J Ind Microbiol Biotechnol
DOI 10.1007/s10295-016-1865-0

FERMENTATION, CELL CULTURE AND BIOENGINEERING - ORIGINAL PAPER

High-throughput FTIR-based bioprocess analysis of recombinant cyprosin production

Pedro N. Sampaio[1] · Kevin C. Sales[2] · Filipa O. Rosa[2] · Marta B. Lopes[2,3,4] · Cecília R. C. Calado[4]

Received: 14 May 2016 / Accepted: 30 October 2016

Abstract To increase the knowledge of the recombinant cyprosin production process in *Saccharomyces cerevisiae* cultures, it is relevant to implement efficient bioprocess monitoring techniques. The present work focuses on the implementation of a mid-infrared (MIR) spectroscopy-based tool for monitoring the recombinant culture in a rapid, economic, and high-throughput (using a microplate system) mode. Multivariate data analysis on the MIR spectra of culture samples was conducted. Principal component analysis (PCA) enabled capturing the general metabolic status of the yeast cells, as replicated samples appear grouped together in the score plot and groups of culture samples according to the main growth phase can be clearly distinguished. The PCA-loading vectors also revealed spectral regions, and the corresponding chemical functional groups and biomolecules that mostly contributed for the cell biomolecular fingerprint associated with the culture growth phase. These data were corroborated by the analysis of the samples' second derivative spectra. Partial least square (PLS) regression models built based on the MIR spectra showed high predictive ability for estimating the bioprocess critical variables: biomass ($R^2 = 0.99$, RMSEP 2.8%); cyprosin activity ($R^2 = 0.98$, RMSEP 3.9%); glucose ($R^2 = 0.93$, RMSECV 7.2%); galactose ($R^2 = 0.97$, RMSEP 4.6%); ethanol ($R^2 = 0.97$, RMSEP 5.3%); and acetate ($R^2 = 0.95$, RMSEP 7.0%). In conclusion, high-throughput MIR spectroscopy and multivariate data analysis were effective in identifying the main growth phases and specific cyprosin production phases along the yeast culture as well as in quantifying the critical variables of the process. This knowledge will promote future process optimization and control the recombinant cyprosin bioprocess according to Quality by Design framework.

Keywords Cultivation · High-throughput analysis · Mid-infrared spectroscopy · Partial least square regression · Principal components analysis · Recombinant cyprosin

Abbreviations

IR	Infrared
LOOCV	Leave-one-out cross validation
MIR	Mid-Infrared
MSC	Multiplicative scattering correction
NIR	Near-Infrared spectroscopy
PCA	Principal component analysis
PLS	Partial least squares
RMSECV	Root mean square error cross validation
RMSEP	Root mean square error prediction
SNV	Standard normal variate

Introduction

The aspartic protease cyprosin, extracted from *Cynara cardunculus* flowers, presents interesting properties for the dairy industry due to its high clotting activity and

✉ Pedro N. Sampaio
pnsampaio@gmail.com

1 Faculty of Engineering, Lusophone University of Humanities and Technology, Campo Grande, 376, 1749-019 Lisbon, Portugal

2 Faculty of Engineering, Catholic University of Portugal, Avenida Otávio Pato, 2635-631 Rio de Mouro, Portugal

3 Institute of Telecommunications, Technical Higher Institute University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

4 ISEL-Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Conselheiro Emídio Navarro Street, 1959-007 Lisbon, Portugal

Published online: 10 November 2016

specific proteolytic properties [1–3]. To obtain an efficient and low-cost production system of this enzyme, the cyprosin gene was cloned in *Saccharomyces cerevisiae* cells and its production was optimized concerning the expression system (i.e, recombinant plasmid and *S. cerevisiae* strain as cell host), the media composition and cultivation strategy [2, 4–6]. To further characterize, optimize, and control the cyprosin production process along *S. cerevisiae* culture, it is now highly relevant to implement rapid, economic, and simultaneously sensitive monitoring techniques.

To optimize the bioprocess and to efficiently control large-scale productions, biotechnological companies are encouraged to increase the knowledge of the cultivation process based on the development of new analytical tools, enabling the acquisition of a large amount of data from different experimental conditions. This knowledge will lead to a faster and more efficient process development as the ones related to scale-up trajectories and, consequently, will decrease the overall costs while enabling efficient control procedures [7–9]. Indeed, for new biotech products, both the Food Drug Administration (FDA) and the European Medicine Agency (EMA) encourage the development and implementation of effective and efficient innovative monitoring technologies, especially along the bioproduct manufacturing. In view of this guidance, process analytical technology (PAT) has been applied to bioprocesses. The PAT initiative is defined as a system that allows monitoring and controlling all variables that affect the process performance or critical quality attributes of the final product, to ensure reproducible of the manufacturing processes and a high quality and regulatory compliant product [10]. Clear value is also given by the possibility of identifying knowledge gaps, experimental flaws, and uncontrolled variances. Therefore, the interest for bioprocess monitoring using sensitive, rapid, and economic technologies, such as infrared spectroscopy, is increasing, comparatively to the classical, labor-intensive, and time-consuming off-line analyses.

Infrared spectroscopy is a highly sensitive vibrational spectroscopic technique that allows acquiring a huge amount of information in a fast, economic, and automatable mode, with minimal or no sample preparation and reagent consumption, enabling an accurate analysis of all critical variables of the bioprocess from a single measurement. Near Infrared (NIR) (14,000–4000 cm^{-1}) and Mid-Infrared (MIR) (4000–400 cm^{-1}) spectroscopies have been applied as key methods for monitoring of cell cultivations (i.e, process monitoring), to evaluate batch reproducibility (i.e, process supervision), overall process, or plant supervision (i.e, PAT) and to generally increase the bioprocess knowledge [8, 9, 11, 12]. Recently, NIR spectroscopy, using an in situ sterilizable fiber optic probe associated with multivariate data analysis, allowed the real-time prediction with high accuracy of all critical variables of the recombinant cyprosin production, allowing reducing the high number of offline techniques needed to fully characterize the key variables of the process while minimizing the risk of culture contamination [13]. However, MIR spectroscopy presents advantages over NIR spectroscopy, as it covers the fundamental vibrations of most common chemical bonds in biomolecules, whereas the NIR region covers combinations and overtones of fundamental vibrations. NIR spectra present lower adsorption coefficients and broader overlapping bands, which turn more difficult the extraction of the sample molecular composition, compared to the MIR spectra [9, 12, 14]. Furthermore, in biological cells, while NIR captures mainly from groups containing the hydrogen atom (C–H, N–H, O–H, and S–H), MIR also captures C–N, P–O, C=C, C=O, and C=N groups [9]. MIR spectra can be acquired in automatic high-throughput mode, using plates with multi-micro wells, which enables the simultaneous analysis of hundred samples at once, requiring only the culture broth dehydration before analysis [15–17]. The possibility of high-throughput spectral acquisition by MIR spectroscopy also represents an advantage for optimization protocols, due to possible space constraints or high costs associated with the use of multi-fiber optic probes for multi-bioreactors.

Independently of the IR spectral region (NIR or MIR), the huge amount of data acquired by IR spectroscopy requires the use of multivariate data analysis to extract as much relevant information as possible from the spectra and to transform these data into important bioprocess information [9, 18]. The most common techniques applied to bioprocess monitoring, due to their efficiency in retrieving information associated with the simplicity of application and interpretability, are principal component analysis (PCA), for qualitative analysis by capturing data trends, and partial least squares (PLS) regression, for quantitative estimation of critical process variables. PCA reduces the dimensionality of the data by creating new variables, called principal components (PC), which are linear combinations of the original variables that can be displayed to reveal patterns of similarity between the observations as points in maps (score plots) [19]. PLS regression predicts a set of dependent variables from a group of independent variables or predictors, by extracting from the predictors a series of orthogonal factors, called latent variables (LV) [20].

