surface 'T'. These results are an indication that Vis-NIRS is able to detect the difference between fresh and frozen/thawed samples.

Acknowledgments

The financial support from Royal Society New Zealand within the New Zealand-EU International Research Staff Exchange Scheme (IRSES)/REPLAY is fully appreciated. Ekaitz Martinez acknowledges the scholarship of the Department of Environment, Territorial Planning, Agriculture and Fisheries of the Basque Country Government "*Aplicación de tecnología analítica de procesos para la evaluación de atributos de calidad de productos alimentarias*" for the development of this work. The authors are also grateful to Basque Government for the financial support through the grant "*CLASISENSOR - Clasificación de alimentos por calidad mediante el desarrollo de sistemas sensóricos*" .

References

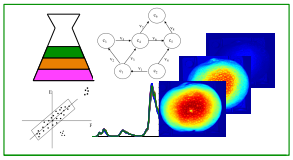
[1] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[2] MLPLSDA, Multilevel Data Analysis http://www.bdagroup.nl/content/Downloads/software/matlab_routines/MLPLSDA.zip (access 18/08/2015).

[3] P. Filzmoser, B. Liebmann, and K. Varmuza.(2009). Repeated double cross validation. Journal of Chemometrics, 23,160-171.

[4] E. Szymaníska, E. Saccenti, A.K. Smilde, J. A. Westerhuis. (2012). Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics, 8, S3–S16.

[5] J. A. Westerhuis, E. J. J. van Velzen, H. C. J. Hoefsloot, A. K. Smilde. (2010). Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6, 119–128.



Chemometrics applied to study the solution equilibria of cytosine-rich and guanine-rich DNA sequences near th promoter region of the SMARCA4

Sanae Benabou[a], Anna Aviñó[b], Ramón Eritja[b], Raimundo Gargallo[a]

[a]Department of Analytical Chemistry, University of Barcelona, Spain, [b] Institute for Advanced Chemistry of Catalonia (IQAC), CSIC, Networking Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain. E-mail: sbenabou_13@ub.edu

Abstract

The SMARCA4 gene is involved in small cell carcinoma of hypercalcemic type, a rare and aggressive type of ovarian cancer (Witkowski et al. 2014). In the promoter region of this gene, there is a wealth of guanine and cytosine bases that could lead to the formation of complex DNA structures such as G-quadruplex and i-motif.

In this context, we have focused our attention on the study of the solution equilibria of two long guanine- and cytosine-rich sequences found near the promoter region of the gene SMARCA4.

The results show that the application of a multivariate approach allows the succesful resolution of systems involving i-motif and G-quadruplex structures.

Keywords: *G-quadruplex, i-motif, DNA, SMARCA4, Multivariate Analysis*

Introduction

Recently, the study of complex structures of DNA, such as G-quadruplex and *i*-motif from guanine- and cytosine-rich regions, respectively, is being subjected to an intensive research. The core of G-quadrupex structure is formed by two or more tetrads, an ensemble of four

guanine bases linked by hydrogen bonds in almost the same spatial plane (Figure 1). This structure is stabilized by intramolecular and intermolecular stacking, and strong electrostatic interactions with cations within the structure (Neidle and Balasubramanian 2006, Shafer and Smirnov 2001). On the other hand, the i-motif structure consists of two parallel-stranded duplexes that are intercalated in an antiparallel manner (Figure 2). The building block of the structure is a base pair involving one neutral cytosine and one protonated cytosine at N3, known as the C-C^+ base pair, bonded by three hydrogen bonds. The C-C^+ base pair needs the protonation of one of the cytosines at N3, the pK_a value of which is around 4.5. For this reason the formation of a relatively stable *i*-motif structure needs a slightly acid environment.

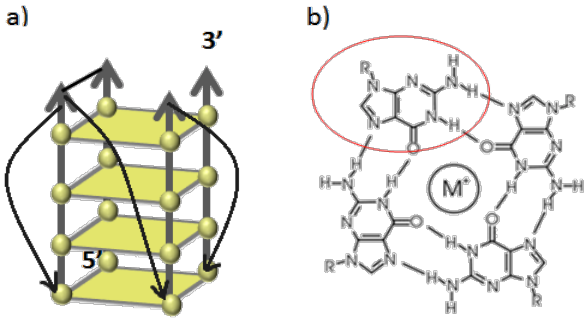


Figure 1. G-quadruplex structure. (a) 3D structure of parallel G-quadruplex showing four tetrads. b) four guanine linked by hydrogen bonds in the same plane (tetrad) in the presence of monovalent ions.

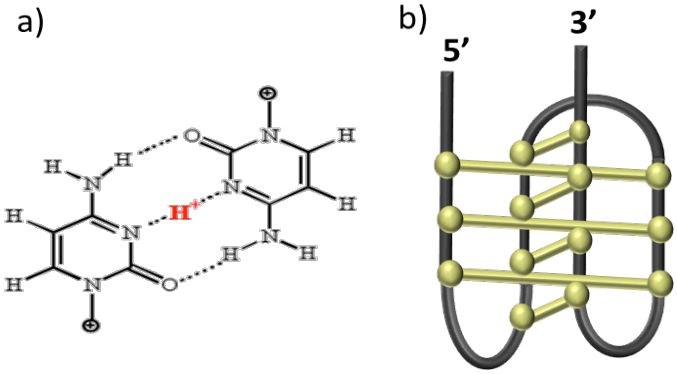


Figure 2. *i*-motif structure. (a) the C·C^+ base pair, (b) 3D structure of an i-motif showing seven C·C^+ base pairs in four strands.

The formation of G-quadruplex by guanine-rich sequences has been widely studied because of their recent discovery *in vivo* and their possible role in biological processes like cancer. On the contrary, the formation of i-motif structures by complementary cytosine-rich sequences has received little attention due to their strong dependence on the pH. Nowadays, the proposal of potential roles *in vivo*, as well as nanotechnological applications has produced an increasing interest in the study of *i*-motif structures. In this context, it has been shown the formation *in vitro* of such structures in DNA sequences corresponding to the end of telomeres and to the promoter regions of several oncogenes, such as *c-kit*, *c-myc* or *bcl-2* (Day et al. 2014).

The main aim of the proposed work is the study of the solution equilibria of two long guanine- and cytosine-rich sequences found near the promoter region of the gene SMARCA4 and the interaction of these with the porphyrin TMPyP4. The interest in the study of this gene lies in the fact that it is important in controling the cell differentiation induced by retinoic acid and glucocorticoids in both lung cancer and others (Witkowski et al. 2014). This ligand may be considered as a model ligand because it has been widely used for the study of G-quadruplex structures.

Circular dichroism and molecular absorption spectroscopies have been used to determine the conditions under which both G-quadruplex and *i*-motif structures are formed (pH, temperature, ionic strength). Multivariate data analysis based on hybrid- and hard-modeling methods has been used to recover qualitative and quantitative information about the species and conformations present in all experiments (Benabou et al. 2014, Bucek et al. 2010). In this study, we have been used hard-modeling such as EQUISPEC for acid-base titration (Figure 3) and DNA-ligand interaction, and hybrid-modeling for melting experiments. Finallly, Size-Exclusion Chromatography (SEC) has been used to complement the results obtained from spectroscopy.

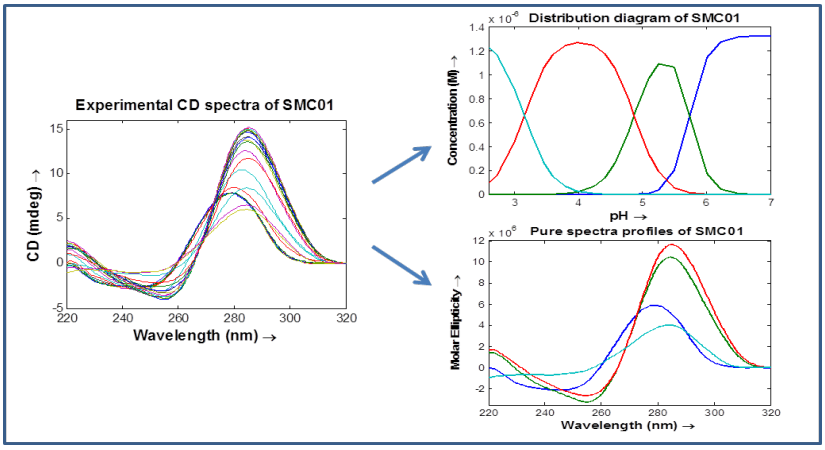


Figure 3. Experimental spectra of CD recorded throughout the acid-base titration of the SMC01 sequence. Pure spectra and proposed distribution diagrams obtained from EQUISPEC analysis.

The results have been shown that the application of a multivariate approach allows the succesful resolution of systems involving *i*-motif and G-quadruplex structures.

Acknowledgement

We acknowledge funding from the Spanish government (CTQ2012-38616-C02-02) and recognition from the Autonomous Catalan government (2014SGR1106).

References

[1] S. Benabou, R. Ferreira, A. Aviñó, C. González, S. Lyonnais, M. Solà, R. Eritja, J. Jaumot and R. Gargallo, Biochim. Biophys. Acta, Gen. Subj., 2014, 1840, 41–52.

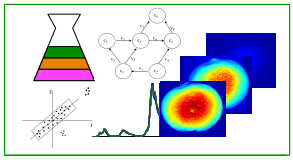
[2] P. Bucek, R. Gargallo and A. Kudrev, Anal. Chim. Acta, 2010, 683, 69–77.

[3] H. A. Day, P. Pavlou, and Z. a E. Waller, Bioorganic Med. Chem., 2014, 22, 4407–4418.

[4] Shafer RH & Smirnov I (2001) Biological aspects of DNA/RNA quadruplexes. Biopolymers 56, 209–227.

[5] Neidle S & Balasubramanian S (2006) Quadruplex Nucleic Acids. Royal Society of Chemistry, London.

[6] L. Witkowski, J. Carrot-Zhang, S. Albrecht, S. Fahiminiya, et al., Nat. Genet., 2014, 46, 438–43.



Second-order calibration using high-performance liquid chromatography with dual UV and fluorimetric detection for the analysis of sex hormones in environmental samples

Pérez, Rocío L.[a] y Escandar, Graciela M.[b]

[a] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario. Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (S2002LRK), Argentina – rpereziquir-conicet.gov.a, [b] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario. Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (S2002LRK), Argentina – escandariquir-conicet.gov.ar

Abstract

The objective of this work was the development of a green method based on non-sophisticated instrumental for the quantification of sex hormones: seven natural and synthetic estrogens, three progestagens and one androgen. The approach involves isocratic high-performance liquid chromatography with dual diode array and fluorescence detection in a single run, coupled to second-order calibration. It takes advantage of: (1) chromatography, which allows total or partial resolution of a large number of compounds, (2) dual detection, which permits selection of the most appropriate signal for each analyte and, (3) second-order calibration, which enables mathematical resolution of incompletely resolved chromatographic bands and analyte determination in the presence of other sample constituents. A marked decrease is achieved in the consumption of organic solvents for cleaning, extraction and chromatographic separation, and experimental and elution times are shortened, with only a single solid-phase extraction with C18-membranes. Outstanding selectivity is attained with the MCR-ALS second-order algorithm, which allowed the green analyte determination in natural waters and sediments. Limits of detection in the ranges 6–20, 14–21, and 18–24 ng L^{-1} for estrogens, progestagens and the studied androgen were respectively achieved in real water samples. In sediments, they were 0.1–0.9, 0.2–0.8, and 0.5–0.9 ng g^{-1}. Relative prediction errors from 2 to 10 % for water samples and from 1 to 8 % for sediments were reached.

Keywords: *sexual hormones-environmental samples-MCR-ALS.*

Introduction

The determination of sexual hormones in aquatic bodies and related environmental samples such as sediments is a very important activity in modern steroid hormone analysis (S. Görög 2011).

Within the past few years, a new set of methods has arisen, the so-called "green analytical chemistry" (GAC) methods. The driving force has been the need to protect the environment, without negative impact on basic analytical properties(S. Armenta, S. Garrigues et al. 2008).

In this context, the main objective of the present work was the development of a GAC method for the analysis of a significant number of sex hormones at part per trillion concentrations in environmental samples such as surface waters, underground waters and sediments. In the present work, single-run dual diode array detection (DAD) and fluorescence dectection (FLD) are applied for the determination of eleven analytes involving natural [estriol (E3), estradiol (E2), estrona (E1)] and synthetic [ethynilestradiol (EE2), diethylstilbestrol (DES), hexestrol (HEX), mestranol (MEST)] estrogens, endogenous [progesterone (PROG)] and synthetic [norethisterone (NOR), levonorgestrel (LEV)] progestagens, and a common precursor of male and female sex hormones, androstenedione (AE). The dual detection allows us to quantify: (1) estrogens, through the intense fluorescence displayed by most of them in the employed mobile phase, and (2) the remaining non-fluorescent hormones by their UV absorption properties. The benefits obtained by combining the applied analytical method with the chemometric algorithm multivariate curve resolution with alternating least-squares (MCR-ALS) (R. Tauler, M. Maeder et al. 2009) are demonstrated. To the best of our knowledge, this is the first time that eleven sex hormones are evaluated in challenging media using a GAC method, and second-order calibration is applied to both high-performance liquid chromatography (HPLC)-DAD and HPLC-FLD matrices measured for a single chromatographic run.

Results

A calibration set was composed by ten samples. Eight samples of the set correspond to the concentrations provided by a semi-factorial for four overlapped analytes (E1, DES, AE and HEX) and equally spaced concentrations for those analytes with resolved bands. The remaining calibration samples were a blank solution and a mixture of all the analytes at intermediate concentrations. For each sample, two matrices (HPLC-DAD and HPLC-FLD) were collected in an only chromatographic run under isocratic mode (ACN:H_2O, 50:50 % v/v). The data matrices were collected each 1.8 s using wavelength from 200 to 330 nm in steps of 1 nm for the DAD, and each 1.5 s from 295 to 350 nm in steps of 1 nm for the FLD, setting the excitation wavelength at 275 nm and the slit widths at 1 nm. Specifically, while low or non-fluorescent compounds were chromatographically quantified through their UV signals (namely, NOR, DES, AE, LEV, PROG and E1), the estrogens E3, E2, EE2, HEX and MEST were determined by fluorescence.

Due to the losses of trilinearity of chromatographic-spectral matrices the algorithm MCR-ALS was chosen for the chemometric analysis and performing matrix augmentation in the temporal direction (R. Tauler, M. Maeder et al. 2009). However, in the system under study, an additional problem must be taken into account: some analytes exhibit very similar absorbance and fluorescence spectra. In this situation, if the full DAD and FLD chromatograms are processed, unsuitable results are obtained because the mathematical pseudorank is smaller than the chemical rank (A. C. Olivieri and G. M. Escandar 2014). Therefore, to overcome this inconvenience, MCR-ALS was applied with matrix augmentation in the temporal direction in various selected time ranges, ensuring that each partial chromatographic region includes analytes with different spectral profiles.

A validation set of ten samples were prepared and were analyzed chemometricly. The number of components in each data matrix was estimated by principal component analysis, and justified taking into account the presence of the corresponding analytes and background signals in each time region. Non-negativity restrictions were applied in both modes; unimodality restriction was applied in the elution time mode to the signals corresponding to the analytes. The selected ALS convergence criterion was 0.01 % (relative change in fit for successive iterations), and in validation samples convergence was achieved in less than 20 iterations. The good recovery results in validation samples in addition to the elliptical joint confidence region (EJCR) (A. G. González, M. A. Herrador et al. 1999) test for the slope and intercept of the plot corresponding to each analyte. Because all ellipses include the theoretically expected values of (1,0) for slope and intercept, respectively, the accuracy of the applied methodology for these compounds in validation samples can be claimed.

With the purpose of testing the applicability of the investigated method, two types of samples (water and sediment samples obtained from different sources) were selected as examples of environmental matrices. The investigated water samples were spiked with all analytes, combining random values for estrogens (except E1 and DES), the ranges were 10-20 ng L^{-1} (low), 25-35 ng L^{-1} (medium) and 40-52 ng L^{-1} (high), whereas for the remaining analytes they were 19-32 ng L^{-1} (low), 46-65 ng L^{-1} (medium) and 81-99 ng L^{-1} (high). Before the inyection, the samples were subjected to a simple pre-concentration with a C18- membrane (pre-concentration factor of 1:500). The sediment samples were spiked with standard methanol solutions in order to obtain concentration levels the range 2.5-24.3 ng g^{-1}, frozen and lyophilized. The fortified samples were subjected to a simple extraction procedure. And after the extraction the samples were subjected to the same pre-concentration procedure as the water samples.

