Universidade Católica Portuguesa
Faculdade de Engenharia

Fourier Transform Infrared Spectroscopy, a powerful tool to monitor biopharmaceuticals production

Kevin Costa Sales

Dissertação para obtenção do Grau de Mestre em Engenharia Biomédica

Mestrado em Engenharia Biomédica
Especialização em Engenharia Biomolecular, de Tecidos e de Órgãos

Júri
Prof. Doutor Manuel José Martinho Barata Marques (Presidente)
Prof. Doutor Luís Joaquim Pina da Fonseca
Doutora Carla Maria Cadete Martins Moita Brites
Prof.ª Doutora Cecília Ribeiro da Cruz Calado (Orientadora)

Abril 2014

To BestGFriend,
my parents and
my grandparents.

Resumo

A *Escherichia coli* é o microorganismo mais usado como hospedeiro para a produção de produtos recombinantes, tais como plasmídeos usados para terapia génica e vacinação de ADN. Desta forma, torna-se importante compreender as relações metabólicas complexas e a bioprodução de plasmídeo, que ocorre em ambientes de cultura dinâmicos, a fim de controlar e optimizar o desempenho do sistema de expressão recombinante. O objectivo principal deste trabalho consiste em avaliar a potencialidade da espectroscopia FT-IR para monitorizar e caracterizar a produção do plasmídeo pVAX-LacZ em culturas recombinantes de *E. coli*, nomeadamente para extrair informação relacionada com as variáveis críticas (biomassa, plasmídeo, fontes de carbono e acetato) e informação metabólica da célula hospedeira *E. coli*. Para tal, culturas de *E. coli* com diferentes concentrações de glucose e glicerol e diferentes estratégias de cultivo (*batch* e *fed-batch*) foram monitorizadas por espectroscopia de infravermelho perto (NIR) e de infravermelho médio (MIR).

Tanto a espectroscopia NIR com a MIR permitiram extrair informação sobre as variáveis críticas do bioprocesso, através da construção de modelos de regressão por mínimos quadrados parciais, que resultaram em elevados coeficientes de regressão e baixos erros de previsão. A abordagem NIR apresenta a vantagem de aquisição em tempo real das variáveis do bioprocesso, já a abordagem MIR permite a leitura simultânea de centenas de amostras de várias culturas ao mesmo tempo através do uso multi-microplacas, sendo muito vantajosa nos casos de micro-bioreactores usados para optimização. Para além disso, como os espectros MIR apresentam mais informação do que os espectros NIR, uma vez que representam os modos de vibração fundamentais das biomoléculas, enquanto que os espectros NIR representam sobreposições e combinações de vibrações, os dados espectrais MIR também permitiram a aquisição de informação bioquímica ao longo das culturas de *E. coli* a partir da análise das componentes principais (PCA) bem como do estudo das características bioquímicas, tais como as reservas de glicogénio e os níveis de transcrição aparente.

Portanto, a espectroscopia FT-IR apresenta assim características relevantes para a compreensão e monitorização do processo de produção de culturas recombinantes, sendo, de acordo com *Quality-by-Design* e *Process Analytical Technology*, muito importante para fins de controlo e optimização.

Palavras-chave: *Escherichia coli*, espectroscopia MIR, espectroscopia NIR, caraterização metabólica, monitorização de bioprocessos.

Abstract

Escherichia coli is the most used microorganism as host for the production of recombinant products, such as plasmids used for gene therapy and DNA vaccination. Therefore, it is important to understand the complex metabolic relationships and the plasmid bioproduction process occurring in dynamic culture environments, in order to control and optimize the performance of the recombinant expression system. The main goal of this work is to evaluate the potential of Fourier Transform Infrared (FT-IR) spectroscopy to monitor and characterize recombinant *E. coli* cultures producing the plasmid model pVAX-LacZ, namely to extract information concerning the critical variables (biomass, plasmid, carbon sources and the by-product acetate) and metabolic information regarding the host *E. coli*. To achieve that cultures of *E. coli* conducted with different mixture of glucose and glycerol and different cultivation strategies (batch and fed-batch) were monitored *in-situ* by a fiber optic probe in near- infrared (NIR) and of the cell pellets in *at-line* in high-throughput mode by mid-infrared (MIR) spectroscopy.

Both NIR and MIR spectroscopy setup enabled to extract information regarding the critical variables of the bioprocess by the implementation of partial least square regression models that result in high regression coefficients and low prediction errors. The NIR setup presents the advantage of acquiring in real time the knowledge of the bioprocess variables, where the *at-line* measurements with the MIR setup presents more advantageous in cases of micro-bioreactors used in optimization protocols, enabling the simultaneously information acquisition of hundreds samples by using multi-microplates. Furthermore, as the MIR spectra presents more information than the NIR spectra, since it represents the fundamental vibration modes of biomolecules while the NIR spectra represents overtones and combinations of vibrations, the MIR data also enabled to acquire biochemical information along the *E. coli* cultures as pointed out in an principal component analysis and by the estimation of biochemical features as glycogen reserves and apparent transcriptional levels.

Therefore, FT-IR spectroscopy presents relevant features towards the understanding and monitoring of the production process of recombinant cultures for control and optimization purposes, in according to the Quality-by-Design and the Process Analytical Technology.

Keywords: *Escherichia coli*, MIR spectroscopy, NIR spectroscopy, metabolic profiling, bioprocess monitoring.

Acknowledgments

I would like to express my sincere gratitude to my advisors, Prof. Cecília Calado, Dr. Marta Belchior Lopes and Prof. Pedro Sampaio, for their continuous support, patience and motivation. It is an honor to work with all of you!

I would also like to thank Professor Manuel Barata Marques, Director of the *Faculdade de Engenharia, Universidade Católica Portuguesa*, for all the understanding and support from the first day of classes to the last days of my Master's Thesis!

To Professor António Mendonça from *Universidade da Beira Interior* thanks for providing technical support.

To *Instituto Nacional de Saúde Dr. Ricardo Jorge*, and within to Dr. Jorge Machado, Dr. João Brandão, Dr.ª Helena Rebelo and Dr.ª Raquel Rodrigues for collaboration and technical support.

I also thank *Fundação para a Ciência e Tecnologia* (FCT) and *Agência de Inovação* for supporting this work by the project PTDC/BIO/69242/2006 and CLARO, respectively.

To Filipa, my colleague, thank you for helping with data interpretation and also for all the support and patience along the Master's Thesis!

To all of my friends, not only who shared this academic journey with me, especially João Mesquita, Sónia Vaz, Catarina Gama, Gonçalo Condeço, André Valério and Maria Raimundo, but only who are always present in my heart, a great "thank you" for supporting me unconditionally, making me to smile and giving me the strength to continue.

To my beloved girlfriend, Vânia, the greatest thanks for all the unconditionally support, I am truly sure that I could not have made it without you! You are my motivation in everything I do and I will be eternally grateful.

I would also like to thank Vânia's parents who believe me and in my own capabilities, and, of course, for supporting me.

Last, an enormous thanks to my family. You supported me despite of anything, specially my bad temper! Parents, thank you for allowing me to have an education.

Table of Contents

List of Figures

List of Tables

Nomenclature and Abbreviations

μ	-	specific growth rate
Ace	-	acetate
ACKA	-	acetate kinase
C-source(s)	-	carbon source(s)
DCW	-	dry cell weight
DOC	-	dissolved oxygen concentration
EMA	-	European Medicines Agency
FDA	-	Food and Drug Administration
FT-IR	-	Fourier Transform Infrared
Glu	-	glucose
Gly	-	glycerol
HPLC	-	high performance liquid chromatography
IR	-	infrared
LOO	-	leave-one-out
lv	-	latent variable
max	-	maximum
MIR	-	mid-infrared
MSC	-	multiplicative scatter correction
NIR	-	near-infrared
OD	-	optical density
PAT	-	Process Analytical Technology
PC	-	principal component
PCA	-	principal component analysis
PLS	-	partial least squares
PTA	-	phosphotransacetylase
QbD	-	Quality-by-Design
R^2	-	coefficient of determination
RMSE	-	root mean square error
RPM	-	rotation per minute
S	-	substrate
SNR	-	signal-to-noise ratio
SNV	-	standard normal variate

v/v	-	volume/volume
w/v	-	weight/volume
$Y_{X/S}$	-	biomass per substrate yield

Chapter I

Thesis Overview

I.1. Objectives

The main goal of the present work was to evaluate the potential of Fourier Transform Infrared (FT-IR) spectroscopy to monitor and characterize recombinant *Escherichia coli* cultures during the production of biopharmaceuticals, namely the production of a plasmid model used for the construction of DNA vaccines. For that, i) first, the critical variables of the bioprocesses studied (e.g., host cell growth, plasmid production, carbon sources consumption and the by-product acetate production and consumption) were monitored based on infrared (IR) spectral data acquired along the cultivation time based on mid-infrared (MIR) spectroscopy of the cell pellets in high-throughput analysis using multi-microplates, and on near-infrared (NIR) spectroscopy by the cultivation *in-situ* analysis using a fiber optic probe; ii) second, metabolic information regarding, e.g., lipids, proteins, nucleic acids, glicids and other chemical species present in cells, was extracted from the MIR spectra for metabolic profiling of the host cell, as the MIR spectra represents the fundamental vibration modes of biomolecules.

I.2. Thesis Outline

The thesis is divided into 5 chapters. A general introduction is presented in chapter II, which contains a brief introduction to the *E. coli* systems and bioprocesses' monitoring, an overview on IR spectroscopy and a short introduction to chemometrics and spectral analysis. The following two chapters, chapters III and IV, describe the experimental work developed. Chapter III presents a comparative study of MIR and NIR spectroscopies for monitoring the critical variables involved in the production of a biopharmaceutical (e.g., the host cell growth, the production of plasmid, the carbon sources consumption (glucose and glycerol) and the by-product acetate production and consumption) by different recombinant *E. coli* cultures producing the plasmid pVAX-LacZ. Chapter IV describes the potential of FT-IR spectroscopy for estimating the metabolic profiles of, e.g., lipids, proteins, nucleic acids, glicids, and other biochemical information from the host cell along the cultures. The last chapter, chapter V, comprises the main conclusions of the previous chapters and presents new research directions for future work.

Chapter II

General Introduction

II.1. *E. coli* recombinant systems and bioprocesses monitoring

The growing interest in biopharmaceutical products calls for a need for developing reproducible, reliable and cost-effective production processes. An example of such products are plasmids, which can be used as vectors for gene therapy and DNA vaccination, as an alternative to viral based vectors [Carnes, 2005; Prather *et al.*, 2003].

Recombinant *E. coli is* the most used microorganism for plasmid production, given its capacity to growth under a wide range of conditions, from rich complex organic media to salt-based chemically defined media, as well as it is ease manipulation by genetic engineering [Moen *et al.*, 2009; Prather *et al.*, 2003; Scholz *et al.*, 2012; Yang, 1999]. As a consequence of the current growing interest on plasmids, their production has to meet the market requirements, i.e., the optimization and efficiency of plasmid production are required, as well as the monitoring of the bioproduction process. Generally, the main goals for an optimization procedure for recombinant *E. coli* cultures are (**Figure II.1**) [Carnes, 2005; Voss *et al.*, 2003]:

- Maximize the plasmid production in the supercoiled conformation, the most efficient conformation in relation to circular and linear conformations for therapeutic applications, according to Food and Drug Administration (FDA) and European Medicines Agency Home (EMA);
- Maximize the plasmid concentration;
- Maximize the productivity;
- Maximize the biomass per carbon source yield, to make the best use of nutritional media;
- Maximize the specific yield, i.e., the quantity of plasmid produced per cell, in order to simplify the purification processes.

Nevertheless, differences in the cultivation strategies adopted (e.g., batch and fed-batch) and environmental conditions and medium composition with respect to the carbon source (glucose or glycerol) influence the stability and expression of the cloned gene product, and consequently the optimization of the plasmid production processes [O'Kennedy *et al.*, 2003; Ow *et al.*, 2007; Ow *et al.*, 2009]. Furthermore, the characteristics of the plasmid and host cell are critical factors that should be carefully evaluated [McNeil and Harvey, 1990].

Therefore, to control and optimize the performance of recombinant systems, the complex interrelationships between these factors and its effects must be well understood towards a more economic and robust process that ensures reproducibility and quality of the final product, in accordance to the Process Analytical Technology (PAT) initiative launched in 2004 by FDA. The PAT initiative encourages biopharmaceutical companies to adopt modern bioprocess monitoring tools based on *at-line* or *in-situ* analyses of critical parameters along the manufacturing processes, thus enabling the formulation of mathematical models through of the complex datasets acquired along of all process stages, towards more robust control and optimization processes [FDA, 2004]. IR spectroscopy is an example of a powerful tool for bioprocesses' monitoring, which perfectly matches the PAT initiative and presents promising capabilities to serve the above purposes, as described next.

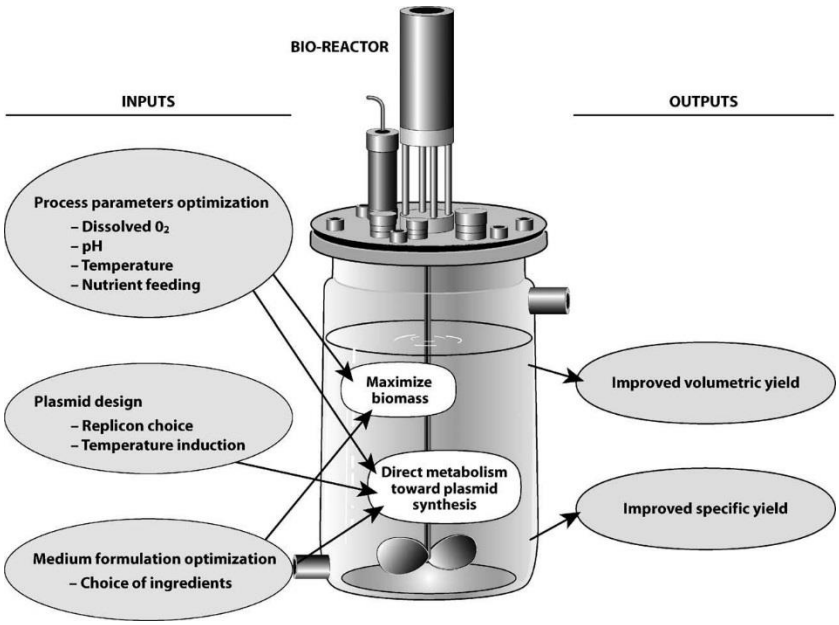


Figure II.1: Important factors in the monitoring and control of plasmid production, in bioreactors. (adapted from Prather *et al.*, 2003)

II.2. Infrared Spectroscopy

II.2.1. Theory of the Infrared Spectroscopy

Originally, spectroscopy was defined the study of the interaction between electromagnetic radiation and matter as a function of wavelength. Afterwards, the concept was expanded to include the measurement of any property, as a function of wavelength or frequency [Lourenço *et al.*, 2012]. All electromagnetic spectroscopic techniques work on the same principle, i.e., under certain conditions, the materials interacting with the radiation, absorb or emit energy. However, some materials can also reflect and/or disperse/diffract radiation. Absorption spectroscopy is based on the measurement of the radiation that is emitted by the light source but attenuated by the sample, while emission spectroscopy is based on the measurement of the radiation that is produced by the sample on excitation. The reflection and diffraction of the radiation essentially depends on the materials' surface and composition, shape and microstructure of the sample, respectively [Nicolaï *et al.*, 2007].

IR spectroscopy is a spectroscopic technique that uses the infrared region of the electromagnetic spectrum. The IR region ranges from 14000 to 4 cm^{-1} (0.7 to 250 µm) and is surrounded by the visible and microwave regions, as shown in the **figure II.2**. The IR region is further subdivided in the near infrared (NIR), the mid infrared (MIR) and the far infrared (far-IR) regions. MIR represents the region of the IR spectrum between 4000 and 400 cm^{-1}, whereas the NIR region is between 14000 and 4000 cm^{-1} (**Figure II.2**). Both regions will be discussed along this work, as they represent the IR radiation that are most used in several applications of spectroscopy [Landgrebe *et al.*, 2010; Smith, 2011].

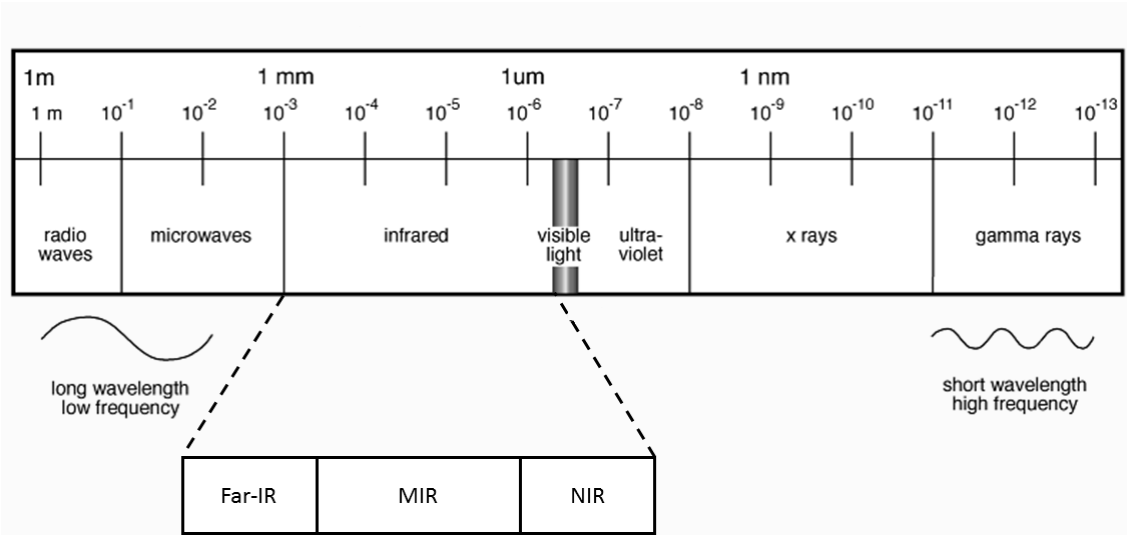


Figure II.2: Electromagnetic spectrum with IR region highlighted.

