Anal Bioanal Chem (2013) 405:1293–1302
DOI 10.1007/s00216-012-6507-4

ORIGINAL PAPER

Exploring in vivo violacein biosynthesis by application of multivariate curve resolution on fused UV–VIS absorption, fluorescence, and liquid chromatography–mass spectrometry data

Clecio Dantas · Romà Tauler · Márcia Miguel Castro Ferreira

Received: 17 July 2012 / Revised: 16 October 2012 / Accepted: 18 October 2012 / Published online: 17 November 2012

Abstract In this work, the application of multivariate curve resolution-alternating least squares (MCR-ALS) is proposed for extracting information from multitechnique fused multivariate data (UV–VIS absorption, fluorescence, and liquid chromatography–mass spectrometry) gathered during the biosynthesis of violacein pigment. Experimental data sets were pretreated and arranged in a row-wise augmented data matrix before their chemometric investigation. Five different chemical components were resolved. Kinetic and spectral information about these components were obtained and their relationship with violacein biosynthesis was established. Three new chemical compounds with molar masses of 453, 465, and 479 u, until now not reported in the literature, were identified and proposed as intermediates in the biosynthesis of other indolocarbazoles. The precursor (tryptophan), one intermediate (deoxyviolacein), and the final product (violacein) of violacein biosynthesis were identified and characterized using the proposed approach. The chemometric procedure based on the MCR-ALS method has proved to be a powerful tool to investigate violacein biosynthesis and its application can be easily extended to the study of other bioprocesses.

Keywords In vivo violacein biosynthesis · Multivariate curve resolution-alternating least squares · Data fusion · UV–VIS · Fluorescence spectroscopy · LC–MS

Electronic supplementary material The online version of this article (doi:10.1007/s00216-012-6507-4) contains supplementary material, which is available to authorized users.

C. Dantas · M. M. C. Ferreira (✉)
Institute of Chemistry, University of Campinas—UNICAMP, Campinas, P.O. Box 6154, 13084-971, Brazil
e-mail: marcia@iqm.unicamp.br

R. Tauler
Department of Environmental Chemistry, IDAEA-CSIC, 08034 Barcelona, Spain

Introduction

Violacein is an indole derivative synthesized by several Gram-negative bacterial strains (mainly by bacteria of the genus *Chromobacterium*) that has attracted increased interest owing to its broad-spectrum of antibacterial activity [1–3] besides other properties such as trypanocide [2, 4, 5], tumoricide [2, 6–8], and antivirus activities [2, 9].

The biosynthesis of this compound has been extensively studied since 1934 from both basic science and biotechnology perspectives [10–26]. These studies allowed proposing the enzymatic pathway for violacein production shown in Fig. 1.

The complexity of this biosynthesis is notorious. It begins with VioA oxidation of L-tryptophan to indole-3-pyruvic acid (IPA) imine. Then the oxidative coupling of two molecules of IPA imine by VioB gives an uncharacterized intermediate proposed to be an IPA imine dimer. The skeleton of this compound is catalytically rearranged by VioE through an intramolecular rearrangement of the indole ring, producing protodeoxyviolaceinic acid. The latter can be converted into protodeoxyviolacein spontaneously by autoxidation or it can undergo hydroxylation catalyzed by VioD at the 5-position of the left side indole ring producing the protoviolaceinic acid. This compound is then hydroxylated by VioC at the 2-position of the right side indole ring to form violaceinic acid. The subsequent conversion to violacein involves a non-enzymatic process of oxidative decarboxylation.

In an earlier work [27], multi-wavelength fluorescence spectroscopy and PARAFAC methods were proposed to detect and identify (through spectral resolution) fluorophores that are consumed and produced by bacteria and a possible relationship of this process with the violacein biosynthesis was established. Using this methodology, six fluorophores were detected but full spectral identification could not be achieved.

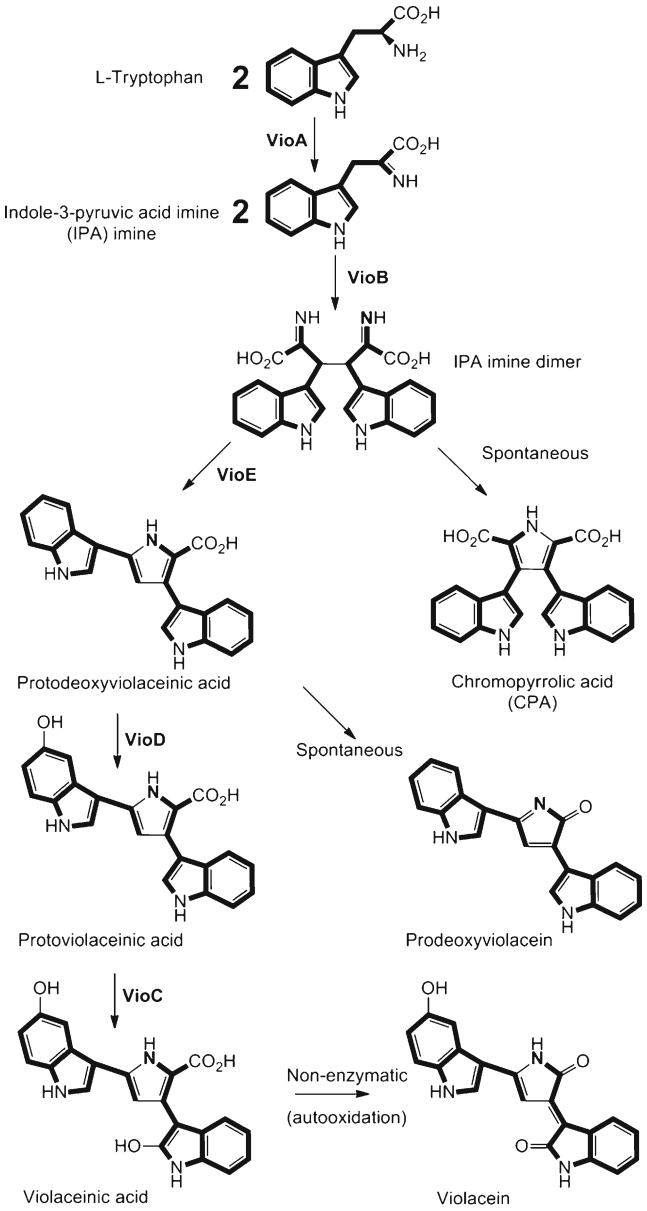

Fig. 1 Violacein biosynthesis pathway

Apart from fluorescence data, UV–VIS absorption spectra and mass chromatograms (LC–MS) of the same samples were obtained. It was then decided to explore the combination of these analytical methods to obtain more reliable and consistent kinetic and spectral information for this complex biochemical reaction. For this purpose, the use of curve resolution methods is proposed.