The aim of the present work is to implement a rapid high-throughput method based on MIR spectroscopy associated with multivariate data tools to analyze the recombinant cyprosin production by a transformed *S. cerevisiae* strain in bioreactor. The MIR spectra were acquired from culture samples at a microliter scale, using plates with

multi-wells, after a short preprocessing step of dehydratation. PCA of MIR spectra was performed to identify trends in the spectral data which are related to the metabolism of the carbon sources (galactose and glucose) and by-products formation (ethanol and acetate) along the cell host culture. PLS regression models were built to accurately predict all critical variables of the bioprocess, namely, biomass, cyprosin activity, glucose, galactose, ethanol, and acetate. This way, the high informative data retrieved along the culture by MIR spectroscopy in a rapid, economic and high-throughput mode may strongly promote cyprosin bioprocess knowledge and consequently optimization and control, according to the PAT initiative.

Materials and methods

Cell culture conditions

The *S. cerevisiae* BJ1991 strain previously transformed with an expression cassette pCAF15 that contains the *CYPRO11* gene that codifies the aspartic protease cyprosin was used [21]. The inoculum was grown in 150 ml (10% of the total volume of the culture medium in bioreactor) in YNB medium [0.67% (w/v) yeast nitrogen base without amino acids and 2% (w/v) glucose, supplemented with auxotrophic markers (70 mg l^{-1} L-tryptophan and 20 mg l^{-1} uracil)] during 16 h at 30 °C and 200 rpm. The inoculum was transferred into the bioreactor containing the expression medium YPGal-Glu [1% (w/v) yeast extract; 2% (w/v) bacto peptone; 2% (w/v) galactose; and 2% (w/v) glucose]. Cultivations were performed in a vessel with 2 l (Biostat MD, B. Braun, Germany) with a working volume of 1.5 l, at pH 5.5 ± 0.1 and 30.0 ± 0.1 °C. The dissolved oxygen concentration (DOC) was controlled at approximately 25% of dissolved oxygen in the culture medium in relation to the air saturation, by controlling the stirring speed of two Rushton turbines, and maintaining a constant flow rate at 1.5 l min^{-1} (equivalent to 1 VVM).

Samples and analytical reference methods

Samples were collected from the bioreactor at regular time intervals (30 min). The biomass concentration was determined by optical density measured at 600 nm ($OD_{600\ nm}$), using a spectrophotometer (Shimadzu UV-1603, Japan). These absorbance values were converted to dry cell weight (dcw) using a previous standard curve relating $OD_{600\ nm}$, and the weight of cell samples dried at 80 °C until constant value. The cyprosin enzymatic activity was determined according to Heimgartner et al. (1990), using fluorescein isothiocyanate-labeled *k*-casein (FITC-casein). One unit of proteolytic activity was defined as the amount of enzyme that increases one unit of emitted fluorescence at 525 nm (excitation at 495 nm), after 60 min of hydrolysis at 37 °C, pH 5.1. The fluorescence intensity was measured using a spectrofluorimeter (Shimadzu RF-1501, Japan). This method specifically measures cyprosin activity according to Sampaio et al. [6] that observed by western blot analysis using antibodies against cyprosin and that the secretion of other proteases by the present *S. cerevisiae* expression system is negligible. The protein concentration was determined according to the Bradford method using Coomassie reagent [22]. Specific cyprosin activity was defined as the cyprosin activity in relation to the total protein. Galactose, glucose, ethanol, and acetate were analyzed by high pressure liquid chromatography (HPLC), using an Aminex (Bio-Rad Laboratories, USA) column assembled to an HPLC system (Merck, Germany), at 50 °C with a 5 mM H_2SO_4 at 0.6 ml min^{-1} as eluent.

IR spectroscopy measurements

IR spectral acquisition

Broth samples taken along the cultivation were centrifuged, and the cell pellets were stored at −20 °C. For analysis, the cell pellets were defrosted and resuspended in 0.9% (w/v) NaCl to obtain an $OD_{600\ nm}$ of 6.0. After that, 30 microliters of the cellular suspension were placed on IR transparent Si microtiter plates with 96 wells (Bruker Optics, Germany), and subsequently dehydrated for 2.5 h in a vacuum desiccator (ME2, Vaccubrand, Germany). Each culture sample was analyzed in triplicate, i.e., three MIR spectra from three independent 30 microliters samples from each cellular suspension were acquired. The MIR spectra were obtained by averaging 40 scans in the spectral window between 4000 and 400 cm^{-1}, using a HTS-XT associated with Vertex-70 spectrometer (Bruker Optics, Germany) at a spectral resolution of 4 cm^{-1}.

Spectral preprocessing

Several spectral preprocessing techniques were evaluated to minimize physical disturbances from spectra or enhance their information content, namely: baseline correction by offset or strait line subtraction, min–max normalization, multiplicative scatter correction, and first and second spectral derivatives [23]. Before spectral derivatives, a Savitzky–Golay smoothing filter was applied (using a filter window of 15 points and a second-order polynomial fit) [24]. Baseline correction was performed using the OPUS® software (Bruker Optics, Germany). All the remaining analyses were performed using MATLAB® 7.9.0.

Multivariate analysis

Principal component analysis (PCA)

Principal component analysis is a linear pattern recognition technique allowing the reduction of the dimensionality of multivariate data to *n* principal components. The first principal component (PC1) represents the direction of the largest variation in the data set, and the nth principal component (PCn) describes the direction of the nth-largest variation in the data set. Each principal component relates to an independent sample variation pattern and an independent variable variation pattern, which can be displayed in score plots and correlation loading plots, respectively [25]. In the score plots, usually, scores of two components are plotted as scatter plots at a time. In the correlation loading plots, correlation between variables and scores is displayed. When scores and correlation loading plots are studied together for the same pair of components, variable variation patterns can be directly related to sample variation patterns for the respective components [26]. All samples replicates were considered for analysis to enable inferring how sample variability may affect possible trends inferred from the direct observation of the scores plot. A data matrix composed of 195 analyzed samples (i.e., spectra) represented by 1815 variables (i.e., wavenumbers) was taken for PCA. PCA was performed using MATLAB® 7.9.0.

Partial least square (PLS) regression

PLS regression computes linear combinations between spectral data and reference values extracting the scores that are most correlated and explain the maximum covariance between them [27]. The PLS method was used for predicting the concentration of the critical variables of the bioprocess from the spectral data, such as biomass, enzymatic activity, glucose, galactose, ethanol, and acetate. The validation of the developed PLS model was conducted using: an independent test set validation for the variables analyzed composed by randomly chosen samples from the entire data set, not used for model calibration; a leave-one-out cross validation (LOOCV) was used for predicting glucose, due to the reduced number of samples available. LOOCV works by taking one sample from the original data set and using it as a validation set, cross-validation method was used to construct the best model.

Several PLS models were built based on the combination of the preprocessing techniques and wavelength selection as follows: the spectral region was divided into ten equal sub-regions, and the best combination of spectral regions and preprocessing techniques providing the best predictive model performance was chosen. The predictive performance of each PLS model was assessed the number of latent variables (LV) and the correspondent data variance explained, regression coefficient (R^2), and the root mean square error of prediction (RMSEP) or the leave-one-out cross-validation error (RMSECV).