Photon energies associated with the infrared region of the electromagnetic spectrum are not large enough to excite electrons, rather they induce vibrational excitation of covalently bonded molecules. At temperatures above absolute zero, all atoms in molecules are in continuous vibration with respect to each other. Therefore, in IR spectroscopy, when a sample is irradiated by IR light, the absorption of this radiation results in changes in the vibrational modes of the molecules, which are sensible to the IR light and are presented in the sample. However, the absorption of IR only occurs when the radiant energy matches the energy of the specific molecular vibration, and the covalent bond of a molecule must undergo a net change in dipolar moment, as a consequence of its vibrational motion. The changes in the vibrational modes of the molecules produce the bands seen in the IR spectrum, with each band being characterized by a frequency and an amplitude [Babrah, 2009; Duygu, 2009].

Considering the changes in the vibrational modes of the molecules, there are essentially two types of vibrations, which can be classified depending on changes on the bond length or angle: stretching and bending vibrations (**Figure II.3**). The stretching is a symmetric or antisymmetric rhythmical movement along the bond length. The bending vibration occurs when there is a change of the angle between two atoms or a group of atoms [Babrah, 2009].

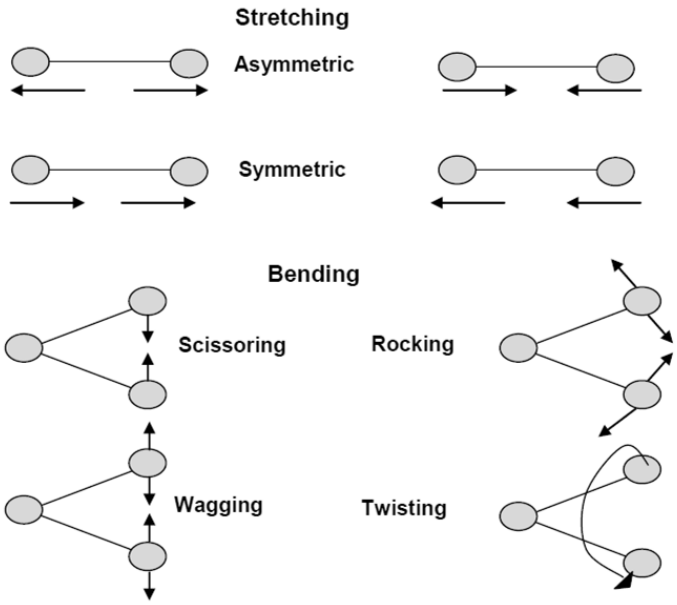


Figure II.3: Main molecular vibrational modes [Babrah, 2009]

Therefore, an IR spectrum is characteristic of each kind of molecule, since it depends mainly on the mass of the atoms, their geometric arrangement and the bound forces between them. Consequently, each molecule presents a distinct IR spectrum, since there are no different molecules that have the same three characteristics previously presented. When that concept is extended two different samples, which have distinct molecular composition, different spectra will also be obtained, thus enabling to distinguish, qualify or quantify virtually any type of sample [Smith, 2011].

The application of the IR spectroscopy in the biological field is possible, because certain regions of the IR spectrum have been attributed to certain molecular bonds and combinations of atoms, and the composition of every biomolecules is known, thus being possible to associate the biomolecules to certain IR region, especially in MIR region. Despite of the complex composition of biological samples and the presence of several biomolecules in the samples, it can be observed that the strongest vibrational frequencies correspond to macro-biomolecules, such as proteins, lipids, carbohydrates and nucleic acids [Smith, 2011], due to its high concentration in the cell, when compared to other biomolecules.

II.2.2. Instrumentation

The instrument used in IR spectroscopy is called infrared spectrometer or, more precisely, spectrophotometer, and consists mainly in a beam source, a monochromator or an interferometer, depending on the type of spectrometer, a sample holder or sample presentation interface and a detector, which detects the radiation that is transmitted or reflected by the sample [Reich, 2005].

Considering the beam source, it may consist on an inert solid thermally heated [Hsu, 1997] or in an incandescent filament, like tungsten or quartz/halogen lamps, for the NIR region, and carbon-silicon bars, for the MIR region [Christian, 1994].

The existing detectors are essentially of two types: thermal detectors, which measures the heat produced by the IR radiation when in contact with the sample, and photon detectors that are based on the interaction of IR light with semiconductor materials, allowing the excitation of electrons and the generation of a small quantifiable electrical current [Hsu, 1997].

Another important component of the spectrometer is the monochromator or the interferometer, which enables the light modulation and defines the type of spectrophotometer. There are mainly two types of spectrometers: Dispersive Infrared Spectrometers and Fourier Transform Infrared Spectrometers. In both configurations the beam source, detectors and sample holders used are essentially the same.

The Dispersive Infrared Spectrometers were the first kind of spectrophotometers developed, using a monochromator in its configurations. A monochromator is a device used to separate a

range of radiations in a certain range of wavelengths or frequencies. The most common monochromator are prism and gratings coupled with systems of mirror and filters [Stuart, 2004].

The introduction of interferometry brought significant improvements to IR spectroscopy and the monochromator has been substituted by the interferometer. An interferometer measures the interference pattern between two light beams. After entering in the interferometer, the IR radiation is divided in two beams that will travel by different paths. Before leaving the interferometer, these two beams will be merged in a single beam again. The development of interferometers opened the window to the Fourier Transform Infrared (FT-IR) spectrometers.

The first spectrometer with interferometer to be developed was a Michelson interferometer and the current interferometers are based on the same principle. The Michelson interferometer consists of four active components: a collimating mirror, a moving mirror, a fixed mirror oriented perpendicularly and a beamsplitter (**Figure II.4**). The collimating mirror collects the IR light from the source and makes its rays parallel to each other, while directing them to the beamsplitter. The beamsplitter splits the radiation from collimating mirror in two beams, with half the IR beam being transmitted to the fixed mirror and the other half reflected to the moving mirror. These beams recombine at the beamsplitter, but the difference in paths lengths creates constructive and destructive interference: an interferogram. The recombined beam passes through the sample, which absorbs all the different wavelengths characteristic of its spectrum, and this subtracts specific wavelengths from the interferogram. A mathematical operation, known as a Fourier transformation, converts the interferogram (a time domain spectrum displaying intensity versus time within the mirror scan) to the final IR spectrum, which is the frequency domain spectrum showing intensity versus frequency [Smith, 2011; Stuart, 2004].

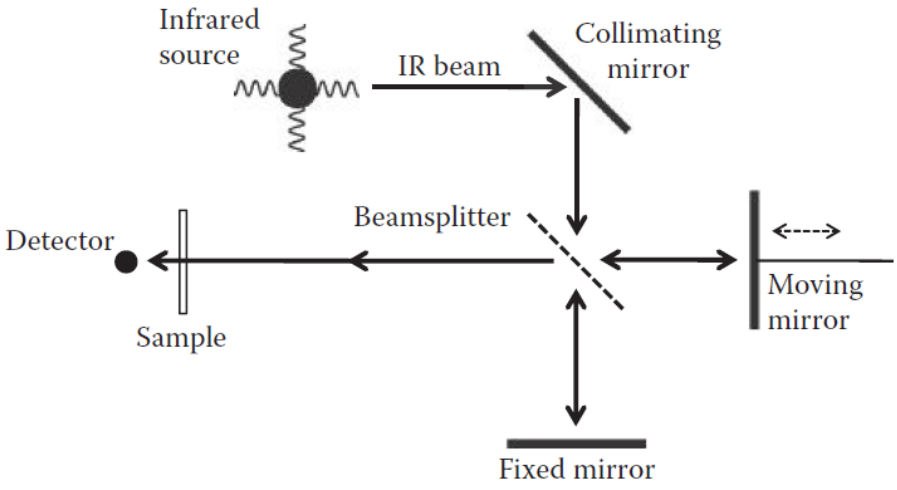


Figure II.4: Scheme of the Michelson interferometer [Smith, 2011].

When compared to dispersive systems, the development of FT-IR spectrometers and their implementation contributed to high reproducibility and low sampling noise, while making sample analysis faster [Hsu, 1997; Pistorius, 1995]. The low sampling noise is maybe the mains contribution of the FT-IR spectrometer, since it allows more sensitive measurements and, consequently, less noisy spectra with smaller peaks becoming evident.

As the amount of signal in a spectrum is highly dependent on the amount of light that reaches the detector, the signal-to-noise ratio (SNR) in FT-IR spectrometers is higher than in the dispersive spectrometers. This is due to the fact that in the dispersive spectrometers the beam needs to travel through prisms, slits and gratings, before reaching the sample and, consequently, the final beam that is detected have a much lower intensity, compared to the beam that leaves the source and, consequently, the final spectrum has a lower SNR [Smith, 2011].

II.2.3. Types of acquisition

Depending on the sample properties, the spectral data can be acquired essentially by two different modes: transmission and reflection.

In transmission mode, IR radiation passes through the sample and the decrease in the radiation intensity due to absorption or scattering by the sample is measured. Therefore, the spectrum obtained is the result of the radiation that passes through the sample (that is proportional to the radiation absorbed by the sample) as function of wavelength, and depends of the radiation's pathlength [Hsu, 1997].

In reflectance mode, the ratio of the intensity of the radiation reflected by a sample to the radiation reflected by a background reflective surface is measured. This acquisition mode is useful when the sample absorbs too much or too less energy, as well as in cases where samples reflect the majority of the incident radiation.

Though transmission and reflectance are the main acquisition modes, transflection has been increasingly being used in NIR spectroscopy applications. It combines the transmittance and reflectance measurements, i.e., the IR radiation is transmitted through a sample and the unabsorbed radiation is reflected back from a mirror or a diffuse reflectance surface placed at the end of the probe [Lourenço *et al.*, 2012].

II.2.4. Mid-Infrared Spectroscopy

MIR spectroscopy is an extremely reliable and widely recognized fingerprinting technique. Many compounds can be characterized, identified and quantified by this method, since it is in the MIR region, between 4000 cm^{-1} and 400 cm^{-1}, where most of the fundamental structural information is produced, therefore presenting enhanced sensitivity and selectivity and more distinctive spectral features, when compared to NIR spectroscopy [Smith, 2011]. Another important particularity of MIR spectroscopy is that it can be applied in an automatable way with high-throughput instruments [Scholz *et al.*, 2012].

Nevertheless, an important disadvantage of the use of MIR spectroscopy is related to the fact that MIR region presents higher interference by water than NIR region, being usually necessary to dehydrate the samples before spectral acquisition [Landgrebe *et al.*, 2010]. Furthermore, MIR radiation has a shorter wavelength than NIR radiation and consequently less energy, so the ability of this kind of radiation to penetrate the sample is reduced. The difficulty of transport and to obtain remote measures is also a disadvantage of the MIR radiation.

MIR spectroscopy allows a rapid acquisition of spectra, no sample preparation is necessary, beside the dehydration step for aqueous samples, and it is a non-invasive method, which is extremely useful when the sample preservation is required. However, spectra can be changed due to fluctuations in the equipment's environment and sometimes chemometric methods are necessary, in order to extract all information contained in a spectrum.

II.2.5. Near-Infrared Spectroscopy

NIR spectroscopy is a spectroscopic method that uses the NIR region of the electromagnetic spectrum from 14000 to 4000 cm^{-1} [Smith, 2011]. This technique is usually applied for aqueous in-situ analyses, given the low adsorption coefficient of NIR radiation and the low interference of water in this IR region, when compared with MIR region. NIR spectroscopy allows the direct analysis of samples that are highly absorbing or strongly light scattering without dilution or extensive preparation. Nevertheless, most bands in NIR region are consequence of overtones and combinations of vibrations from different chemical elements and functional group, which makes NIR spectroscopy less sensitive and informative than MIR spectroscopy. Since NIR spectroscopy is less sensitive and its spectra are visually poor, it is often necessary to apply chemometric methods to extract meaningful information from the data [Hall *et al.*, 1996; Lourenço *et al.*, 2012; Shenk *et al.*, 2001].

Given the low interference of water in the NIR region, NIR spectroscopy can be used non-destructively for monitoring bioprocesses, by placing of a fiber optic probe inside the bioreactor. Other particularities of this technique are the fact that the NIR radiation presents a greater penetration power, since it is little absorbed by sample, and can be easily transported by optical fibers, which makes possible a remote acquisition of spectra [Lourenço *et al.*, 2012].

In sum, NIR spectroscopy is a non-destructive fast technique, it does not need any sample's preparation and it can measure several samples' properties at once [Smith, 2011].

II.3. Chemometrics

Chemometrics is the application of statistical, mathematical and computational methods to analyze chemical data and to extract information from certain chemical systems. These methods allow the extraction of the relevant information concerning the analytes of interest, which otherwise would be very difficult [Lourenço *et al.*, 2012; McGovern *et al.*, 2002]. Chemometrics was first introduced in the chemical field, although today is a widely used tool in several other areas such as spectroscopy [Geladi, 2003].

The successful implementation of the spectroscopic techniques, essentially NIR spectroscopy, which produces broad and overlapping spectral bands, was only possible due to the development of chemometric methods. Beside NIR spectroscopy, MIR spectroscopy, normally producing well defined spectral bands, also rely on chemometric methods for easy of interpretation and handling of large data sets, as well as to reduce the noise that is often present in spectra.

The most widely chemometric methods used for spectral data analysis in spectroscopy are mathematical pre-processing techniques and multivariate data analysis, which are mainly divided into qualitative and quantitative methods.

In the present work some pre-processing techniques are reviewed, namely those studied along the work: multiplicative scatter correction (MSC), standard normal variate (SNV), baseline correction, normalization and derivatives. Principal component analysis (PCA) and partial least squares (PLS) regression are the choice for multivariate data analysis of spectral data acquired during this work, and will be presented next.

II.3.1. Mathematical Pre-Processing Techniques

The application of pre-processing techniques is a very important step in the analysis of spectral data, since they enable the elimination of physical phenomena due to undesired variations, such as noise, differences along the sample thickness, differences in the number cells across the sample and scattering events [Rinnan *et al*., 2009; Sharaf *et al.*, 1986]. This procedure has as goal of minimizing the irrelevant information present in the final spectra.

Multiplicative Scatter Correction (MSC)

MSC is a pre-processing method used to eliminate the effect of physical phenomena like the light scattering effect of particles of different sizes and shapes [Helland *et al.*, 1995]. The goal is to find the "ideal" spectrum of the group. For that, it is necessary a reference spectrum, which is usually the mean spectrum of all available spectra or the mean spectrum of replicate spectra. MSC works by fitting each spectrum to the average spectrum, which is supposed to be the ideal, performing a transformation where the spectral data x_i is converted into new values z_i, where i = 1, …, p, with p being the wavelengths [Fearn *et al.*, 2009]. The following equation describes the transformation from x_i to z_i:

$$z_i = \frac{x_i - a}{b},$$

where a represents the intercept and b the slope of a least squares regression of x_i on the values r_i coming from the reference spectra.

Standard Normal Variate (SNV)

The SNV transformation centers each spectrum and then scales it by its own standard deviation. The resulting spectra have always zero mean and variance equal to one, and are thus independent of original absorbance values. Dhanoa et al. (1994) and Helland et al. (1995) observed that MSC and SNV transformed spectra are closely related and that the difference in prediction ability between these methods is very small.

Hence, SNV eliminates the interference of scatter events by individually transforming the spectral data x_i into new values z_i, where $i = 1, \ldots, p$ (p are the wavelengths), according to the following equation:

$$z_i = \frac{x_i - m}{s},$$

where m corresponds to the mean and s to the standard deviation of x_i values in the original spectrum [Fearn *et al.*, 2009].

Baseline Correction

Since the obtained spectra are not always grounded at zero, methods for baseline correction are usually necessary to remove both baseline offset and slope from a spectrum. The type of algorithm used depends on the baseline correction needed. Spectra which are dislocated from zero by a constant value are the simpler cases and, consequently, subtracting the value in question from the spectrum is usually enough. However, there are cases where the baseline presents a slope or even spectra with curvatures, which makes baseline correction more difficult. In these cases, an algorithm generating a function, a linear or polynomial function, can bring the spectrum to zero [Otto, 1999; Smith, 2011].

Baseline correction has a limited utility as a spectral pre-processing, since it is difficult to find a function that exactly adjusts to the spectrum curvature. Although there are algorithms that automatically determine the best parallel function, they do not always work properly and may add variance to the data. Furthermore, the slope and curvature along the spectrum is not always the same, so a unique function will hardly correctly adjust to the entire spectrum.

Considering the disadvantages related to baseline correction methods, derivatives for offset correction may be preferred. But the problem of derivatives' application is that the resulting spectra will be noisier than the raw one. In cases where there is a low SNR, baseline correction must be applied instead [Smith, 2011].

Normalization

There are many possible ways to normalize the data. Normalization involves multiplying all spectra by a different scaling factor for each wavenumber. The goal is to remove differences between the samples that are related with factors, such as differences in the samples' number of cells, and not with the property of interest. It should be noted that a careful design of the experience is still a critical factor that must be always taken into account before pre-processing the data. There are several methods for normalizing spectral data and a full review on this topic may be found at Randolph (2006).

Spectral Derivatives

Spectral derivatives can be used to eliminate offset and background slope variations among spectra. The first derivative removes baseline offset variations in spectral profiles and the second derivative removes both baseline offset differences and differences in baseline slopes between spectra [Otto, 1999].