Multivariate curve resolution (MCR) methods perform the bilinear decomposition of the mixed multicomponent information present in an experimental data matrix, **D**, into chemically meaningful pure contributions of each individual component. Among MCR methods, the one based on alternating least squares (MCR-ALS) has been shown to be a powerful and flexible chemometric method able to investigate complex data sets obtained by combination of several detector systems [28–30].

Hence, in this work we present the results obtained in the simultaneous (fused) analysis of data coming from three different sources (UV–VIS absorption, fluorescence, and liquid chromatography–mass spectrometry) using the MCR-ALS method to explore the kinetic and spectral information associated with violacein biosynthesis.

Material and methods

Organisms, culture media, and cultivation conditions

A colony of *Chromobacterium violaceum* CCT 3496 was inoculated into a 250-mL Erlenmeyer flask containing 50 mL of sterilized culture medium (0.50 % D-glucose, 0.50 % bacteriologic peptone, 0.25 % yeast extract, and 0.03 % L-tryptophan) and grown for 24 h at 33 °C on an orbital shaker at 200 rpm. Afterwards, 10 mL of this bacterial culture was transferred to a 1,500-mL BioFlo bioreactor (New Brunswick Scientific) containing 1,000 mL of sterilized culture medium and the parameters: temperature, stirring, and air flow were adjusted as described in [27].

Sample collection

Aliquots of 10 mL were withdrawn from the bioreactor every 2 h with the aid of a glass syringe, stored in Falcon tubes of 15 mL and frozen (−20 °C). At a later stage, each aliquot was thawed in a thermostated bath at 30 °C for 5 min and then centrifuged at 7,000 rpm for 10 min. The supernatant was eliminated and 5 mL absolute ethanol was added to extract intracellular compounds. Then, it was centrifuged at 7,000 rpm for 10 min for cell removal and the new supernatant was collected. The crude ethanolic extract was passed through a 0.45-μm pore-size filter. Altogether, 18 samples corresponding to the crude ethanolic extract of *C. violaceum* at different stages of biosynthesis were collected.

Data acquisition

Fluorescence measurements were performed using a Varian Cary Eclipse fluorescence spectrophotometer. Aliquots of 0.5 mL of the samples were diluted with 2.5 mL of absolute ethanol in the quartz cuvette to reduce effects of quenching, inner filter and energy transfer processes. The emission spectra were collected for each sample at spectral ranges of 270–800 nm when excited at 280, 310, and 350 nm. The scan rate was 1,200 nm$\,$min^{-1}. The increment for emission wavelengths was 2 nm. The photomultiplier tube (PMT) voltage was set to 600 V. The measurements were carried

out with 90° geometry. Rayleigh scatter peaks were removed and areas affected replaced with interpolated values.

UV–VIS absorption spectra were recorded on a HP8453 diode array spectrophotometer (Hewlett-Packard) over the wavelength range of 205–800 nm with spectral resolution of 1 nm. Aliquots of 0.5 mL of the samples were diluted with 2.5 mL of absolute ethanol in the quartz cuvette.

The samples were also analyzed by a Waters Acquity UPLC system coupled to a MS detector. Electrospray-ionization (ESI) in the positive mode was chosen as the method of ionization at the following conditions: cone voltage, 45 V; capillary voltage, 2.5 kV; source temperature, 150 °C; and desolvation temperature, 400 °C. Mass spectra were recorded every 0.01 s, between 100 and 1,000 mass units. Chromatographic separations were obtained with a C18 BEH column (50×2.1 mm, 1.7 μm). A gradient elution was carried out using the following solvent systems: phase A (Milli-Q water + 0.1 % formic acid) and phase B (acetonitrile + 0.1 % formic acid). The solvent gradient programming was (95:5) to (30:70) in 8 min at a flow rate of 350 $\mu L min^{-1}$, returning to the initial conditions in 1 min. The run time was 9 min per sample.

Data preprocessing and arrangement

The analysis of each sample obtained by UV–VIS absorption and fluorescence (at a single excitation wavelength) spectroscopies provides one data vector per sample, with the number of elements equal to the respective number of wavelengths. In contrast, data obtained by LC–MS provides a data matrix per sample, with the number of rows equal to the number of elution times and with the number of columns equal to the number of mass-to-charge ratios.

The simultaneous analysis of data sets obtained by different instrumental methods requires that they have at least one common direction, mode, or order. In the case under study this common direction will be the sample mode or direction. Thus, it is necessary to transform the LC–MS data matrix into a data vector per sample. There are two ways of doing this: (a) summing the values of every row of the data matrix producing a new single vector of total mass spectra (TMS) data or (b) summing the values of every column of the data matrix producing a new single vector of total ion chromatograms (TIC) data.