Results and discussion

MIR spectroscopy was applied to monitor recombinant cyprosin production process along *S. cerevisiae* BJ1991 culture in bioreactor, in an automatic mode with high-throughput instrumentation (using multi-microwell with 96 well silicon plate), enabling this way a high-frequency acquisition of data containing cellular metabolic information, such as transient signals, that cannot be captured otherwise. The cultivation of this expression system was previously optimized in terms of inoculum, culture medium composition, expression system, pH, and aeration [4–6]. To better understanding how PCA of spectral data enables capturing in a highly sensitive way the cell metabolic shifts along the culture, a brief description of the culture is provided next.

The culture was subdivided in six main growth phases, all characterized by the consumption of a specific carbon source, specific growth rate, and specific cyprosin production yield per biomass (Fig. 1; Table 1). During the first 3 h (defined as growth phase I, and corresponding to samples S1–S7), a growth lag phase was observed, corresponding to the cells' adaptation period to the new culture environment and, consequently, with no cell growth and glucose consumption (Fig. 1a, b; Table 1). In the presence of two main carbon sources, glucose was preferentially consumed over galactose (growth phase II, corresponding to samples S8–S18; 3:20–8:20 h), leading to a maximum specific growth rate on glucose of 0.338 h^{-1} (Fig. 1a, b). Glucose was metabolized first, due to carbon catabolic repression of *GAL* genes transcription involved in galactose uptake [28]. The high glucose uptake produced fundamentally biomass (3 g dcw l^{-1}), with a residual proteolytic activity per biomass of 80.3 U mg^{-1} dcw being registered (Fig. 1d), which is not cyprosin but other endogenous proteases in accordance with previous studies [6, 29]. Glucose was completely depleted after 7 h of cultivation, and despite constant aeration (1 VVM) and high DOC (>25%), ethanol was produced reaching a maximum concentration of 11.0 g l^{-1} (Fig. 1a, b), mostly due to overflow metabolism (i.e, Crabtree effect) [30]. Only when glucose reached a critical level, the *GAL* gene cluster was desrepressed and the galactose started to be consumed (growth phase III, corresponding to samples S19–S29 and 8.50–13.50 h), (Fig. 1b; Table 1). During this period, a significant increase of ethanol production was also observed, reaching the maximum value of 22.9 g l^{-1} (corresponding

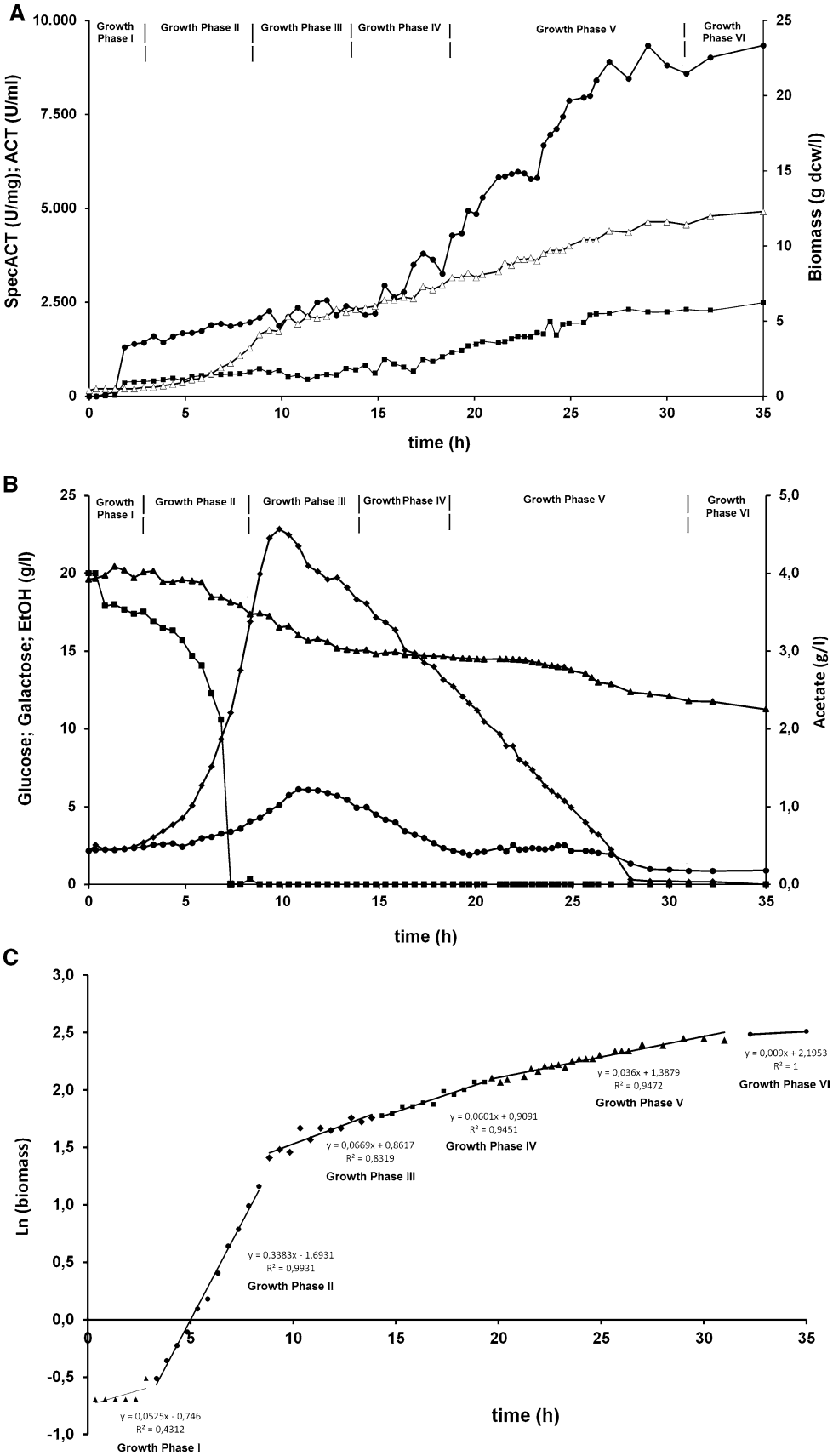
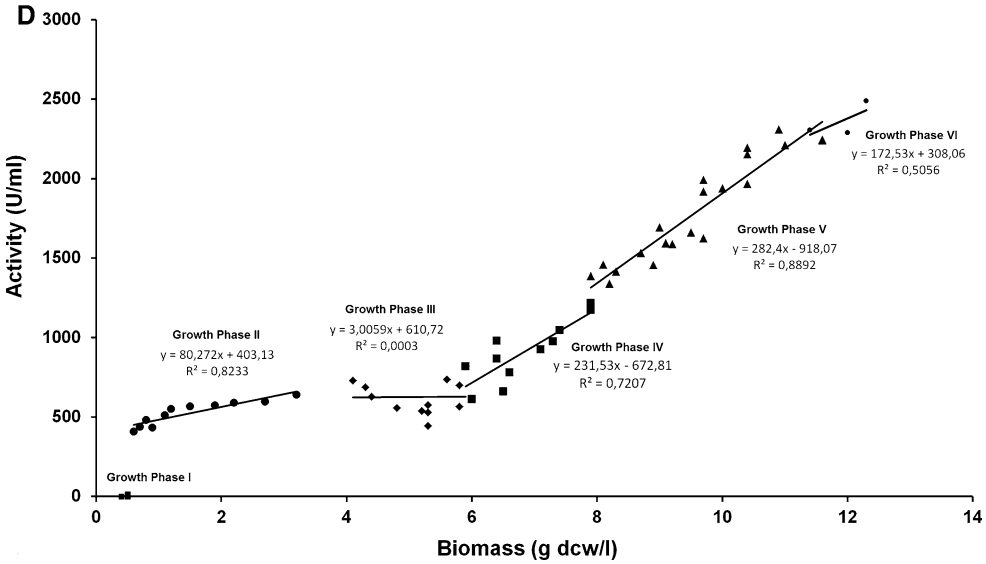