First and second derivatives also enable the resolution of overlapping peaks, being the second derivative the most used for this purpose. However, before applying derivatives, it is important to have in mind that the derivative spectra will have more noise than the initial spectra and, consequently, a decrease in the SNR will be observed. In order to avoid the SNR decreasing, the smoothing has to be incorporated when applying derivative. The Savitzky-Golay smoothing is the most common algorithm used to avoid the decrease of the SNR. Its principle is the same of an average filter, i.e., each point of the dataset is replaced by the average of itself and *n* points before and after [Lourenço *et al.*, 2012; Scholz *et al.*, 2012].

II.3.2. Multivariate data analysis

The most widely used chemometric techniques are principal component analysis (PCA) and partial least-squares (PLS) regression.

Principal Component Analysis (PCA)

The PCA is a data-reduction method extensively used for qualitative spectral analysis that works by reducing the dimension of a dataset to a simpler representation in the space of the new variables, called principal components (PCs). PCs are ordered in terms of variance explained in the data set, with the first PCs representing the major variance in the data. Sometimes the variance in data can be distributed by more PCs, so it may be more difficult to select those which are relevant to extract some useful information [Jollife, 2002]. This kind of method is a very useful tool for chemometricians, not only for data compression but also information extraction, allowing the identification of major trends in the data [Naes *et al.*, 2002].

The PCA model can be described in matrix notation as:

$$X = TP^T + E$$

where X is the spectral data matrix, T is the matrix containing the scores of the PCs, P the matrix containing the loadings and E the matrix that contains the model residuals and represents the noise or irrelevant variability in X. The scores in T are linear combinations of the original variables of X (wavelengths). The loadings in P are estimated by regressing X on to T and the residual matrix E is calculated by subtracting the estimated TP^T from X [Naes *et al.*, 2002].

Data evaluation and qualification can be generally achieved by plotting different combinations of PC's scores, since it is easier to visualize and evaluate the samples in a smaller dimensional space. As the fraction of variance can be covered by one, two or three PCs, it is possible to visualize almost the entire data by plotting these PCs against each other [Otto, 1999]. In theory the samples with closer scores will be more similar to each other.

Partial Least Squares (PLS) regression

The PLS regression is a quantitative method that establishes a relationship between spectra and the quantifiable properties of samples. It works by determining a small number of latent variables (lv) that allow predicting the sample properties by using the spectral data as efficiently as possible [Naes *et al.*, 2002].

Let X be the mean-centered $n \times p$ matrix composed of the n sample vectors x_i, $i = 1, \dots, p$ containing the spectral measurements at p wavelengths and let y be the mean-centred vector containing the reference values for the variable of interest. With this information, PLS finds new variables t_i, $i = 1, \dots, p$, which will be used to estimate the lv, and determines the loadings matrix P and y-loadings vector q by maximizing the correlation between those variables t_i found, as described below:

$$X = TP^T + E$$

$$y = Tq^T + f,$$

where E and f are the X and y residuals, which are the difference between the observed and the modelled variable [Naes *et al.*, 2002].

The PLS regression coefficients β are given by:

$$\beta = W(P^T W)^{-1}(T^T T)^{-1} T y,$$

where W is the PLS weights matrix and can be used to obtain the predictions:

$$\hat{y} = X\beta$$

In order to evaluate the performance of the developed models to predict the samples' properties, the root mean squared error (RMSE) was used, which is based on the squared differences between real and predicted *y*-values. The RMSE is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(y_i^{predicted} - y_i{}^{real}\right)^2},$$

where s is the number of spectra. Besides RMSE, the coefficient of determination (R^2) was also calculated in order to evaluate the robustness of the PLS models.

The validation of the developed PLS models is usually performed using two different approaches: external validation, where a set of external samples not used for calibration was used for validating the model developed; and the leave-one-out (LOO) cross-validation, where the calibration and validation are done by successively excluding a sample from the calibration set and using it as validation set, until all samples have been used for calibration and validation. Generally, the choice of the approach to be used is dependent on the number of samples available.

Chapter III

In-situ near-infrared (NIR) versus high-throughput mid-infrared (MIR) spectroscopies to monitor biopharmaceuticals bioproduction

Abstract

The process development for biopharmaceuticals presents a number of relevant constraints, being the major one the fact that they are synthesized by living cells with inherent variability, further enhanced by sensitivity to the manufacturing environment. To monitor and consequently control the cultivation processes it is therefore relevant to develop *at-line* and/or *in-situ* monitoring techniques. The versatility presented by FT-IR spectroscopy, both in the near (NIR) and mid infrared (MIR) regions, makes it a relevant tool towards this goal as it enables an economic, rapid, sensitive and simultaneous measurement of all critical variables of the bioprocess. In the present work the high-throughput *at-line* MIR spectral analysis from dehydrated cell pellets and the *in-situ* analysis of the whole culture broth using a NIR fiber optic were compared for monitoring the same cultures of recombinant *E. coli* DH5α producing a plasmid model, conducted over different media compositions and on different cultivation modes (batch and fed-batch). For that, several Partial Least Square (PLS) regression models for MIR and NIR spectra data were built to estimate the host cell growth, the production of plasmid, the carbon sources consumption (glucose and glycerol) and the by-product acetate production and consumption. Robust calibration PLS models were developed, that are valid through different cultivation processes, presenting a range of final biomass concentrations between 5.6 to 12.1 g DCW/L, of final plasmid concentrations between 14 to 142 mg/L and of plasmid productions per biomass from 1.8 to 12.4 mg/g DCW. The PLS models developed are valid for control purposes in cases of possible industrial environment fluctuations and for optimizations purposes. The PLS models developed, both for MIR and NIR regions, presented very high and similar correlation coefficients and low predictive errors. The predictive errors for models based on MIR data, were 0.71, 8.55, 0.29, 0.4 and 0.4 concerning biomass, plasmid, glucose, glycerol and acetate, respectively. For NIR data, the predictive errors were 0.39, 7.86, 0.3, 0.23 and 0.41, respectively.

NIR spectral data is in general less informative as it results from combinations and overtones of the fundamental vibrations. Therefore, the slightly better PLS models based on NIR spectra could result from the fact that in the NIR setup an *in-situ* probe was used, whereas in the MIR setup it was necessary to extract the pellet sample from the bioreactor and subsequently dehydrate it, which could input an error associated to the biomass acquisition. Moreover, the concentration of glucose, glycerol and acetate were directly analyzed from the culture broth in the NIR setup while in the MIR setup this information was indirectly estimated obtained from the biomass.

In conclusion, NIR and MIR spectroscopy represents valuable approaches for bioprocess monitoring. The use of a NIR fiber optic probe enables to extract *in-situ*, i.e, in real time, information concerning the critical variables of the bioprocess. In cases of cultivation optimization, where multi-bioreactors of small size are used, the use of a NIR fiber optic probe may not be possible, or due to economic limitations of having a large number of these probes. The MIR setup can therefore represent an efficient alternative, as it can be conducted in high-throughput mode by using multi-microplates that enable the simultaneously reading of hundreds of samples from several cultures at once.

Keywords: Biopharmaceuticals, Bioprocess monitoring, Cultivation, MIR spectroscopy, NIR spectroscopy, PAT, PLS models, QbD

III.1. Introduction

The biopharmaceuticals products like plasmids have becoming more appealing due it is potential for advanced medical therapies (e.g. DNA vaccines and gene therapy), being the bacterium *Escherichia coli* the most used host microorganism for their production, since it presents capacity to grow under a wide range of conditions, from rich complex organic media to salt-based chemically defined media, as well as ease of manipulation by genetic engineering [Carnes, 2005; Coban *et al.*, 2013; Coban *et al.*, 2011; Kalams *et al.*, 2013; Moen *et al.*, 2009; Prather *et al.*, 2003; Scholz *et al.*, 2012; Shibui *et al.*, 2013; Yang, 1999].

However, the process development for biopharmaceuticals presents a number of relevant constraints, being the major one the fact that they are synthesized by living cells with inherent variability, further enhanced by sensitivity to the manufacturing environment. To monitor and consequently control the cultivation processes it is therefore relevant to develop monitoring techniques.

Currently, online information about a bioprocess concerning its critical variables (*e.g.*, biomass, products, nutrients and metabolites) is possible mainly through offline analyses, which are labor-intensive and time-consuming, and imply removing samples from the bioreactor. However, in order to better understanding the bioprocesses, and to reach a more economic and robust process regarding reproducibility, and consequently quality of the final product, the adoption of modern bioprocess monitoring tools based on *in-situ* analyses is essential, in accordance to the Process Analytical Technology (PAT) initiative, introduced by the Food and Drug Administration (FDA), in 2004 [FDA, 2004]. This is especially relevant in heterologous products used as medicines, i.e. in the case of biopharmaceuticals.

The introduction of the PAT initiative in the biopharmaceutical industry opened the window to the implementation of spectroscopic techniques, namely Fourier transform infrared (FT-IR) spectroscopy, to monitor bioprocesses. FT-IR spectroscopy is a physicochemical method that measures vibrations of the functional groups of molecules, providing therefore information about the biochemical composition of a biological sample. It is rapid, requires minimal sample preparation or no preparation at all and is multi-parametric, *i.e.*, it enables the determination of the concentration of multiple compounds at once, from a single spectroscopic measurement [Huang *et al.*, 2006; McGovern *et al.*, 2002; Schenk *et al.*, 2006].

The versatility presented by FT-IR spectroscopy, both in the near (NIR) and mid infrared (MIR) regions, makes it a potential tool in many applications, either in the laboratory or in industrial plants. Nevertheless, it is in the domain of monitoring and optimization of bioprocesses that this technique has increasingly been applied, as it enables a rapid, sensitive and simultaneous measurement of all critical variables of the bioprocess, namely, the host cell growth, the

production of the heterologous product, the carbon sources consumption and by-products (*e.g.*, acetate and ethanol) production and consumption [Di Egidio *et al.*, 2010; Roychoudhury *et al.*, 2006; Scholz *et al.*, 2012]. Depending on the infrared (IR) region used, MIR or NIR, IR spectroscopy presents specific characteristics, and therefore specific advantageous and limitations, that at the end may complement each other.

While MIR spectroscopy reflects the fundamental vibrations of the molecular bonds, NIR spectroscopy reflects overtones and combinations of vibrations, which makes MIR spectra more informative concerning the samples' biomolecular composition. However, due to the high absorption of water in the MIR region, it is usually necessary to take the samples from the bioreactor and subsequently dehydrate the samples, which increases the risk of bioreactor contamination and inputs a time delay in the analysis [Arnold *et al.*, 2002; Cimander and Mandenius, 2002; Guillen and Cabo, 1997; Tamburini *et al.*, 2003]. An advantageous of MIR-spectroscopy is that it is possible to *at-line* conduct the MIR spectral acquisition in a high-throughput mode, using micro-plates, which is particularly important if hundreds of samples are to be analyzed in a short period of time, as is the case of bioprocess optimization protocols [Scholz *et al.*, 2012].

In spite of being theoretically less informative, NIR spectroscopy is not so affected by the water present, and combined with chemometric techniques, also allows the construction of calibration models for the prediction of the critical variables of the bioprocess. Moreover, NIR fiber optic probes, that can be immersed directly in the culture broth and steam sterilized with it, enable the acquisition of information *in-situ*, *i.e.*, in real time [Arnold *et al.*, 2002; Cimander and Mandenius, 2002; Lopes *et al.*, 2013; Navrátil *et al.*, 2005; Shenk *et al.*, 2001; Tamburini *et al.*, 2003; Tosi *et al.*, 2003]. Nevertheless, the use of this kind of probes in optimization protocols in microbioreactors may be impaired, due to space constraints and low biomass concentrations.

The use of chemometric techniques is crucial in IR spectroscopy, as it allows extracting quantitative information from the IR spectra [Huang *et al.*, 2006; McGovern *et al.*, 2002; Moen *et al.*, 2009]. Chemometrics is the application of statistical or mathematical methods to analyze chemical data and to extract information from certain chemical systems. These methods allow the extraction of the relevant information concerning the analytes of interest enclosed in the spectral data [Lourenço *et al.*, 2012; McGovern *et al.*, 2002]. The application of pre-processing techniques is also a very important step in the analysis of spectral data, since they enable the elimination of physical phenomena, thus improving the subsequent multivariate analysis [Rinnan *et al*., 2009; Sharaf *et al.*, 1986]. The classical spectrum pre-processing methods include multiplicative scatter correction (MSC), standard normal variate (SNV) and derivatives.

When scatter effects are the dominating sources of spectral variability, a MSC or a SNV transformation can be used to remove those effects. MSC is a pre-processing method eliminates the light scattering effect due to particles of different sizes and shapes [Helland *et al.*, 1995], by calculating a reference spectrum, which is usually given by the mean of all samples, and each spectrum is then fitted to this reference spectrum. With this method, each spectrum is corrected and all samples appear to have the same scatter level as the ideal. On the other hand, the SNV transformation centers each spectrum and then scales it by its own standard deviation. The resulting spectra have always zero mean and variance equal to one, and are thus independent of the original absorbance values. Dhanoa *et al.* (1994) and Helland *et al.* (1995) observed that MSC and SNV transformed spectra are closely related and the difference in models predictive ability using these pre-processing techniques is very small. Derivatives can be used to eliminate offset and background slope variations among spectra. The first derivative removes baseline offset variations in spectral profiles, whereas the second derivative removes both baseline offset differences and differences in baseline slopes between spectra.

For spectral data analysis, the most widely used chemometric techniques are principal component analysis (PCA) and partial least-squares (PLS) regression. The PCA is a data-reduction method extensively used for qualitative spectral analysis that reduces the dimension of a dataset to a simpler representation by creating new variables, called principal components. This kind of method is a very useful tool for chemometricians, not only for data compression but also for information extraction, allowing the identification of major trends in the data [Naes *et al.*, 2002]. However, for quantification purposes the most used multivariate data analysis is the PLS regression, used to establish a relationship between the spectra and the quantifiable properties of samples, by determining a small number of latent variables that allow predicting sample properties, using the spectral data as efficiently as possible [Naes *et al.*, 2002]. When a calibration model is developed from the full spectra, the prediction results can be affected by wavelengths that do not provide relevant information about the metabolite of interest. Wavelength selection is therefore very useful, as it allows eliminating the uninformative wavelengths [Triadaphillou *et al.*, 2007].

It is intended in the present work to compare the MIR and NIR spectroscopy in monitoring in high-throughput and *in-situ* mode, respectively, a heterologous product production over a recombinant culture. As expression system model, a recombinant *Escherichia coli* DH5α producing the plasmid pVAX-LacZ (Invitrogen, USA) was chosen, since *E. coli* is the most widely used expression host, and the production of plasmids has also gained considerable attention as a safer vector for gene therapy and DNA vaccination. The use of FT-IR spectroscopy for plasmid production monitoring has previously been studied for example, Lopes *et al.* (2013)

used *in-situ* NIR spectroscopy and Scholz *et al.* (2012) a high-throughput *at-line* MIR spectroscopy to monitor a plasmid bioproduction process in *Escherichia coli* cultures. However, the efficiency of the two techniques for monitoring plasmid bioprocesses, provided by these two studies, cannot be compared, as different cultures conditions were used, and in the case of the MIR spectroscopy monitoring few samples were used for model building that could have impaired the model prediction capability. In this regard, in the present work the high-throughput *at-line* MIR spectral analysis and the *in-situ* analysis using a NIR fiber optic were compared for monitoring the same cultures of recombinant *E. coli* DH5α producing pVAX-LacZ. For that, several PLS regression models for MIR and NIR data were built to estimate the critical variables of the bioprocess, such as the host cell growth, the production of plasmid, the carbon sources consumption (glucose and glycerol) and the by-product acetate production and consumption. As it was also intended to compare the robustness of the predicting models over different cultivation conditions, several PLS were also built based on cultures conducted over different media conditions and on different cultivation modes (batch and fed-batch).

III.2. Materials and Methods

III.2.1. Cultivation

Escherichia coli DH5-α containing the plasmid model pVAX-LacZ (Invitrogen, USA) was used. The stock cultures, grown on 2% (w/v) Luria-broth (Sigma, UK) and 30 µg/ml kanamycin (Sigma-Aldrich, Germany), were maintained in 40% (v/v) glycerol solution (Panreac Quimica SA, Spain) with 10 mM Tris-HCl (Sigma-Aldrich, Germany) buffer pH 8.0 at -80 °C. An aliquot of 10 µl of stock culture was inoculated into 1 L shake flask containing 300 mL with 20 g/L bactotryptone (BD, USA), 10 g/L yeast extract (Difco, USA), 10 g/L sodium chloride (Merck, Germany) and 30 µg/mL kanamycin), grown to mid-exponential phase, and then used to inoculate a batch culture to an initial optical density at 600 nm (OD_{600}) of approximately 0.5.

The cultivation was performed in a 2 L bioreactor (Biostat MD, B. Braun, Germany) with a 1.8 L working volume, in absence of antibiotic. Cultivation was maintained at pH 7.0 ± 0.1 by automatic control through 1 M NaOH (Fluka, Switzerland) or 1 M HCl (Sigma-Aldrich, Germany) addition, and at 37 ± 0.1 °C with a minimal dissolved oxygen concentration (DOC) of 30 ± 5% of air saturation, by automatic adjustment of the agitation rate, while adjusting the air flow rate range between 1.0 and 1.5 vvm (volume of air/volume of medium/minute). The initial batch cultivation media of the three cultures studied contained 10 g/L of yeast extract (Difco, USA), 20 g/L bactotryptone (BD, UK) and 7 g/L of glycerol (culture A), 7 g/L of glucose (culture

B), and 6 g/L of glycerol and 8 g/L of glucose (culture C). An exponential feeding phase was started on cultures B and C with a feeding of 0.3 L medium, containing 22.5 g yeast extract, 22.5 g bactotryptone and 45 g glucose, and considering a maximum specific growth rate of 0.18 h^{-1} and a constant yield of biomass per glucose of 0.6 g/g. Samples were taken from the bioreactor along the culture, and subsequently used for offline reference analysis of biomass, glucose, glycerol, acetate and plasmid.