In the real samples, the resolution of these systems represents a real analytical challenge. However, MCR-ALS, as other second-order algorithms, achieves the so-called “second-order advantage”, which avoids the major obstacle of traditional zeroth-order calibration methods applied to complex mixtures: the requirement of interference removal before the quantitative

analytical method is applied (A. C. Olivieri 2008). MCR-ALS data processing was similar to that for validation samples, but in addition to non-negativity in both modes and unimodality in the time mode restrictions, the correspondence restriction was applied to most samples, which fixes the sequence and the presence or absence of components in specific matrices (R. Tauler, M. Maeder et al. 2009). In real samples, with an unknown number of constituents, the number of components was estimated as in validation and varied between 6 and 10, depending on the sample and analyzed time region. The number of ALS iterations in these complex samples was less than 30 in most cases, with residual fits in the order of the expected instrumental noise associated with each detector.

It is necessary to make a distinction between the presently proposed strategy, that only needs to remove suspended particles in some natural waters (e.g. river and underground water) from more strict extraction and/or clean-up protocols usually employed in chromatographic analysis coupled to MS or tandem MS for the determination of sex hormones in natural waters (Streck 2009; S. Görög 2011). In our case, because of the second-order advantage, soluble sample constituents injected in the chromatographic column with the analytes do not interfere in the analysis, as is demonstrated with the successful MCR-ALS predictions. Regarding this latter issue, it is also remarkable how the amount of organic solvents is presently decreased using the proposed strategy, in comparison with that currently employed in sample pre-treatments for the analysis of the studied hormones in sediments (P. Labadie and E. M. Hill 2007; I. Matić, S. Grujić et al. 2014).

The obtained results for the real water and sediments samples, in terms of the elliptical joint confidence region test with ellipses for each type of sample including the theoretically expected values of (1,0), indicate the accuracy of the used methodology. The relative errors of prediction are very acceptable (smaller than 10 %) taking into account the complexity of the studied samples.

Conclusions

Eleven sex hormones included in the group of endocrine disruptors have been analyzed by LC-DAD-FLD under an isocratic regime, in a short elution time, and applying a minimal sample pre-treatment. The flexibility of the multivariate algorithm MCR-ALS allowed the successful resolution of coeluted peaks belonging to analytes and interferents in challenging scenarios, such as those formed by natural waters and sediments. Since the length of the chromatographic run, the solvent consumption, the waste generation and the operator time are significantly reduced, while the frequency of sample processing is notably increased, the proposed method meets the criteria defined in the framework of green analytical chemistry principles and may allow to substitute more complex analytical methods.

References

[1] S. Armenta, S. Garrigues, et al. (2008). Green analytical chemistry. TrAC Trend Anal. Chem. 27(6): 497-511.

[2] A. G. González, M. A. Herrador, et al. (1999). "Intra-laboratory testing of method accuracy from recovery assays." Talanta 48(3): 729-736.

[3] S. Görög (2011). Advances in the analysis of steroid hormone drugs in pharmaceuticals and environmental samples (2004–2010). J. of Pharmaceut. Biom. 55(4): 728-743.

[4] P. Labadie and E. M. Hill (2007). Analysis of estrogens in river sediments by liquid chromatography–electrospray ionisation mass spectrometry: Comparison of tandem mass spectrometry and time-of-flight mass spectrometry. J. Chrom. A 1141(2): 174-181.

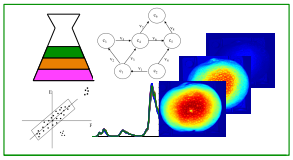
[5] I. Matić, S. Grujić, et al. (2014). Trace analysis of selected hormones and sterols in river sediments by liquid chromatography-atmospheric pressure chemical ionization–tandem mass spectrometry. J. Chrom. A 1364: 117-127.

[6] A. C. Olivieri (2008). Analytical Advantages of Multivariate Data Processing. One, Two, Three, Infinity? Anal. Chem. 80(15): 5713-5720.

[7] A. C. Olivieri and G. M. Escandar, Eds. (2014). Practical Three-Way Calibration. Amsterdam, Elsevier.

[8] Streck, G. (2009). Chemical and biological analysis of estrogenic, progestagenic and androgenic steroids in the environment. TrAC Trend Anal. Chem. 28(6): 635-652.

[9] R. Tauler, M. Maeder, et al. (2009). 2.24 - Multiset Data Analysis: Extended multivariate curve resolution. Comprehensive Chemometrics. S. D. Brown, R. Tauler and B. Walczak. Oxford, Elsevier: 473-505.



A workflow based on MCR-ALS for feature detection in untargeted metabolomic experiments

Elena Ortiz-Villanueva, Joaquim Jaumot y Romà Tauler
Departamento de Quimica Ambiental, IDAEA-CSIC, Barcelona, España.

Abstract

In the last few years, the use of chemometric tools plays a crucial role to gain new knowledge in the –omics field. The inherent complexity of –omics data requires the application of multivariate data analysis tools. In this work, the potential of multivariate curve resolution alternating least squares (MCR-ALS) as a feature detection tool is demonstrated for the analysis of untargeted -omics data. The proposed workflow consists of the preliminary analysis of total ion current chromatograms (TICs) for exploratory purposes using principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). Then, MCR-ALS is applied to selected regions of the multiple full-scan MS data sets. This strategy permits the resolution of a large number of elution profiles characterized by their chromatographic peaks and mass spectra. Finally, in the last step of the workflow, these resolved profiles allows the identification of the detected features (considering the resolved mass spectra) and the assessment of their statistical significance (considering the resolved elution profiles). Biological interpretation of the system under study can be gathered considering the final list of identified features. The advantages of the application of this method are shown for an untargeted LC-MS metabolomic study related to bisphenol-A effects on zebrafish embryos.

Keywords: *Liquid chromatography-mass spectrometry, Metabolic profiling, Multivariate data analyses, Untargeted analysis*

Resumen

En los últimos años, los métodos quimiométricos han jugado un papel muy importante para obtener nuevos conocimiento en el campo de la ciencias ómicas. La complejidad de los datos ómicos hace necesaria la aplicación de métodos de análisis multivariante con el fin de extraer la máxima información relevante. En este trabajo, el potencial del método de resolución multivariante de curvas por mínimos cuadrados alternados (MCR-ALS) como una herramienta de selección de variables se ha demostrado para el caso de datos ómicos no dirigidos. El flujo de trabajo propuesto comienza con el análisis exploratiorio preliminar de los cromatogramas de corriente iónica total (TIC) mediante análisis de componentes principales (PCA) y análisis discriminante por mínimos cuadrados parciales (PLS-DA). Seguidamente, se aplica MCR-ALS al conjunto de los datos LC-MS. Esta estrategia permite la resolución de un gran número de perfiles de elución caracterizados por sus picos cromatográficos y sus correspondientes espectros de masas. Finalmente, en el último paso del proceso, a partir de los componentes resueltos se puede obtener la identificación de los variables (a partir del espectro de masas resuelto) y la evaluación de su significación estadística (a partir de los perfiles cromatográficos resueltos). Esto permite obtener una interpretación biológica a los cambios observados en el sistema estudiado a partir de la lista de metabolitos identificados. Las ventajas de la aplicación de este flujo de trabajo se muestran para un estudio LC-MS metabolómico no dirigido basado en la exposición de embriones de peces cebra a bisfenol-A.

Palabras clave: *Cromatografía de líquidos acoplada a espectrometría de masas, Evaluación perfil metabólico, Análisis de datos multivariantes, Análisis no dirigido.*

Introducción

La metabolómica es un campo relativamente moderno que tiene como objetivo obtener una amplia cobertura de los compuestos de bajo peso molecular de los sistemas biológicos (Villas-Bôas et al. 2005, Hirayama et al. 2014). En los últimos años, las respuestas metabólicas a estímulos externos se han investigado ampliamente teniendo en cuenta los cambios en los niveles de concentración detectados por diferentes plataformas de análisis de alto rendimiento. Entre todas estas plataformas, destaca la espectroscopia de resonancia magnética nuclear (RMN) (Puig-Castellví et al. 2015) y, sobretodo, la espectrometría de masas, comúnmente acoplada a una técnica de cromatografía líquida (LC-MS) (Bedia et al. 2015), de cromatografía de gases (GC-MS) (Parastar et al. 2012)o a la electroforesis capilar (CE-MS) (Ortiz-Villanueva et al. 2015). Actualmente, se ha puesto mucha atención a la aplicación de LC con columnas de interacción hidrófilica (HILIC) debido a su capacidad para llevar a cabo análisis a pequeña escala de compuestos pequeños y polares. Las herramientas de análisis de datos juegan un papel crucial para lograr extraer información de los datos y obtener una buena interpretación biológica. Por lo tanto, la necesidad de analizar grandes conjuntos de datos complejos procedentes de estudios de ómica con técnicas analíticas de alto rendimiento han alentado a los investigadores a desarrollar y aplicar herramientas de análisis de datos avanzadas (Trygg et al. 2007).

En el caso concreto de los datos LC-MS, existen diferentes aproximaciones para el procesamiento de los datos como, por ejemplo, el ampliamente utilizado XCMS (Tautenhahn et al. 2008). Sin embargo, la utilización de herramientas quimiométricas durante el proceso de análisis facilita la obtención de más información. Por un lado, se puede llevar a cabo un análisis exploratorio preliminar sobre los cromatogramas de corriente iónica total (TIC) con el fin de establecer la posibilidad de diferenciar las muestras mediante métodos como el análisis por componentes principales. Otros métodos como el análisis por mínimos cuadrados parciales discriminante (PLS-DA) pueden ser también utilizados con el fin de distinguir las muestras al mismo tiempo que se pueden identificar las regiones del cromatograma causantes de estas diferencias mediante el uso de métodos de selección de variables como los parámetros Variables Importance in Projection (VIPs) o la Selectivity Ratio (Trygg et al. 2007, Fasoula et al. 2105). Sin embargo, el estudio de los TICs podría conllevar la pérdida de la información más relevante debido a la presencia de picos superpuestos en los datos LC-MS originales. En consecuencia, la aplicación de métodos de resolución como, por ejemplo, la resolución multivariante de curvas por mínimos cuadrados alternados (MCR-ALS) (Tauler 1995) se puede proponer como una herramienta muy poderosa para llevar a cabo un análisis más profundo de los datos de LC-MS. Este método se ha demostrado especialmente útil para analizar múltiples tipos de sistemas multicomponentes y, recientemente, en casos de datos metabolómicos obtenidos mediante LC-MS (Gorrochate-

gui et al. 2015, Farrés et al. 2105). Además, este enfoque propuesto proporciona la solución a problemas adicionales que se encuentran habitualmente en los datos LC-MS, como las contribuciones del ruido de fondo, una baja relación señal-ruido, la presencia de picos asimétricos y los leves desplazamientos en los tiempos de retención entre inyecciones.

El objetivo de este trabajo es demostrar la utilidad de la aproximación basada MCR-ALS en el análisis de datos LC-MS para obtener información fiable en el estudio de datos metabolómicos. La idoneidad de la aplicación de este método se demuestra en el caso de perfiles metabólicos de embriones de pez cebra (*Danio rerio*) expuestos a bisfenol-A (BPA) a diferentes concentraciones (0, 1, 2 y 4 µg/mL de BPA). En la Figura 1 se muestra, como ejemplo, uno de los componentes resueltos. En la parte superior de la figura se observan los perfiles cromatográficos resueltos para el componente que permiten evaluar el efecto que tienen los diferentes niveles del contaminante en las muestras estudiadas. En la parte inferior de la figura, se muestra el espectro de masas resuelto para el componente a partir del cual se puede llevar a cabo la identificación del metabólito y su interpretación biológica.

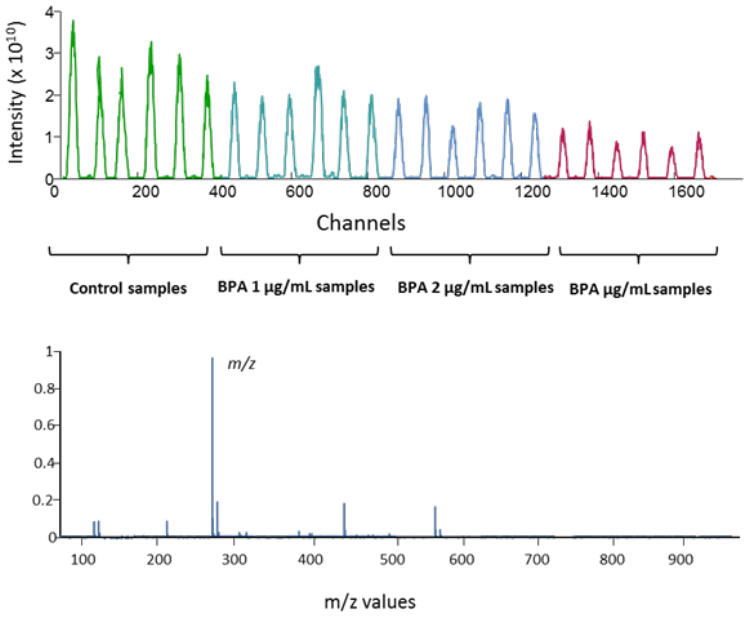


Figura 1. Ejemplo de resolución de MCR-ALS: (A) Perfil de elución para cada mues-tra y (B) espectro de masas resuelto.

Referencias

[1] Bedia, C., Dalmau, N., Jaumot, J.,Tauler, R., Environ Res.,2015, 140,18-3.

[2] Farrés, M., Piña, B., Tauler, R., Metabolomics, 2015, 11, 210-224.

[3] Gorrochategui, E., Casas, J., Porte, C., Lacorte, S., Tauler, R., Anal.Chim. Acta, 2015, 854, 20-33.

[4] Fasoula, S., Zisi, C., Sampsonidis, I., Virgiliou, C., Theodoridis, G., Gika, H., Nikitas, P., Pappa-Louisi, A., J. Chromatogr. A, 2015, 1387,49-52.

[5] Hirayama, A., Wakayama, M., Soga, T., TrAC Trend Anal Chem. 2014, 61, 215-222.

[6] Ortiz-Villanueva, E., Jaumot, J., Benavente, F., Piña, B., Sanz-Nebot,V., Tauler,T., Electrophoresis, 2015, DOI: 10.1002/elps.201500027.

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

[8] Puig-Castellví, F., Alfonso, I., Piña, B., Tauler, R. Metabolomics.2015, DOI 10.1007/s11306-015-0812-9.

[9] Tauler, R., Chemometr Intell Lab 1995, 30, 133-146.

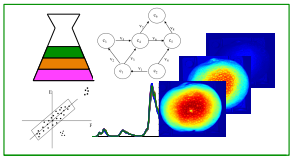
[10] Trygg, J., Holmes, E., Lundstedt, T., J Proteome Res., 2007, 6(2), 469-479.

[11] Tautenhahn, R., Böttcher, C., Neumann. S., BMC Bioinformatics, 2008, 9, 504-520.

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

Agradecimientos

Este trabajo ha recibido financiación del Consejo Europeo de Investigación dentro del Marco del Séptimo Programa de la Unión Europea (FP / 2007-2013) / ERC Advanced Grant n. 320737.



Estudio sobre un pre-procesado alternativo de datos ómicos para análisis no dirigidos

Núria Dalmau, Carmen Bedia y Romà Tauler
Departamento de Química Ambiental (IDAEA-CSIC), Jordi Girona 18-26, 08034, Barcelona
nuria.dalmau@idaea.csic.es

Abstract

In this work an alternative procedure for the untargeted analysis of LC-MS data sets in omics field is presented. A preliminary pre-processing of data sets was based on the Regions of Interest (ROI) strategy. This methodology allows an important reduction of size without loss of resolution and accuracy on m/z measure and an easy data manipulation of mass data sets. This strategy is based on the search of significant mass traces regions with high mass densities. The adjustment of ROIs parameters has been performed by the use of different dilutions of standard stock solutions mixtures injected in UHPLC-ToF-MS instrument. The optimization of ROIs parameters resulted in new data matrices containing the same m/z accuracy as original data which contain all valuable information in a reduced size. These data matrices were subjected to multivariate curve resolution-alternating least squares (MCR-ALS) (Tauler 1995), which is a valid method for proper resolution of chromatographic profiles and mass spectra profiles with the same accuracy. Also, ROI parameters were optimized for the construction of augmented matrices from individual matrices containing information of different standard mixtures dilutions, in order to enable its further comparative analysis through MCR-ALS.

Altogether, the optimization of ROI pre-processing parameters described in this work enabled the reduction of consuming times in the untargeted analysis of large LC-MS datasets, without loss of information.