LC–MS TIC data were calculated from the LC analysis of every sample and they were together assembled into the new data matrix $\mathbf{D}_{TIC}$ (18×794) following the evolving sample order of the reaction. UV–VIS absorption and fluorescence data were also assembled into matrices, $\mathbf{D}_{UV-VIS}$ (18×595) and $\mathbf{D}_{emission}$ (18×798), respectively, following the same reaction time order. In the case of UV–VIS absorption spectra a single data matrix, $\mathbf{D}_{UV-VIS}$, was obtained for all samples collected during the bioprocess. For the fluorescence spectra, three different data matrices were collected, each at a different excitation wavelength for all analyzed samples, i.e., three data matrices were obtained and arranged in a single one by matrix row-augmentation, $\mathbf{D}_{emission} = [\mathbf{D}_{em1}\ \mathbf{D}_{em2}\ \mathbf{D}_{em3}]$ where the colon symbol stands for the row-wise matrix augmentation, of the emission matrices at the three excitation wavelengths, set one beside the other. Before the simultaneous analysis of fusion of the different types of data matrices ($\mathbf{D}_{UV-VIS}$, $\mathbf{D}_{emission}$, and $\mathbf{D}_{TIC}$), some data pretreatments were applied. $\mathbf{D}_{TIC}$ and $\mathbf{D}_{emission}$ matrices were converted to a common intensity scale similar to the $\mathbf{D}_{UV-VIS}$ matrix by appropriate scaling factors to give similar weight to all of them. Preliminary baseline background signal correction was performed on the $\mathbf{D}_{TIC}$ matrix using the asymmetric least squares algorithm [31]. Finally, data fusion of the three type of data matrices (UV–VIS, emission, and TIC) was carried out using a row-wise matrix augmentation procedure to give matrix $\mathbf{D}_{aug}$ (18×2,187)=$[\mathbf{D}_{UV}\ \mathbf{D}_{emission}\ \mathbf{D}_{TIC}]$.

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

MCR-ALS is an iterative curve resolution method that focuses on the bilinear decomposition of mixed multicomponent information stored in a data matrix into chemically meaningful pure contributions of each of the components in the mixture [32, 33]. The bilinear decomposition of the experimental data matrix is performed using the model equation:

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

where the dimensions of the matrices are $\mathbf{D}(nr \times nc)$, $\mathbf{C}(nr \times n)$, $\mathbf{S}^{\mathbf{T}}(n \times nc)$, and $\mathbf{E}(nr \times nc)$; *nr* is the number of rows of data/matrix **D**, related in our case to the number of ethanolic extract samples collected at the different stages of the biosynthesis; *nc* is the number of columns of data matrix **D**, related to the number of spectroscopic channels (wavelengths, *m*/*z* ratios) used during the monitoring of the biosynthesis process); *n* is the number of chemical components contributing to the signal measured in the data matrix **D**; **C** is the matrix describing how the contributions of the *n* components or chemical species change along the *nr* rows of the data matrix (concentration profiles), i.e., along process time, and $\mathbf{S}^{\mathbf{T}}$ is the matrix describing how the spectral responses of these *n* component change along the *nc* columns of the data matrix (pure spectra); and **E** is the residual matrix describing the variance not explained by the model, $\mathbf{C}\mathbf{S}^{\mathbf{T}}$.

The first step in the resolution procedure is the determination of the number of chemical components, *n*, causing the observed variance in **D**. In the second step, an initial estimation of the (row) concentration profiles of the components, i.e., matrix **C**, is obtained, for instance using evolving factor analysis (EFA) [34, 35]. Given an initial estimation of **C**, Eq. 1 is solved for $\mathbf{S}^{\mathbf{T}}$ by linear least squares. At each new

iterative cycle of the alternating least squares optimization procedure, both matrices **C** and $\mathbf{S}^{\mathbf{T}}$ are recalculated under constraints derived from the physical nature of the system and from prior knowledge of the problem under study. This process is finished when the convergence criterion is fulfilled and a reproduced data matrix **D** using **C** and $\mathbf{S}^{\mathbf{T}}$ is obtained. Convergence criterion is based on the comparison of the lack of fit (LOF) values obtained in two consecutive iterations. The lack of fit values are calculated according to the expression:

$$\%\mathrm{LOF} = 100 \times \sqrt{\frac{\sum_{ij}\left(d_{ij} - d_{ij}^{*}\right)^{2}}{\sum_{ij} d_{ij}^{2}}} \tag{2}$$

where d_{ij} is the experimental data matrix element from *i*-th ethanolic extract collected in the violacein biosynthesis at the *j*-th spectroscopic channel; d_{ij}^{*} is the corresponding element obtained by the MCR-ALS model. Another parameter used to indicate the fit quality of the MCR-ALS results is the percentage of explained variance (R^2), calculated according to the equation:

$$R^{2} = 100 \times \frac{\sum_{ij} d_{ij}^{*2}}{\sum_{ij} d_{ij}^{2}} \tag{3}$$

An outstanding feature of the MCR-ALS method consists in performing the simultaneous analysis of multiple data sets obtained either by monitoring the same experiment using several analytical techniques or monitoring several experiments with the same analytical technique [32, 33, 36]. This type of analysis reduces or even eliminates the possible rotational ambiguities associated with the obtained profiles, which can be especially critical in kinetic studies, where concentration profiles lack selectivity and have similar overlapped shapes [37]. The final model equation for MCR-ALS bilinear decomposition was:

$$\left[\mathbf{D}_{\mathrm{UV}}\,\mathbf{D}_{\mathrm{em1}}\,\mathbf{D}_{\mathrm{em2}}\,\mathbf{D}_{\mathrm{em3}}\,\mathbf{D}_{\mathrm{TIC}}\right] = \mathbf{C}\left[\mathbf{S}_{\mathrm{UV}}^{\mathbf{T}}\,\mathbf{S}_{\mathrm{em1}}^{\mathbf{T}}\,\mathbf{S}_{\mathrm{em2}}^{\mathbf{T}}\,\mathbf{S}_{\mathrm{em3}}^{\mathbf{T}}\,\mathbf{S}_{\mathrm{TIC}}^{\mathbf{T}}\right] + \mathbf{E} \tag{4}$$

$$\mathbf{D}_{\mathrm{aug}} = \mathbf{C}\,\mathbf{S}_{\mathrm{aug}}^{\mathbf{T}} \tag{5}$$

All calculations have been performed using the MCR-ALS computer program [38] implemented in the MATLAB 7.0 (The Mathworks Inc, USA) environment.

Results and discussion

UV–VIS absorption data matrix ($\mathbf{D}_{\mathrm{UV-VIS}}$), fluorescence data matrix and ($\mathbf{D}_{\mathrm{emission}}$) and liquid chromatography mass spectrometry TIC data matrix ($\mathbf{D}_{\mathrm{TIC}}$) of the 18 crude ethanolic extracts collected during the different stages of violacein biosynthesis were simultaneously analyzed in detail. The corresponding data matrices were pretreated and arranged in the wide row-wise augmented fused data matrix $\mathbf{D}_{\mathrm{aug}} = [\mathbf{D}_{\mathrm{UV}}\,\mathbf{D}_{\mathrm{em1}}\,\mathbf{D}_{\mathrm{em2}}\,\mathbf{D}_{\mathrm{em3}}\,\mathbf{D}_{\mathrm{TIC}}]$ (see previous section for data preprocessing and arrangement, Eqs. 4 and 5). Figure 2 shows the plot of this fused data matrix.