III.2.2. Reference analyses

Biomass in units of dry cell weight (DCW) per volume of culture medium (g/L) was determined by centrifuging the cultivation samples, washing the pellet with 0.9% (w/v) sodium chloride and drying at 80 ºC until constant weight. The bacterial cell pellet and the supernatant obtained from sample centrifugation (Hermle Z160M, Germany) were frozen at -20 ºC. Glucose, glycerol and acetate were determined by HPLC with a L-6200 Intelligent Pump (Merck-Hitachi, UK), a L-7490 LaCrom-Ri-detector (Merck, Germany), a D-2500 Chromato-integrator (Merck-Hitachi, Germany) and an Aminex® Fermentation Monitor HPLC column (Bio-Rad, USA) maintained at 50 °C, and by using H_2SO_4 at 0.6 mL/min as eluent. Plasmids were extracted from the bacteria cell by the alkaline cell lysis method, and subsequent plasmid concentration and purity degree were determined by hydrophobic interaction HPLC, as described in Scholz *et al.* (2012).

III.2.3. MIR spectroscopy

The cell pellet obtained from the centrifugation of each 1 mL sample taken from the bioreactor was resuspended with NaCl 0.9% (w/v), so that an equivalent optical density of 6.0 (at 600 nm) in all samples was achieved. Triplicates of 25 µL of this suspension were placed on IR-transparent ZnSe microtiter plates with 96 wells (Bruker Optics, Germany) and subsequently dehydrated for 2.5 h in a vacuum desiccator (ME2, Vaccubrand, Germany). The MIR spectra were recorded in transmission mode by a HTS-XT associated to Vertex-70 spectrometer (Bruker Optics), using a spectral resolution of 4 cm^{-1} and 40 scans per sample.

III.2.4. NIR spectroscopy

NIR spectra were obtained using an NIR transflection fiber optic probe IN-271P (Bruker Optics, Germany), with a pathlength of 2 mm, coupled to a Vertex-70 spectrometer (Bruker Optics, Germany) with a TE-InGaAs detector. The fiber optic probe was submerged in the bioreactor and stem sterilized simultaneously with the cultivation medium. NIR spectra were

collected every 2 minutes in the 12500-5400 cm^{-1} (800-1851 nm) range, consisting of 32 coadded scans with 8 cm^{-1} resolution (2 nm steps). The scanner velocity was set to 20 kHz and the aperture setting defined was 6 mm.

III.2.5. Chemometric Methods

MIR data consisted of mean spectra of triplicates in each well of the ZnSe plate, while NIR data consisted of the spectra correspondent to the samples taken from the bioreactor and analyzed by offline reference methods.

Pre-processing

The following data pre-processing methods were studied: constant offset elimination, straight line subtraction, first and second derivatives, multiplicative scatter correction (MSC) and standard normal variate (SNV), and a combination between them.

While constant offset elimination shifts the spectra in order to set the y-minimum to zero through the subtraction of the spectra by a certain constant, straight line subtraction fits a straight line to the spectra and subtracts it, enabling the shift of the spectra to zero [Otto, 1999; Smith, 2011]. Spectral first and second derivatives were also employed to remove baseline offsets. As derivatives usually broaden spectra noise, a Savitzky-Golay smoothing was applied, where each point of the dataset is replaced by the average of itself and *n* points before and after.

SNV eliminates the interference of scatter events by individually transforming the spectral data x_i into new values z_i, where $i = 1, \ldots, p$ (p are the wavelengths), according to the following equation:

$$z_i = \frac{x_i - m}{s},$$

where m corresponds to the mean and s to the standard deviation of x_i values in the original spectrum [Fearn *et al.*, 2009].

MSC was also used to eliminate changes in spectra due to radiation scattering, by determined the mean spectrum of replicate spectra, by performing a transformation where the spectral data x_i is converted into new values z_i, where $i = 1, \ldots, p$, with p being the wavelengths [Fearn *et al.*, 2009]. The following equation describes the transformation from x_i to z_i:

$$z_i = \frac{x_i - a}{b},$$

where a represents the intercept and b the slope of a least squares regression of x_i on the values r_i coming from the reference spectra.

Multivariate data analysis

For multivariate calibration the PLS method, also known as projection to latent structures, was applied, by determining a small number of latent variables (lv) that can predict sample properties by using the spectral data.

Let X be the mean-centered $n \times p$ matrix composed of the n sample vectors x_i, $i = 1, \dots, p$ containing the spectral measurements at p wavelengths and let y be the mean-centred vector containing the reference values for the variable of interest. With this information, PLS finds new variables t_i, $i = 1, \dots, p$, which will be used to estimate the lv, and determines the loadings matrix P and y-loadings vector q by maximizing the correlation between those variables t_i found, as described below:

$$X = TP^T + E$$

$$y = Tq^T + f,$$

where E and f are the X and y residuals, which are the difference between the observed and the modelled variable [Naes *et al.*, 2002].

The PLS regression coefficients β are given by:

$$\beta = W(P^T W)^{-1}(T^T T)^{-1} T y,$$

where W is the PLS weights matrix and can be used to obtain the predictions:

$$\hat{y} = X\beta$$

In order to evaluate the performance of the developed models to predict the samples' properties, the root mean squared error (RMSE) was used, which is based on the squared differences between real and predicted *y*-values. The RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(y_i^{predicted} - y_i^{real}\right)^2},$$

where s is the number of spectra. Besides RMSE, the coefficient of determination (R^2) was also calculated in order to evaluate the robustness of the PLS models.

The validation of the developed PLS models was performed using two different approaches: external validation, where a set of external samples not used for calibration was used for validating the model developed; and the leave-one-out (LOO) cross-validation, where the

calibration and validation are done by successively excluding a sample from the calibration set and using it as validation set, until all samples have been used for calibration and validation. The choice of the technique to be used is dependent on the number of samples available.

Wavelength selection

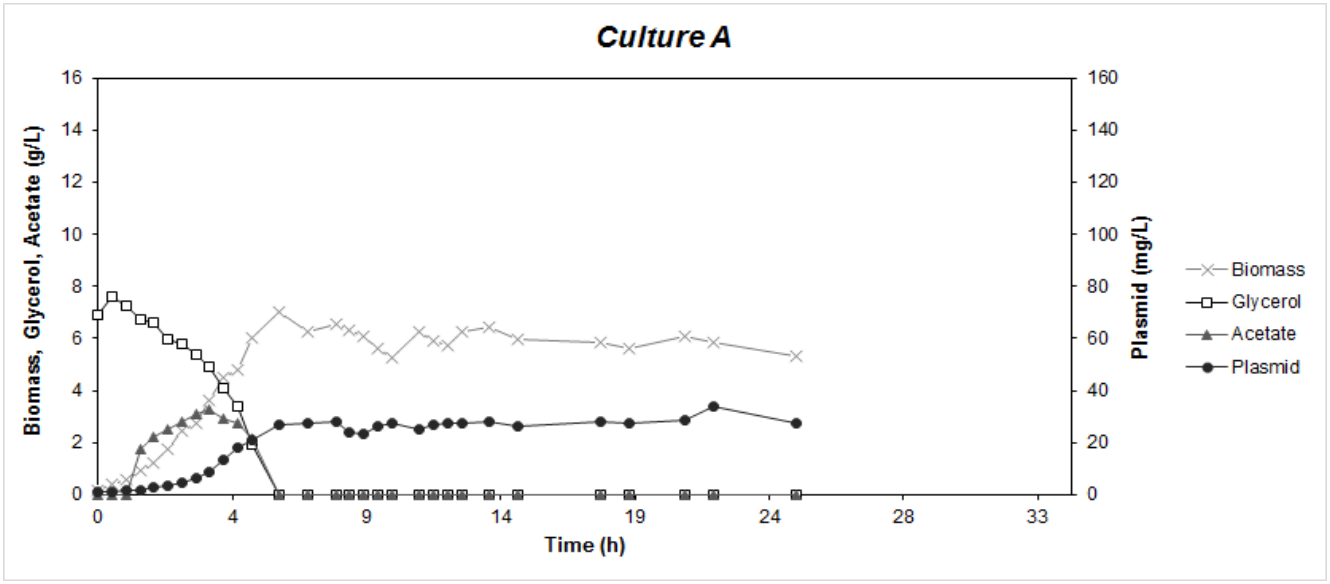
The wavelength selection was performed by dividing the spectral region into 10 equal subregions and finding the best combination of spectral regions providing the best predictive performance. The calculation starts with one subregion and after the best subregion has been found the next subregions will be individually added after the best combination of regions has been found. This procedure was repeated for data pre-processed using the techniques described above. The best PLS model was assessed by picking the wavelength regions and pre-processing technique providing the smallest RMSE. Both wavelength selection and PLS model building were performed using software OPUS Ver. 7.2 (Bruker, Germany).

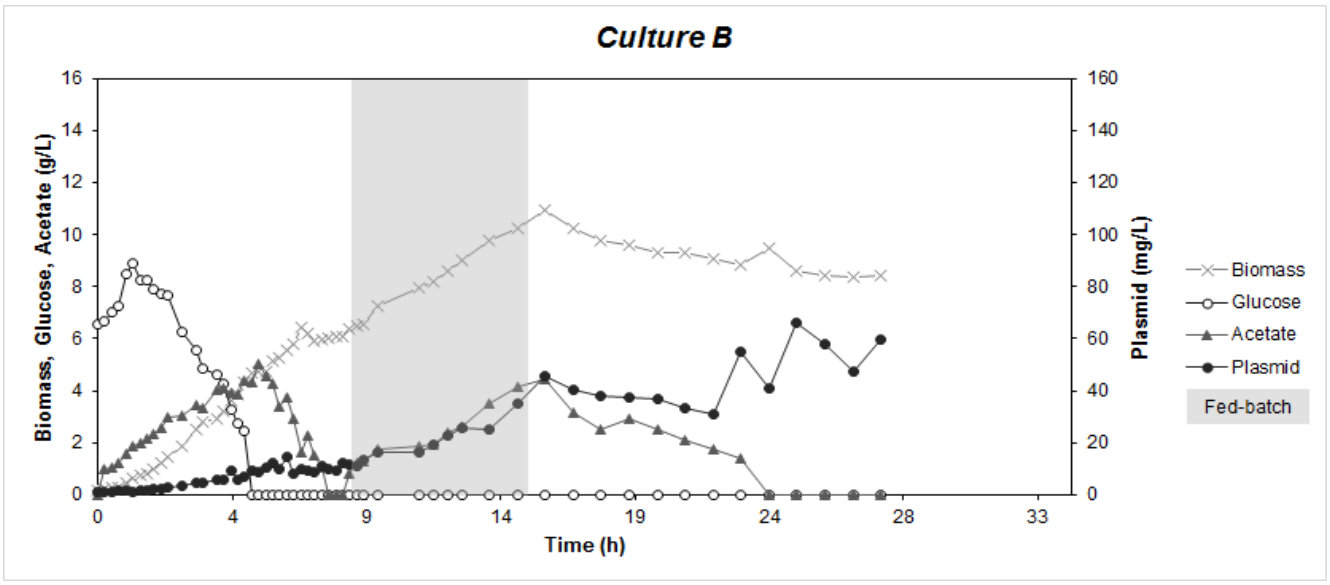
III.3. Results and Discussion

One of the major challenges associated to the production of biopharmaceuticals is the development of methods to *at-line* or *in-situ* monitoring the production of the recombinant product, this way promoting process control to ensure high quality products and optimization towards a more economical bioprocess. As a biopharmaceutical product, plasmids have become appealing due to its potential for advanced medical therapies like DNA vaccines and gene therapy [Carnes, 2005; Coban *et al.*, 2013; Coban *et al.*, 2011; Kalams *et al.*, 2013; Prather *et al.*, 2003; Shibui *et al.*, 2013]. Plasmids are usually produced in recombinant *E. coli* cultures, which as living cells present inherent variability that is further enhanced by the cell sensitivity to the manufacturing environment. Therefore, it is crucial the development of *in-situ* bioprocess monitoring tools along the culture time so that the plasmid bioproduction could be controlled in real-time, as described in the present work by using a NIR fiber-optic spectroscopy probe stem sterilized with the bioreactor. In cases where the fiber-optic probe cannot be used, due to limitations of bioreactor dimensions as in the case of optimization protocols using microbioreactors, the use of high-throughput analysis using microplates based in MIR spectroscopy could represent a solution. In both cases (in NIR and in MIR spectroscopy) the ideal calibration models developed should be valid for a wide range of cultivation conditions, that will cover perturbations of the cultivations conditions naturally occurring at industrial scale, or that will cover the cultivation conditions evaluated under optimization protocols.

To develop robust PLS models, enabling to predict the key variables of the plasmid bioproduction from a IR-spectrum, three *E. coli* cultures conducted under different mixtures of glucose and glycerol as carbon sources on the batch phase and over different cultivation strategies (batch and fed-batch) were prepared. The batch phase of cultivation A to C were conducted with glycerol (Culture A), with glucose (Culture B) and with a mixture of glucose and glycerol (Culture C). After the batch phase, a feeding phase with glucose was started on cultures B and C. The three cultures were monitored by high-throughput mode in MIR spectroscopy and *in-situ* NIR spectroscopy.

Considering the batch phases of cultures A and B, it was possible to observe that culture A (conducted only on glycerol) produced 2 times more plasmid than culture B (conducted only on glucose), most probably as a result of the lower specific growth rate and lower acetate productions, that however resulted in an also lower volumetric productivity. Indeed, glycerol has being used as an alternative C-source in relation to glucose, in order to minimize overflow metabolism, due to a lower glycerol transport to the cell, that consequently will increase the energetic metabolism efficiency while reducing the acetate production [Korz *et al.*, 1995; Hansen and Eriksen, 2007; Scholz *et al*., 2012]. The acetate production, besides its direct consequence of decrease biomass yield, may also reduce product yield per biomass [Smirnova and Oktyabrskii, 1985]. Therefore, the use of glycerol instead of glucose will implies a lower specific growth rate and consequently a lower plasmid productivity, that however results in a lower acetate production and consequently on a slight higher biomass and a much higher plasmid production per biomass, and consequently on a much higher final plasmid concentration in relation to the culture conducted on glucose. Since the goal of the biopharmaceutical companies is to obtain simultaneously maximum plasmid final concentration, plasmid yield per biomass and plasmid productivity, a mixture of glucose and glycerol as carbon source should be therefore used (**Table Figure III.5**, **III.1**) [Scholz *et al.,* 2012]. Indeed, it was observed that the batch culture C, conducted with a mixture of glucose and glycerol, presented the highest plasmid productivity of 4.4 mg/L/h, plasmid concentration of 42 mg/L and plasmid production per biomass of 7.23 mg/g (**Table III.1**), when compared with the other two batches phases.





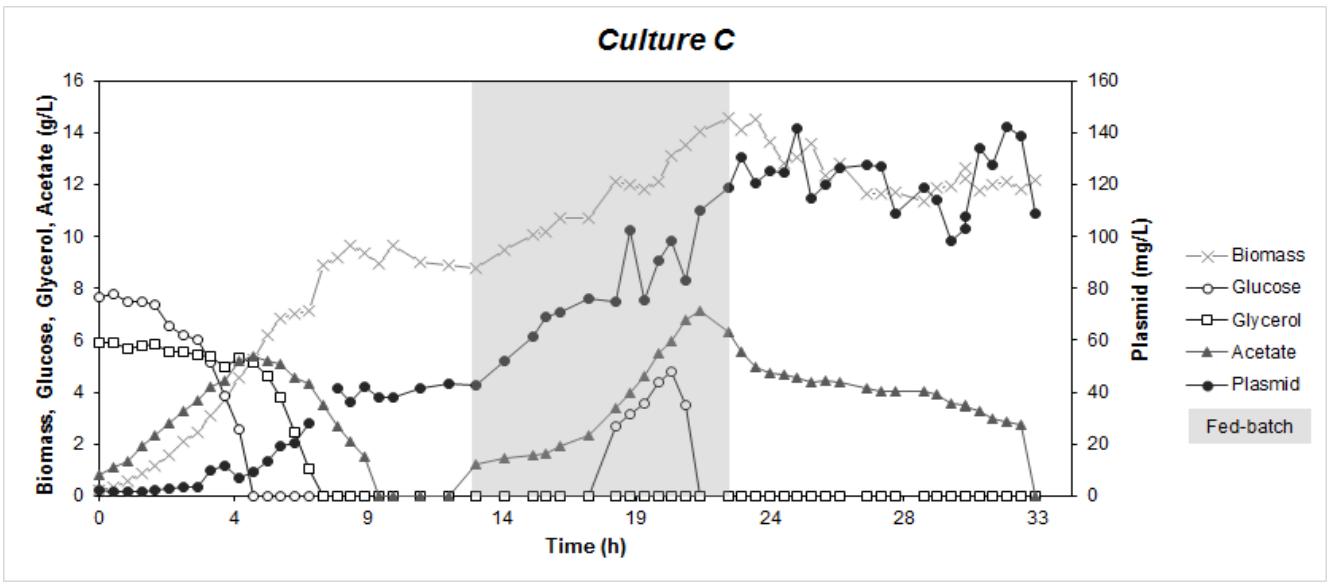


Figure III.5: Evolution along the time of the biomass, glucose, glycerol, acetate and plasmid concentrations for the three cultures (A to C), conducted with a C-source composition on the batch phase of glycerol (culture A), glucose (culture B) and glucose and glycerol (culture C). After all acetate produced during the batch phase was consumed on culture B and C an exponential feeding phase with glucose was started considering a $\mu=0.18h^{-1}$, $Y_{X/S}=0.6$ and S=150g/L. The feeding phase is represented in the graph by the grey area.