Keywords: *Regions of interest (ROI), multivariate curve resolution-alternating least squares (MCR-ALS), untargeted omic analysis.*

Resumen

En este trabajo se presenta un procedimiento alternativo de análisis no dirigido de datos LC-MS para el análisis de datos ómicos. Se trata de un pre-procesado preliminar de los datos a partir de la estrategia de las Regiones de Interés (ROI). Esta metodología permite una reducción importante del tamaño de los datos sin pérdida de resolución y exactitud en la medida de las m/z. Además, este pre-tratamiento permite una manipulación más sencilla de los conjuntos de datos de masas. Esta estrategia está basada en la búsqueda de regiones con densidades altas de masas con intensidades significativas. El ajuste de los parámetros de los ROIs se ha realizado a partir de los datos de mezclas de patrones a diferentes concentraciones conocidas, inyectadas en un instrumento UHPLC-ToF-MS. La optimización de los parámetros de los ROIs da como resultado una nueva matriz de datos con la misma exactitud m/z que los datos originales que contienen toda la información de interés en un tamaño mucho menor, lo que posibilita su análisis directo. Estas matrices de datos pueden entonces analizarse por el procedimiento de resolución multivariante de curvas por mínimos cuadrados alternados (MCR-ALS) (Tauler 1995), el cual permite la resolución correcta de los perfiles cromatográficos y sus respectivos espectros de masas con la máxima exactitud. También, se han optimizado los parámetros de los ROIs para la construcción de matrices aumentadas formadas a partir de matrices individuales de muestras de mezclas de patrones a diferentes concentraciones para su posterior análisis comparativo por MCR-ALS.

En resumen, la optimización de los parámetros para el pre-procesado ROI descrito en este trabajo posibilita, sin perder información, una reducción del tiempo de procesado de los datos para los análisis no dirigidos.

Palabras clave: *Regiones de interés (ROI), resolución multivariante de curvas por mínimos cuadrados alternados (MCR-ALS), análisis ómico no dirigido.*

Introducción

En los últimos años la cromatografía de líquidos acoplada a espectrometría de masas (LC-MS) ha evolucionado mucho permitiendo un gran avance en muchos campos de investigación, incluidas las disciplinas ómicas como la metabolómica y la lipidómica.

Su uso más común y destacado es la identificación directa de compuestos (aproximación dirigida), posible para los analizadores de masas exactas como el tiempo de vuelo (ToF) o el Orbitrap. Estos permiten una elevada resolución de picos cromatográficos y la posibilidad de realizar determinaciones a bajas concentraciones de muestra. Por otro lado, el análisis no dirigido de compuestos permite una interpretación más global de los perfiles ómicos. Esta aproximación no dirigida evita la focalización en unas especies determinadas y abre la puerta a la interpretación general de los datos ómicos sin seguir hipótesis previas que puedan hacer perder información.

Los archivos de datos masivos (*big data*) obtenidos generan varias dificultades para su análisis. En primer lugar, se trata de archivos que requieren mucha memoria de almacenamiento en el ordenador, lo que afecta a la velocidad general de análisis. Además, en algunos casos existen dificultades para exportar los datos a otras plataformas de trabajo que no sean las propias del equipo. Por último, se debe conocer cómo mide y adquiere los datos el instrumento, su precisión y su exactitud; parámetros importantes para la identificación posterior de compuestos.

En nuestro grupo de investigación, el estudio de datos masivos obtenidos por LC-MS o por GC-LC y RMN, se ha basado en el método de resolución multivariante de curvas por mínimos cuadrados alternados (MCR-ALS). Este método permite la resolución de los compuestos en este tipo de datos sin la necesidad de alinear u ordenar los picos, como se ha demostrado en estudios anteriores (Farrés et al. 2015, Bedia et al. 2015, Gorrochategui et al. 2015). Debido a las enormes dimensiones de las matrices a analizar y a la naturaleza multivariante de los datos, la aplicación de MCR-ALS se había realizado mediante el análisis por separado de las submatrices de la matriz original divididas por ventanas de tiempo y con compresión (*binning*) de las unidades m/z. Este hecho suponía una pérdida de exactitud de *m/z*, un procedimiento laborioso de preselección de ventanas y un tiempo de análisis computacional muy largo. En este trabajo se presenta una posible mejora en el preprocesado de los datos que permite reducir la dimensión de las matrices sin pérdida de exactitud de *m/z* y agilizar así su análisis por MCR-ALS. Esta estrategia, llamada Regiones de Interés (ROI), está basada en la búsqueda de las intensidades de masas más relevantes que constituyen un pico cromatográfico entre los datos LC-MS obtenidos, lo que permite una mayor compresión de los datos originales sin perder la exactitud de las medidas de *m/z* en los espectros de masas. Los ROIs son las regiones con mayor densidad de puntos conse-

cutivos con una intensidad de masas superior al límite de ruido establecido según el tipo de muestra y de equipo, como ejemplifica la Figura 1.

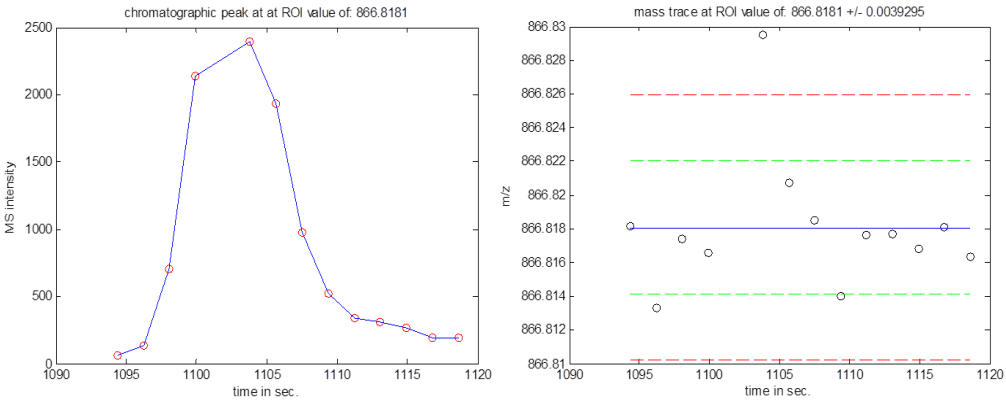


Figura 1. Perfil de elución obtenido siguiendo el pre-procesado de Regiones de Interés (ROI) y sus correspondientes masas con intesidades significativas.

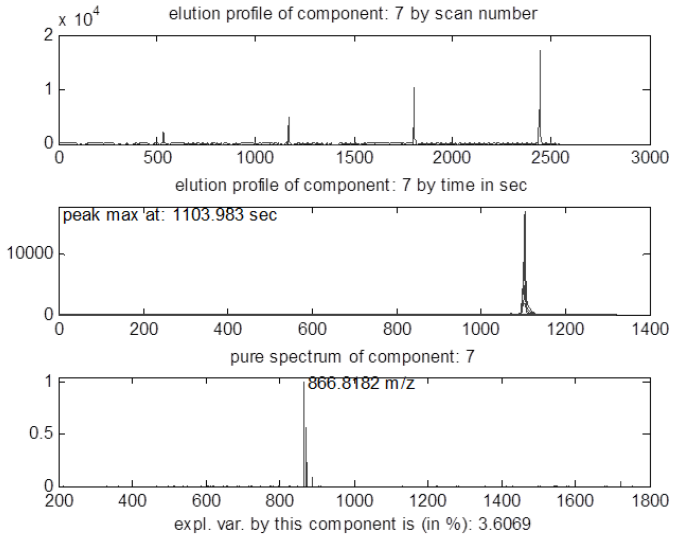


Figura 2. Componente 7 obtenido del análisis MCR-ALS de la matriz aumentada (2.5, 5, 10 y 20 ppm) dónde se muestran el perfil de elución anterior para cada muestra de la matriz. También se representan los perfiles de elución y de masas del componente en segundos y *m/z* respectiva-mente.

En este trabajo se muestra la optimización de estos parámetros de ROI utilizando datos LC-MS de mezclas de soluciones de patrones a diferentes concentraciones. Además se ha estudiado la formación de matrices aumentadas y posterior análisis con el método MCR-ALS para obtener los componentes más relevantes (Figura 2).

Referencias

[1] Bedia, C., et al., Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors. Environ Res, 2015. 140: p. 18-31.

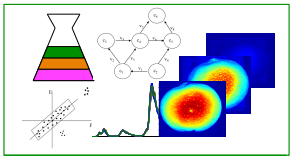
[2] Farres, M., et al., Chemometric evaluation of metabolic profiles using LC-MS. Metabolomics, 2015. 11: p. 210-224.

[3] Gorrochategui, E., et al., Chemometric strategy for untargeted lipidomics: biomarker detection and identification in stressed human placental cells. Anal Chim Acta, 2015. 854: p. 20-33.

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

Agradecimientos

Este trabajo ha recibido financiamiento del Consejo Europeo dentro del marco del séptimo programa de la Unión Europea (FP/2007-2013) / ERC Advanced Grant n. 320737.



A program for variable recognition from liquid chromatograms with diode array detection and application to protein classification

Clara Burgos-Simón[a], Enrique J. Carrasco-Correa[b], Miriam Beneito-Cambra[b], Guillermo Ramis-Ramos[b]

[a]Department of Applied Mathematics, University of Valencia, 46100 Burjassot, Spain, claburssi@alumni.uv.es, [b]Department of Analytical Chemistry, University of Valencia, 46100 Burjassot, Spain, enrique.carrasco@uv.es, miriam.beneito@uv.es, guillermo.ramis@uv.es

Abstract

A recurrent issue in Chemometrics is the recognition of variables in complex records, such as those obtained when natural products are chromatographed. This problem has been addressed by methods mainly based on peak alignment. Protein digestion with trypsin leads to complex mixtures of peptides and therefore also to complex chromatograms. In this work, a classification method for enzymes was developed. For this purpose, the enzymes were digested with trypsin, followed by liquid chromatography with spectrophotometric detection at several wavelengths. The training set was constructed with enzymes of three classes, belonging to industrial enzymes that are usual components of cleaners. The resulting complex chromatograms were treated by a program which was based on six identifiers (indices of peak identity), and on multiple comparisons at the local level between the chromatograms of the training set. The final model, constituted by the pooled chromatograms of the classes, and whose elements were vectors containing the six peak identifiers, was used to classify the enzymes of the training set (cross-validation by leave-one-out).

Keywords: *Peak recognition, fingerprinting, liquid chromatography, diode array detection, protein analysis, trypsin digests, classificatory analysis*

Resumen

Un problema recurrente en Quimiometría es el reconocimiento de variables en registros complejos, tales como los que se obtienen cuando se cromatografían productos naturales. Este problema se ha abordado mediante métodos basados principalmente en el alineamiento de los picos. La digestion de

proteínas con tripsina da lugar a complejas mezclas de péptidos, y por tanto, también se tienen cromatogramas complejos. En este trabajo se ha desarrollado un método de clasificación de enzimas basado en su digestión con tripsina, seguida de cromatografía líquida con detección espectrofotométrica a varias longitudes de onda. El conjunto de entrenamiento se construyó con enzimas industriales de tres clases, utilizadas como componentes habituales de productos de limpieza. Los complejos cromatogramas resultantes se trataron mediante un programa basado en seis índices de identidad o identificadores de pico, y mediante comparaciones múltiples a nivel local de los cromatogramas del conjunto de entrenamiento. El modelo final, constituido por los cromatogramas conjuntos de las clases, y cuyos elementos son vectores conteniendo los seis identificadores de pico, se utilizó para clasificar las enzimas del conjunto de entrenamiento (validación cruzada mediante leave-one out).

Palabras clave: *Reconocimiento de picos, huella dactilar, cromatografía líquida, detector de fila de diodos, análisis de proteínas, digestos de tripsina, análisis clasificatorio*

Introduction

In the detergent industry, the identification and quantification of enzymes is frequently demanded. This information is important for the quality control of some raw materials and manufactured products. In previous work, two methods capable of identifying the type of enzyme present in cleaners were proposed. These methods were based on total protein hydrolysis, followed by either direct infusion of the hydrolysate into a mass spectrometer (MS, Beneito-Cambra, 2008) or its injection in a liquid chromatograph with UV-Vis detection (Beneito-Cambra, 2009). Hydrolysis provides the amino acid profile of the samples, which can be used as a fingerprint to establish the enzyme class.

However, certain proteases are capable of cutting the proteins by specific peptide bonds, yielding complex mixtures of peptides, much richer in information than the aminoacid profiles obtained by total hydrolysis of proteins. For this purpose, trypsin, an endoprotease, is widely used. Trypsin mainly cleaves proteins at the carboxyl side of the amino acids lysine and arginine. After digestion with trypsin, the application of a suitable analytical technique, followed by data treatment of the resulting peptide profile, should made possible the classification of the original protein. Using calibration standards, information leading to the identification of the enzyme would be also provided. High-performance liquid chromatography (HPLC) and mass spectrometry (MS) are most frequently used to analyze the complex peptide mixtures resulting from trypsin digestion. However, due to the high price

of MS detectors and HPLC-MS interfaces, HPLC with UV-Vis spectrophotometric detectors are much more commonly found in industrial laboratories for quality control.

In this work, a method capable of classifying the enzymes present in household and industrial cleaners and related raw materials is presented. Enzymes are precipitated with acetone, and the precipitate is digested with trypsin. Next, a chromatogram is obtained in reversed phase mode (HPLC-RP) using a UV-Vis diode array detector. Data treatment of the chromatogram is then used to retrieve the relevant information concerning the enzyme class. For this purpose, a program written in MatLab was used. Essential parts of the program were the algorithms capable of recognizing characteristic peaks of each enzyme class, in a process leading to establish a model of each class. These models were used to classify the enzymes of the training set (with leave-one-out cross-validation).

Materials and methods

The training set was constructed with three classes of enzymes typically found in the detergent industry: proteases, amylases and cellulases. For each class, between 6 and 8 industrial enzyme concentrates from Novozymes (Bagsvaerd, Denmark), Biocon (Bangalore, India), ChemWorld (Barcelona), Enmex (Tlalnepantla, Mexico) and Genencor (Rochester, NY, USA) were collected and treated. The enzymes, initially in aqueous solution, were precipitated with acetone. The precipitates were redissolved in water and treated overnight with trypsin at 37 ° C. Chromatograms of the digests were obtained with a Kinetex column (core-shell 2.6 microns, C18, 100 A, 100x3 mm, Phenomenex, USA), using a water-acetonitrile gradient in the presence of 0.1% trifluoroacetic acid, at 25 ° C and at a flow rate 0.4 mL/min. The chromatograms were recorded at three wavelengths: 214, 260 and 280 nm. The program for data treatment was written in MatLab vs. 7.12.0 (R2011a).

Results and discussion

Within the useful time region of ca. 60 min, the chromatograms contained about 72,000 points each. The program contained the necessary routines to establish the baseline, locate peaks, establish their limits and location of the maximum on the time scale, and complete the peak description by measuring up to six parameters per peak (peak “identifiers”, see below). The derivative of the chromatogram, with a working window of 90 points, was used to locate peaks and establishing their limits. Then, a couple of blank chromatograms (obtained with trypsin and without enzymes) were used to establish the trypsin peaks. The presence of a large and intense trypsin peak at the beginning of the chromatogram, and a characteristic group of less intense peaks at the end, largely facilitated this task. These peaks, common to all samples, were also used for alignment purposes. This was made by

establishing a reduced retention time, beginning at the first trypsin peak and finishing at the last one.

After discarding the trypsin peaks, further data treatment was performed with the other peaks which contained the required information. Six indexes or "identifiers" containing information about the identity of each one of these peaks were established, namely the peak reduced retention time, peak relative area (percentage of total peak area for the recording at 214 nm), reduced width of the peak base, peak area ratio at different wavelengths (concretely, area ratios for 260/214 nm and 280/214 nm), and asymmetry factor. To measure the asymmetry factor, A1 was taken as the area from the left end of the peak to the position of the maximum, and A2 as the area from the maximum to the right end of the peak; finally, the asymmetry factor was calculated as A1/A2. For partially overlapped peaks, the asymmetry factor was calculated for the peak area which surpassed the merging point with the adjacent peaks.

For each chromatogram, and using the six peak identifiers, a matrix constituted by six rows and as many columns as identified peaks (typically between 70 and 150 peaks per chromatogram, this number mainly depending on the enzyme class) was then created. Next, the matrices of chromatograms belonging to a same enzyme class were compared in order to identify characteristic peaks of each class. For this purpose, the rows of the matrices were first normalized. In this way, the six peak identifiers were made equally important among them. However, after normalization, weigths were assigned to the identifiers. The weigths were established according to the relative relevance of each identifier, e.g. more weigth was given to the reduced retention time than to other identifiers; however, attemps to optimize the weigths according to any quantitative criterion were not made.