Fig. 1 Cultivation of transformed *S. cerevisiae* BJ1991 strain: **a** Off-line analysis concerning biomass (continuous *line triangle*); cyprosin activity (continuous *line square*) and specific cyprosin activity (continuous *line circle*); **b** Off-line analysis concerning carbon source taken up [glucose (continuous *line square*) and galactose (continuous *line triangle*)] and the metabolites produced along the culture [ethanol (*continuous line—diamond*) and acetate (*continuous line circle*)]; **c** Maximum specific growth rate observed along the cultivation as indicated by the slope of the linear regression; **d** Cyprosin production along biomass growth

Fig. 1 continued



to sample S21). High glucose and galactose concentrations lead to a high cellular glycolytic flux, which results in a setup of pyruvate that exceeds the oxidative capacity of the strain and consequently leads to the overflow metabolism and the ethanol formation for re-oxidizing the cofactor NADH relevant to ATP production [31]. This implied a low biomass yield (0.087 h^{-1}) and a lower recombinant protein expression per biomass (3 U mg^{-1} dcw) (Fig. 1a, c, d; Table 1). The high ethanol production was followed by acetate production that reached its maximum concentration (of 1.2 g l^{-1}) at sample S23, due activation of biosynthetic pathways in yeast [32]. The main carbon sources of growth phase IV were ethanol and acetate (corresponding to samples S30–S40, 14:20 h–19:20 h) (Fig. 1a, b; Table 1). On that period, a low cellular growth rate (0.060 h^{-1}) and a high specific activity production (231.5 U mg^{-1} dcw) were observed (Fig. 1b, c, d; Table 1). During the growth phase V (defined by samples S41–S63; 19.40–30:00 h), the remaining ethanol was totally consumed, being associated with a significant specific cyprosin production per biomass (282.4 U mg^{-1} dcw) (Fig. 1a–d; Table 1). In the following 5 h (designated as growth phase VI), which correspond to the group of samples (S64 and S65), the cultivation reached the stationary phase and a maximum of the proteolytic activity (2489 U ml^{-1}), as consequence of the exhaustion of the unique carbon source then available, the inductor (galactose).

PCA to capture the molecular fingerprint of samples associated with cultivation phases

A PCA of spectra acquired along cultivation was performed. To minimize spectral data noise, while highlighting spectral features, two preprocessing techniques were evaluated: multiplicative scatter correction (MSC), for minimizing the effect of physical phenomena, such the light scattering resulting from the existence of particles with different sizes and shapes; and derivatives, to remove baseline artifacts, while highlighting overlapping peaks. Based on the analysis of the PCA score plots, it was observed that replicate samples were more close to each other and spatially separated from the remaining samples when first derivative was used instead of MSC (data not shown), resulting in two principal components (PCs) accounting for 96.3% of the variance in the data explained (Fig. 2a). As in general, replicates were together and apart from other samples, and major data sets composed by samples collected in similar culture times were observed (as represented by bold circles in Fig. 2a), and it may be concluded that MIR spectroscopy captured both the culture samples' molecular fingerprint and the biomolecular similarities between samples. The PCA (Fig. 2a) score plots also showed clusters based on the six growth phases, as summarized in Table 1.

Besides enabling the visualization of general data patterns, PCA is also useful to identify sample outliers, when samples are detached from the cluster composed of the remaining sample replicates and other samples taken in similar culture phase, as isolated samples S10, S11, S23, S35, and S42 (Fig. 2a). The main consequence of these observations was that those replicates were classified as outliers and were not considered on the subsequent PLS model building, in contrary to the remaining duplicates of these samples that were used in the PLS model building and validation.

As mentioned before, the PCA (Fig. 2a) score plots showed clusters based on the six growth phases, as follows: The group constituted by samples from the lag phase (S1–S7) (growth phase I) is localized in the bottom left

Table 1 Characteristics of the six growth phases of the recombinant S. *cerevisiae* batch culture

	Lag growth phase	Glucose consumption phase	Galactose consumption phase with ethanol and acetate production	Simultaneous galactose, ethanol and acetate consumption phase	Galactose and final ethanol consumption phase	Stationary growth phase (galactose consumption)
Growth phase	I	II	III	IV	V	VI
Culture period (h)	0–2:50	3:20–8:20	8:50–13:50	14:20–20:00	20:25–30:00	31:00–35:00
Culture samples	S1–S7	S8–S18	S19–S29	S30–S40	S41–S62	S63–S65
Maximum specific growth rate (h^{-1})	0.053	0.338	0.067	0.060	0.041	0.017
Cyprosin production per biomass (U mg^{-1} dcw)	–	80.3	3	231.5	282.4	172.5

side of the score plot (Fig. 2a; Table 1). The growth phase II (S8–S18) showed positive PC1 values as opposed to the previous group (phase I), corresponding to a high glucose consumption rate with a consequent high specific growth rate (0.338 h^{-1}), and decreases PC2 values. The cluster constituted by samples (S19–S29) (growth phase III), corresponds to the galactose consumption period as the main carbon source, with maximum ethanol (22.9 g l^{-1}) (sample S21) and acetate concentrations (1.2 g l^{-1}) (sample S23) being registered. Samples corresponding to ethanol and acetate consumption (S30–S40, growth phase IV) showed increasing PC1 values along the culture time (Fig. 2a). The sample group S41–S63 (growth phase V), associated with a high activity-specific production (282.4 U mg^{-1}), is also grouped in a defined score plot region (Fig. 2a; Table 1). The cluster defined by (S64–S65; 31:00–35:00 h) (growth phase VI) represents the stationary phase of growth, characterized by high biomass and cyprosin activity (high PC2 value). The following patterns of samples' clustering were also observed: cells preferentially up taking glucose are positioned in the score plot quadrant with PC1 <0 and PC2 <0; while cells up taking other carbon source are presented at the quadrant with PC1 >0; samples with a low enzymatic activity are positioned in the score plot with PC2 <0, while cells characterized by high activity are localized in the quadrant with PC2 >0.

The PC1 and PC2 loading vectors represent the MIR spectral regions that mostly contributed to the relations between the original variables and the new PC variables (Fig. 2b). A high loading value (positive or negative) corresponds to a high contribution of the spectral region to the respective PC, while a loading value close to zero corresponds to a low contribution. The main absorbance regions of the MIR spectra represent the fundamental vibration modes of molecules, mainly from bond stretching and bond deformations. Other minor vibrational modes are twisting, wagging, rocking, and scissoring motions [33, 34]. The regions showed high contribution on the PCA were: between 3000 and 2830 cm^{-1}, particularly at wavenumbers ~2937, 2912, 2864, and 2842 cm^{-1} that are dominated by the absorption modes of aliphatic chains (mainly from fatty acids and phospholipids), and in particular by asymmetric stretching at ~2960 and 2920 cm^{-1} and symmetric stretching at ~2864 and 2842 cm^{-1} of CH_3 and CH_2 groups, respectively. This high contribution most probably reflects the cell membrane relevance to metabolic exchanges between the host cell and its environment along diverse carbon-source uptakes and metabolites secretions. The relevance of this spectral region on the cell dynamic metabolism was, for example, observed in cells presenting high specific growth rate as the one associated with carcinogenic processes [35]. The fingerprint region (between 1500 and 500 cm^{-1}), reflecting a variety of overlapped vibrations