Assuming that chemical compounds have a larger contribution to data variance than other variance sources like noise, background, and baseline after data pretreatment, the number of chemical components evolving during the biosynthesis process was initially estimated from the number of significant singular values of this augmented data (chemical rank) in its singular value decomposition (SVD) analysis. From the singular value plot of the data matrix ($\mathbf{D}_{\mathrm{aug}}$), it was possible to infer that the more probable number of distinguishable chemical components evolving during the violacein biosynthesis was five, since singular value sizes sharply decreased after the fifth singular value was extracted and then slowly stabilize for the remaining ones that mostly should contain only experimental noise

An initial estimation of the concentration profiles for these five proposed components formed during the bioprocess can be estimated by evolving factor analysis [34, 35] of the fused data matrix $\mathbf{D}_{\mathrm{aug}}$ and then used to initialize the alternating least squares optimization in the MCR-ALS procedure. Applied constraints during the ALS optimization were non-negativity (for both concentration and spectra) and spectra normalization. The final model with five resolved components produced lack of fit and percentage of explained variance values of 8.94 % and 99.20 %, respectively.

Although these results were obtained without using any other prior information about the chemical system under study, they already described nicely the main trends followed by the experiment. However, additional knowledge, when existing, could be used to tailor more precisely the sought after pure contribution profiles according to certain known features [39]. This allows improving the shapes of the resolved concentration and spectral profiles and reduces possible rotational ambiguities associated with multivariate curve resolution bilinear decompositions.

In fluorescence spectroscopy, it is known that the shape of the emission spectra of a chemical species is not dependent on the excitation wavelength and vice versa. This feature is consistent with the premises of the fulfillment of the trilinear model [40] for three-way data (multiple data matrices or slices analyzed simultaneously). Therefore, in this work, the application of the MCR-ALS method with the additional constraint of trilinearity is proposed to improve the resolution of emission spectra and the corresponding concentration profiles. This trilinearity constraint was specifically applied only to the augmented emission data matrix

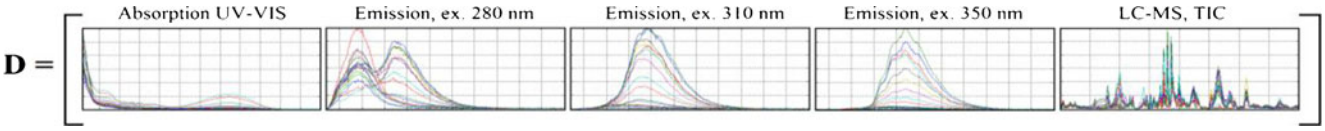

Fig. 2 Arrangement of the fused row-wise augmented data matrix containing the experimental data obtained by UV–VIS absorption, $\mathbf{D}_{\mathrm{UV-VIS}}$, emission at three excitation wavelengths, $\mathbf{D}_{\mathrm{em1}}$, $\mathbf{D}_{\mathrm{em2}}$, $\mathbf{D}_{\mathrm{em3}}$, and LC–MS TIC analysis, $\mathbf{D}_{\mathrm{TIC}}$

formed by the three data matrices with the emission spectra obtained at the three different excitation wavelengths (see Fig. 2), $\mathbf{D}_{\mathrm{emission}} = [\mathbf{D}_{\mathrm{em1}}\ \mathbf{D}_{\mathrm{em2}}\ \mathbf{D}_{\mathrm{em3}}]$. The MCR-ALS resolution of this augmented data matrix aided by the trilinearity constraint assures that the shape of the pure emission spectra of a given component would be the same at the three different excitation wavelengths.

Following the proposed strategy, different MCR-ALS models with three, four, and five components were tested for the simultaneous analysis of the augmented emission data matrix. A model with three components was the most adequate, showing a lack of fit of 5.1 % and a percentage of explained variance of 99.73 %. Basically, from this data analysis, it was possible to infer that the first component, represented by the blue line (in Figs. 3 and 4), showed effective fluorescence when excited at 280 nm. The second component, represented by the green line (in Figs. 3 and 4), displays fluorescence when excited at 280 and 310 nm. And the third component represented by the red line (in Figs. 3 and 4) showed effective fluorescence when excited at 310 and 350 nm. Additionally, the independence of the shape of emission spectra with respect to excitation wavelength was verified.

Using this information, an additional constraint on the spectral profiles was then implemented for improvement of the resolution of all the species profiles during the ALS optimization of the wide row-wise augmented (fused) data matrix, $\mathbf{D}_{\mathrm{aug}}$. This additional constraint implied the elimination of the fourth and fifth spectral components from the emission data submatrices and allowed describing the first three emission spectra with shapes independent of the excitation wavelength. When this model was applied, the lack of fit increased up to 10.77 % and the total explained variance decreased to 98.84 %. Resolution results obtained by application of these spectral constraints on the three emission spectra were more satisfactory and reliable from the chemical point of view. The concentration profiles first obtained by MCR-ALS were then smoothed using a Savitzky–Golay filter with a three-point window and a polynomial of degree one. These concentration profiles are shown in Fig. 3.