Then, the chromatograms were compared by pairs, taking all possible pairs within each class. For each pair, a pooled chromatogram was created. For instance, to compare two chromatograms, A and B, each peak of A was first compared with all the peaks of B located within a reduced time window centered around the peak of A. The size of the window used was 5% of the reduced time total range. For each pair of peaks of A and B, the six peak identifiers were compared. For each peak identifier, a score was given to the candidate peak of B. High or low scores were given according to the similarity between the identifiers of the peaks of A and B. For each peak of A, the peak of B with the highest sum of scores, also meeting the condition of overcoming 80% of the maximal possible sum of scores (e.g. 20 points out of a maximum of 25), was provisionally chosen as a characteristic peak of the class.

Then, the same system of comparison of peaks by pairs using scores was applied to each one of the peaks of B. As indicated above, each one of them was compared to the peaks of A located within a window of A equal to the 5% of the reduced time total range. The peaks

finally accepted as characteristics of the class were those recognized as the same peak in both comparisons, A with B and B with A. The pooled chromatogram M_{AB}, containing the accepted peaks and the average values of the six peak identifiers, was then created. It should be noted that the pooled chromatogram M_{AB} was a matrix containing six rows (the identifiers, with average values of the paired peaks of A and B) and as many columns as accepted peaks.

After treating all the possible pairs of chromatograms within a class, the resulting pooled chromatograms, M_{AB}, M_{AC}, M_{BC}, etc., were further processed as follows. First, the pooled chromatogram having the largest number of accepted peaks (e.g. M_i) was successively compared to the other pooled chromatograms of the class (M_j being $j \neq i$). This comparison was made again using scores, following a procedure similar to that previously used to compare the original chromatograms of the standards. In this way, new pooled chromatograms which were representative of sets of four original ones were obtained. This process was repeated to obtain pooled chromatograms representing a larger number of original chromatograms, until a single pooled chromatogram for each enzyme class was obtained. It should be noted that each chromatogram was in fact a matrix having six rows (the six average values of the peak identifiers) and as many columns as peaks have been accepted for the class.

This process of comparisons by pairs of chromatograms and successive clustering of the pooled chromatograms had a shortcoming. Some peaks which were truly representatives of the class were excluded when they were not recognized as such in at least one of the comparisons. To reduce this limitation, the pooled chromatogram representing the N chromatograms of a class was compared with the N pooled chromatograms representing N – 1 chromatograms of the same class. This comparison was made reciprocally, but with some differences with respect to the procedure explained above. First, the distances between the six peak identifiers were used to recognize peaks as the same one; second, peaks which were not recognized as the same one were not excluded, but added to the final pooled chromatogram of the class.

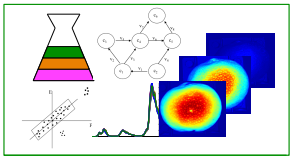
The pooled chromatogram of each class differed from that of the other classes in the number and location of the peaks. Therefore, an obvious classification method resulted from the comparison of the chromatogram of any new sample with the pooled chromatograms of the classes. For this purpose, cross validation (leave-one-out) was used. Thus, after excluding a chromatogram, the pooled chromatograms of the classes were obtained again. These were compared with the excluded chromatogram. Correct predictions were made in all cases by simply using the number of recognized peaks as the predictor. A 100% of correct predictions was also achieved by using Euclidean distances in the space of the recognized variables.

References

[1] Beneito-Cambra M., Herrero-Martínez J.M., Simó-Alfonso E.F., Ramis-Ramos, G. *Rapid classification of enzymes in cleaning products by hydrolysis, mass spectrometry and linear discriminant analysis*. Rapid Commun. Mass Spectrom. 2008; 22: 3667–3672.

[2] M. Beneito-Cambra, V. Bernabé-Zafón, J.M. Herrero-Martínez, E.F. Simó-Alfonso, G. Ramis-Ramos, *Enzyme class identification in cleaning products by hydrolysis followed by derivatization with o-phthaldialdehyde, HPLC and linear discriminant analysis*,Talanta 2009; 79: 275–279.

Acknowledgements: Project CTQ2014-52765-R (MEC of Spain and FEDER funds). E.J.C-C. thanks the MINECO of Spain for an FPI grant.



Identifying and retrieving common and distinctive components underlying two object-wise linked data blocks

Raffaele Vitale[a], Johan A. Westerhuis[b], Onno E. de Noord[c], Age K. Smilde[b], Alberto Ferrer[a]

[a]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain, [b]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, 1018 WV Amsterdam, The Netherlands and [c]Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN Amsterdam, The Netherlands.

Abstract

A novel method for disentangling common and distinctive sources of variability, which underlie two sets of data sharing the object dimension, is proposed.

Keywords: *singular value decomposition, canonical correlation analysis, permutation tests, multiset data analysis, common and distinctive components.*

Resumen

En este trabajo, se propone un nuevo método de análisis multivariante simultáneo de dos conjuntos de datos, que comparten el mismo número de observaciones, para discriminar de manera directa la variabilidad común y específica de cada uno de ellos.

Palabras clave: *descomposición en valores singulares, análisis de la correlación canónica, pruebas de permutación, análisis simultáneo de distintos conjuntos de datos, variabilidad común y específica*

Introduction

In many research and practical domains, it has recently become quite frequent to exploit multiple analytical platforms to comprehensively study the same system of interest. In these cases, an intriguing and challenging task is to distinguish the common and distinctive sources of variability (or *components*) underlying the various blocks of data, resulting from

the application of the different characterisation techniques and sharing the same number of objects (thus defined *object-wise linked*). Here, a novel method to achieve such disentanglement when two sets of measurements are dealt with, namely $\mathbf{X}_1$ ($N \times J_1$) and $\mathbf{X}_2$ ($N \times J_2$), is proposed.

Modelling strategy

The developed approach comprises five different steps:

1. First, the total number of components underlying $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, is determined by a Singular Value Decomposition (SVD)-based permutation test;
2. The number of common components shared by $\mathbf{X}_1$ and $\mathbf{X}_2$, A_{com}, is identified resorting to a Canonical Correlation Analysis (CCA)-based procedure;
3. The profile of the common components is retrieved by applying SVD to the between-block covariance matrix $\mathbf{X}_1^T\mathbf{X}_2$ and retaining the eigenvectors associated to the first A_{com} eigenvalues:

$$\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{U}_{com}\mathbf{S}_{com}\mathbf{V}_{com}^T + \mathbf{E}_{com} \quad (1)$$

4. The common sources of variability are filtered out by deflation:

$$\mathbf{X}_{1,dis} = \mathbf{X}_1 - \mathbf{X}_1\mathbf{U}_{com}\mathbf{U}_{com}^T \quad (2)$$

$$\mathbf{X}_{2,dis} = \mathbf{X}_2 - \mathbf{X}_2\mathbf{V}_{com}\mathbf{V}_{com}^T \quad (3)$$

5. The distinctive components of $\mathbf{X}_1$ and $\mathbf{X}_2$ are extracted by performing SVD on $\mathbf{X}_{1,dis}$ and $\mathbf{X}_{2,dis}$ and retaining the eigenvectors associated to the first $A_{1,dis}$ and $A_{2,dis}$ eigenvalues[1]:

$$\mathbf{X}_{1,dis} = \mathbf{U}_{1,dis}\mathbf{S}_{1,dis}\mathbf{V}_{1,dis}^T + \mathbf{E}_{1,dis} \quad (4)$$

$$\mathbf{X}_{2,dis} = \mathbf{U}_{2,dis}\mathbf{S}_{2,dis}\mathbf{V}_{2,dis}^T + \mathbf{E}_{2,dis} \quad (5)$$

The potential of this method will be assessed in simulated and real case-studies and its possible extension for the analysis of data blocks sharing the variable dimension will be discussed.

References

[1] Cherry S. (1997). *Some Comments on Singular Value Decomposition Analysis*. J. Climate, 10, 1759-1761.

[2] DelSole T., Tippett M.K. (2014). *Canonical correlation analysis.* Available online at ftp://cola.gmu.edu/pub/delsole/clim763/chapters/week3_cca/ch.cca2.pdf

1 The number of distinctive components underlying $\mathbf{X}_1$ and $\mathbf{X}_2$, $A_{1,dis}$ and $A_{2,dis}$, can be determined by executing the same SVD-based permutation test, as in step 1, on $\mathbf{X}_{1,dis}$ and $\mathbf{X}_{2,dis}$, respectively.

[3] Dobrić V. (1986). *On a class of robust methods for multivariate data analysis.* COMPSTAT, 211-216

[4] Endrizzi I., Gasperi F., Rødbotten M., Næs T. (2014). *Interpretation, validation and segmentation of preference mapping models.* Food Qual. Prefer., 32, 198-209.

[5] Hardoon D., Szedmak R., Shawe-Taylor J. (2004). *Canonical correlation analysis : An overview with application to learning methods.* Neural Comput., 16, 2639-2664.

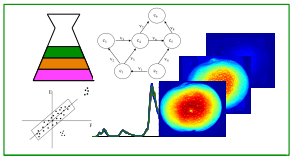
[6] Lock E.F., Hoadley K.A., Marron J.S., Nobel A.B. (2013). *Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types.* Ann. Appl. Stat., 7, 523-542

[7] Löfsted T. (2012). *Orthogonal Projections to Latent Structures in Multiblock and Path Model Data Analysis*. Ph.D. thesis, Department of Chemistry, Umeä University, Sweden.

[8] Momirović K., Radaković J., Dobrić V. (1988). *An expert system for the interpretation of results of canonical covariance analysis.* COMPSTAT, 135-141.

[9] Press W.H. (2011). *Canonical correlation clarified by Singular Value Decomposition.* Available online at http://www.nr.com/whp/notes/CanonCorrBySVD.pdf.

[10] Van Deun K., Smilde A.K., Thorrez L., Kiers H.A.L., Van Mechelen I. (2013). *Identifying common and distinctive processes underlying multiset data.* Chemometr. Intell. Lab., 129, 40-51.



Dynamic elementary modes modeling of non-steady state flux data

Abel Folch-Fortuny[a], Henk A. L. Kiers[b], Huub C. J. Hoefsloot[c,d], Age K. Smilde[c,d], Alberto Ferrer[a]

[a]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, Spain. [b]Heymans Institute for Psychology, University of Groningen, Groningen, The Netherlands. [c]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands. [d]Netherlands Metabolomics Center, Leiden, The Netherlands.

Abstract

New methods are presented here aiming at decomposing non-steady state metabolic flux distributions into a meaningful set of dynamic elementary modes and its biological activation for data compression and discrimination purposes.

Keywords: *elementary mode, metabolic network, fluxes, concentrations, (non-)steady state.*

Background

Principal component analysis (PCA) and multivariate curve resolution (MCR) models have been proposed to obtain a set of key pathways in metabolic networks, assuming steady state conditions (González-Martínez *et al.* 2014, Folch-Fortuny *et al.* 2015). These pathways or modules in the network are identified using the existing relationships between fluxes, measured experimentally. Recently, a new method called principal elementary modes analysis (PEMA) (Folch-Fortuny *et al.* submitted) has been proposed to model this kind of data. The methodology is based on the projection of the fluxes into a reduced set of elementary modes (EMs) of the metabolic network. The EMs are the simplest representations of pathways crossing the metabolic network. Basically, each EM connects substrates with end-products concatenating reactions in a thermodynamically feasible way.

For non-steady state conditions, *e.g.* when measuring the concentrations of the metabolites at early stages after perturbation, different methodologies have been proposed, such as kinetic modeling (Teusink *et al.* 2000), ^{13}C-metabolic flux analysis (MFA) (Wiechert 2001), dynamic flux balance anlysis (FBA) (Mahadevan *et al.* 2002), the Goeman's global

test (Hendrickx *et al.* 2012), and a recently proposed approach combining time-resolved metabolomics and dynamic FBA (MetDFBA) (Willemsen *et al.* 2015).

Here we define a new framework to model non-steady state metabolic fluxes. This methodology is based on adapting the PEMA model to work with deformed or dynamic EMs (dynEMs), *i.e.* EMs that are used partially at each time point. In this way we propose two methods: dynamic elementary modes analysis (dynEMA) to compress the data into a set of reduced explanatory dynEMs, and dynamic elementary modes regression discriminant analysis (dynEMR-DA) to identify the set of dynEMs whose activation pattern allows discriminating between different biological conditions.

Methods

PEMA

PEMA is described here for the sake of understanding the elementary modes modeling in steady and non-steady state cases. PEMA uses the set of EMs from a given metabolic network as the candidates of the principal components in a classical PCA.

From stoichiometric modeling, any stady state flux distribution $\mathbf{x} = (x_1, \ldots, x_K)$ can be decomposed as a positive linear combination of EMs (Llaneras and Picó 2008):

$$\mathbf{x} = \sum_{e=1}^{E} \lambda_e \cdot \mathbf{p}_e$$

where K is the number of fluxes (matching the number of reactions in the network), $\mathbf{p}_e = (p_{e1}, \ldots, p_{eK})$ is the EM e, λ_e is the positive weighting factor of EM e, and E is the number of EMs needed to reconstruct the flux distribution $\mathbf{x}$.

When N flux distributions are considered, a PEMA model can be built as follows:

$$\mathbf{X} = \mathbf{\Lambda} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{F}$$

where $\mathbf{X}$ is the $N \times K$ flux data matrix, $\mathbf{P}$ is the $K \times E$ principal elementary modes matrix, formed by a subset of E EMs; $\mathbf{\Lambda}$ is the $N \times E$ weightings matrix; and $\mathbf{F}$ is the $N \times K$ residual matrix. It is worth noting that the values in $\mathbf{\Lambda}$ are forced to be non-negative.

In the PEMA algorithm, the principal EMs are chosen from the complete set of EMs in a step-wise fashion, including at each step the EM explaining more variance in flux data. The weightings associated to the principal EMs are obtained by solving:

$$\mathbf{\Lambda} = \mathbf{X} \cdot \mathbf{P} \cdot (\mathbf{P}^{\mathrm{T}} \cdot \mathbf{P})^{-1}$$

Unlike the loadings in PCA, the principal EMs are not orthonormal, so the previous equation usually requires the computation of the pseudo-inverse of $\mathbf{P}^{\mathrm{T}} \cdot \mathbf{P}$.

dynEMA

Non-steady state flux distributions cannot be decomposed as linear combinations of elementary modes, as in steady state. However, the EMs are indeed the simplest pathways along which the non-steady state fluxes have to flow, but not in a stable or constant fashion. Following this rationale the EMs can be deformed to fit this unstability. This is what we call a dynamic elementary mode (dynEM). To deform an EM we simply have to assign not a single coefficient multiplying the EM (Λ in PEMA) but a coefficient to each reaction activated by the EM.

Thus, a single non-steady state flux distribution **x** can be decomposed as:

$$\mathbf{x} = \sum_{e=1}^{E} \boldsymbol{\alpha}_{\mathrm{e}} * \mathbf{p}_{\mathrm{e}}$$

where $\boldsymbol{\alpha}_{\mathrm{e}} = (\alpha_{e1}, \ldots, \alpha_{eK})$ are the coefficients that deform reactions 1 to *K* in the selected dynamic EM *e* to reproduce the fluxes in **x**, and * is the Hadamard element-wise product.

Let us consider now a set of non-steady state flux distributions, which are usually obtained from single experiment, measuring the concentration of the metabolites at different time points. The set of active dynEMs are obtained from the dynEMA model:

$$\mathbf{X} = (\mathbf{I}_{\mathrm{N}} \otimes \mathbf{1}_{\mathrm{E}}^{\mathrm{T}}) \cdot [\mathbf{A} * (\mathbf{1}_{\mathrm{N}} \otimes \mathbf{P}^{\mathrm{T}})] + \mathbf{F}$$

where **A** is the $EN \times K$ coefficients matrix, $\mathbf{I}_{\mathrm{N}}$ is the $N \times N$ identify matrix, $\mathbf{1}_{\mathrm{E}}$ is a column vector of *E* ones, and $\otimes$ is the kronecker product. The other matrices are the same as in the PEMA model.

The coefficients matrix **A** in the previous equation is indeed a $E \times K \times N$ three-way matrix unfolded variable-wise, and each entry in the matrix α_{ekn} represents the coefficient multiplying reaction *k* of EM *e* to reconstruct the flux x_k at time point *n*.

Using this modeling it is possible to study the time evolution of a dynEM, *i.e.* how the dynEM is deformed or dynamically used along all measured time points.

dynEMR-DA

When the aim is to establish differences between conditions, *e.g.* presence/absence of a compound or case/control studies, a discriminant model is needed. dynEMR-DA focuses on

finding which are the dynEMs with a strongly different time evolution or performance between conditions. For this, different experiments are combined in a single **X** matrix defined as multiset data, being the *K* fluxes the common mode.

The algorithm of dynEMR-DA is as follows:

1) For each EM:

1.1) Calculate the coefficients matrix **A** using dynEMA.

1.2) Reconstruct the flux data using **A** and **P**.

1.3) Fit a PLS model between the reconstructed data and the **Y** data.

2) The EM whose PLS model explains most variance in **Y** is classified as the first dynEM.