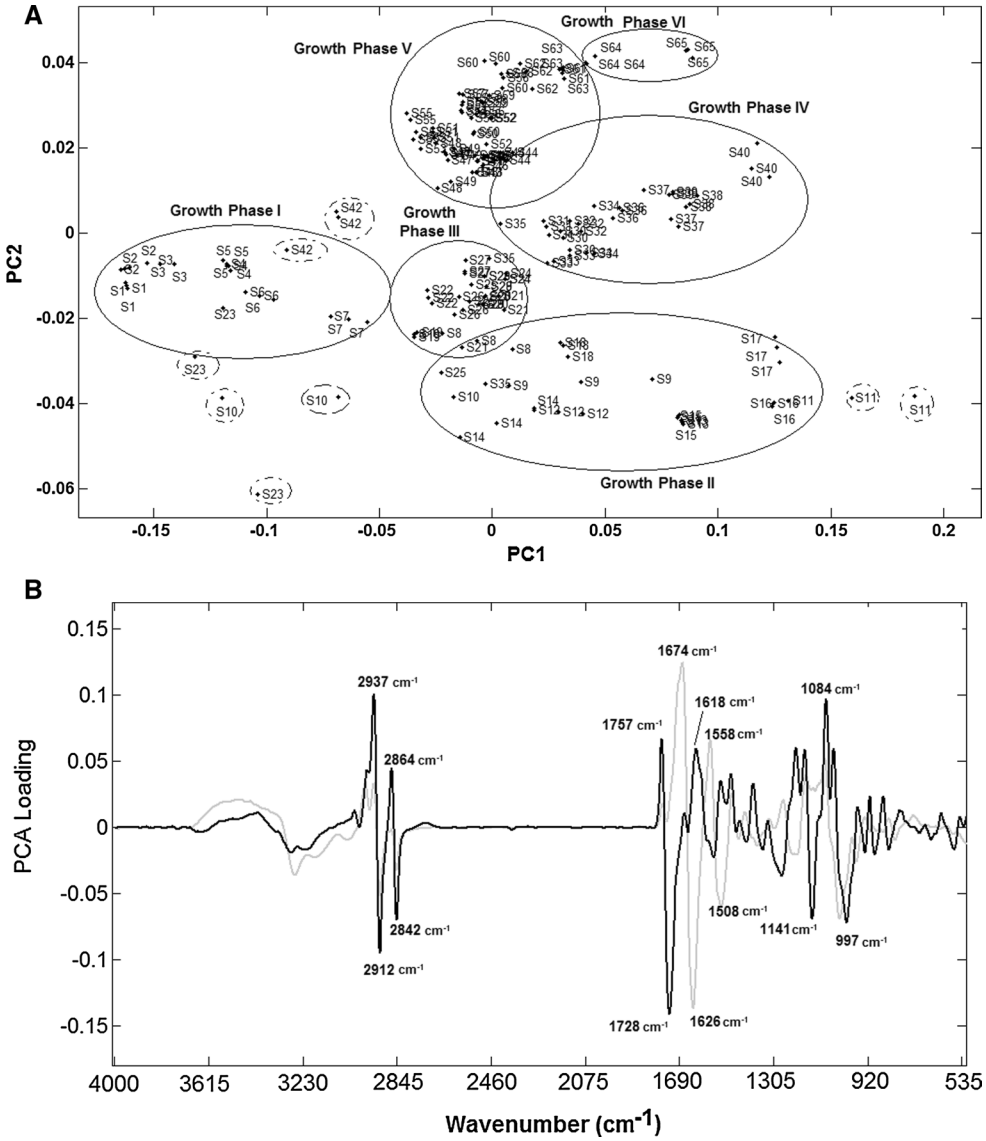


Fig. 2 PCA of first derivative MIR spectra acquired along the culture of recombinant *S. cerevisiae*: **a**-PC1 versus PC2 score plot; continuous line represented the six different growth phases, as described in Table 1; *dashed line* represents the outliers data. **b** Loading vectors of PC1 (*grey continuous line*) and PC2 (*black continuous line*)

due to proteins, glicids, lipids, and nucleic acids also contributed to PC1 and PC2. Interestingly, the region around 1740 cm^{-1} (that includes, for example, C=O vibrations of ester bonds on phospholipids esters) shows a high contribution for PC1, probably reflecting the capturing of PC1 of the C-source consumption, whereas the regions which are characteristic of amide I (~1655 cm^{-1}) and amide II (~1545 cm^{-1}) of proteins mostly contributed for PC2, most probably reflecting the caption of the cyprosin production behavior by PC2, as previously pointed.

The second derivative of the MIR spectra was also investigated, as it enables better definition of the peak bands while minimizing the effect of baseline drifts and the effect of the biomass when comparing different spectra in relation to non-derivative spectra. For simplicity of analysis, Fig. 3 represents spectral second derivative from critical samples related to defined growth phases, as resumed in Table 1. The spectral replicates, also represented in Fig. 3, appear superimposed, highlighting again the technique reproducibility to acquire in a high sensitive mode the molecular fingerprint of the yeast cell along the culture. Samples at the final growth phases (Phases V and VI), represented by continuous lines and present distinct second derivative spectra in relation to samples from early growth phases (Fig. 3a–c). The three main regions, where different second derivative values were observed along the yeast culture, were the same as identified by the PCA loadings: the region between 3000 and 2830 cm^{-1}, reflecting X–H (where X may be N, O, or C) vibrations, and particularly that included the absorption modes of aliphatic chains; the region between 1800 and 1500 cm^{-1}, reflecting stretching vibrations of double bonds (e.g., C=O, C=C, and C=N), where the 1749 cm^{-1} band results from $–CH_2–CCOR$ vibrations from phospholipids esters mostly from DNA