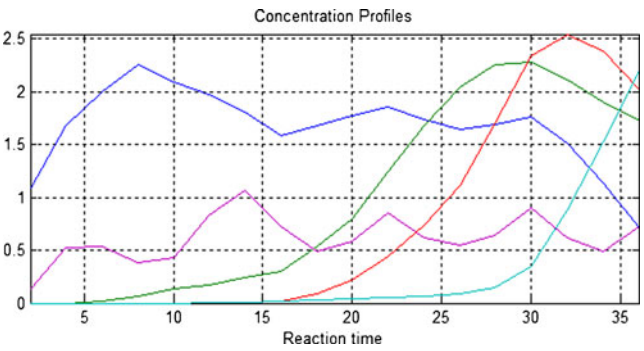

Fig. 3 Smoothed MCR-ALS concentration profiles, **C** (Eqs. 4 and 5), using non-negativity and selectivity constraints

The profile of the first species represented by the blue line displays a rather slow decreasing trend with reaction time. The green line indicates that this second species starts to be produced after 6 h and reaches its maximum concentration after approximately 28 h. The concentration profile represented by the red line shows that a third chemical species starts to be produced after approximately 15 h and reaches its maximum concentration after 32 h. The cyan line profile shows that the concentration of a fourth species increases steadily with reaction time, starting at about 15 h. The unusual shape of the concentration profile of a fifth resolved component represented by the violet line describes a probable background spectral contribution in the samples. The first three concentration profiles (blue, green, and red lines) are similar to those reported in an earlier work using multi-wavelength fluorescence spectroscopy and PARAFAC [27]. The pure spectra resolved for them are shown individually in Fig. 4.

Resolved absorption spectra in Fig. 4a–d should correspond to chemical species having indole groups into their chemical structure because of the presence of the absorption band maximum around 270 nm. Although the absorption spectrum of the third species, 4c, presents two absorption maxima at 270 and 570 nm, its excitation spectrum reported previously [27] is only in agreement with the broad band with vibrational structure located between 310 and 360 nm. Hence, the resolved spectrum for this third component seems to be a sum of pure spectra of two different chemical species. This fact could be justified by the presence of linear dependencies in the concentration profiles of the corresponding species [41]. The absorption spectrum represented in Fig. 4d displays three absorption band maxima at 270, 370, and 570 nm and it was confirmed to be similar to the violacein spectrum [42]. The fifth resolved absorption background spectrum, in Fig. 4e, has an absorption band maximum at 225 nm.

Resolved emission spectra in the analysis of the fused augmented data matrix are consistent with those previously

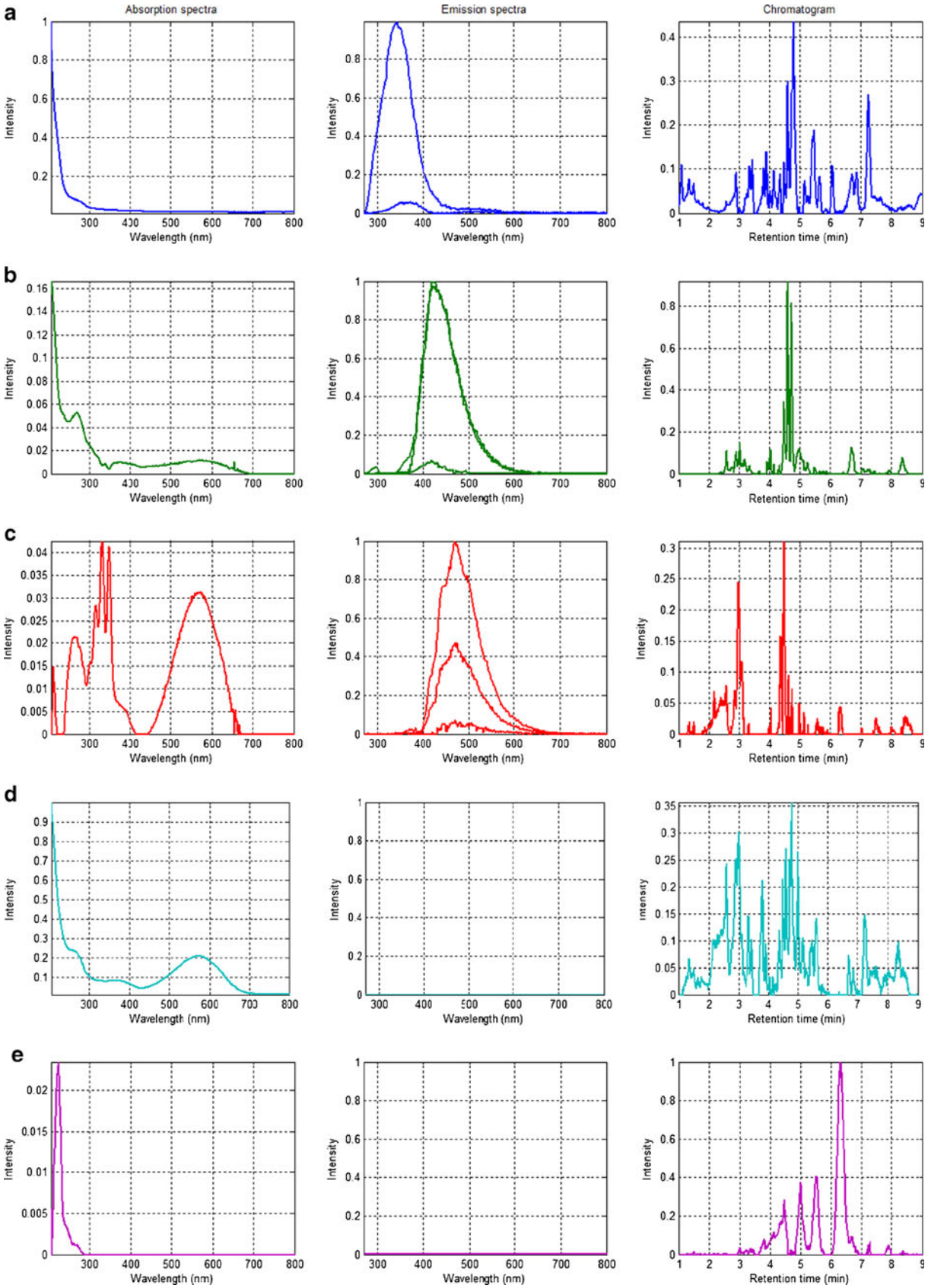

Fig. 4 MCR-ALS resolved pure spectra ($\mathbf{S}^{\mathbf{T}}_{\mathrm{UV-VIS}}$, UV-VIS, $\mathbf{S}^{\mathbf{T}}_{\mathrm{em1}}$, $\mathbf{S}^{\mathbf{T}}_{\mathrm{em2}}$, $\mathbf{S}^{\mathbf{T}}_{\mathrm{em3}}$ emission, and $\mathbf{S}^{\mathbf{T}}_{\mathrm{TIC}}$ TIC chromatograms) of individual components in $\mathbf{S}^{\mathbf{T}}_{\mathrm{aug}}$ augmented spectra matrix (Eqs. 4 and 5)

resolved independently for the augmented emission data matrix with the trilinearity constraint. The emission spectrum of the first component, represented in Fig. 4a, is similar to the pure emission spectrum of tryptophan [27]. The fourth (violacein) and fifth (background) components (Fig. 4d, e) are not fluorescent when excited at 280, 310, and 350 nm (no emission spectra).