3) Check the predictions of the PLS model. If the current model discriminates perfectly, stop. If not, fix the first dynEM and repeat steps 1-3 to extract the second dynEM using, for more than one EM, a multiblock PLS model. And so on.

It is worth noting that a perfect discrimination using dynEMs is achieved when all time points of one condition are perfectly classified and *some* time points of the other condition are assigned to the other class. This is due to the dynamic activation of the EMs in non-steady state fluxes. The dynEM that discriminate between conditions may be activated not from the begining to the end of the culture but only at some time period in the middle, *e.g.* when a particular metabolite is produced/consumed.

Results

A dense in silico network is created to study the dynamic approaches presented here. Using dynEMA we are able to identify a few set of active dynEMs and study their evolution, identifying which reactions of the relevant EMs are used at each time point. To test the regression model, a small EM, artificially introduced in the network, and responsible of the changes in the discrete **Y** variable, is clearly identified by dynEMR-DA.

Actual non-steady state flux data from *Saccharomyces cerevisiae* is also tested using this methodology to differentiate between aerobic/anaerobic conditions, usage/not usage of glucose, and high/low usage of glucose.

Conclusion

The framework proposed here permits decomposing non-steady state flux distributions into a set of active dynEMs. These techniques allows to create reduced dynamic models of flux

data while preserving biological and thermodynamical meaning. dynEMA and dynEMR-DA have potential applications in bioprocess engineering to understand the small changes in cell metabolism at early stages of the culture.

References

[1] Folch-Fortuny A., Tortajada M., Prats-Montalbán J.M., Llaneras F., Picó J., Ferrer A., MCR-ALS on metabolic networks: Obtaining more meaningful pathways, *Chemometrics and Intelligent Laboratory Systems* 142 (2015) 293-303

[2] Folch-Fortuny A., Marques R., Isidro I.A., Oliveira R., Ferrer A., Principal Elementary Modes Analysis (PEMA), *submitted*

[3] González-Martínez J.M., Folch-Fortuny A., Llaneras F., Tortajada M., Picó J., Ferrer A., Metabolic flux understanding of Pichia pastoris cultures on heterogeneous culture media, *Chemometrics and Intelligent Laboratory Systems* 134 (2014) 89-99

[4] Hendrickx D.M., Hoefsloot H.C.J., Hendriks M.M.W.B., Canelas A.B., Smilde A.K., Global test for metabolic pathway differences between conditions, *Analytica Chimica Acta* 719 (2012) 8-15

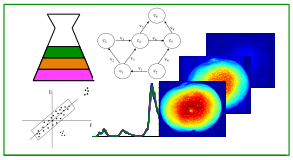
[5] Llaneras F., Picó J., Stoichiometric Modelling of Cell Metabolism, *Journal of Bioscience and Bioengineering* 105(1) (2008) 1-11

[6] Mahadevan R., Edwards J.S., Doyle F.J. 3rd, Dynamic flux balance analysis of diauxic growth in Escherichia coli, *Biophysical Journal* 83 (2002) 1331-1340

[7] Teusink B., Passarge J., Reijenga A., Esgalhado E., ven der Weijden C.C., Schepper M., Walsh M.C., Bakker B.M., van Dam J.C., Westerhoff H.V., Snoep J.L., Can yeast glycolysis be understood in terms of in vivo kinetics of the constituent enzymes? Testing biochemistry, *European Journal of Biochemistry* 267(17) (2000) 5313-5329

[8] Wiechert W., 13C metabolic flux analysis, *Metabolic Engineering* 3(3) (2001) 195-206

[9] Willemsen A.M., Hendrickx D.M., Hoefsloot H.C.J., Hendriks M.M.W.B., Wahl S.A., Teusink B., Smilde A.K., van Kampen A.H.C., MetDFBA: incorporating time-resolved metabolomics measurements into dynamic flux balance analysis, *Molecular Biosystems* 11 (2015) 137-145



Combined LASSO-N-PLS for variable selection in metabolomic data

David Hervás[1], Agustín Lahoz[2], Alberto Ferrer[3] and José M. Prats-Montalbán[3]

[1]Unidad de Bioestadística, IIS La Fe, Valencia, Spain. [2]Unidad de Hepatología Experimental, IIS La Fe, Valencia, Spain. [3]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Valencia, Spain.

Abstract

A combination of N-PLS with LASSO penalization is presented here as a novel method for variable selection in N-way metabolomics data. Results of the method using different simulated data structures and also a real dataset are also presented..

Keywords: *variable selection, N-way data, LASSO, metabolomics.*

Background

N-PLS is a useful tool to analyze multi-way data, reducing the inclusion of noise in the models and obtaining more robust parameters and, at the same time, producing easy-to-understand plots. The LASSO (Tibshirani 1996) is a regularization method for linear regression which reduces variance by imposing a L1 penalty constrain to the least squares fit. This constrain shrinks the coefficients of the model, causing some of them to be exactly zero and thus performing variable selection at the same time.

In some situations supervised projection methods such as N-PLS can produce models with very low bias in p>>n settings, but at the cost of a high variance (Stoica *et al.* 1998). On the other hand, the variance reduction performed by LASSO comes at a cost of a noticeable increase in bias (Hastie *et al.* 2009). Squared error is a function of bias, variance and irreducible error:

$$Squared\ error\ =\ Irreducible\ Error\ +\ Bias2\ +\ Variance$$

This relation between bias and variance is also known as the bias-variance tradeoff, and is depicted in **Figure 1.** Since both low bias and high variance models and high bias and low variance models can actually be very inaccurate, we propose introducing L1 penalization in N-PLS as a way to obtain a robust variable selection method which has the flexibility to smoothly adjust its bias-variance tradeoff by changing the amount of L1 penalization imposed on the model.

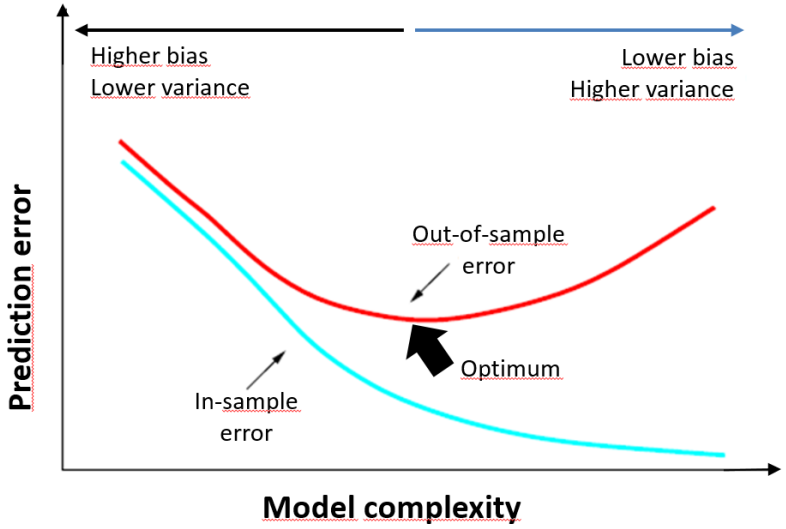


Figure 1: Bias-variance tradeoff. High bias models show high prediction errors due to poor fitting to the sample data. High variance models show high prediction errors due to overfitting to the sample data and failing to generalize to new data.

Methods

The Lasso (Least Absolute Shrinkage and Selection Operator) is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients (L1 penalization). It shrinks some coefficient and sets others to 0, and hence tries to retain the good features of both subset selection (interpretation) and ridge regression (stability and precision in estimations). The original LASSO for least squares is as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Subject to the restriction (penalization): $\sum_{j=1}^{p}|\beta_j| \leq s$

Increasing the penalization by reducing *s* forces the parameters to zero, producing a simpler model by deselecting some features. Thus, assuming data are standardized, Lasso automatically selects the most relevant features and discards the others.

To introduce the L1 penalization in the N-PLS algorithm we make use of soft-thresholding which can be derived as a solution of the LASSO lagrangian form:

$$\hat{\beta}_i^{\text{lasso}} = \text{sgn}(\hat{\beta}_i^{\text{LS}})(|\hat{\beta}_i^{\text{LS}}| - \lambda)^+$$

Our final LASSO-N-PLS algorithm is based on the sparse-PLS algorithm (Le Khao *et al. 2008)* as follows:

Center X and Y, and unfold X (and Y when necessary) into a two-way matrix.

Let u be some column of Y, and set f=1

1. $w^T=u^TX/u^Tu$
2. Build **Z** by refolding **w** according to the modes dimensions
3. Determine $\mathbf{w}^J$ y $\mathbf{w}^K$ by *SVD*
4. *LASSO inclusion*
    a. Apply soft-thresholding on $\mathbf{w}^J$: $f(y) = \text{sgn}(y)(|y| - \lambda_j)^+$
    b. Apply soft-thresholding on $\mathbf{w}^K$: $f(y) = \text{sgn}(y)(|y| - \lambda_k)^+$
    c. Input the new **w** as kron($\mathbf{w}^K$, $\mathbf{w}^J$)
5. $t=Xw/w^Tw$
6. $q=Y^Tt/\text{norm}(Y^Tt)$
7. $u=Yq$
8. Check for convergence. If it is achieved, continue; otherwise, go to 1
9. $\mathbf{b} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{u}$; where $\mathbf{T}=[\mathbf{t}_1\ \mathbf{t}_2 \ldots\ \mathbf{t}_f]$
10. $\mathbf{X} = \mathbf{X}-\mathbf{t}\mathbf{w}^T$ and $\mathbf{Y} = \mathbf{Y}-\mathbf{T}\mathbf{b}\mathbf{q}^T$
11. $f = f+1$. Continue from step 1 until a good description of **Y**

Results

We have tested our implementation of the L1 penalized NPLS using simulation. Amount of penalization, λ, was determined by cross-validation. We compared the results provided by LASSO-N-PLS with the creation of random null distributions of VIP's and weights, with posterior calculation of the statistical significance. Our method was also tested in a real dataset of liver regeneration metabolomics data. The method showed good ability for the reliable selection of those important variables comprised in large -omic data sets (**Figure 2**).

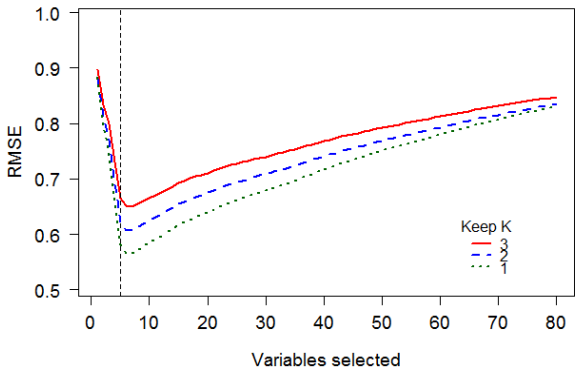


Figure 2: RMSE profile for different number of selected variables. In this specif case, lowest cross-validated RMSE was achived by selecting only 8 of the 80 variables in the data.

Conclusion

LASSO-N-PLS simplified data interpretation and variable selection, which is of utmost importance in the later development of targeted analysis focus on the determination of the biomarkers in a clinical scenario. In our real dataset, N-PLS combined with variable selection allowed us to select those metabolites that showed a higher association with liver regeneration.

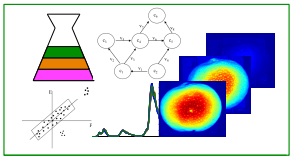
References

[1] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. New York: Springer (2009).

[2] Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, Philippe Besse. A Sparse PLS for Variable Selection when Integrating Omics Data. *Statistical Applications in Genetics and Molecular Biology. Volume 7, Issue 1* (2008)

[2] Stoica, P. and Söderström, T. Partial Least Squares: A First-order Analysis. *Scandinavian Journal of Statistics*, 25 (1998): 17–24. doi: 10.1111/1467-9469.00085

[4] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B., Vol. 58, No. 1 (1996), pages 267-288*.



Correction of intra-batch effects in UPLC-TOF-MS data using quality control samples and Support Vector Regression (QC-SVRC)

Julia Kuligowski[a], Ángel Sánchez-Illana[a], Daniel Sanjuan[b], Máximo Vento[a,c], Guilermo Quintás[b,e*]

[a]Neonatal Research Unit, Health Research Institute La Fe, Valencia, Spain, [b]Safety and Sustainability Division, Leitat Technological Center, Valencia, Spain, [c]Division of Neonatology, University & Polytechnic Hospital La Fe, Valencia, Spain, [e]Analytical Unit, Health Research Institute La Fe, Valencia, Spain

Abstract

Instrumental developments in sensitivity and selectivity boost the application of high resolution liquid chromatography – mass spectrometry (LC-MS) in metabolomics for biomedical research. Uncontrolled intra-batch effects in LC-MS are gradual changes in the instrumental response that reduce the repeatability and reproducibility of the analysis, decrease the power to detect biological responses and hinder the interpretation of the information provided, specially when a high number of samples are analyzed. Because of that, there is an interest in the development of chemometric techniques for the post-acquisition correction of the batch effect. In this work, the use of quality control samples and Support Vector Regression with a radial base function kernel (QC-SVRC) is proposed to correct intra-batch effects. The QC-SVRC method is compared to a recent reference algorithm based on robust cubic smoothing splines (QC-RSC). Initial results from the correction of data obtained from the repeated analysis of a plasma sample showed that QC-SVRC improved data quality and slightly outperformed QC-RSC.

Keywords: *Intra-batch effect, Support Vector Regression (SVR), high resolution liquid chromatography – mass spectrometry, Metabolomics.*

Introduction

High resolution liquid chromatography - mass spectrometry (LC-MS) is rapidly becoming the method of choice in metabolomics for biomedical research because of its increased

sensitivity, higher throughput and better metabolite coverage as compared to other techniques such as nuclear magnetic resonance (NMR) or hyphenated gas chromatography (GC-MS). Besides biological variability among subjects due to e.g. age, sex, medical conditions, drugs, food or the environment, high resolution LC-MS data includes unwanted instrumental variation. This variation can arise from e.g. minor changes in the injection volume, ideally as a normal white noise process with mean zero and constant variance, but also from e.g. gradual inlet interface contamination, drifts in detector sensitivity, temperature, ionization efficiency or column performance that modify the instrumental response (i.e. intra-batch effect). An accurate estimation of the variation in the instrumental response for each detected variable (metabolite) over the batch would allow an effective correction of the intra-batch effect and the shrinkage of the instrumental error. However, this estimation is troublesome as data is typically noisy. The intra-batch effect can be seen as a stationary or non-stationary process depending on the position within the batch (e.g. it can be negligible at the beginning and very significant at the end of the batch) and the size of the effect varies across metabolites. Because of that, the use of algorithms whith high generalization capabilities is required. The use of the response in pooled quality control (QC) samples dispersed evenly throughout the batch and Robust Splines fhas been recently proposed by Broadhurst et al. [1] or the fit and correction of the intra-batch effect (QC-RSC). Support Vector Regression (SVR) is a non-parametric and distribution free model developed by Vapnik that provides high generalization capabilities at a low estimation cost [2,3]. In the present study, SVR using a radial basis function kernel (RBF) was tested to model batch effects using data acquired from QC samples. The proposed QC-SVRC approach was evaluated and compared to the QC-RSC method using the repeated analysis of a single plasma sample as a model example.

Results

The use of QC-SVRC is evaluated using the repeated analysis (n=150) of a single plasma sample in a single batch by UPLC-ESI(+)-TOF-MS as a model example. This model example facilitated the evaluation of the correction accuracy.