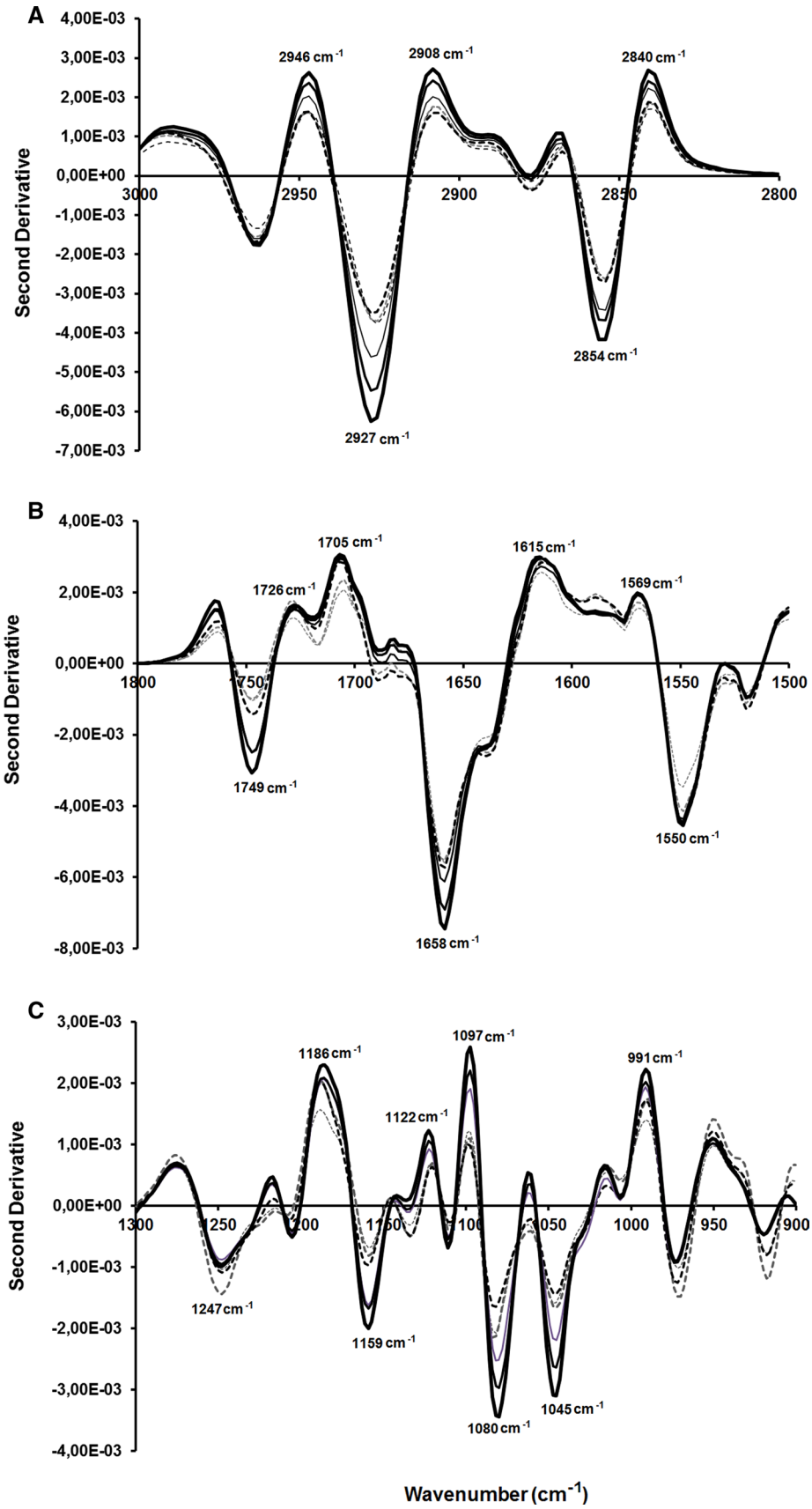


Fig. 3 Second derivative MIR spectra along the different growth phases of *S. cerevisiae* culture, as represented in Table 1. Growth phase I—*grey thin-dashed line*; growth phase II—*grey medium-dashed line*; growth phase III—*grey thick-dashed line*; growth phase IV—*black thin continuous line*; growth phase V—*black medium continuous line*; growth phase VI—*black thick continuous line*

Table 2 Best PLS models obtained for predicting the bioprocess variables

Variable	R^2	RMSEP	Percentage of error (%)	LV	Preprocessing	Spectral region used (cm^{-1})
Biomass	0.99	0.34 g dcw l^{-1}	2.8	8	NSD processing	2948–2248; 1199–499
Cyprosin activity	0.98	87 U ml^{-1}	3.9	8	Min–max normalization	1550–1197
Glucose	0.93	0.68 g l^{-1}	7.2	6	Constant offset elimination	3998–3647; 3299–2946; 2248–1548; 850–499
Galactose	0.97	0.39 g l^{-1}	4.6	8	2nd derivative	3299–2248; 1899–848
Ethanol	0.97	1.17 g l^{-1}	5.3	9	1st derivative	3998–3647; 2948–2598; 2248–1897; 1550–848
Acetate	0.95	0.07 g l^{-1}	7.0	1	Straight line subtraction	3998–3647; 3299–2597; 1899–1198

LV latent variables, *Min–max normalization* minimum–maximum normalization, *MSC* multiplicative scattering correction, *NSDP* no spectral data processing, *RMSEP* root mean square error of prediction, *RMSECV* root mean square error of cross validation

and RNA molecules [14, 36]; a large contribution of bands 1658 cm^{-1} (from amide I vibrations) and 1550 cm^{-1} (from amide II vibrations) and the fingerprint region between 1500 and 500 cm^{-1} [33]. From the fingerprint region, the following peaks contributing to the PC1 and PC2 showed differences in the second derivative spectra along the culture: 1247, 1159, 1080, and 1045 cm^{-1} (Figs. 2b, 3c). The absorbances at 1085 and 1240 cm^{-1} are attributed to PO_2^- symmetric and asymmetric stretching vibrations of phosphodiesters, respectively, presented in nucleic acids, phospholipids, and phosphorylated proteins. The absorbances at 1159 cm^{-1} have been attributed to carbonyl vibrations due to RNA ribose, while 1045 cm^{-1} has been attributed to both DNA and RNA molecules [33].

PLS regression models to predict critical bioprocess variables

The most usual application of MIR spectroscopy for monitoring recombinant cultures is the estimation of all critical variables by PLS regression [37]. In the present work, several PLS models were built to predict the critical variables of the cyprosin bioprocess, based on the combination of preprocessing techniques and specific spectral regions providing the best predictive ability as follows: the whole spectral region (between 400 and 4000 cm^{-1}) was divided into ten equal sub-regions, and the best combination of spectral regions and preprocessing techniques (baseline correction by offset or strait line subtraction, min–max normalization, multiplicative scatter correction, and first and second spectral derivatives) providing the best predictive model performance was chosen. The best PLS model was selected regarding a high regression coefficient (R^2), a low root mean square error of prediction (RMSEP) (or a low RMSECV), and a low number of latent variables (LVs) but covering enough data variance (always higher than 90%). For all the bioprocess critical variables, except for glucose, the RMSEP was based on an independent test validation data set due to the larger number of samples available. For glucose, an RMSECV was determined, as few samples were available. In general, good PLS models to predict all critical variables of the cyprosin bioprocess were obtained, yielding high predictability, a high R^2 ($\geq$0.93), and low errors of prediction ($\leq$7.2%).

The PLS regression model to predict the cyprosin activity from the MIR spectra was built using samples randomly selected throughout the culture. It is relevant to highlight that, and as pointed before, the yeast cell presented along the culture highly different metabolic states (Fig. 1) that resulted in a wide range of cyprosin production per biomass (between 0 and 283 U mg^{-1} dcw, Fig. 1d) and a wide range of cyprosin specific activities (between 0 and 10,000 U mg^{-1} protein, Fig. 1a). The use of this high diversity of samples on PLS model building disrupted relationships between cyprosin production and biomass growth and cyprosin production and total protein production that consequently resulted in PLS model that specifically predicts cyprosin. Furthermore, the PLS model was also validated using samples that were not used for model building and that were randomly selected along the culture and consequently also presented a wide range of cyprosin activity per biomass and specific cyprosin activity.

The best PLS model to predict cyprosin activity was achieved using as preprocessing min–max normalization method and the spectral region between 1550 and 1197 cm^{-1} (Table 2). For PLS models, using, e.g., this wavelength and preprocessing, it was observed, as expected, that as the number of LV increases the R^2 increases and the RMSEP decreases (Fig. 4a), being selected as the best PLS model the one using 8 LV that resulted in a model with R^2 of 0.98 and an RMSEP of 87 U ml^{-1} equivalent to 3.9% error. Figure 4b represents the validation of the best model based on 8 LV, using an independent data set of samples randomly selected along the culture that consequently presented a wide range of cyprosin activity per biomass and specific cyprosin activity.