Direct interpretation of the resolved TIC chromatographic profiles in Fig. 4 is difficult since multiple peaks are allocated for each of the MCR-ALS resolved components. This leads to the question about the presence of more than one chromatographic peak in the elution profiles resolved for each of the component. It is usually considered that mass spectrometry detection allows for the determination of a larger number of species than molecular absorption and fluorescence spectroscopies, including isomers, especially at lower concentrations. Thus, the distinction between these chemical species can be difficult if these species do not contribute differently and significantly to the observed spectral signal or when they show linear dependencies in their concentration profiles. In these situations, it is understandable that several chromatographic peaks can be allocated somewhere during resolution of each species.

In an attempt to relate chromatographic peaks obtained in the TIC data matrix with the previously resolved five components, a correlation analysis was performed between the whole $\mathbf{D}_{TIC}$ chromatograms matrix (independent variables) and the concentration profiles obtained by MCR-ALS (dependent variables), $\mathbf{C}$ matrix, using MATLAB's corrcoef function. This procedure generated five correlograms (see Electronic Supplementary Material—ESM) that show visually the correlation coefficients between each chromatographic peak in $\mathbf{D}_{TIC}$ and each of the five MCR-ALS resolved concentration profile.

At this point, it was necessary to propose a criterion to identify the MCR-ALS resolved chemical species taking into account both the information on the resolved chromatograms and the previously calculated correlograms. The criterion proposed in this work is based on the cross product (element-by-element) between the MCR-ALS resolved TIC chromatograms (in the last columns of the augmented fused $\mathbf{S}^{\mathbf{T}}_{aug}$ matrix, in Eq. 5, and $\mathbf{S}^{\mathbf{T}}_{TIC}$ in Eq. 4) and the previously calculated correlograms. Thus, the most likely species correspondence is found to be the one that has the highest absolute values in this product. In Fig. S1a, several chemical compounds could be found in the first resolved TIC chromatogram, but using the criterion previously mentioned, the most likely compound is located between 1.08 and 1.12 min. This first species was assigned to be tryptophan according to its emission spectrum (Fig. 4a), and also because it is the initial precursor reagent in Fig. 3 (blue line). In order to confirm whether this species is indeed tryptophan, the mass spectrum of a sample at this retention time was compared with the mass spectrum of a standard solution of L-tryptophan (Fig. 5a, b). As can be seen from this figure, the comparison of the major ions generated by the sample and by the tryptophan standard confirmed the identity of this species, corroborating the mechanism presented in Fig. 1 where this compound is the precursor of violacein biosynthesis

In Fig. S1b, it is possible to see that the pure TIC chromatogram resolved for the second component and the corresponding correlogram with the TIC raw data matrix, indicate that the most likely retention times for this component are located between 4.55 and 4.78 min. Two compounds are observed in this range and their mass spectra are shown in Fig. 5c, d. Major ions at *m/z* 313, 339, 392, 436, 454, 480, and 907 can be detected in the spectrum represented in Fig. 5c. The monitoring of the ion fragmentation at *m/z* 454 and 907 with a second MS analyzer allowed inferring that the ion at *m/z* 907 is the dimeric form of the protonated molecular ion ($[M+H]^+$) at *m/z* 454. Hence, the molar mass of this species is deduced to be 453 u, indicating an odd number of nitrogen atoms in its chemical structure. Additionally, the ion at *m/z* 313 is derived from $[M+H]^+$ and the ions at *m/z* 436 and 392 were assigned to $[M+H-H_2O]^+$ and $[M+H-H_2O-CO_2]^+$, respectively. The ions at *m/z* 339 and 480 are not derived from $[M+H]^+$ suggesting that another compound is coeluting with this one. All compounds shown in Fig. 1 have molar masses less than 400 u. Since the compound identified above has a molar mass of 453 u, it seems to have originated from another biosynthetic pathway. Ryan and Drennan [43], in fact, have shown that divergent pathways are possible in the biosynthesis of bisindole natural products, such as violacein. Therefore, it is proposed that this compound is a staurosporine analog, derived from chromopyrrolic acid, with a structure similar to the K252b alkaloid. The *m/z* spectrum shown in Fig. 5d presents a major ion at *m/z* 339. The monitoring of the fragmentation of ions 480 and 959 with a second MS analyzer allowed to infer that the ion at *m/z* 959 was a dimeric form of the protonated molecular ion ($[M+H]^+$) at *m/z* 480. The molar mass of this species is 479 u. The ion at *m/z* 462 was assigned to be $[M+H-H_2O]^+$. Additionally, the ion at *m/z* 339 is derived from $[M+H]^+$.

In Fig. S1c, it is possible to see that retention times most likely for the third component are located between 4.44–4.54 and 2.94–3.04 min. In these regions, two compounds are detected. The first one, represented by the mass spectrum obtained at 4.47 min, is shown in Fig. 5e, and the major ions are at *m/z* 313, 325, 425, 466, and 932. The monitoring of the fragmentation of ions 466 and 931 with a second MS analyzer allowed inferring that the ion at *m/z* 931 is a dimeric form of a protonated molecular ion ($[M+H]^+$) at *m/z* 466. Hence, the molar mass of this species is 465 u. The major ion at *m/z* 313 is not derived from $[M+H]^+$, suggesting that it coelutes with the species having mass 453 (Fig. 5c). MS species spectra