For a training dataset $\{(x_{QC(i)}, y_{QC(i)}) | x_{QC(i)} \in R,\ y_{QC(i)} \in R, i = 1, \dots, n\}$ (n = number of QCs in the training set) the SVR initially maps the sample data into a high dimensional feature space by means of $\varphi(x)$. Then, a linear regression function is defined in the feature space as $\hat{y_i} = f(x_i, w) = w' \cdot \varphi(x_i) + b$, where w = weight vector and b = constant threshold. As in Supporting Vector Machines (SVM), optimization in SVR is solved by quadratic programming through the use of Lagrange multipliers and a kernel function $K(x_i \cdot x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ to replace the inner product operation in the high dimensional feature space. The solution is then expressed as:

$$\hat{y} = f(x) = \sum_{x_i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

where α_i, α_i^* are Lagrange multipliers, and the samples with non zero weights α_i, α_i^* are called the Support Vectors (SVs) used for the construction of the SVR function. In this study, the kernel function used was the Radial Basis Function (RBF): $K(x_i, x) = \exp(-\gamma \|x - x_i\|^2)$, where γ is the width of the RBF kernel. The accuracy of a RBF-SVR is determined by the ε-insensitive loss parameter, the error penalty parameter C and the RBF kernel parameter γ [3,4]. The ε-insensitive loss parameter is used to ignore training errors lower than a threshold value during model development and limit model overfitting. C is the cost associated with the training error. Large C values may lead to over- and small C values may lead to underfitting. Large RBF kernel γ values reduce the area of influence of the SVs and lead to model overfitting. To trade off training error against model complexity using cross-validation (CV), different strategies for the selection of the optimal set of C , γ and ε values can be found like e.g. grid search. However, this method is computationally intensive. Alternatively, the range of output values has been previously proposed to select the C parameter. However, to reduce the impact of outliers, the C value was defined as the difference between the 10^{th} and 90^{th} quartile of the output values in QC samples. The ε-insensitive loss parameter was defined, for each variable, as ±5% of the observed value in the first QC because the precision of the UPLC-MS system used in this work falls within the 10-15% range and so lower ε-insensitive value would overfit the model. Finally, the optimum kernel parameter γ was selected by LOO-CV in the $[2^{-6}, 2^{-5}, \ldots, 2^{6}]$ range. As an example, Figure 1(top) shows the drift in the intensity of an endogenous metabolite (tryptophan, $[M+H]^+$ $C_{11}H_{13}N_2O_2^+$, m/z = 205.0971, retention time = 3.1 min) across the batch measurement. The intra-batch effect was modelled using 18 evenly distributed samples as QC for the modeling of the intra-batch effect. Results from the QC-SVRC approach showed that a SVR function using 5 support vectors provided an accurate estimate of the intra-batch effect (see SVR-curve in Figure 1(top)). Figure 1(middle) shows mean square errors of leave-one-out cross validation as a function of γ, used for the selection of the optimal γ value and Figure 1(bottom) shows the corrected intensity of tryptophan. The generalization capabilities of the SVR function lead to an effective removal of the intra-batch effect. Residual variation around the mean value in 'non-QC' samples could be attributed to a poor injection volume precision and the automatic integration of the chromatographic peaks.

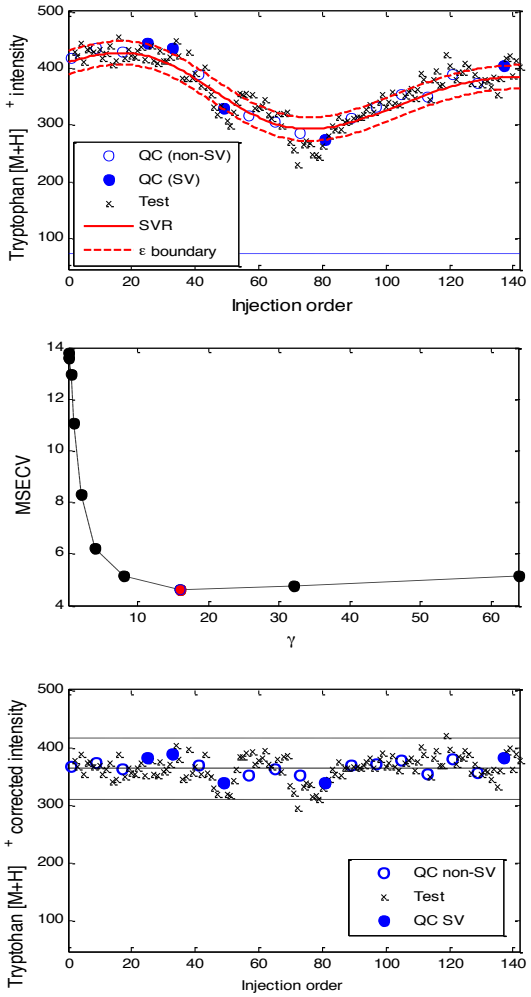


Figure 1. Top) Intensity of tryptophan monitored throughout the batch. The red line de-picts the calculated SVR function. The ε boundaries are given with the red dotted line; Middle) Mean Square Error of Cross Validation (MSECV) for QCs obtained as a function of the γ value. Red dot indicates the selected value; Bottom) Intensity of tryptophan after intra-batch effect correction.

The QC-SVRC correction approach was applied for each variable across the data set. The correction accuracy was evaluated using the number of variables showing an RSD in non-QC samples <15% as a figure of merit. Results depicted in Figure 2 showed that the precision of the analysis was greatly improved. Finally, results were compared to those provided by QC-RSC for the same data set. In this case, the accuracy of the QC-SVRC approach

slightly outperformed QC-RSC. However, it is difficult to estimate whether this difference is statistically significant or not. Further work will be carried out to evaluate the effect of the QC distribution within the batch on the correction accuracy using both, QC-RSC and QC-SVRC.

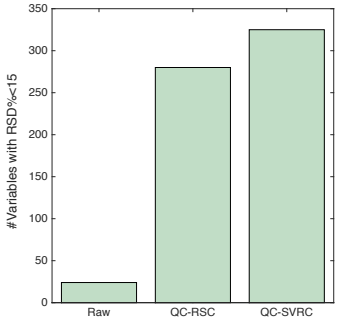


Figure 2. Number of variables showing $RSD_{samples}$<15% before and after correction of the intra-batch effect using QC-RSC and QC-SVRC. Note: Total number of variables in the data set = 552.

Acknowledgements. JK acknowledge the 'Sara Borrell' grant (CD12/00667) from the Instituto Carlos III (Ministry of Economy and Competitiveness, Spain). ASI acknowledges the support of the Red de Salud materno Infantil (SAMID). MV acknowledges the FIS-PI14/0433 grant from the Instituto Carlos III (Ministry of Economy and Competitiveness, Spain). GQ acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (SAF2012-39948).

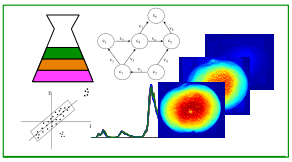
References

[1] S J.A. Kirwan, D.I. Broadhurst, R.L. Davidson, M.R. Viant, Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow, Anal. Bioanal. Chem. 405 (2013) 5147–5157.

[2] V. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. Publ. IEEE Neural Netw. Counc. 10 (1999) 988–999.

[3] V. Vapnik, S.E. Golowich, A. Smola, Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, in: Adv. Neural Inf. Process. Syst. 9, MIT Press, 1996: pp. 281–287.

[4] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.



Optimization of the extraction and conservation conditions of a green tea sample maximizing the information obtained in fingerprints

Tamara Álvarez-Segura[a], Elsa Cabo-Calvet[b], José Ramón Torres-Lapasió[c] and María Celia García-Álvarez-Coque[d]

Department of Analytical Chemistry, University of Valencia, C/Dr. Moliner 50. 46100-Burjassot-Valencia (Spain). [a]talse@alumni.uv.es, [b]elcacal@alumni.uv.es, [c]jrtorres@uv.es, [d]celia.garcia@uv.es

Abstract

The analysis of highly complex samples for which there are no standards available, such as medicinal herbs, is based on the comparison of chromatographic fingerprints. In this work, a strategy is reported for high-performance liquid chromatography to measure the level of information in fingerprints through the concept of peak prominence, which is the protruding part of each visible peak with regard to the valleys that delimit it. Next, the peaks in the fingerprints are ranked according to the areas of the peak prominences, and the number of peaks exceeding an established threshold are discriminated to differentiate between peaks corresponding to real (significant) compounds and those severely affected by non-significant components or noise. Plackett-Burman designs were applied to evaluate the impact of several extraction conditions on the number of significant peaks found in the fingerprints. The methodology was applied to green tea samples analyzed using acetonitrile, ethanol or methanol as extraction solvents, and a linear gradient where the acetonitrile content was raised from 5.0 to 42.5% (v/v) in 45 min. Maximal information in the fingerprints was obtained using methanol as extraction solvent, and high ultrasonication time and temperature.

Keywords: Medicinal herbs; Green Tea; Fingerprint analysis; Reversed-phase liquid chromatography; Extraction yield; Plackett-Burman designs; Number of significant peaks.

Introduction

The analysis of medicinal herbs is troublesome because of the amount of its components and the disparity of their concentrations, the difficulty in knowing the nature of the components, and the absence of standards in the market that allow for their qualitative and quantitative analysis. Chromatographic fingerprint analysis appears as a good alternative for the analytical control of such samples. Chromatographic techniques maximize the information content along the time domain, which is particularly valuable for analyzing fingerprints. To facilitate the recognition, chromatographic fingerprints containing a large number of peaks are desirable.

A main objective in these analyses is to achieve maximal resolution between the chromatographic peaks, which is achieved by adjusting the separation conditions. Also, the extraction protocol can influence the results. It was found necessary to investigate the effect of the extraction conditions on the quality of the information obtained with fingerprints. A wide literature survey highlighted the disparity of extraction conditions reported in the literature by different analysts to process samples of medicinal herbs for fingerprint analysis. Given the number of factors involved, the study in this work was conducted based on factorial designs. A commercial green tea was used as sample.

In previous work, a peak prominence approach was developed, based on the automatic measurement of the protruding part of the chromatographic peaks to characterize samples with unknown compounds. Peak prominences were shown as a useful tool to recognize which peaks are significant for quantifying the information in a chromatographic fingerprint, without the need of standards. In this work, a protocol is described to evaluate the extraction yield of the components of a medicinal herb that is translated into a greater number of significant peaks in chromatographic fingerprints.

In the working protocol, the baseline was first subtracted using smoothing cubic splines. Then, the chromatographic peaks were searched through the standard function of MATLAB “FINDPEAKS”. With the default parameters, this function detects a very high number of peaks (often associated with noise). However, this way of operation was found preferable for relying on the analyst the decision of the selection of the significant peaks, starting from comprehensive primary information.

After normalizing the peaks, software developed in our laboratory called "CHROMSCAN" was applied, which processes automatically, in significant information, the raw vector of maxima indexes given by FINDPEAKS. “CHROMSCAN” locates the optimal tangent points defining the limits for peak integration (those delimiting the two valleys at the sides of each prominence) (Figure 1). It also measures the peak areas and calculates the resolution level.

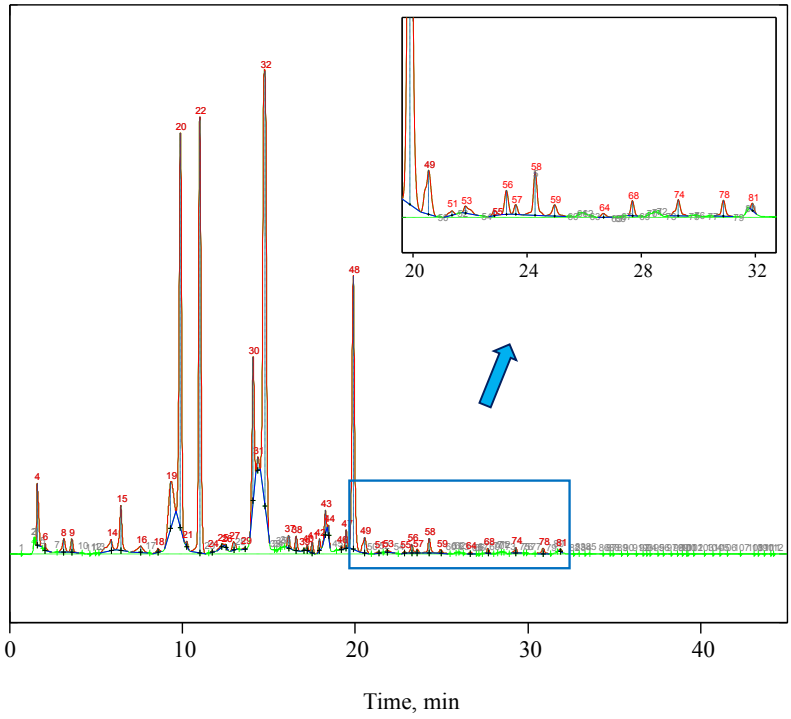


Figure 1. Screenshot of a MATLAB chromatogram corresponding to a green tea ex-tract. Tangents that define the prominence region for each peak are depicted. The peaks are numbered according to their elution order and those that exceeded a rela-tively peak area of 0.025% are marked in red.

The peaks in the fingerprints are ranked according to the areas of the peak prominences, and the peaks exceeding an established threshold are discriminated to differentiate between those corresponding to real (significant) compounds and those severely affected by non-significant components or noise. The peaks considered as significant are discriminated from those that can be assigned to noise assisted by a plot that represents the relative peak area of the chromatographic peaks for several replicated chromatograms, in descending order.

Seven experimental factors that affect the extraction of the samples were studied: type and concentration of the extraction solvent, sonication time in an ultrasonic bath, treatment temperature, sample weight, and conservation time and temperature of the extracts. Due to the high number of factors, factorial analysis using Plackett-Burman designs was applied. These designs have the advantage of allowing the exploration with a small number of experiments. For each experimental factor, a high and a low value were established.

The incompatibility among some of the involved factors did not allow a comprehensive study using designs that considered all factors simultaneously, in order to set the most favorable levels. Therefore, these were divided in two groups of three factors:

(i) First group: Concentration of the extraction solvent, and time and temperature of the treatment in the ultrasonic bath, which were considered altogether in a first step. This study was carried out in blocks for three types of hydro-organic mixtures with containing either acetonitrile, ethanol or methanol. Other blocks with pure water and the three pure organic solvents were also considered.

(ii) Second group: Amount of sample, and time and temperature of conservation of the extracts, which were studied in a second step.

The Plackett-Burman designs used for this study can be represented as a cube with 8 vertices, where each edge represents an experimental factor and each vertex an experiment. Only 4 of the 8 experiments were inspected. Despite this partial inspection, it was possible to establish models to predict how many peaks would be expected using experimental conditions non-assayed in the design, based on the data from the assayed experiments.

The proposed methodology to describe the information content of fingerprints, based on the measurement of peak prominences, showed conclusive results. The best solvent to carry out the extraction of the components in the analyzed sample was methanol, used in 30:70 mixtures with water at high temperature (80°C) and applying a long ultrasonication time (60 min).

The sample weight contributes very significantly to the number of peaks, and this was also larger when the conservation temperature increased (4°C against –10°C), probably due to the degradation of some compounds. In general, the analyses should be performed immediately after obtaining the extracts, due to the degradation of the extracts , which contributes to the number of peaks and should be avoided.

Although the described approach was developed using a green tea sample, it is suitable for finding the best extraction and conservation conditions for other types of medicinal herbs, in order to get fingerprints with the maximal information. Currently, an approach is being developed in our laboratory to optimize the gradient program to analyze fingerprints of medicinal herbs or from other sources, by applying the proposed chromatographic objective function based on the measurement of peak prominences.

References

[1] Alvarez-Segura T., Gómez-Díaz A., Ortiz-Bolsico C., Torres-Lapasió J.R., García-Alvarez-Coque M.C. (2015), A new chromatographic objective function to evaluate the information in fingerprinting analysis, *J. Chromatogr. A* 1409, 79–88.

[2] Dejaegher B., Alaerts G., Matthijs N. (2010), Methodology to develop liquid chromatographic fingerprints for the quality control of herbal medicines, *Acta Chromatographica* 22, 237–258.

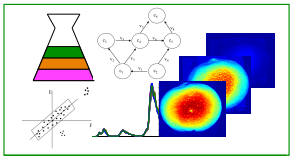
[3] Fan X.H., Cheng Y.Y., Ye Z.L., Lin R.Ch., Qian Z.Z. (2006), Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines, *Anal. Chim. Acta* 555, 217–224.

[4] García-Alvarez-Coque M.C., Torres-Lapasió J.R., Baeza-Baeza J.J. (2006), Models and objective functions for the optimisation of selectivity in reversed-phase liquid chromatography, *Anal. Chim. Acta* 579, 125–145.

[5] Soleo Funari C., Lajarim Carneiro R., Marques Andrade A., Frances Hilder E., Cavalheiro A.J. (2014), Green chromatographic fingerprinting: an environmentally friendly approach for the development of separation methods for fingerprinting complex matrices, *J. Sep. Sci.* 37, 37–44.

[6] Torres-Lapasió J.R., García-Alvarez-Coque M.C. (2006), Levels in the interpretive optimisation of selectivity in high-performance liquid chromatography: A magical mystery tour, *J. Chromatogr. A* 1120, 308–321.

[7] Xie P.S., Chen S.B., Liang Y.Z., Wang X.H., Tian R.T., Upton R. (2006), Chromatographic fingerprint analysis: a rational approach for quality assessment of traditional Chinese herbal medicine, *J. Chromatogr. A* 1112, 171–180.



Assessment of the potential of the combination of hyperspectral Raman images and chemometric methods in metabonomic studies of zebrafish tissues

Laura Benítez[1], Víctor Olmos[1], Pablo Loza[2], Mònica Marro[2], Benjamí Piña[3], Marta Casado[3], Romà Tauler[3] and Anna de Juan[1]

[1]Departament de Química Analítica, Universitat de Barcelona. Barcelona., [2]Institut de Ciències Fotòniques (ICFO). Parc Tecnològic de Castelldefels. Castelldefels., [3]Departament de Química Ambiental, Institut de Diagnòstic Ambiental i Estudis de l'Aigua (IDAEA-CSIC). Barcelona.