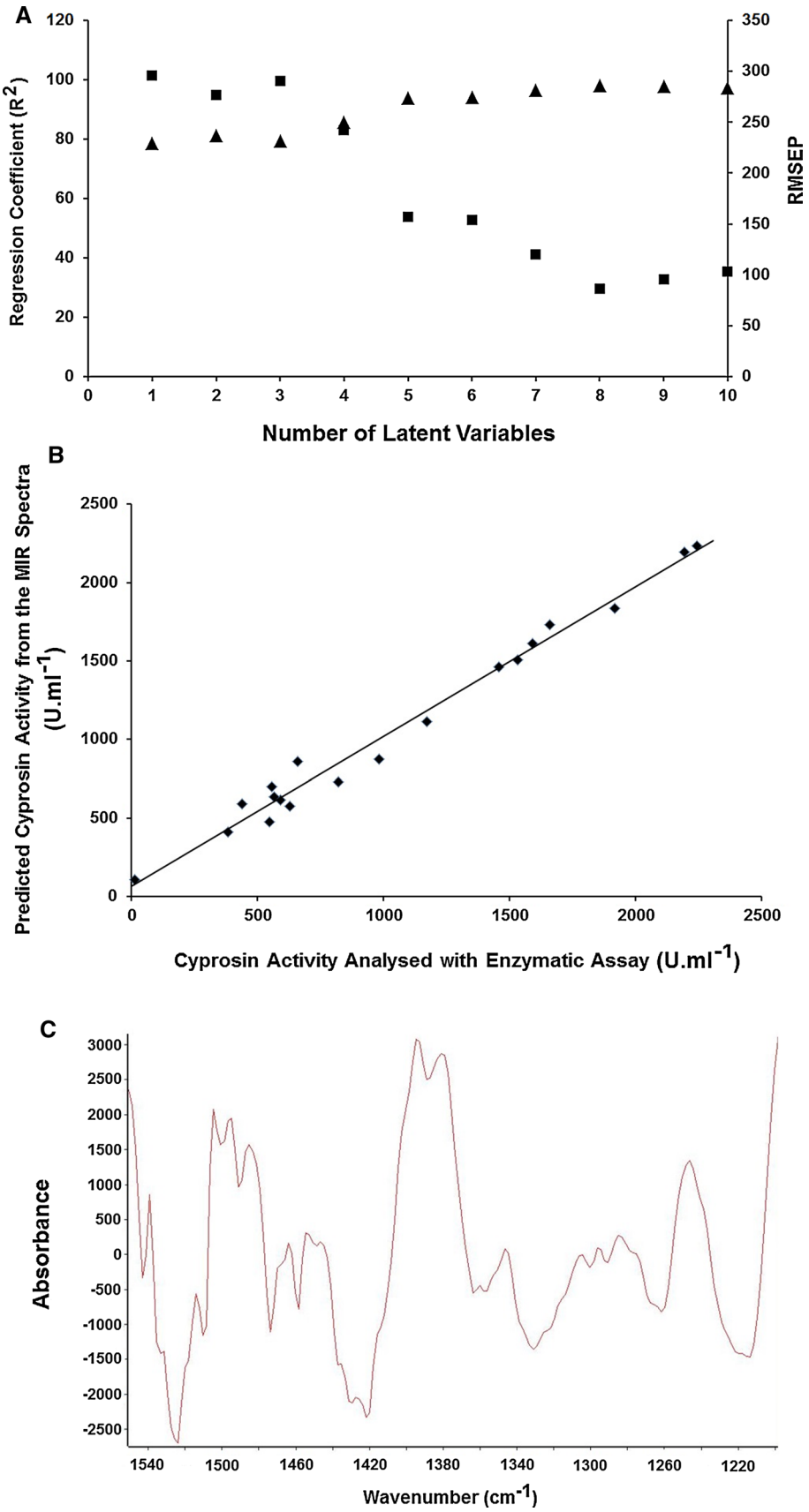


Fig. 4 PLS regression model to predict cyprosin activity from the MIR spectra. **a** Relationships between the number of latent variables and regression coefficient (*triangles*) and root mean square error of prediction (RMSEP) (*squares*) for PLS models based on min–max normalization and the spectral region between 1550 and 1197 cm^{-1}; **b** Prediction of cyprosin activity in relation to cyprosin activity measured by an enzymatic conventional assay and using the PLS model represented in **a** with eight latent variables; **c** Regression vector for the best PLS model using eight latent variables, preprocessing of min–max normalization and the spectral region between 1550 and 1197 cm^{-1}

The best PLS model to predict cyprosin activity used the spectral region between 1550 and 1197 cm^{-1} that contains the amide II region (at 1550 cm^{-1}) and the fingerprint region (between 1500 and 1197 cm^{-1}) that reflects the contribution of a variety of overlapped vibrations. The use for the PLS model of cyprosin activity of one of several amide bands and the fingerprint region reflects the relevance of the cell general metabolic status to predict cyprosin activity (Fig. 4c).

The best PLS model for biomass showed a high predictive capability (R^2 = 0.99) and a low RMSEP (0.34 g dcw l^{-1}), which corresponds to an error as percentage of 2.8%, considering the biomass concentration range used for model calibration (Table 2). The model was based on the spectral regions 2948–2248 and 1199–499 cm^{-1} that represent vibrations of a high diversity of chemical functional groups.

Glucose represents another critical variable relevant to control and minimize the overflow metabolism and, consequently, the ethanol production. The best PLS model for glucose was characterized by R^2 of 0.93 and the highest RMSECV (7.2%), which may be due to data limitation, as glucose was taken by the cells very sharply and in a short-time period. During the cultivation, the galactose, which is a strong inducer for recombinant cyprosin biosynthesis, was slowly consumed. The best PLS regression model for the galactose was characterized by a high R^2 (0.97) and a low RMSEP (0.39 g l^{-1}), representing an error of 4.6%. Ethanol is another critical variable, whose production results from a less efficient energetic metabolism and its accumulation in the broth may also inhibit the yeast growth and cyprosin production. The best PLS regression model for ethanol yielded a high R^2 (0.97) and a low RMSEP (1.17 g l^{-1}), corresponding to an error of 5.3% (Table 2). The best acetate PLS model was characterized by R^2 of 0.95 and an RMSEP of 0.07 g l^{-1}.

It should be noted that MIR spectra were acquired from dehydrated cell pellets and, consequently, the extracellular cyprosin activity, glucose, and acetate predictions from the spectra were only possible due to the correlations between the cell general metabolism and the concentration of these compounds in the extracellular medium. The PLS models obtained were built based on some similar spectral regions, but also different combinations of spectral regions, pointing out the specificity of each model. In spite of the cellular metabolic variations along the process, as highlighted by PCA, which could negatively affect the model, accurate PLS models were obtained. In general, by high-throughput MIR spectroscopic analysis, it was possible to obtain PLS models with similar predictabilities to the ones developed based on in situ NIR analysis [13], with exception for the glucose model which a slightly better model with the NIR setup was obtained (R^2 = 0.97; RMSEP = 0.41 g l^{-1}) in relation to the MIR setup (R^2 = 0.93; RMSEP = 0.97 g l^{-1}).

Conclusion

The development of monitoring techniques for cyprosin bioprocess production in *S. cerevisiae* that are simultaneously sensitive, economic, rapid, and simple to apply is highly relevant towards a better cyprosin process optimization and control. MIR spectroscopy offers all the above capabilities plus by the possibility of acquiring spectra in high-throughput mode, which makes possible the development of such high-standard bioprocess monitoring tool. This work presents MIR spectroscopy combined with multivariate data analysis as a valuable tool for accurately predicting the critical variables involved in the cyprosin production, as well as for extracting metabolic information from the host cell that will significantly increase the knowledge on the bioprocess.