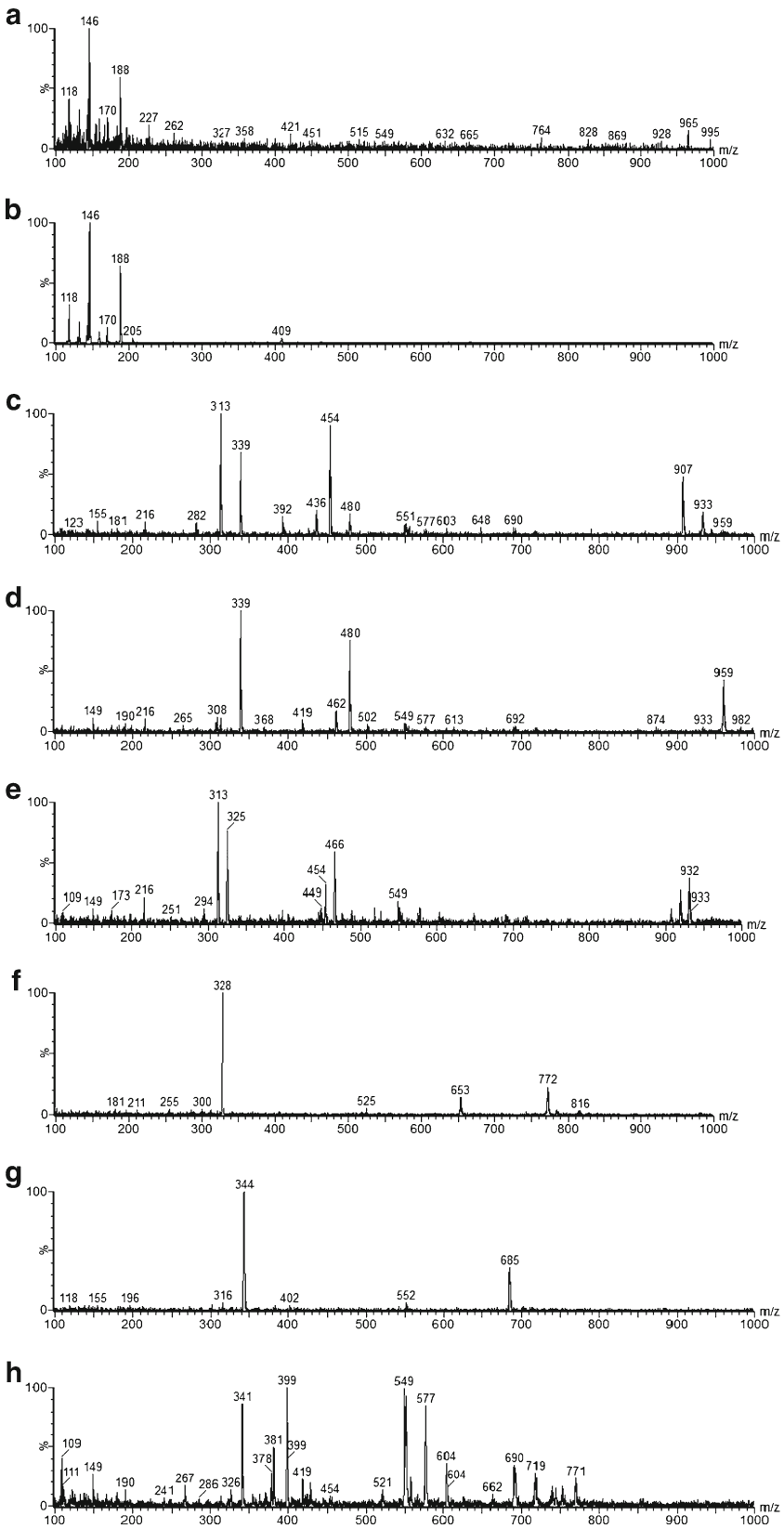

Fig. 5 Electrospray ionization mass spectra: **a**) crude ethanolic extract of *Chromobacterium violaceum* at 1.1 min, **b**) standard solution of L-tryptophan, **c**) at 4.60 min, **d**) at 4.73 min, **e**) at 4.47 min, **f**) at 3.01 min, **g**) at 2.61 min, and **h**) at 6.33 min

represented in Fig. 5c–e differ by 26, 14, and 12 mass units, and the loss of a fragment with *m*/*z* 141 is common for the three species, explicitly, (454–313 *m*/*z*, 480–339 *m*/*z*, and 466–325 *m*/*z*). This may be indicative that these three compounds are intermediates in the biosynthesis of other indolocarbazoles. The second compound represented by the mass

spectrum obtained at 3.01 min, is presented in Fig. 5f. The ion at *m*/*z* 328 was identified as the protonated molecular ion ($[M+H]^+$) and the respective molar mass is 327 u. This species is assigned to be deoxyviolacein.

This result corroborates the discussion about the resolved UV–VIS absorption spectrum in Fig. 4c, which was interpreted as a linear combination of different pure species spectra having linearly dependent concentration profiles. Hence, the band with vibrational structure located between 310 and 360 nm can be associated to a compound with molar mass 465 u and the peaks at 280 and 570 nm can be related to a compound with molar mass 327 u, with a structure similar to violacein.

In Fig. S1d, it is possible to see that both the resolved TIC profile and the correlogram indicate that the retention times most likely for the fourth resolved component are located between 2.94–3.04 and 2.54–2.66 min. Two compounds are observed in this range. Their mass spectra are shown in Fig. 5f, g. The first compound represented by the mass spectrum obtained at 3.01 min and presented in Fig. 5f is the same compound discussed above with molar mass 327. The second compound, characterized by the mass spectrum obtained at 2.61 min, is shown in Fig. 5g. The ion at *m*/*z* 344 is recognized as the protonated molecular ion ($[M+H]^+$) of violacein that has a mass of 343 u and molecular composition $C_{20}N_3O_3H_{13}$. The structural difference between them is 16 mass units, which corresponds to the absence of an oxygen atom in the first compound.

In Fig. S1e, it is possible to infer that the retention time for the fifth MCR-ALS resolved component is most likely located at 6.1–6.5 min. In this region, one peak can be visualized. Several ions at *m*/*z* 341, 381, 399, 549, and 577 can be seen in the mass spectrum (Fig. 5h) of this compound which hampers the identification of the protonated molecular ion.