Table III.1: Description of the three batches cultures conducted with mixtures of glucose and glycerol as carbon source. The following parameters are relative to the time at which the maximum plasmid production was achieved: time, biomass, maximum plasmid and final plasmid productivity.

	Culture A	Culture B	Culture C
[glucose] (g/L)	-	7.0	8.0
[glycerol] (g/L)	7.0	-	6.0
maximum [acetate] (g/L)	3.3	5	5.4
time (h)	22	7	9.5
[biomass] (g/L]	5.9	5.6	9.4
maximum [plasmid] (mg/L)	34	14.4	42
plasmid/biomass (mg/g)	4.8	1.8	7.2
final plasmid productivity (mg/L/h)	1.5	2.1	4.4
specific growth rate in glucose (h^{-1})	-	0.78	0.59
specific growth rate in glycerol (h^{-1})	0.66	-	0.31

Comparing the two fed-batch cultures, culture C produced about 2 times more plasmid in relation to the culture B, and an increase of approximately 60% in the final plasmid productivity (**Table III.2**). The high plasmid productivities observed in the fed-batch phase of culture C can be related to its batch phase, which was conducted with mixtures of glucose and glycerol as carbon source, which might have contributed to maximize plasmid concentrations, plasmid yield per biomass and plasmid productivities.

Table III.2: Description of the two fed-batches cultures conducted with mixtures of glucose and glycerol as carbon source. The following parameters are relative to time where the maximum plasmid production was achieved: time, biomass, maximum plasmid and final plasmid productivity.

	Culture B	Culture C
time (h)	25	32.5
maximum [biomass] (g/L]	8.6	12.1
maximum [plasmid] (mg/L)	66	142
plasmid/biomass (mg/g)	8.7	12.4
final plasmid productivity (mg/L/h)	2.6	4.4
maximum [acetate] during feeding (g/L)	4.1	7.2

These data clearly show that slight differences concerning types and concentrations of carbon sources, as well as the cultivation strategy, have a relevant impact on the culture performance, therefore announcing the need to monitor the bioprocess towards more reproducible processes and to understand how the above factors affect the entire production process. For that, MIR and NIR spectral data (**Figure III.6**) from the cultures described above (three batches phase and two fed-batches phases) were used to build PLS models for predicting the variables of interest in the plasmid bioprocess, namely, glucose, glycerol, acetate, biomass and plasmid concentrations, as for optimization purposes it is very important that PLS models cover a wide range of cultivation conditions.

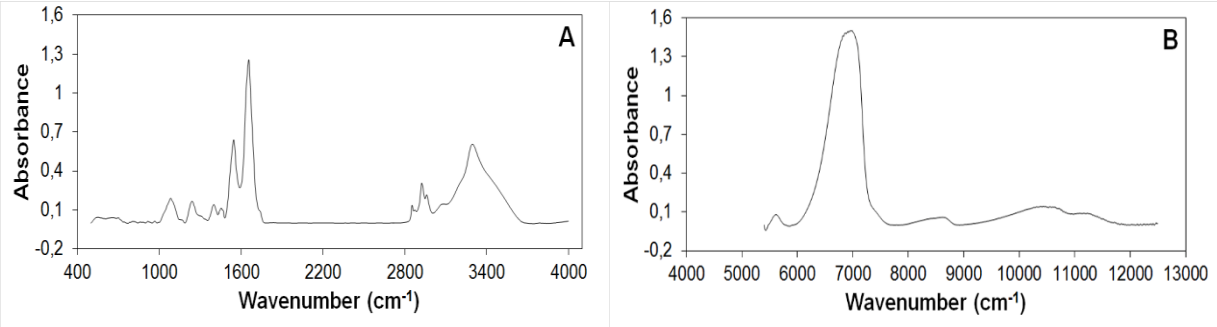


Figure III.6: Examples of MIR (A) and NIR (B) spectra acquired during bioprocess monitoring.

III.3.1. PLS modeling of MIR spectra

PLS models using the MIR spectral data from the three cultures (A to C) were built for biomass, plasmid, glucose, glycerol and acetate, and subsequently evaluated concerning its accuracy and robustness. It should be noted that the prediction of the concentration of the glucose, glycerol and acetate in the culture broth was possible based on metabolism-induced correlations between the spectra and the concentration of the nutrients and metabolites in the extracellular medium.

Several PLS models were built, presenting combinations of the following pre-processing techniques: constant offset elimination, straight line subtraction, multiplicative scatter correction (MSC), standard normal variate (SNV) and first and second derivatives. The PLS models were also optimized using a strategy for wavenumber selection for identifying the spectral regions that best relate with the metabolite, as the prediction results can be improved by excluding spectral regions that do not contain metabolite specific information (Kansiz *et al.*, 2001). The best PLS model was assessed by picking the wavelength regions and several pre-processing technique providing the smallest RMSE. For the biomass, plasmid and acetate models, the RMSE was

calculated based on an independent test validation set due to the larger number of samples available. For the glucose and glycerol models, the RMSE was obtained by LOO cross-validation, as fewer samples were available, provided by the batch consumption phase. The acetate model was also built based on samples from the batch consumption phase.

High accurate PLS regression models were obtained for biomass, plasmid and glucose concentrations, with a R^2 of 0.97 and a RMSE of 0.71, 8.55 and 0.29, respectively, that represented perceptual errors in relation to the range of units of the variables of 4.8, 6.0 and 3.2%, respectively (**Table III.3**; **Figure III.7**). All models produced better results concerning accuracy and prediction errors when compared to the results obtained by Scholz *et al.* (2012), who predict the metabolites concentration of five batch cultures with different initial medium compositions. These five cultures presented a distinct culture behavior, with maximum biomass concentrations between 6.7 and 12.8 g/L and maximum amounts of plasmid produced between 11 and 95 mg/L. Despite the large variability present in the present cultures, being even higher concerning the plasmid range, a great improvement in model performance was seen in these study, which may be partially explained by the use of a larger number of samples that were taken along the time of the bioprocess considered for PLS model building.

Table III.3: Best MIR PLS regression models for biomass, plasmid, glucose, glycerol and acetate concentrations concerning the R^2, the RMSE, the number of latent variables (lv) used, the pre-processing technique and the selected spectral regions for culture A, B and C (*LOO cross-validation).

	***At-line* monitoring by MIR spectroscopy**							
	R^2	**lv**	**RMSE**	**Percentage of error (%)**	**No. calibration samples**	**No. validation samples**	**Pre-processing**	**Wavelength selection**
Biomass (g/L)	0.97	7	0.71	4.8	116	27	Second Derivative	3299,8 - 2946,9 ; 1199,6 - 499,5
Plasmid (mg/L)	0.97	8	8.55	6.0	116	27	First Derivative + MSC	3299,8 - 2597,8 ; 2248,7 - 1897,7
Glucose (g/L)*	0.97	8	0.29	3.2	28	-	First Derivative	3998 - 3297,9 ; 2948,8 - 2597,8 ; 2248,7 - 1897,7
Glycerol (g/L)*	0.92	8	0.40	5.2	24	-	Second Derivative	3998 - 3647 ; 2948,8 - 1897,7 ; 850,5 - 499,5
Acetate (g/L)	0.91	7	0.40	7.4	43	12	Straight line subtraction	2948,8 - 2597,8 ; 2248,7 - 1897,7

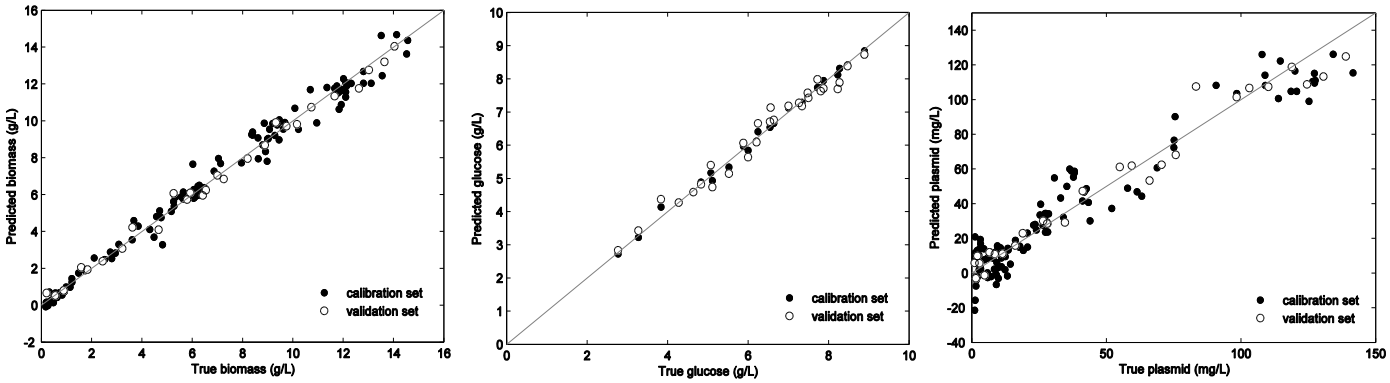


Figure III.7: True and predicted biomass, glucose and plasmid concentrations obtained by the PLS regression model based on the MIR spectra, considering data from three different cultures (A, B and C).

Regarding the plasmid model, the high RMSE achieved can be explained by the wide range of plasmid concentration, thus making the plasmid prediction fairly acceptable (6.0% of the maximum plasmid concentration). Moreover, the experimental errors in the determination of plasmid concentration (**Figure III.5)**, may also influence the prediction error, as this analysis is based on plasmid HPLC analysis after plasmid cell extraction by cell alkaline lysis. The cell plasmid extraction step presents serious concerns, as during plasmid extraction, cells at different metabolic states may present different contents in nucleases, and consequently the efficiency of the plasmid extraction will widely vary, leading to analytical errors between 5 and 8%.

Less accurate models were obtained for glycerol and acetate, with the acetate model providing the highest percentage of prediction error. This error, however, might still be considered reasonable, when compared to the errors provided by the conventional methods for the determination of acetate, as the most used method for HPLC analysis is based on a non-specific HPLC column that presents a broad range of applicability but lower specificity and sensitivity. On the other hand, the good result for the glycerol model regarding the prediction error may be due to a very specific model built based on few samples within a narrow concentration range.

Figure III.8 summarizes the regression coefficients of the PLS models developed for all metabolites. It can be seen that for each variable studied specific spectral windows were selected for model building. Although some overlapping might be expected, each model was developed based on specific spectral regions, with distinct intensities observed, which ensure that one metabolite is not being predicted by another.

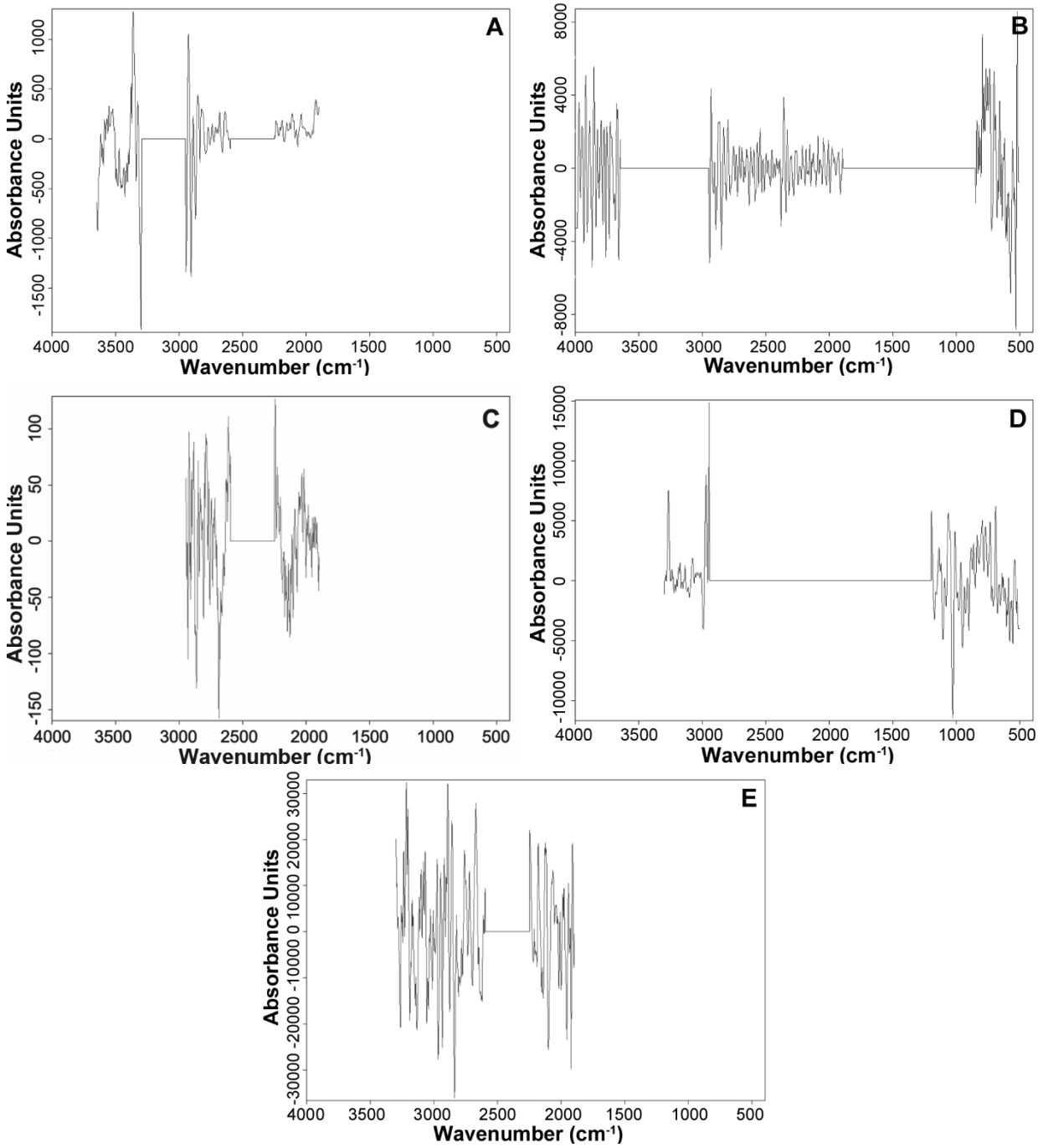


Figure III.8: PLS regression vectors obtained from MIR models for A) glucose, B) glycerol, C) acetate, D) biomass and E) plasmid concentrations.

III.3.2. PLS modeling of NIR spectra

PLS models were built for glucose, glycerol, acetate, biomass and plasmid concentrations using NIR data from the three cultures studied. For the biomass, plasmid and acetate models, the RMSE was calculated based on an independent test validation set, given the larger number of samples available. For the glucose and glycerol models, the RMSE was obtained by LOO cross-validation, as fewer samples were available, provided by batch consumption phase. The acetate model was also built based on samples from the batch consumption phase. As for the PLS models based on MIR data, the best PLS model was assessed by picking the wavelength regions and several pre-processing technique providing the smallest RMSE. The PLS regression vectors (**Figure III.9**) of all PLS models, layout the models specificity for each variable, as different spectra regions contribute to each model building. As expected the calibration models based on NIR presents more overlapping spectra regions among each other, as NIR spectroscopy reflects overtones and combinations of vibrations, where MIR spectroscopy reflects fundamental vibrations modes. It was observed that good PLS models were achieved for all variables studied (**Table III.4**). High accurate PLS regression models were achieved for biomass and glucose, with a $R^2 \geq 0.98$ and a low RMSE of 0.39 and 0.30, respectively (**Table III.4**; **Figure III.10**). The biomass model yielded a similar R^2 compared to previous reports on *E. coli* cultures [Arnold *et al.*, 2002; Cimander and Mandenius, 2002], but lower prediction errors. Accurate PLS models were also obtained for plasmid (**Table III.4**; **Figure III.10**), yielding a R^2 of 0.96 and a RMSE of 7.86. Although the RMSE of plasmid model seems high, when compared to the error associated to the prediction of the other metabolites, it is indeed a low RMSE (5.6% of the maximum plasmid concentration), if taken into account the range of plasmid concentrations, between 0 and 42 mg/L.

In the case of glycerol, an accurate PLS model with a R^2 of 0.96 and a RMSE of 0.23 was obtained, however, as for the PLS model for glycerol based on MIR data, a low number of samples was used, which might have produced a very specific model. A less accurate PLS model was also obtained for the acetate production, compared to the previous models, and as previously observed in the MIR region, which can be related to the distinct level of production of acetate achieved in cultures A and B, therefore contributing to greater complexity and consequently influencing the acetate prediction.

Considering that PLS models were developed based on cultures accounting for a high variability concerning the cultivation conditions and strategies, along with wide concentration ranges of the metabolites, *in-situ* monitoring of the main variables of the plasmid bioprocess, with a high predictive ability, was possible through NIR spectroscopy.