Acknowledgements We thank Dr. Filomena Calixto for providing the *S. cerevisiae* BJ1991 strain transformed with *CYPRO11* gene. We also thank Professor António Mendonça from the University of Beira Interior by availability of laboratory equipment and facilities related to spectra acquisition. We also thank the Professor Fernando Baltazar Duarte (Faculty of Engineering, ULHT) for the support in MATLAB. M.B Lopes acknowledges financial support from the Portuguese Foundation for Science and Technology (SFRH/BPD/73758/2010). This work was supported by CLARO research project funded by ADI agency (Portugal). The present work was partly conducted in the Health Engineering Laboratory resulting from the Protocol between Universidade Católica Portuguesa and the Instituto Politécnico de Lisboa.

References

1. Fernandez-Salguero J, Prados F, Calixto F, Vioque M, Sampaio PN, Tejada L (2003) Use of recombinant cyprosin in the manufacture of ewe's milk cheese. J Agric Food Chem 51:7426–7430
2. Sampaio PN, Fortes AM, Cabral JMS, Pais MS, Fonseca LP (2008) Production and characterization of recombinant cyprosin B in *Saccharomyces cerevisiae* W303-1A strain. J Biosci Bioeng 105:305–312
3. Heimgartner U, Pietrzak M, Geertsen R, Brodelius P, Silva Figueiredo AC, Pais MS (1990) Purification and partial characterization of milk clotting proteases from flowers of *Cynara cardunculus*. Phytochemistry 29:1405–1410
4. Sampaio PN, Calado CRC, Sousa L, Bressler DC, Pais MS, Fonseca LP (2010) Optimization of the culture medium composition using response surface methodology for new recombinant cyprosin B production in bioreactor for cheese production. Eur Food Res Technol 231:339–346
5. Sampaio PN, Sousa L, Calado CRC, Pais MS, Fonseca LP (2011) Use of chemometrics in the selection of a *Saccharomyces cerevisiae* expression system for recombinant cyprosin B production. Biotechnol Lett 33:2111–2119
6. Sampaio PN, Pais MS, Fonseca LP (2014) A novel fed-batch based strategy for enhancing cell-density and recombinant

cyprosin B production in bioreactor. Bioproc Biosyst Eng 37:2515–2527
7. Streefland M, van de Waterbeemd B, Happe H, van der Pol LA (2007) PAT for vaccines: the first stage of PAT implementation for development of a well-defined whole-cell vaccine against whooping cough disease. Vaccine 25:2994–3000
8. Vojinovic V, Cabral JMS, Fonseca LP (2006) Real-time bioprocess monitoring. Part I: in situ sensors. Sens Actuators B 114:1083–1091
9. Landgrebe D, Haake C, Höpfner T, Beutel S, Hitzmann B, Scheper T, Rhiel M, Reardon KF (2010) On-line infrared spectroscopy for bioprocess monitoring. Appl Microbiol Biotechnol 88:11–22
10. Food and Drug Administration (FDA), Guidance for Industry, PAT: a Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance
11. Lopes MB, Gonçalves GAL, Felício-Silva D, Prather KLJ, Monteiro GA, Prazeres DMF, Calado CRC (2015) In situ NIR spectroscopy monitoring of plasmid production processes: effect of producing strain, medium composition and the cultivation strategy. J Chem Tech Biotechnol 90:255–261
12. Roychoudhury P, Harvey LM, McNeil B (2006) The potential of mid-infrared spectroscopy (MIRS) for real time bioprocess monitoring. Anal Chim Acta 571:159–166
13. Sampaio PN, Sales KC, Rosa FO, Lopes ML, Calado CRC (2014) *In situ* near infrared spectroscopy monitoring of cyprosin production by recombinant *Saccharomyces cerevisiae* strains. J Biotechnol 188:148–157
14. Sivakesava S, Irudayaraj J, Ali D (2001) Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. Process Biochem 37:371–378
15. Sales KK, Rosa F, Sampaio PN, Fonseca LP, Lopes MB, Calado CRC (2015) In situ near-infrared (NIR) versus high-throughput mid-infrared (MIR) spectroscopy to monitor biopharmaceuticals production. J Appl Spectrosc 69(6):760–772
16. Gross-Selbeck S, Margreiter G, Obinger C, Bayer K (2007) Fast quantification of recombinant protein inclusion bodies within intact cells by FT-IR spectroscopy. Biotechnol Progr 23:762–766
17. Scholz T, Lopes VV, Calado CRC (2012) High-throughput analysis of the plasmid bioproduction process in *Escherichia coli* by FTIR spectroscopy. Biotechnol Bioeng 109:2279–2285
18. Wold S, Cheney J, Kettaneh N, McCready C (2006) The chemometric analysis of point and dynamic data in pharmaceutical and biotech production (PAT)—some objectives and approaches. Chemom Intell Lab Syst 84:159–163
19. Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Rev Comput Stat 2:433–459
20. Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip Rev Comput Stat 21:97–106
21. Pais MS, Calixto F, Planta F (2002) Production by yeasts, of aspartic proteinases from plant origin with sheep's, cow's, goat's, etc. milk clotting and proteolytic activity. International Patent, WO 0075283
22. Bradford MM (1976) A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248–254
23. Dhanoa MS, Lister SJ, Sanderson R, Barnes RJ (1994) The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. J Near Infrared Spectrosc 2:43–47
24. Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1639
25. Jackson JE (1991) A user's guide to principal components. Wiley, New York
26. Kohler A, Hanafi M, Bertrand D, Quannari M, Oust Janbu A et al (2008) Interpreting several types of measurements in bioscience. In: Lasch P, Kneipp J (eds) Biomedical vibrational spectroscopy. Wiley, New Jersey, pp 333–356
27. Martens H, Naes T (1996) Multivariate calibration. Wiley, Chichester
28. Adams B (1972) Induction of galactokinase in *Saccharomyces cerevisiae*: kinetics of induction and glucose effects. J Bacteriol 111:308–315
29. Lopes MB, Martins G, Calado CRC (2014) Kinetic modeling of plasmid bioproduction in *Escherichia coli* cultures conducted in different media compositions. J Biotechnol 186:38–48
30. Käpelli O (1986) Regulation of carbon metabolism in *Saccharomyces cerevisiae* and related yeasts. Adv Microb Physiol 28:181–208
31. Gancedo JM (1998) Yeast carbon catabolite repression. Microbiol Mol Biol Rev 62:334–361
32. Ferreira BS, Calado CRC, van Keulen F, Fonseca LP, Cabral JMS, Fonseca MR (2004) Recombinant *Saccharomyces cerevisiae* strain triggers acetate production to fuel biosynthetic pathways. J Biotech 109:159–167
33. Bellisola G, Sorio C (2012) Infrared spectroscopy and microscopy in cancer research and diagnosis. Am J Cancer Res 2(1):1–21
34. Maquelin K, Kirschner C, Choo-Smith LP, van den Braak N, Endtz HP, Naumann D, Puppels GJ (2002) Identification of medically relevant microorganisms by vibrational spectroscopy. J Microbiol Methods 51:255–271
35. Huleihel M, Salman A, Erukhimovitch V, Ramesh J, Hammody Z, Mordechai S (2002) Novel spectral method for the study of viral carcinogenesis in vitro. J Biochem Biophys Methods 50:111–121
36. Matthäus C, Bird B, Miljković M, Chernenko T, Romeo M, Diem M (2008) Infrared and Raman microscopy in cell biology. Methods Cell Biol 89:275–308
37. Wold S (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab Syst 58:109–622