Conclusions

The fermentation of *C. violaceum* conducted in a bioreactor was investigated by analytical and chemometric techniques to explore the information associated to violacein biosynthesis. In this context, different chemical compounds were detected during the evolution of the biofermentation reaction and some of them were identified by the proposed methodology. Five components could be resolved independently by MCR-ALS, although more chemical species could be formed during the reaction and not be directly resolved due to possible linear dependencies among their concentration profiles. Three of these chemical compounds were related to the violacein biosynthesis, namely, tryptophan, deoxyviolacein, and violacein, which were clearly identified from their UV–VIS absorption and/or emission spectra and from their mass spectra. Another three additional chemical species, with molar masses of 453, 465, and 479 u, were assigned as probable intermediates in the biosynthesis of indolocarbazoles, indicating the existence of divergent pathways in the biosynthesis of the bisindole natural product. This last compound presented an UV absorption maximum at 225 nm and its mass spectrum was not conclusive to infer its identity.

Finally, it should be emphasized that the proposed methodology of MCR-ALS analysis of fused spectrometric and chromatographic data has been shown to be suitable for the in vivo exploration of biosynthetic pathways of natural products. This methodology is a powerful tool able to achieve deep insights of complex fermentation bioprocesses, like the one investigated herein about the violacein biosynthesis.

Acknowledgments The authors acknowledge CNPq's sandwich doctorate program for financial support and Prof. Carol H. Collins for English revision.

References

1. Duran N, Erazo S, Campos V (1983) An Acad Bras Ciencias 55:231–234
2. Duran N, Menck CF (2001) Crit Rev Microbiol 27:201–222
3. Lichstein HC, Van de Sand VF (1945) J Infect Dis 76:47–51
4. Duran N, Antonio RV, Haun M, Pilli RA (1994) World J Microbiol Biotechnol 10:686–690
5. Leon LL, Miranda CC, de Souza AO, Duran N (2001) J Antimicrob Chemother 48:449–450
6. Melo PS, Justo GZ, de Azevedo MB, Duran N, Haun M (2003) Toxicology 186:217–225
7. Saraiva VS, Marshall JC, Cools-Lartigue J, Burnier MN Jr (2004) Melanoma Res 14:421–424
8. Ferreira CV, Bos CL, Versteeg HH, Justo GZ, Duran N, Peppelenbosch MP (2004) Blood 104:1459–1464
9. Andrighetti-Frohner CR, Antonio RV, Creczynski-Pasa TB, Barardi CR, Simoes CM (2003) Mem Inst Oswaldo Cruz 98:843–848
10. Tobie WC (1934) J Bacteriol 29:223–227
11. DeMoss RD, Evans NR (1959) J Bacteriol 78:583–788
12. DeMoss RD, Evans NR (1960) J Bacteriol 79:729–733
13. Hoshino T, Kondo T, Uchiyama T, Ogasawara N (1987) Agric Biol Chem 51:965–968
14. Hoshino T, Takano T, Hori S, Ogasawara N (1987) Agric Biol Chem 51:2733–2741
15. Hoshino T, Ogasawara N (1990) Agric Biol Chem 54:2339–2346
16. Pemberton JM, Vicent KM, Penfold RJ (1991) Curr Microbiol 22:355–358
17. Hoshino T, Kojima Y, Kayashi T, Uchiyama T, Kaneko K (1993) Biosci Biotechnol Biochem 57:775–781
18. Hoshino T, Hayashi T, Odajima T (1995) J Chem Soc Perkin Trans 1:1565–1571
19. Hoshino T, Yamamoto M (1997) Biosci Biotechnol Biochem 61:2134–2136
20. Ruhul Momen AZM, Hoshino T (2000) Biosci Biotechnol Biochem 64:539–549
21. August PR, Grossman TH, Minor C, Draper MP, MacNeil IA, Pemberton JM, Call KM, Holt D, Osburne MS (2000) J Mol Microbiol Biotechnol 2:513–519
22. Howard-Jones AR, Walsh CT (2005) Biochemistry 44:15652–15663

23. Sánchez C, Braña AF, Méndez C, Salas JA (2006) ChemBioChem 7:1231–1240
24. Balibar CJ, Walsh CT (2006) Biochemistry 45:15444–15457
25. Shinoda K, Hasegawa T, Sato H, Shinozaki M, Kuramoto H, Takamiya Y, Sato T, Nikaidou N, Watanabe T, Hoshino T (2007) Chem Commun 28:4140–4142
26. Jiang P-X, Wang H-S, Zhang C, Lou K, Xing X-H (2010) Appl Microbiol Biotechnol 86:1077–1088
27. Dantas C, Volpe PLO, Duran N, Ferreira MMC (2012) J Braz Chem Soc (in press)
28. Perá-Trepat E, Tauler R (2006) J Chromatogr A 1131:85–96
29. Navea S, Tauler R, de Juan A (2006) Anal Chem 78:4768–4778
30. Cutler P, Gemperline PJ, de Juan A (2009) Anal Chim Acta 632:52–62
31. Eilers PHC (2004) Anal Chem 76:404–411
32. Tauler R, Kowalski B (1993) Anal Chem 65:2040–2047
33. Tauler R, Smilde A, Kowalski B (1995) J Chemom 9:31–58
34. Gampp H, Maeder M, Meyer CH, Zuberbuhler AD (1985) Talanta 32:1133–1139
35. Gampp H, Maeder M, Meyer CH, Zuberbuhler AD (1986) Talanta 33:943–951
36. Tauler R, Izquierdo-Ridorsa A, Casassas E (1993) Chemom Intell Lab Syst 18:293–300
37. de Juan A, Maeder M, Martínez M, Tauler R (2000) Chemom Intell Lab Syst 54:123–141
38. Jaumont J, Gargallo R, de Juan A, Tauler R (2005) Chemom Intell Lab Syst 76:101–110
39. de Juan A, Tauler R (2003) Anal Chim Acta 500:195–210
40. Tauler R, Marques I, Casassas E (1998) J Chemom 12:55–75
41. Amrhein M, Srinivasan B, Bonvin D, Schumacher MM (1996) Chemom Intell Lab Syst 33:17–33
42. Rettori D, Durán N (1998) World J Microbiol Biotechnol 14:685–690
43. Ryan KS, Drennan CL (2009) Chem Biol 16:351–364