Abstract

Hyperspectral imaging can become a very useful tool in metabonomic studies because of the capability to provide chemical (spectra interpretation) and morphological (distribution maps) information about the sample and the environmental conditions.

The aim of this study is to explore the potential of the combination of Raman hyperspectral images and chemometric methods for metabonomic studies on zebrafish, a model organism showing gene homology with humans. A methodological protocol including zebrafish breeding, tissue cryosection, image acquisition (600 - 1800 cm^{-1} fingerprint region) and image data analysis and interpretation has been proposed. In order to assess the effect of chlorpyrifos-oxon (CPO) on zebrafish, images of control and CPO exposed fish samples have been acquired and analyzed.

Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) (Jaumot et al. 2015) was used as a chemometric technique to analyze all images. MCR-ALS provides pure spectra and distribution maps of the sample constituents (de Juan et al. 2014). Multiset analysis has been used to analyze different images simultaneously because of the presence of common components in some images (de Juan et al. 2014).

The results allowed us to distinguish different tissue regions in zebrafish. The interpretation of Raman spectral signatures was easy due to the specific fine bands that helped to identify differences in tissues coming from control and CPO exposed samples. Based on the comparison of resolved signatures, we have been able to distinguish and characterize differences in regions of the tissues studied, such as the melanin in the pigmented areas of the eye, iridophores in tail tissue or highly proteic regions related to the crystalline lens.

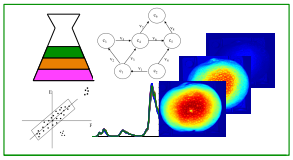
Keywords: Hyperspectral imaging, Raman spectroscopy, metabonomic, zebrafish, chemometric techniques

References

[1] Jaumot J., de Juan A. and Tauler R. (2015) "MCR-ALS GUI 2.0: New features and applications" Chemom. Intell. Lab. Syst., vol. 140, pp. 1–12.

[2] de Juan A., Maeder M., Hancewicz T., Duponchel L. and Tauler R.. (2014) 'Chemometric Tools for Image Analysis' in 'Infrared and Raman Spectroscopic Imaging'. (Salzer R.and Siesler, H.W.eds) Wiley-VCH. Chapter 2,65-106.

[3] de Juan A., Jaumot J. and Tauler R. (2014). "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem". Analytical Methods, vol. 6(14), pp 4964-4976.



Study of the effect of chlorpyriphos-oxon on zebrafish embryos by hyperspectral imaging and chemometric techniques

Víctor Olmos[a], Laura Benítez[a], Mónica Marro[b], Jordi Navarro[b], Pablo Loza[b], Benjamí Piña[c], Romà Tauler[c], Anna de Juan[a]

[a]*Universitat de Barcelona*, Facultat de Química, Departament de Química Analítica, Barcelona; [b]*Institut de Ciències Fotóniques* (ICFO), Parc Tecnologic de Castelldefels, Barcelona; [c]*Institut de Diagnosi Ambiental i Estudis de l'Aigua* (IDAEA-CSIC), Barcelona

Keywords: *Metabonomics, Hyperspectral imaging, MCR-ALS*

Abstract

Environmental –omics consists of the characterization and quantification of biological molecules of an organism related to the exposure to an environmental stress. Analytical techniques used in –omics(e.g. immunoassays, NMR, HPLC-MS…) (Bedia et al. 2015) are usually destructive. Hyperspectral images may be a potentially useful methodology in metabonomics because they provide spatial and chemical information and preserve the natural morphology of the samples. The aim of this work is to asses the potential of hyperspectral images acquired with differents platforms in order to carry out metabonomic studies on zebrafish embryos.

Zebrafish embryos have been obtained by natural breeding. The age of the embryos in the experiment may vary from 5 to 8 days post-fertilization. Some of the embryos have been exposed to a pollutant stress (e.g. chlorpirypfos-oxon) during 24h and the rest have been used as control samples. The embryos have been frozen and embedded in Optimal Cutting Temperaure Compund (OCT) to perform cryosections of the parts of interest. Cryosectioned tissues have been stored at -10º until the measurements have been performed.

Hyperspectral images have been acquired using different spectroscopic platforms (FT-IR, Raman and Fluorescence). Multivariate curve resolution- alternating least squares (MCR-ALS) (Jaumot et al. 2015, de Jaun et al. 2014, de Juan et al. 2014) has been used to analyse hyperspectral images. This method is appropriate to resolve multicomponent systems and provides the distribution maps and pure spectra of the image constituents.

Raman hyperspectral images in combination with MCR-ALS have shown some variations in the retinal pigmented epithelium and the crystalline lens after chlorpiryphos-oxon exposure. FT-IR hyperspectral images permitted to clearly differentiate some areas of the

zebrafish tissues, but more sample replicates will be performed in order to improve the results obtained. Fluorescence images were acquired using different excitation modes (linear and two-photon) to detect different kinds of fluorophores.

MCR results from the analysis of the different hyperspectral image techniques will be presented. Pure spectral signatures and distribution maps of the different zebrafish tissues contributions will be described. MCR spectral signatures of control and contaminated samples will be compared to asses the effect of chlorpiriphos-oxon on zebrafish embryos.

References

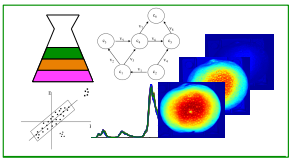
[1] Bedia C., Dalmau N., Jaumot J., Tauler R. (2015) "Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors" Environmental Research, vol. 140, p. 18

[2] Jaumot J., de Juan A. and Tauler R. (2015) "MCR-ALS GUI 2.0: New features and applications" Chemom. Intell. Lab. Syst., vol. 140, p. 1

[3] de Juan A., Maeder M., Hancewicz T., Duponchel L. and Tauler R.. (2014) 'Chemometric Tools for Image Analysis' in 'Infrared and Raman Spectroscopic Imaging'. (Salzer R.and Siesler, H.W.eds) Wiley-VCH.Chapter 2,65-106.

[4] de Juan A., Jaumot J. and Tauler R. (2014). "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem". Analytical Methods, vol. 6(14), p. 4964.

.



Optimization of the chromatographic selectivity of *o*-phthalaldehyde amino-acid derivatives using diode array detection

Tamara Álvarez Segura[a], José Ramón Torres Lapasió[b] and María Celia García Álvarez-Coque[c]

Department of Analytical Chemistry, University of Valencia, C/Dr. Moliner 50. 46100-Burjassot-Valencia (Spain)., [a]talse@alumni.uv.es, [b]jrtorres@uv.es, [c]celia.garcia@uv.es

Abstract

Nowadays chromatographers are assisted by multichannel detectors, giving rise to second-order signals. It results thus surprising that the search of optimal separation conditions is still being focused on reaching the best time selectivity, neglecting the potential benefits of the spectral order. In this work, we use first and second-order multivariate selectivity as chromatographic objective functions to separate a mixture of 19 primary amino-acid derivatives. However, the use of second-order objective functions implies accepting a certain degree of overlap in the best separation conditions. Thus, for obtaining the pure profiles, orthogonal projection approach and alternating least-squares were applied for cluster deconvolution.

Keywords: Chromatographic objective function; Multivariate selectivity; Peak purity; Amino acids; Orthogonal projection approach; Alternating least-squares

Introduction

Current HPLC instruments are able to yield two-way signals, where full spectra are collected as a function of time. With such rich signals readily available, highly specific columns are not so imperative for resolving complex samples at routine level. Hence, the usual strategy of finding the best separation conditions giving prevalence to the chromatographic resolution in the time order seems not too logical nowadays.

With so-rich-in-information signals, the usual practice of focusing the search of the best separation conditions in the time order results rather surprising. However, a price should be accepted: using spectral information implies peak deconvolution. As far as the analyte contributions can be correctly retrieved by deconvolution from the spectrochromatogram, the found optimized separation condition will be acceptable. In this concern, two-way signals may be resolved without any peak shape assumption by self-modelling techniques, such as the orthogonal projection approach followed by alternating least-squares (OPA-ALS). This kind of deconvolution is more desirable than its one-way counterparts, which are based on forcing the accommodation of the overall signal to a linear combination of peak models, and consequently, they are more subjected to larger uncertainties.

Another reason for supporting the active use of spectral information is the reduction of analysis time. By increasing the elution strength, the retention times decrease, but yielding overlapping among the peaks (*i.e.*, shorter separation times imply minor peak capacity, which makes coelution more likely). As a consequence, deconvolution is needed to complete the time selectivity. In this sense, a chromatographic objective function (COF) sensitive to spectral differences would help to find out the separation condition giving rise to the most favorable overlap taking into account the spectral differences. Even in situations involving two peaks with poor resolution in both data orders, nearly selective wavelength windows may exist, and consequently, the spectrochromatogram can be rich enough in analytical information to retrieve the underlying contributions.

Several COFs have been considered for quantifying simultaneously time and spectral information. The easiest ones consisted of extending one-way COFs (COF_1) to include spectra. A COF_1 with particularly good performance is the one-way peak purity (P_1), which can be defined as the analyte area fraction (time×absorbance) free of overlapping with regard to the chromatogram of its interferences. This COF can be easily extended to its two-way counterpart (the COF_2 two-way peak purity, P_2), by replacing area fraction by volume (time×wavelength×absorbance).

An alternative to conventional COFs is using assessments derived from the net analyte signal concept (NAS), which has given rise to a number of associated figures of merit. One-way NAS (NAS_1) can be defined for single wavelength chromatograms as the fraction of

the analyte vector signal that cannot be explained as a linear combination of the vector signals of its interferences. The importance of NAS in Chemical Analysis is indeed extreme. For instance, it can be demonstrated that multivariate calibration models build implicit relationships between NAS and the concentration of the analytes, without an explicit calculation of NAS, and the IUPAC recommends using derived concepts to build and validate models producing good predictions from highly unselective data. However, the usefulness of NAS reaches other ambits different from calibration. Thus, when the data are the underlying signals of the compounds in a chromatogram, NAS (and derived measurements) may help to rank the difficulties of a deconvolution, and its magnitude correlates well with the chromatographic resolution.

In principle, the NAS definition is applicable to both single-wavelength chromatograms and spectrochromatograms, but it is sensitive to the signal size (*i.e.*, analyte concentration and detector sensitivity) and data dimensions (*i.e.*, matrix or vector size), and for these reasons, less sensitive NAS-derived figures of merit are better choices. An interesting one, identically applicable to measure the resolution, is the multivariate selectivity (SEL), which is related to the level of orthogonality of NAS with regard to the space spanned by multiple interferences. Multivariate selectivity ranges between 0 (in case of complete overlap) and 1 (when there is no overlap). In previous work, we reported that SEL_1 correlated with the deconvolution error for single wavelength chromatograms, being thus a useful tool to appraise the difficulty of signals including overlapped peaks. Thus, SEL_2 assessments can be expected to be good candidates to COFs.

All these resolution expressions describe signal overlaps with different perspectives. The resolution measurements only accounting the separation in the time direction (P_1 and SEL_1) tend to find optimal conditions where the peaks are well separated, since they qualify negatively any peak overlap (especially P_1). Meanwhile, those measurements that also include spectral information score situations of excessive overlap as valid. This can be risky, since predictions are affected by errors in peak position up to a certain extent (introduced in the modelling step). Also, the impact of eluent mispreparation, or gradient generation, is more severe when the overlap is significant. This does not invalidate the conclusions of the study: better models or more careful experimentation is just required. These possible errors constitute another level in the optimization problem. In addition, the presence of heterocedastic noise, a poorly subtracted baseline, and any non-linearity, would lead to less favorable deconvolutions. These issues are also beyond the scope of this work, which focuses on the selection of the most favorable gradients.

When spectral information is available, the most economical way of outlining the separation problem is performing first the optimization attending only to the time order, using P_1. If the resolution is satisfactory, we have arrived in a short time to the optimal separation conditions and the spectral order is not needed. In case the resolution was not satisfactory,

we should calculate SEL_2 for the experimental conditions that offered the largest P_1 scores. If one or more of these conditions gives a SEL_2 score for the critical peak pair exceeding 0.80, then the deconvolution of the overlapped peaks can be expected to be favorable. On the contrary, in case the 0.80 threshold for the critical pair is not exceeded in any separation condition, or the analysis time is too large, then the SEL_2 calculation should be extended to the whole grid of experimental conditions. The Pareto plot of SEL_2 scores against the analysis time will assist in the selection of the best separation conditions.

In previous work, the proposed approach was applied to the separation of a mixture of 25 phenolic compounds, which remained unresolved in the chromatographic order using linear and multi-linear gradients of acetonitrile-water. An application is here reported for a mixture of the OPA-NAC derivatives of the 19 proteic amino-acids derivatives, which constitute a critical separation example owing to the spectra similarities. Several separation conditions are examined: isocratic elution, linear and multi-linear gradients, and multi-isocratic separations. When full separation in the time order is demanded, isocratic and gradient elution could resolve the mixture but at too high analysis times. However, using COFs considering spectral information, experimental conditions could be found with acceptable analysis times at the cost of some coelution, which nevertheless could be fully resolved by multivariate deconvolution. In all instances, the analysis times could be considerably reduced. Several signal-to-noise levels were examined to appraise the influence of non-idealities. The results were correct in a wide range of conditions, up to SNR ca. 100.

References

[1] Carda-Broch S., Torres-Lapasió J. R., García-Alvarez-Coque M. C. (1999), Evaluation of several global resolution functions for liquid chromatography, *Anal. Chim. Acta* 396, 61–74.

[2] Cuesta Sánchez F., van den Bogaert B., Rutan S. C., Massart D. L. (1996), Multivariate peak purity approaches, *Chemom. Intell. Lab. Sys.* 34, 139–171.

[3] Faber K., Lorber A., Kowalski B. R. (1997), Analytical figures of merit for tensorial calibration, *J. Chemom.* 11, 419–461.

[4] de Juan A., Tauler R. (2003), Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution, *Anal.Chim. Acta* 500, 195–210.

[5] Lorber A., Faber K., Kowalski B. R. (1997), Net analyte signal calculation in multivariate calibration, *Anal. Chem.* 69, 1620–1626.

[6] Ortiz-Bolsico C., Torres-Lapasió J. R., García-Alvarez-Coque M. C. (2014), Optimization of gradient elution with serially-coupled columns. Part I: Single linear gradients, *J. Chromatogr. A* 1350, 51–60.

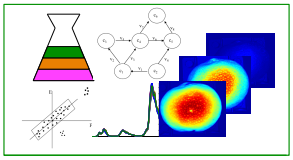
[7] Ortiz-Bolsico C., Torres-Lapasió J. R., García-Alvarez-Coque M. C. (2015), Optimization of gradient elution with serially-coupled columns. Part II: Multi-linear gradients including isocratic steps, *J. Chromatogr. A* 1373, 51-60.

[8] Tauler R. (1995), Multivariate curve resolution applied to second order data, *Chemom. Intell. Lab. Syst.* 30, 133–146.

[9] Torres-Lapasió J. R., García-Alvarez-Coque M. C. (2006), Levels in the interpretive optimisation of selectivity in high-performance liquid chromatography: A magical mystery tour, *J. Chromatogr. A* 1120, 308–322.

[10] Torres-Lapasió J. R., Pous-Torres S., Ortiz-Bolsico C., García-Álvarez-Coque M. C. (2015), Optimisation of chromatographic resolution using objective functions including both time and spectral information, *J. Chromatogr. A* 1377, 75–84.

[11] Vivó-Truyols G., Torres-Lapasió J. R., García-Alvarez-Coque M. C. (2003), Net analyte signal as a deconvolution-oriented resolution criterion in the optimisation of chromatographic techniques, *J. Chromatogr. A* 991, 47–59.



Prostate Diffusion Weighted-Magnetic Resonance Image Analysis using Multivariate Curve Resolution Methods

Eric Aguado-Sarrió, José Manuel Prats-Montalbán, Alberto Ferrer

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, DEIOAC-UPV, Valencia, España.

Abstract

Multivariate Curve Resolution (MCR) has been applied on prostate Diffusion Weighted-Magnetic Resonance Images (DW-MRI). Different physiological-based modeling approaches of the diffusion process have been submitted to validation by sequentially incorporating prior knowledge on the MCR constraints. Results validate the biexponential diffusion modeling approach and show the capability of the MCR models to find, characterize and locate the behaviors related to the presence of an early prostate tumor.