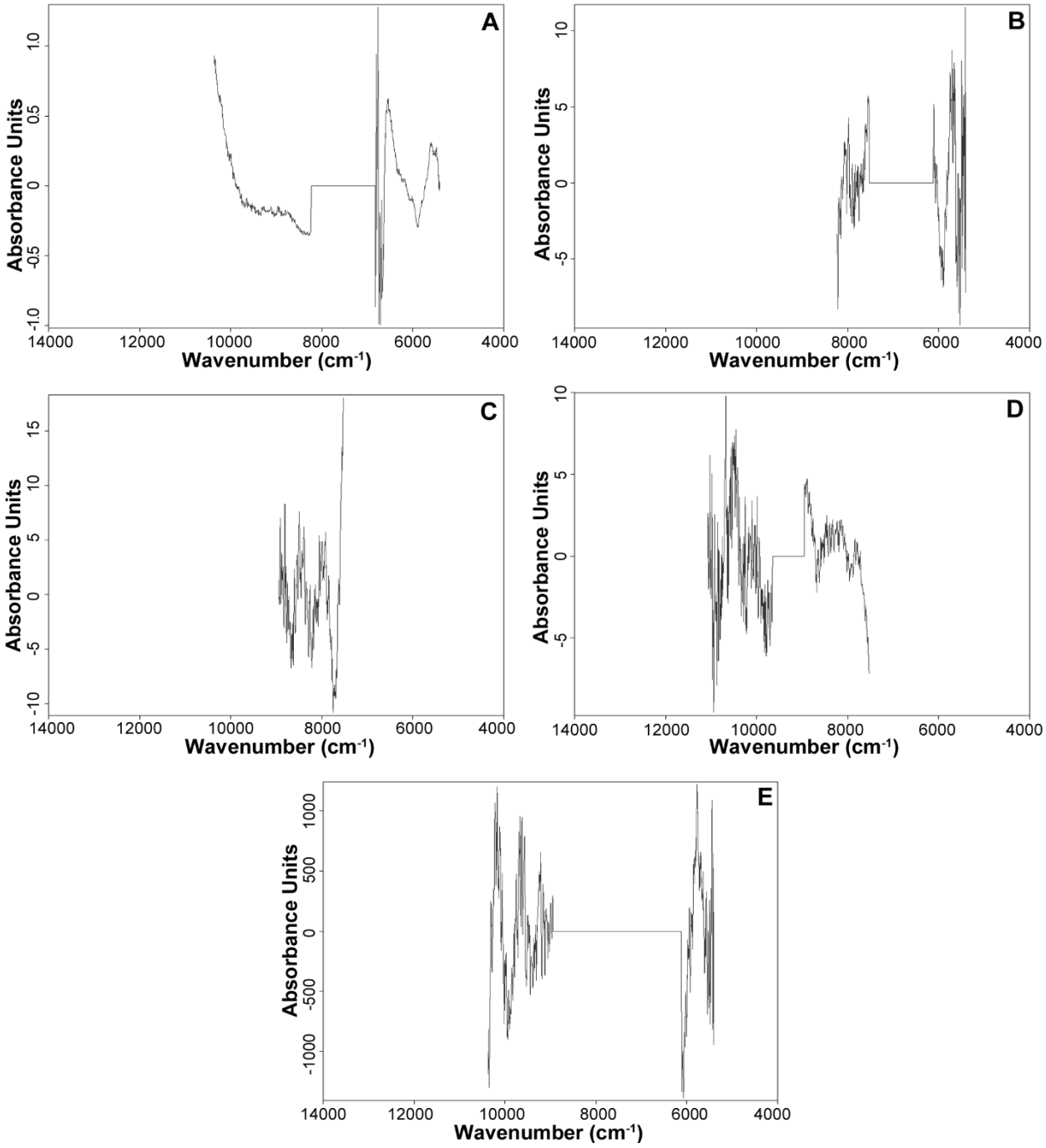


Figure III.9: PLS regression vectors obtained from NIR models for A) glucose, B) glycerol, C) acetate, D) biomass and E) plasmid concentrations.

Table III.4: Best NIR PLS regression models for biomass, plasmid, glucose, glycerol and acetate concentrations concerning the R^2, the RMSE, the number of latent variables (lv) used, the pre-processing technique and the selected spectral regions for culture A, B and C (*LOO cross-validation).

In-situ monitoring by NIR spectroscopy								
	R^2	**lv**	**RMSE**	**Percentage of error (%)**	**No. calibration samples**	**No. validation samples**	**Pre-processing**	**Wavelength selection**
Biomass (g/L)	0.99	7	0.39	2.6	116	27	Constant offset elimination	11077,7 - 9654,4 ; 8948,5 - 7525,3
Plasmid (mg/L)	0.96	7	7.86	5.6	116	27	SNV	103368 - 8944,7 ; 6109,7 - 5400
Glucose (g/L)*	0.98	6	0.30	3.4	27	-	None	10368 - 8235 ; 6819,4 - 5400
Glycerol (g/L)*	0.96	7	0.23	3.3	23	-	None	8238,8 - 7525,5 ; 6109,7 - 5400
Acetate (g/L)	0.90	4	0.41	7.6	44	15	Constant offset elimination	8948,5 - 7525,3

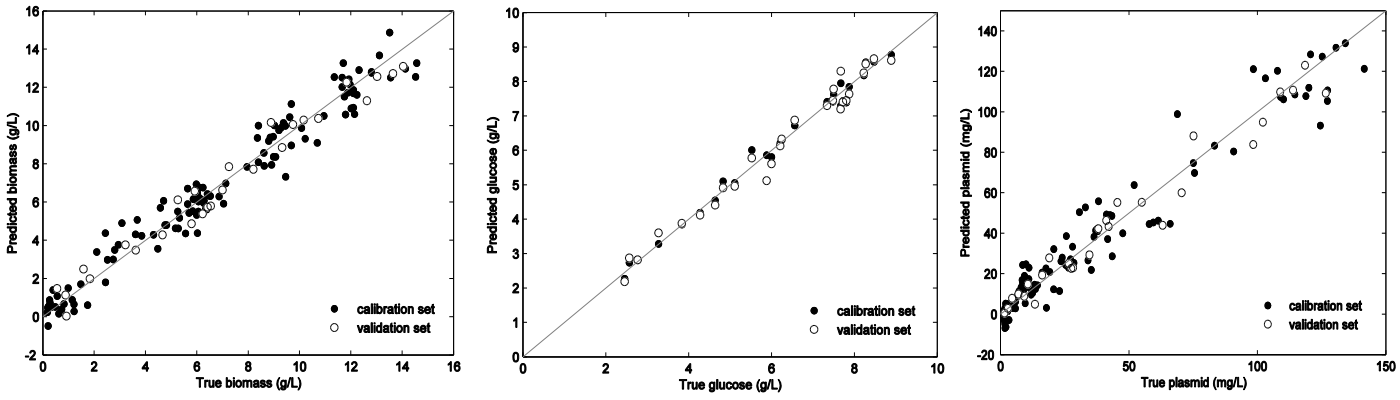


Figure III.10: True and predicted biomass, glucose and plasmid concentrations obtained by the PLS regression model based on the NIR spectra, considering data from three different cultures (A, B and C).

III.3.3. MIR versus NIR models

Generally, PLS models with MIR spectra present better results than models with NIR spectra, as shown by previous studies [Sandor *et al.*, 2013; Sivakesava *et al.*, 2001], as MIR spectroscopy reflects the fundamental vibration of the molecular bonds, therefore becoming more informative, in relation to NIR spectroscopy that reflects overtones and combinations vibration modes. However, using exactly the same data from 3 batch cultures and 2 feeding phases, very similar PLS regression models were obtained to predict the critical variables of the bioprocess as the concentrations of biomass, plasmid, glucose, glycerol and acetate. If it is taken into account the final prediction errors, most PLS models built based on NIR spectra are even slightly better than those built on MIR data. A possible reason for this result might be the use of a different NIR probe in this study that works in transflectance mode and presents a mirror with a conical shape that avoids the accumulation of solids and air bubbles in the pathlength, therefore improving the final results for NIR spectra. Furthermore, the transflectance mode most probably presents a wider range of applicability from low to high biomass concentrations. Indeed, comparing MIR and NIR data pre-processing for models' construction, PLS models built on NIR data did not require the use of derivatives as pre-processing, as reported by other authors using *in-situ* NIR probes [Arnold *et al.*, 2002; Cimander and Mandenius, 2002; Lopes *et al.*, 2013; Lourenço *et al.*, 2012; Navrátil *et al.*, 2005; Shenk *et al.*, 2001; Tamburini *et al.*, 2003; Tosi *et al.*, 2003]. For example, for glucose and glycerol models, no pre-processing was necessary, and only a constant offset elimination was applied for biomass and acetate models. The best PLS models for MIR data required data pre-processing using derivatives for most variables studied, except for the acetate model, for which a straight line subtraction was applied to the spectral data.

The results achieved for NIR and MIR data clearly show that both NIR and MIR spectroscopies represent valuable approaches for bioprocess monitoring, however, they must be chosen regarding the final purpose. For example, if the goal is to monitor the bioprocess along time, NIR spectroscopy may be chosen, since a NIR fiber-optic probe (stem sterilized with the bioreactor vessel) can be placed inside the bioreactor and extract information in real time. On the other hand, if several hundred of samples from several cultures are to be analyzed for optimizing cultures' conditions and strategies, high-throughput MIR spectroscopy could be the choice. Nevertheless, taking into account the necessary sample dehydration for MIR analysis, *in-situ* NIR spectroscopy, when available and there is not a minimum bioreactor volume, is still more promising, as spectral data acquired online, with no sample extraction/preparation, can be directly used also for optimization purposes.

Therefore, for bioprocess monitoring of biopharmaceutics, as plasmid production in recombinant *E. coli* hosts, MIR and NIR spectroscopies are techniques that present specific characteristics and therefore advantageous and limitations associated, were can be seen as complementary and together represent a powerful tool for bioprocess monitoring.

Chapter IV

Metabolic profiling of recombinant cell cultivations based on high-throughput FT-IR spectroscopy analysis

Abstract

The increasing interest in biopharmaceuticals products like plasmids that have becoming more appealing due it is potential for advanced medical therapies (e.g. DNA vaccines and gene therapy), calls for the need of developing economic ways for their production. However, genetic, physiological and environmental factors influence the expression of the cloned gene product with a high degree of complexity. Therefore, in order to control and optimize the performance of recombinant expression systems, it is very important to understand the complexity of the interrelationships between cultivation conditions and the genetic and physiological characteristics of the expression system. For that, the metabolic profile of two recombinant *E. coli* cultures producing plasmid pVAX-lacZ were evaluated based on FT-IR spectra collected in a high-throughput mode along the cultivation time.

The principal component analysis (PCA) method enabled to capture the metabolic state of the cell in both cultivations, as identifying the different C-sources consumption phases. It was also possible by direct analysis of the FT-IR spectra to acquire biochemical and metabolic information along the cultivation process: it was observed a decreasing of glycogen levels at high specific growth rate, namely during the carbon sources consumption; it was also possible to observe the RNA concentrations and transcriptional levels increase before the beginning of a new carbon sources consumption, most probably due to the need of new genes transcription, to enable the new carbon source metabolism; it was also observed an increase of the translational level (estimated as the ratio between the amide II spectral bands and the nucleic acids total) during the consumption of the carbon source, most probably as a result from a higher protein expression; it was also possible to identify protein conformational changes in the cell proteome.

In summary, FT-IR spectroscopy enables to acquire along the cultivation process of recombinant *E. coli* several features of the biochemical and the metabolic status of the cell, which could strong contribute to understand the complex interrelationships between the recombinant cell metabolism and the bioprocess towards the design of more economic and robust processes according to the PAT initiative.

Keywords: Bioprocess monitoring, FT-IR spectroscopy, Metabolic Profiling, PCA

IV.1. Introduction

The bacterium *Escherichia coli* is the most used host microorganism for the production of recombinant products, such as heterologous proteins and plasmids. The main reason for that is its capacity to grow under a wide range of conditions, from rich complex organic media to salt-based chemically defined media, as well as its ease manipulation by genetic engineering [Moen *et al.*, 2009; Prather *et al.*, 2003; Scholz *et al.*, 2012; Yang, 1999]. However, differences in the cultivation strategies (e.g., batch and fed-batch), environmental conditions and medium composition, are known to affect the stability and expression of the cloned gene product [O'Kennedy *et al.*, 2003; Ow *et al.*, 2007; Ow *et al.*, 2009]. The characteristics of the plasmid and the host cell, i.e., the cell expression system, are also important factors that should be carefully evaluated [McNeil and Harvey, 1990]. The combination of the above genetic, physiological and environmental factors influence the expression of the cloned gene product with a high degree of complexity. Therefore, in order to control and optimize the performance of recombinant systems, the effects of these factors and their interrelationships must be well understood.

To help understanding the complex relationships between the media composition, cultivation strategy and condition, and the characteristics of the cell expression system, the effect of these variables on recombinant cultures must be studied. This can be done by simply following the evolution along the time of critical variables of the process, namely, biomass, recombinant product, carbon source and acetic acid [Xiong *et al.*, 2008]. To further understand the complexity of the interrelationships between cultivation general conditions and the genetic and physiological characteristics of the expression system, other metabolic information from the host recombinant cell along the culture would be also highly useful. Understanding the complex interrelationships between cultivation conditions and the expression system characteristics would therefore bring valuable insight on the bioprocess, thus promoting control and optimization protocols towards a more economic and robust process regarding reproducibility and consequently quality, in accordance to the Process Analytical Technology (PAT) initiative launched in 2004 by the Food and Drug Administration (FDA) [FDA, 2004].

Currently, the extraction of metabolic information from the host recombinant cell along the bioprocess is performed by conventional cellular and molecular biology methods, which are limited, time-consuming and labor-intensive. Alternative techniques, like Fourier Transform Infrared (FT-IR) spectroscopy, a promising tool in the biomedical and pharmaceutical sciences, have emerged in the last decade and shown to be a powerful tool to obtain information about all stages of production's process in a simpler, rapid and high-throughput mode [Card *et al.*, 2008; Orsini *et al.*, 2000; Scholz *et al.*, 2012].

As a physicochemical method measuring vibrations of the functional groups of molecules, FT-IR spectroscopy is able to provide information about the structural and biochemical composition of a biological sample [Huang *et al.*, 2006; McGovern *et al.*, 2002; Schenk *et al.*, 2006; Scholz *et al.*, 2012]. Examples of biomedical and pharmaceutical applications of FT-IR spectroscopy include biodiagnostics (e.g., to detect inflammatory and precancerous cell states) [Gaigneaux *et al.*, 2004; Gazi *et al.*, 2006; Lee *et al.*, 2009; Lewis *et al.*, 2010; Maziak *et al.*, 2007] and screening the "mode of action" of new drugs [Gasper *et al.*, 2009]. FT-IR spectroscopy has also become important for bioprocess monitoring and control [Gasper *et al.*, 2009].

Generally, direct information can be obtained from a given IR spectrum, however, chemometric techniques enable further extraction of qualitative and quantitative information. The most common methods for these purposes are principal component analysis (PCA) and partial least squares (PLS) regression models [Huang *et al.*, 2006; McGovern *et al.*, 2002; Moen *et al.*, 2009]. The application of spectral pre-processing techniques is also an important step in multivariate spectral analysis, since they enable the elimination of physical phenomena, thus improving the extraction of quantitative and qualitative information [Rinnan *et al.*, 2009; Sharaf *et al.*, 1986].

The studies on bioprocess monitoring by FT-IR spectroscopy generally apply PLS regression methods to estimate from the FT-IR spectra critical variables of the bioprocess, i.e., biomass growth, the consumption of the main carbon sources as glucose and glycerol, the production and consumption of by-products as acetate and ethanol, and the recombinant product production as proteins and plasmids [Arnold *et al.*, 2002; Cimander and Mandenius, 2002; Lopes *et al.*, 2013; Navrátil *et al.*, 2005; Scholz *et al.*, 2012; Shenk *et al.*, 2001; Tamburini *et al.*, 2003; Tosi *et al.*, 2003]. However, besides this kind of information, it would be highly useful to extract from the FT-IR spectra other information that enables the biochemical and metabolic profiling of the host cell, e.g. the energetic level (i.e. the glycogen contents), total quantities of nucleic acids, proteins and lipids as well the apparent transcription and translation rate, as conducted by other authors in human cells and in carcinogenic studies [Baran *et al.*, 2013; Gaigneaux *et al.*, 2007; Gazi *et al.*, 2003; Maziak *et al.*, 2007; Lewis *et al.*, 2010].

Thus, the main goal of the present work is to evaluate the potential of FT-IR spectroscopy to characterize the biochemical and metabolic status of recombinant *E. coli* DH5-α cultures producing the plasmid model pVAX-lacZ (Invitrogen). Due to the relevance of using glucose as the main carbon-source to promote the bacterial growth and glycerol to minimize the production of acetate, two *E coli* cultures were conducted with different mixtures of glucose and glycerol and different cultivation strategies (batch and fed-batch). A PCA of the spectral data was first performed in order to evaluate the ability of FT-IR spectroscopy to reveal relationships between

spectral data and cellular events. The biochemical and metabolic profiling of the cell host along the cultivation process was evaluated by estimating from the spectral data for example lipids, proteins, nucleic acids and glicids and translational level.

IV.II. Materials and Methods

IV.2.1. Cultivation

Escherichia coli DH5-α containing the plasmid model pVAX-LacZ (Invitrogen, USA) was used. The stock cultures, grown on 2% (w/v) Luria-broth (Sigma, UK) and 30 µg/ml kanamycin (Sigma-Aldrich, Germany), were maintained in 40% (v/v) glycerol solution (Panreac Quimica SA, Spain) with 10 mM Tris-HCl (Sigma-Aldrich, Germany) buffer pH 8.0 at -80 °C. An aliquot of 10 µl of stock culture was inoculated into 1 L shake flask containing 300 mL with 20 g/L bactotryptone (BD, USA), 10 g/L yeast extract (Difco, USA), 10 g/L sodium chloride (Merck, Germany) and 30 µg/mL kanamycin (Sigma-Aldrich, Germany), and grown to mid-exponential phase (resulting in an optical density at 600 nm of 0.5).

The cultivation was performed in a 2 L bioreactor (Biostat MD, B. Braun, Germany) with a 1.8 L working volume, in absence of antibiotic. Cultivation was maintained at pH 7.0 ± 0.1 by automatic control through 1 M NaOH (Fluka, Switzerland) or addition of 1 M HCl (Sigma-Aldrich, Germany), and at 37 ± 0.1 °C with a minimal dissolved oxygen concentration (DOC) of 30 ± 5% of air saturation, by automatic adjustment of the agitation rate and the air flow rate range between 1.0 and 1.5 vvm (volume of air/volume of medium/minute). The initial batch cultivation media of the two cultures studied contained 10 g/L of yeast extract (Difco, USA), 20 g/L bactotryptone (BD, UK) and 7 g/L of glycerol (culture A) and 6 g/L of glycerol and 8 g/L of glucose (culture B).

After the batch phase of the culture B, an exponential feeding phase was started with 0.3 L medium, containing 22.5 g yeast extract, 22.5 g bactotryptone and 45 g glucose, and considering a maximum specific growth rate of 0.18 h^{-1} and a constant yield of biomass per glucose of 0.6 g/g. Samples were taken from the bioreactor along the culture, and subsequently used for offline reference analysis of biomass, glucose, glycerol, acetate and plasmid.

IV.2.2. Reference analyses

Biomass in units of dry cell weight (DCW) per volume of culture medium (g/L) was determined by centrifuging the cultivation samples, washing the pellet with 0.9% (w/v) sodium chloride and drying at 80 ºC until constant weight. The bacterial cell pellet and the supernatant obtained from sample centrifugation (Hermle Z160M, Germany) were frozen at -20 ºC. Glucose, glycerol and acetate were determined by HPLC with a L-6200 Intelligent Pump (Merck-Hitachi, UK), a L-7490 LaCrom-Ri-detector (Merck, Germany), a D-2500 Chromato-integrator (Merck-Hitachi, Germany) and an Aminex® Fermentation Monitor HPLC column (Bio-Rad, USA) maintained at 50 °C, and by using H_2SO_4 at 0.6 mL/min as eluent. Plasmids were extracted from the bacteria cell by the alkaline cell lysis method, and subsequent plasmid concentration and purity degree were determined by hydrophobic interaction HPLC, as described in Scholz *et al.* (2012).

IV.2.3. FT-IR spectroscopy

The cell pellet obtained from the centrifugation of each 1 mL sample taken from the bioreactor was resuspended with NaCl 0.9% (w/v), so that an equivalent optical density of 6.0 (at 600 nm) in all samples was achieved. Triplicates of 25 µL of this suspension were placed on IR-transparent ZnSe microtiter plates with 96 wells (Bruker Optics, Germany) and subsequently dehydrated for 2.5 h in a vacuum desiccator (ME2, Vaccubrand, Germany). The FT-IR spectra were recorded in transmission mode by a HTS-XT associated to Vertex-70 spectrometer (Bruker Optics), using a spectral resolution of 4 cm^{-1} and 40 scans per sample.

IV.2.4. Chemometric Methods

Pre-processing

Different data pre-processing methods were studied, namely baseline correction, first and second derivatives and multiplicative scatter correction (MSC), and a combination between them.

The baseline correction was performed in OPUS Ver. 7.2 (Bruker, Germany) and it allows to subtract baselines from spectra by getting spectra with band edges of up to theoretical baseline, i.e., 0. The remaining pre-processing and processing techniques were performed in MATLAB 7.8.0 (MathWorks, USA).

While first derivative allowed offset elimination, as the offset represents a constant value added to the entire spectrum and the derivative of a constant is zero, second derivative, besides offset elimination, enabled removing the slope from the spectral data set. As derivatives usually

broaden spectral noise, a Savitzky-Golay smoothing was applied, with each point of the dataset being replaced by the average of itself and 15 points before and after.

MSC was also used to eliminate changes in spectra due to radiation scattering, by determined the mean spectrum of replicate spectra, and performing a transformation where the spectral data x_i is converted into new values z_i, where i = 1, …, p, with p being the wavelengths [Fearn *et al.*, 2009]. The following equation describes the transformation from x_i to z_i:

$$z_i = \frac{x_i - a}{b},$$

where a represents the intercept and b the slope of a least squares regression of x_i on the values r_i coming from the reference spectra.

In the present work, the MSC was applied to each group of replicates.

Spectral deconvolution

The software OriginPro Ver. 7.0 (OriginLab, USA) was used for the deconvolution of specific spectral bands. The goal of this operation is to resolve the underlying and overlapping peaks present in an IR spectrum.

Before spectral deconvolution, baseline correction, MSC and the normalization to the amide II peak (at wavenumber 1550 cm^{-1}) were applied. The normalization strategy becomes important, since it enables highlighting differences in the spectra, which are not related to biomass. For that, spectra are divided by a constant value, which was chosen as the maximum height of the amide II peak [Maquelin *et al.*, 2002], as it is proportional to the cell quantities present in the sample.

Multivariate data analysis

PCA is a data-reduction method extensively used for qualitative spectral analysis that reduces the dimension of a dataset to a simpler representation by creating new variables, called principal components (PCs). This kind of method is very useful tool for chemometricians not only for data compression but also for information extraction, allowing the identification of major trends in the data [Naes *et al.*, 2002].

The PCA model can be described in matrix notation as:

$$X = TP^T + E$$

where X is the spectral data matrix, T is the matrix containing the scores of the PCs, P the matrix containing the loadings and E the matrix that contains the model residuals and represents the noise or irrelevant variability in X. The scores in T are linear combinations of the original variables of X (wavelengths). The loadings in P are estimated by regressing X on to T and the residual matrix E is calculated by subtracting the estimated TP^T from X [Naes *et al.*, 2002].

IV.3. Results and Discussion

To help understanding the complexity of the interrelationships between cultivation general conditions and the genetic and physiological characteristics of the expression system, the metabolic profiling of the host recombinant cell along the bioprocess becomes therefore very important [Cash, 2014; Guernec *et al.*, 2013; Lenahan *et al.*, 2013; McQuillan *et al.*, 2014; Moen *et al.*, 2009; Trauchessec *et al.*, 2014]. One useful and potential technique that enables the screening of changes in the total biomolecular composition is FT-IR spectroscopy, namely between 1800 and 800 cm^{-1}. This spectral region corresponds to the fundamentals vibration of molecules and presents therefore biological distinctive spectral features, which would allow extracting more detailed information about the biochemical composition of the cell, namely lipids, glicids, proteins, nucleic acids and other chemical species. Considering these purposes, two recombinant *E. coli* cultures producing plasmid were evaluated based on FT-IR spectra collected in high-throughput mode along the cultivation time.

Culture A was conducted on glycerol and culture B was conducted on a mixture of glycerol and glucose. Glucose is usually the main carbon source used to promote the bacterial growth. However, in recombinant *E. coli* cultivation high glucose concentrations lead to the production of acetate , which reduces the cellular energetic yield and can inhibit growth, while decreasing the recombinant product yield [Johnston *et al.*, 2003; Luli and Strhol, 1990; MacDonald and Neway, 1990; Xu *et al.*, 2005]. The production of acetate arises from two different mechanisms: when the maximum oxygen transfer capacity of the reactor is reached, anaerobiosis occurs, leading to mixed-acid fermentation; when acetate is formed aerobically in the presence of high concentrations of the primary carbon source that leads to the uptake of the carbon substrate greater than a critical value. This latter process is known as the overflow metabolism and it has been associated either to the saturation of tricarboxylic acid cycle [Fox *et al.*, 1986], or the electron

transport phosphorylation process, or both [El-Masi and Holms, 1989; Majewski and Domaach, 1990].

Since the use of glucose as carbon source leads to high levels of acetate production, glycerol can be used as an alternative, as it usually results in a lower acetate production. Furthermore, the use of glycerol presents the advantage that it does not have to be heat sterilized apart from other media components, as opposed to glucose, which simplifies the preparation of the bioreactor in large-scale operations. Nevertheless, there is evidence of a high production of acetate from glycerol, which can be related to the high product yields per biomass observed [Scholz *et al.*, 2012; Silva *et al.*, 2009].

Besides the carbon sources, economic media based on complex nitrogen sources (as yeast extract and bactotryptone) were also used in both cultures. Rich and complex media, such as media containing yeast extract and/or hydrolyzed proteins, are often chosen over defined media because they are relatively simple to prepare and generally lead to higher biomass yields and high specific growth rates [Durland and Eastman, 1998]. Therefore, complex and rich nitrogen source constitutes a good choice to ensure an economic recombinant product production process in large scale [Danquah and Forde, 2007; Durland and Eastman, 1998].

Comparing the batch phases of cultures A and B, it was possible to observe that in the presence of glucose and glycerol, glucose is the first carbon source to be consumed, followed by glycerol (**Figure IV.11**). In both cultures, acetate was produced during the consumption of the carbon sources consumption, achieving its highest concentration (5.4 g/L) in culture B (**Table IV.5)**.

Considering the production plasmid efficiency on these batch phases, the batch culture B, conducted on a mixture of glucose and glycerol, presented the highest plasmid productivity of 4.4 mg/L/h, plasmid concentration of 42 mg/L and plasmid production per biomass of 7.2 mg/g (**Table IV.5**), when compared with the batch culture A, conducted on glycerol. These data clearly show the advantages the using a mixture of glucose and glycerol, since glucose promotes high productivities, due to the high specific growth rate on glucose, and the glycerol contribute to high plasmid yields, associated to its lower specific growth rates. To further improve the plasmid production, after the batch phase of cultivation B, a feeding phase with glucose was conducted, resulting in 3.4 fold higher plasmid concentration and 1.7 fold higher plasmid yield per biomass, while maintaining the plasmid high productivity in relation to the batch phase.

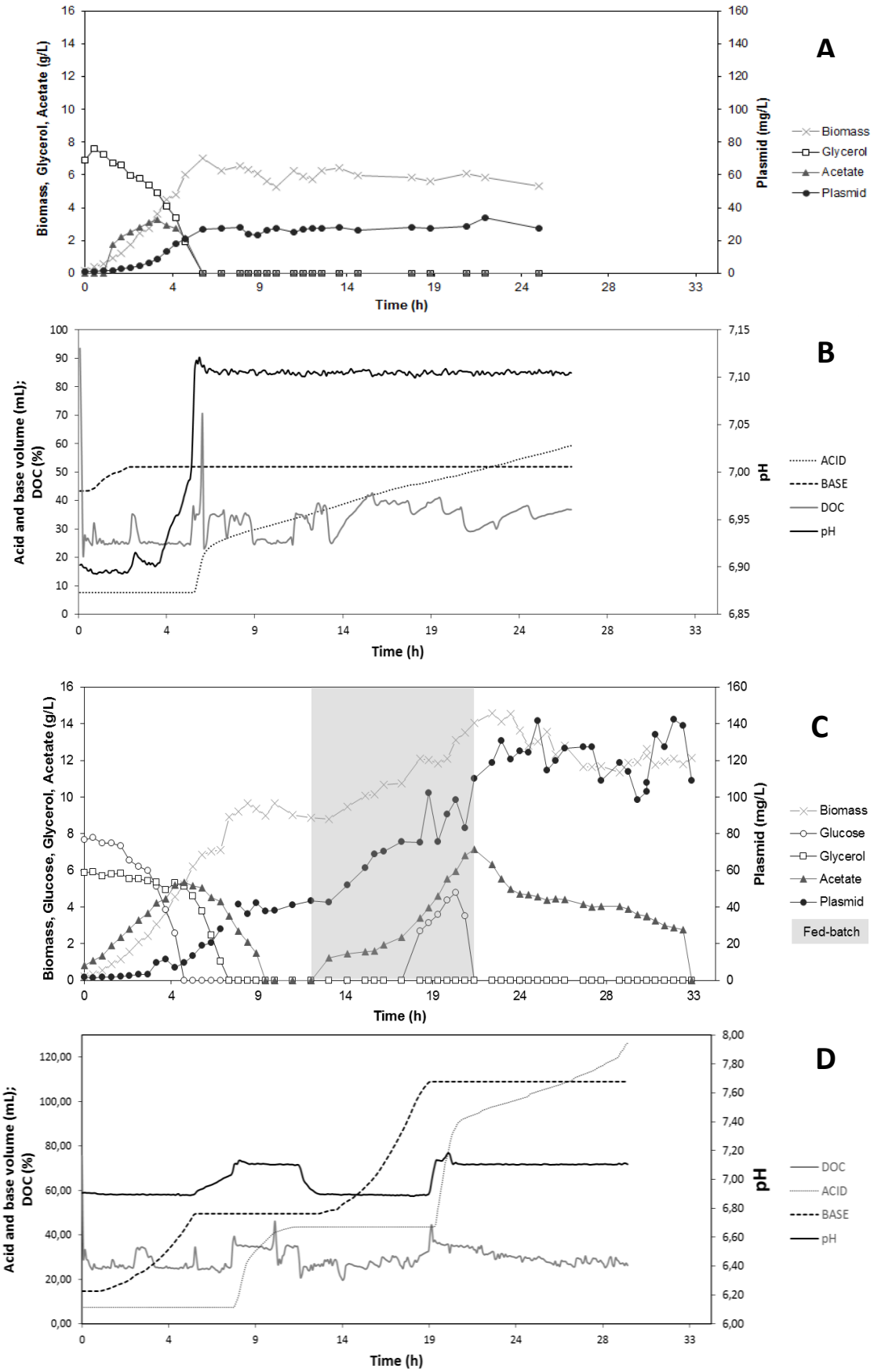


Figure IV.11: Evolution along the time of the biomass, glucose, glycerol, acetate and plasmid concentrations for the two cultures (A to B), conducted with a C-source composition on the batch phase of glycerol (culture A) [**A**] and glucose and glycerol (culture B) [**C**]. After all acetate produced during the batch phase was consumed on culture B an exponential feeding phase with glucose was started considering a μ=0.18h^{-1}, $Y_{X/S}$=0.6 and S=150g/L. The feeding phase is represented in the graph by the grey area. The plots **B** and **D** represent the evolution along the time of the online parameters: base, acid, pH and DOC, for the cultures A and B, respectively.

Table IV.5: Description of the two batches cultures conducted with mixtures of glucose and glycerol as carbon source. The following parameters are relative to the time at which the maximum plasmid production was achieved: time, biomass, maximum plasmid and final plasmid productivity.

	Culture A	**Culture B**
[glucose] (g/L)	-	8.0
[glycerol] (g/L)	7.0	6.0
maximum [acetate] (g/L)	3.3	5.4
time (h)	22	9.5
[biomass] (g/L]	5.9	9.4
maximum [plasmid] (mg/L)	34	42
plasmid/biomass (mg/g)	4.8	7.2
final plasmid productivity (mg/L/h)	1.5	4.4
specific growth rate in glucose (h^{-1})	-	0.68
specific growth rate in glycerol (h^{-1})	0.75	0.31

In order to direct extract information from the FT-IR spectra, the following pre-processing techniques were applied with goal of reducing data noise, while highlighting spectral features: baseline correction, i.e., all spectra have the same baseline; MSC, which was applied to reduce the physical interferences, such as light scattering resulting from irregularities on the samples' surface or particles with different sizes and shapes; and normalization, applied in order to minimize the effect of the biomass concentration, as pointed out in **figure IV.12.**