Keywords: *MCR, Multivariate Image Analysis, Diffusion, Magnetic Resonance, Prostate, Tumor.*

Resumen

Se han aplicado los métodos de resolución multivariante de curvas (MCR) a imágenes de resonancia magnética de difusión en próstata (DW-MRI). Diferentes aproximaciones basadas en fenómenos fisiológicos se han aplicado con el objetivo de validar los modelos de forma secuencial, incorporando el conocimiento "a priori" que se tiene del proceso en las restricciones del modelo MCR. Los resultados permiten validar la aproximación biexponencial en difusión, además de mostrar la capacidad de los modelos MCR para encontrar, caracterizar y localizar los comportamientos relacionados con la presencia de tumores precoces en la próstata.

Palabras clave: *MCR, Análisis multivariante de imágenes, Difusión, Resonancia Magnética, Próstata, Tumor.*

Introducción

En el estudio de carcinomas precoces, dos de los principales indicadores de la presencia de un proceso tumoral son la vascularización y el incremento de la densidad celular. Cuando un grupo de células en crecimiento presenta demandas de oxígeno y nutrientes anormalmente altas, el tejido reacciona creando nuevos vasos sanguíneos (angiogénesis) o desarrollando los ya existentes (neovascularización). Por otro lado, el proceso biológico asociado a elevadas densidades celulares que conlleva a la aglomeración celular en los tejidos se denomina celularización. La combinación de ambos procesos es lo que normalmente determina la presencia de un tumor precoz como un primer paso en oncogénesis. Una manera de analizar esta combinación de procesos es por medio del estudio del proceso de difusión en los tejidos (Charles-Edwards and De Souza 2006), el cual es un proceso físico que sucede debido a la agitación térmica de las moléculas de agua en el interior del cuerpo humano. Estos movimientos translacionales dependen, además de otros factores, de la estructura del tejido según su organización celular. Cuando el tejido está altamente celularizado, las moléculas de agua presentan mayor restricción al movimiento debido a que disminuye el espacio intersticial y a la presencia de un mayor número de interfases de membrana celulares. Sin embargo, cuando el tejido se encuentra altamente vascularizado, las moléculas se encuentran en un espacio no restringido dentro de los vasos, y los movimientos son aleatorios, con menor restricción, en todas las direcciones espaciales.

El proceso de difusión se puede evaluar mediante técnicas de imagen de resonancia magnética ponderada (DW-MRI). Esta técnica no invasiva permite proporcionar imágenes de alta resolución que son sensibles a los movimientos de las moléculas de agua dentro de los tejidos. Dependiendo de la configuración del equipo de resonancia magnética y basándose

en la duración y la amplitud del campo magnético aplicado, a la adquisición de imágnes se la asocia a un parámetro conocido como valor-**b** (Le Bihan 1991, Lemke 2011). La señal de la imagen asociada a cada píxel disminuye a medida que aumenta el valor de **b**. Esta atenuación de la señal depende de las características del tejido, siendo más rápida si el tejido se encuentra vascularizado y mucho más lenta si el tejido está altamente celularizado. El rango de las diferentes atenuaciones de señal entre estos dos tipos de tejido para el mismo valor de b es la base del estudio de los diferentes comportamientos presentes en el proceso de difusión.

Con el objetivo de modelizar la caída de la señal, los espectros se pueden ajustar con diferentes expresiones ó modelos matemáticos. El modelo más ampliamente utilizado en el ambiente clínico es el modelo monoexponencial de difusión (Le Bihan 1991) con el coeficiente de difusión aparente (ADC) como parámetro, este modelo supone que existe una única caída exponencial para el modelado de la señal. El principal problema del modelo monoexponencial es que no tiene en cuenta los diferentes mecanismos del proceso de difusión. Actualmente, una manera de solventar estos problemas es empleante un modelo biexponencial. Este es un modelo más complejo, ya que considera dos comportamientos, difusión rápida y lenta, ponderados mediante un nuevo parámetro llamado fracción vascular (f), que se corresponde con la proporción de tejido vascular en un vóxel. Este modelo también es conocido como "intra-voxel incoherent motion" (IVIM) (Le Bihan 1986), debido a los dos tipos de movimientos considerados, relacionados con celularización (difusión lenta) y vascularización (difusión rápida). A pesar de que el modelo IVIM es teóricamente más apropiado según el criterio fisiológico, el modelo monoexponencial es, actualmente, el más utilizado en la práctica médica para modelar el proceso de difusión.

Una posible alternativa para analizar los comportamientos de difusión es mediante la aplicación de modelos estadísticos multivariantes, con los que es posible aprovechar la relación entre píxeles. Cuando se trabaja con imágenes, la aplicación de este tipo de modelos se la conoce como Análisis Multivariante de Imágenes (MIA) (Geladi and Grahn 1996, Prats-Montalbán et al. 2011). La principal característica de este tipo de modelos es la capacidad de estudiar el conjunto completo de píxeles al mismo tiempo, extrayendo las fuentes de variabilidad causadas por las estructuras latentes presentes en las imágenes. De esta forma, estos modelos pueden ayudar a proporcionar nuevos modelos no-paramétricos que permitan explicar los principales comportamientos de difusión extraídos a partir de las imágenes DW-MRI. También pueden ser útil para comprobar la adecuación de las diferentes aproximaciones propuestas en la literatura (mono y biexponencial).

La principal técnica de análisis multivariante es PCA (Análisis de Componentes Principales) (Jackson 1991). Sin embargo, existen dos problemas cuando se aplica PCA a los datos de DW-MRI:

1) No se puede introducir información a priori en el modelo.
2) La ortogonalidad de los componentes principales es una limitación para modelar los diferentes comportamientos de difusión que no son necesariamente ortogonales.

Con el objetivo de evitar estos problemas, es posible emplear técnicas más flexibles, como es el caso del modelo "Resolución Multivariante de Curvas" ó MCR, el cual ha sido aplicado anteriormente a imágenes dinámicas de resonancia magnética (Dynamic Contrast Enhanced-MRI) (Prats-Montalbán et al. 2014).

Los objetivos de este trabajo son:

1) explorar la capacidad de los métodos MCR para modelar los diferentes comportamientos asociados al proceso de difusión a partir de imágenes DW-MRI, ayudando a los especialistas a detectar y caracterizar tumores precoces en la próstata.
2) Validar la adecuación de los diferentes modelos teóricos más comúnmente aplicacdos en la práctica clínica, mediante la incorporación secuencial de restricciones en el algoritmo MCR empleando el conocimiento "a priori" que se tiene sobre el proceso de difusión.
3) Proporcionar nuevas imágenes de biomarcadores que permitan complementar a los más comúnmente usados en el diagnosis clínico.

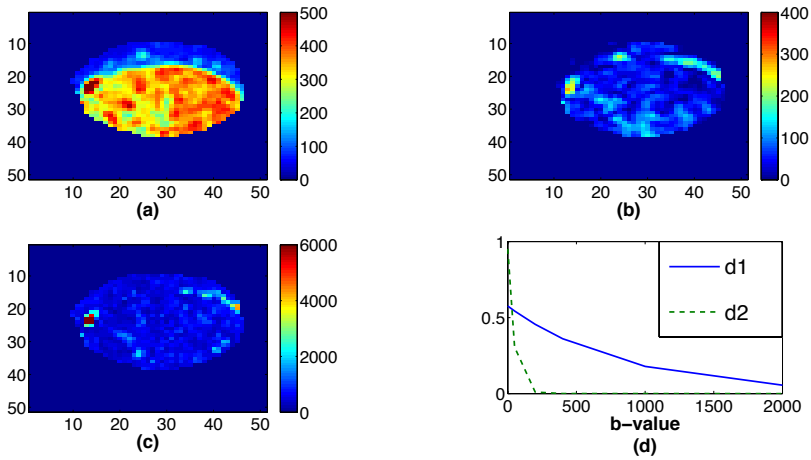


Figura 1. Modelo MCR (99% de variabilidad explicada). (a) mapa de scores asociado a d1 (difusión lenta, línea sólida azul). (b) Mapa de scores asociado a d2 (difusión rápida, línea verde punteada). (c) mapa de RSS (residuos al cuadrado). (d) Comportamientos proporcionados por el modelo MCR.

Referencias

[1] D. Le Bihan, Molecular diffusion nuclear magnetic resonance imaging, Magn. Reson. Q. 7 (1991) 1–30.

[2] D. Le Bihan, et al., MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders, Radiology. 161 (1986) 401–407.

[3] E.M. Charles-Edwards, N.M. De Souza, Diffusion-weighted magnetic resonance im- aging and its application to cancer, Cancer Imaging. 6 (2006) 135–143..

[4] P. Geladi, H. Grahn, Multivariate Image Analysis, Wiley, Chichester, England, 1996.

[5] J.E. Jackson, A User's Guide to Principal Components, Ed. Wiley, New York, 1991.

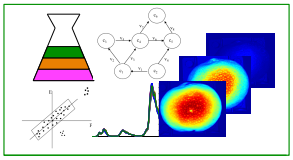
[6] A. Lemke, B. Stieltjes, L.R. Schard, F.B. Faun, Toward an optimal distribution of b values for intravoxel incoherent motion imaging, Magn. Reson. Imaging. 29 (2011) 766–776.

[7] J.M. Prats-Montalbán, A. Ferrer, A. de Juan, Multivariate image analysis: a review with applications, Chemom. Intell. Lab. Syst. 107 (2011) 1–23.

[8] J.M. Prats-Montalbán, R. Sanz-Requena, L. Martí-Bonmatí, A. Ferrer, Prostate functional magnetic resonance image analysis using multivariate curve resolution methods, J. Chemometrics 28 (2014) 672–680.

Agradecimientos

Los autores de este trabajo quieren agradecer a la profesora Anna de Juan por sus comentarios y ayuda en el uso del software para este estudio. Este trabajo fue parcialmente financiado por el Ministerio de Economía y Competitividad bajo el proyecto DPI 2011-28112-C04-02.



Optimal Design of Experiments in the Original Space and in the Latent Space in Mixture Design

Daniel Palací López[a], Alberto Ferrer[b] and Peter Goos[c]

[a] Universitat Politècnica de València, Valencia (Spain), dapalpe@etsii.upv.es, [b] Universitat Politècnica de València, Valencia (Spain), aferrer@eio.upv.es, [c] Katholieke Universiteit Leuven, Leuven (Belgium), Peter Goos, peter.goos@biw.kuleuven.be

Abstract

Optimal Design Of Experiments (DOE) in Mixture Design problems, in which the ratios of the different ingredients being blended are at least as important as their absolute quantities to achieve the final product desired properties, is here approached for both the situation in which data analysis will be made using classical tools such as Ordinary Least Squares (OLS) or Generalized Least Squares (GLS) as well as when it is to be made using methods based on Projection to Latent Structures (PLS). A comparison of the results achieved in both cases is made in order to evaluate to which extent building an optimal DOE in the latent space or in the original space influences the performance of a predictive model when using PLS-based methods for data analysis.

Keywords: *Mixture Design, Optimal Design of Experiments, Design of Experiments, DOE, Projection to Latent Structures, Partial Least Squares, PLS, Ordinary Least Squares, OLS*

Introduction

Mixture design problems are those in which the ratios (r_1, r_2 … r_J) of the J different components - or raw materials- of a blend are at least as relevant as their absolute quantities in terms of their influence on the final product properties of interest (Y). Different processes from the chemical, pharmaceutical or bioprocess sector can be addressed as mixture design problems where the process conditions (Z) and - in some cases - the raw materials properties (x_1, x_2 … x_J) are also considered.

Due to the fact that the raw materials rates must sum up to 1 - or 100% - perfect collinearity among these variables is always present in this kind of problem. Traditionally a reparametrization of the classical polynomials - the Scheffé models - has been used in order to cope with this issue. As an example, the usual interpretation of α_0, α_i, α_{ij} in the polynomial

$$E(Y) = \alpha_0 + \sum_{i=1}^{J} \alpha_i \cdot r_i + \sum_{i=1}^{J} \sum_{j \geq i}^{J} \alpha_{ij} \cdot r_i \cdot r_j$$

makes no sense since α_0 would correspond to the expected value for *Y* for a mixture with no ingredients at all, α_i would be the change in such expected value when the ratio r_i of the i^{th} ingredient is increased in one unit without changing the ratios of the rest of the ingredients, and so on. Instead, when the constrain $\sum_{i=1}^{J} r_i = 1$ is taken into account the previous polynomial can be reparametriced to the Scheffé polynomial

$$E(Y) = \sum_{i=1}^{J} \beta_i \cdot r_i + \sum_{i=1}^{J-1} \sum_{j>i}^{J} \beta_{ij} \cdot r_i \cdot r_j$$

where β_i would be the expected value of *Y* when only one ingredient of the mixture is present and β_{ij} is related to deviations from the ideal mixture for a mixture with only the i^{th} and j^{th} component.

Although this reparametrization allows model fitting using techniques such as Ordirary Least Squares (OLS) or Generalized Least Squares (GLS), classical Design of Experiments (DOE) cannot be used. Extensive analysis on this matter has been made in the literature, pointing at standardized optimal designs of experiments depending on the degree of the polynomial, as long as the shape of the mixture space remains a simplex - a triangle if there are 3 ingredients, tetrahedron for 4 ingredients… (Cornell 2002, Snee 1976).

However, when restrictions imposed on the rates of the ingredients - lower and upper bounds other than 0 and 1, and possible linear constrains - change the shape of the space in a way that it stops being a simplex, these standardized DOEs are no longer usable- furthermore, different sources of correlation may appear which the Scheffé polynomials don't

take into account. Even if no linear constrains are present, a highly restricted mixture space may lead to unreliable stimations of the coefficients of the fitted model, a problem that can be dealt by using methods based on Projection to Latent Structures (PLS), which also offer the possibility to fit models in the form of the classical polynomials, though with an interpretation equivalent to that of the Cox models, a reparametrization of the scheffé ones (Kettaneh-Wold, 1992, Eriksson et al. 1998).

By using PLS-based methods, the perfect collinearity among raw materials rates is no longer a problem and models built become much simpler, since there is no need to treat mixture variables and process variables differently. This also leads to DOEs with a - sometimes significantly - lower number of required experiments when process variables are taken into account, compared to the case where OLS-based methods are used for data analysis.

On the other hand, OLS-based methods cannot be used if the raw material properties are to be considered when building the corresponding model. Instead, a number of PLS-based algorithms have been proposed, such as the L-PLS (Martens et al. 2005, Muteki and MacGregor 2007)], WS-PLS (García-Muñoz and Polizzi 2012) or the JR-PLS or TPLS (García-Muñoz 2014), that can be used in such circumstances depending on the complexity of the data.

In spite of all the aforementioned PLS-based methods and algorithms having been proposed, it is seldom mentioned in the literature how the data was collected or which DOE was used - if any - although the results presented seem to be generally better, or at least as good, as those achieved using OLS-based methods, whenever a comparison is made. However, some work has already been done in this field and methods for selecting a good candidate set for calibration purposes or model building have been proposed [Wold et al. 1986, Ferré and Rius 1996, Wole et al. 2004).

While it has been made clear that only PLS-based methods permit the analysis of data from the most complex mixture design problems, building a DOE in the latent space also requires previous knowledge about it. Because of this and the fact that results in the literature seem to generally show that better results can be achieved using PLS-based methods compared to OLS-based ones, despite the presented DOE having been built in the space of the original variables, here a comparison between the optimality of a DOE built in the original space and that of a DOE built in the latent space will be made, when data analysis is to be performed using PLS-based methods as well as OLS-based ones.

References

[1] Cornell, J.A. *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. ISBN: 978-0-471-39367-2

[2] Eriksson L.; Johansson E.; Wikström C. *Chemometrics and Intelligent Laboratory Systems* **1998**, *43*, 1-24.

[3] Ferré, J.; Rius, F. X. *Analytical Chemistry* **1996**, *68*, 1565-1571.

[4] García-Muñoz, S.; Polizzi, M. A. *Chemometrics and Intelligent Laboratory Systems* **2012**, *114*, 116-121.

[5] García-Muñoz, S. *Chemometrics and Intelligent Laboratory Systems* **2014**, *133*, 49-62.

[6] Kettaneh-Wold N. *Chemometrics and Intelligent Laboratory Systems* **1992**, *14*, 57-69.

[7] Martens, H.; Anderssen E.; Flatberg, A.; Gidskehaug, L.H.; Hoy, M.; Westad, F.; Thybo, A.; Martens, M. *Computational Statistics and Data Analysis* **2005**, *48 (1)*, 103-123

[8] Muteki, K.; MacGregor, J. F. *Chemometrics and Intelligent Laboratory Systems* **2007**, *85*, 186-194.

[9] Snee R.D.; Marquardt D.W. *Technometrics* **1976**, *18*, 26-28.

[10] Wold, S.; Sjöström, M.; Carlson, R.; Lundstedt, T.; Hellberg, S.; Skagerberg, B.; Wikström, C.; Öhman, J. *Analytica Chimica Acta* **1986**, *191*, 17-32 .

[11] Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A. *Journal of Chemometrics* **2004**, *18*, 156-165.