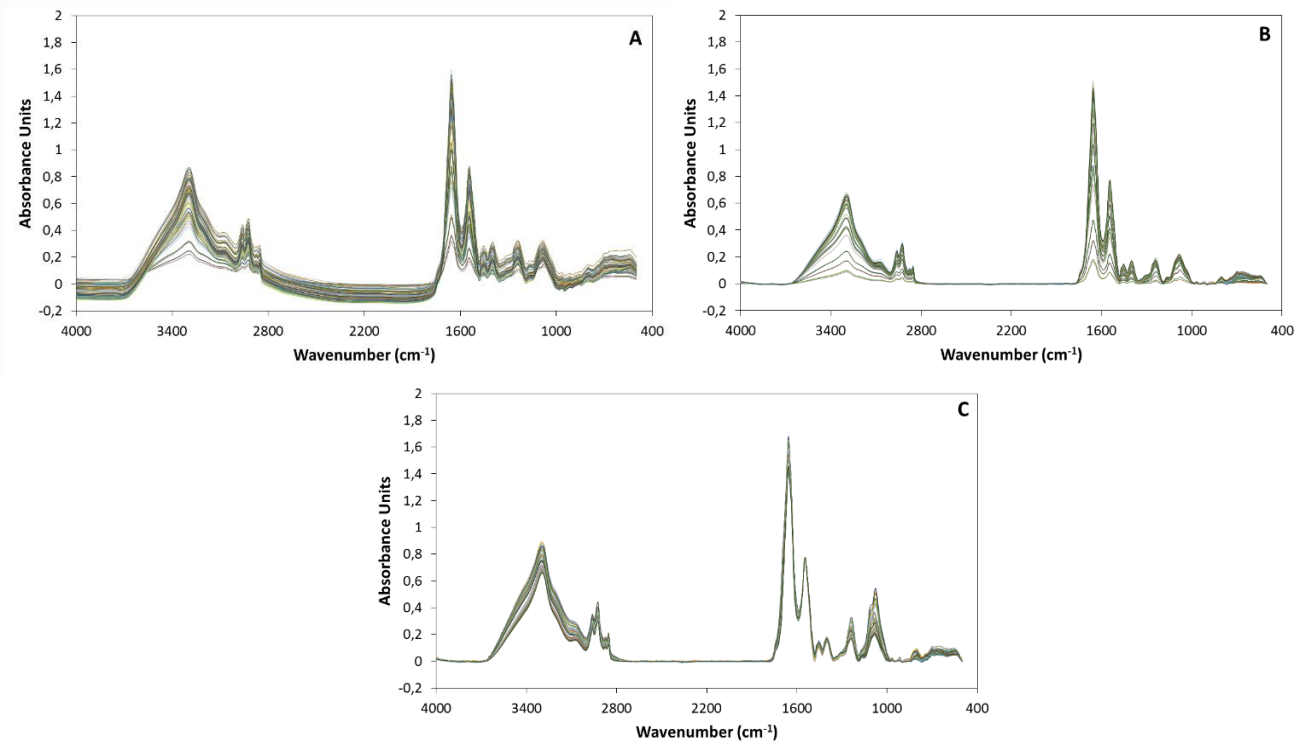
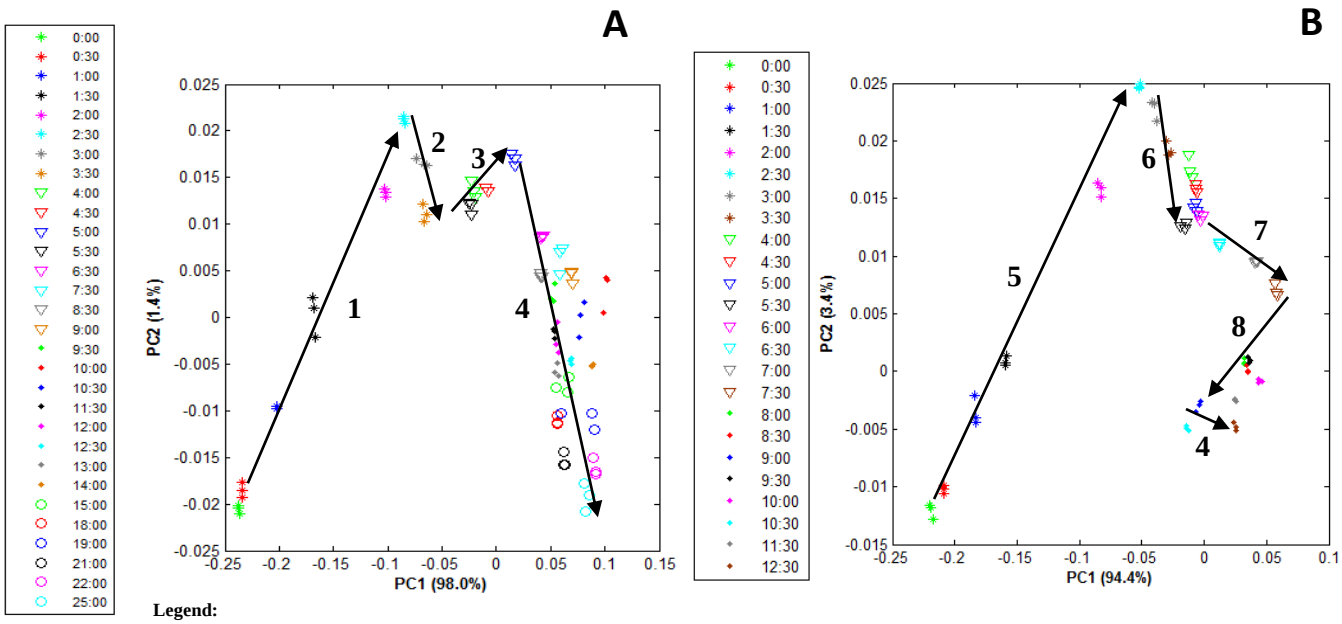


Figure IV.12: IR spectra from different samples in different stages of the bioprocess: without pre-processing (**A**); with baseline correction and MSC (**B**); and with baseline correction, MSC and normalization to amide II band (**C**).

As previously described, principal component analysis (PCA) is a data reduction method often used for qualitative spectral data analysis that decompose the spectral data into new variables, called principal components (PCs), which capture most variance in data [Jollife, 2002]. Consequently, PCA models will enable to find meaningful relationships between the spectral data and cellular events, such as different consumption's phases of the cell. Indeed, PCA applied to the spectral data obtained from cultures A and B, captured the metabolic state of the cell cultivation, as a separation of the samples in the score plots according to the C-source consumption phase could be observed. For example, in the batch phase of cultures A and B, samples at the stationary growth phase were separated from the remaining samples, as pointed out in **figures IV.13A** and **B**, where the samples mentioned are identified by the line 4. It was also observed that the PC2 scores increase as the first carbon source consumption occurs, as pointed by the line 1 presented in the score plots of **figure IV.13A** and **B**. Samples with a high acetate concentration trend to present higher PC2 scores in both cultivations A and B. As acetate starts to be consumed, the PC2 values also decrease in both cultivations A and B (**Figure IV.13A** and **B**).



1 – glycerol consumption and acetate production; **2** – starting acetate consumption; **3** – final phase of glycerol and acetate consumption; **4** – stationary growth phase; **5** – glucose consumption and acetate production; **6** – starting glycerol consumption and final phase of the glucose consumption; **7** – starting acetate consumption and final phase of glycerol consumption; **8** – final phase of the acetate consumption

Figure IV.13: Principal components analysis of the batches cultures A (**A**) and B (**B**). The legend presents the culture time of each sample and in each axis legend it is presented the PC and the respective explained variance. The spectra have been pre-processed with MSC and the first derivative using Savitzky-Golay smoothing.

Due to biochemical complexity of a living cell, the majority of the spectra peaks represents combinations of vibrations of different chemical bonds. To resolve these peaks in relation to individual contributions, several spectral regions were deconvoluted based on the second derivative, as represented in **figure IV.14**, where the negative part of the second derivative spectra corresponds to the peaks of the IR spectrum.

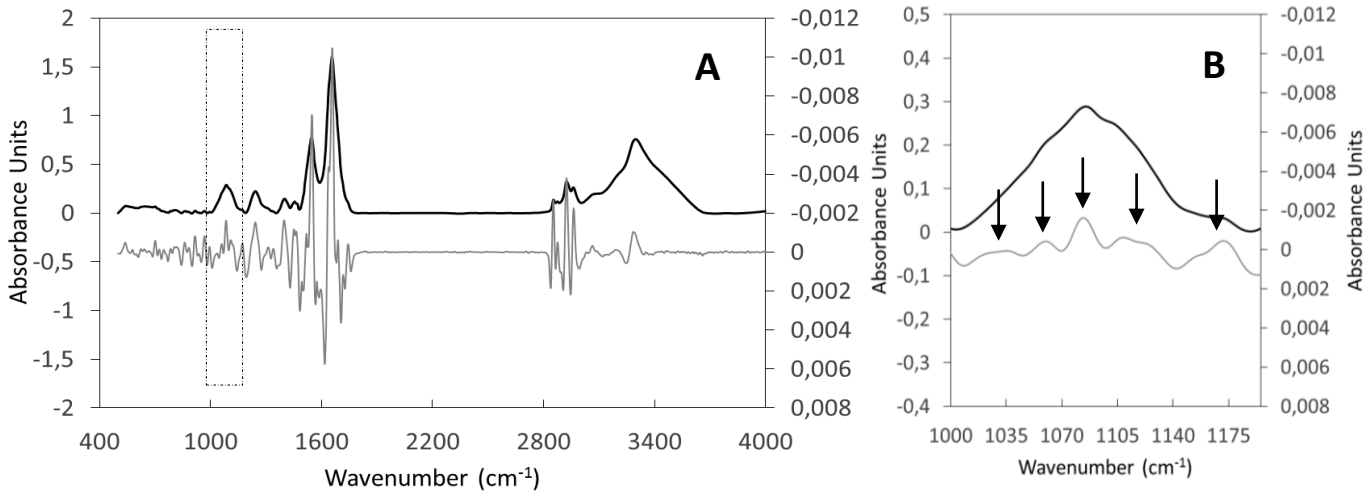


Figure IV.14: IR spectrum of a sample of the culture A (black line) and the reversed second derivative spectrum of the same sample (grey line) (**A**), and an amplification of the spectral region between 1000 and 1185 cm^{-1} (**B**).

The deconvolution also enables to estimate the absorbance contribute of each chemical species, as the sum of their areas corresponds to the total area of this spectrum region. **Figure IV.15** presents examples of the deconvolution results of two distinct spectral regions. For example, the region between 1000 and 1195 cm^{-1}, according to the second derivative spectra, presents at least five underlying bands. Consequently, the deconvolution of this spectral region accounted for five deconvoluted peaks, as highlighted in **figure IV.15A**. The only region that did not need deconvolution was the spectral region between 2800 and 3000 cm^{-1}, due to a high peak definition.

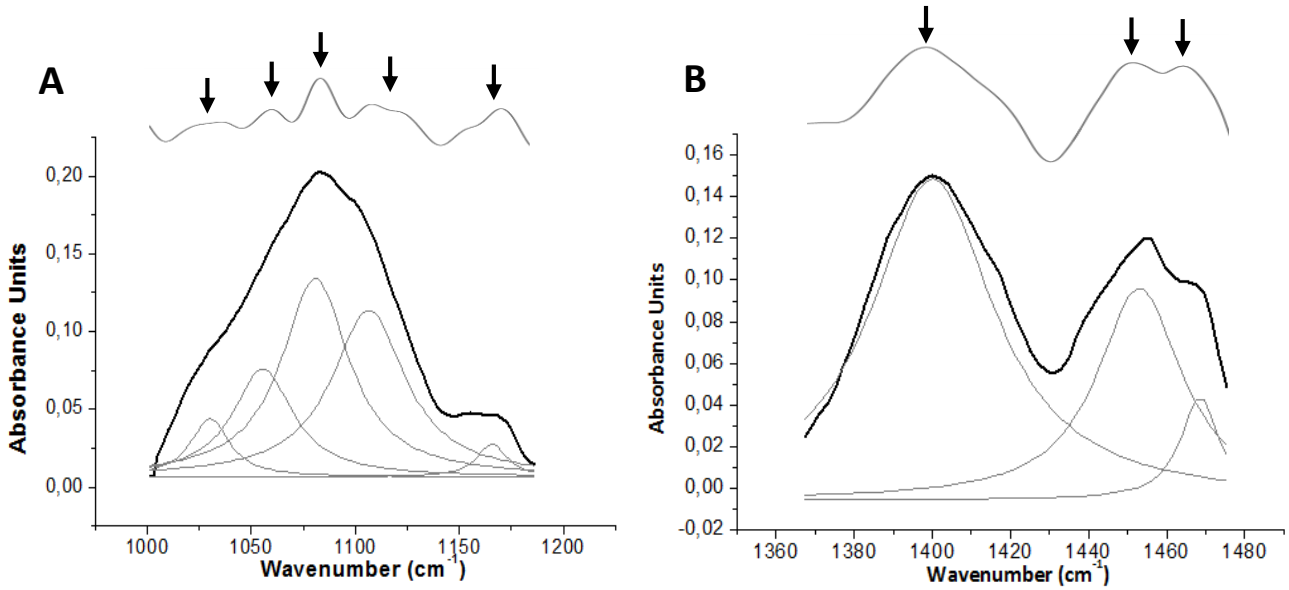


Figure IV.15: The reversed second derivative spectrum of a given sample with the presentation of the peaks identified, followed by IR spectrum of the same sample with the deconvoluted peaks, after the deconvolution process. This representation includes the following spectral regions: (**A**) 1000 – 1195 cm^{-1} and (**B**) 1360 – 1480 cm^{-1}.

Table IV.6 presents the proposed meaning of the several spectral bands identified. This information was essentially obtained from studies related to early cancer diagnosis, where the authors try to find biochemical changes between carcinogenic cells and non-carcinogenic cells [Baran *et al.*, 2013; Gaigneaux *et al.*, 2007; Gazi *et al.*, 2003; Maziak *et al.*, 2007; Lewis *et al.*, 2010; Wang *et al.*, 2010], studies in areas related to natural tissues and cell biology [Movasaghi *et al.*, 2008] and studies related to the identification of bacteria [Garip *et al.*, 2009; Maquelin *et al.*, 2002].

Table IV.6: The identified bands and its proposed assignment according to the literature.

Wavenumber (cm^{-1})	Assignment
~ 1032	C–O str and C–O bend: glycogen
~ 1057	C–O str deoxyribose: DNA
~ 1082	PO_2^- sym str: nucleic acids
~ 1111	C–O str vibration of C–OH group of ribose: RNA
~ 1168	C–O str: protein side chains
~ 1240	PO_2^- asym str: mainly nucleic acids with the little contribution from phospholipids
~ 1255	Amida III
~ 1304	-
~ 1339	-
~ 1400	**COO– sym str: aminoacid side chains and fatty acids; CH_3 sym bend: methyl groups of proteins**
~ 1450	CH_2 bend: mainly lipids with little contribution of the proteins; CH_3 asym bend: methyl groups of proteins
~ 1468	-
~ 1522	-
~ 1550	Amide II: proteins, mainly N–H bend and C–N str
~ 1638	Amide I: proteins
~ 1655	
~ 1685	
~ 2850	CH_2 sym str: mainly lipids with the little contribution from proteins, nucleic acids and carbohydrates
~ 2870	CH_3 sym str: protein side chains and some contribution from lipids, proteins and carbohydrates
~ 2920	CH_2 asym str: mainly lipids with the little contribution from proteins and carbohydrates
~ 2960	CH_3 asym str: mainly lipids and protein side chains, with the little contribution from proteins and carbohydrates
~ 3070	Amide B: C–N and N–H str of proteins
~ 3185	-
~ 3300	Amide A: mainly N–H str of proteins
~ 3442	-
str=stretching ; bend=bending ; def=deformation ; sym=symmetric ; asym=antisymmetric	

Considering the glycogen content, corresponding to the spectral band at 1032 cm^{-1}, along both cultivations, a decrease in glycogen levels was observed, especially along the consumption of the C-sources of the batch phase (**Figure IV.16**), which can be related to the high specific growth rate of the host cell at the beginning of the cultures. This evidence is in accordance to several studies related to early cancer diagnosis, which state that cells with a higher cell division, as carcinogenic cells, present lower glycogen contents [Gazi *et al.*, 2003; Yano *et al.*, 1996]. It was observed that as the specific growth rate along the C-source consumption diminishes, the decrease

in glycogen is less accentuated (**Table IV.7)**. Furthermore, a slight increase of the glycogen contents at the beginning of the C-source consumption was also observed, as in the case of acetate.

With the beginning of the feeding in the culture B, the glycogen levels increased, as opposed to the batch phase, as during the fed-batch phase the bacteria use the carbon source to produce plasmid instead of growing. Indeed, during the feeding phase a biomass per C-source yield of 0.23 g/g and a plasmid production per biomass of 10.9 mg/g were achieved, against 0.83 g/g and 4.4 g/g in the batch phase, respectively. The lower cell growth observed during the feeding phase can be associated to a nutritional limitation, which is advantageous in this case, since a greater plasmid production per biomass was obtained.

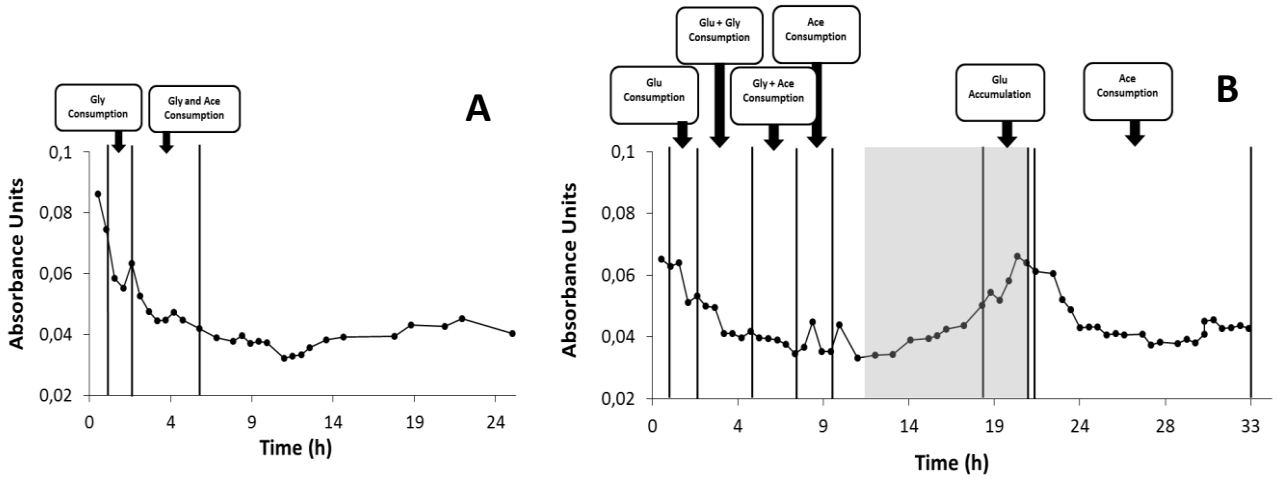


Figure IV.16: Glycogen levels, corresponding to 1032 cm^{-1} band, along the cultivations A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

Table IV.7: Specific growth rates in the different consumption phases of the cultures A and B.

	Culture A	**Culture B**
specific growth rate in glucose (h^{-1})	-	0.68
specific growth rate in glycerol (h^{-1})	0.75	-
specific growth rate in glucose and glycerol (h^{-1})	-	0.41
specific growth rate in glycerol and acetate (h^{-1})	0.46	0.04
specific growth rate in acetate (h^{-1})	-	0.08

The RNA content in the host cell, which is mainly RNA messenger (mRNA) [Ciccolini *et al.*, 2002], was estimated by the spectral band at 1111 cm^{-1} (**Figure IV.17**). An increase of the RNA concentration in both cultures during the consumption of the first C-source in batch phase occurred. A slight increase was also observed in both cultures immediately before the beginning of other carbon source consumption, i.e. acetate, in culture A or glycerol in culture B. This initial

increase in the mRNA concentration is most probably related to increase gene expression due to cell adaptation to new environmental conditions as media composition. Some examples of enzymes genes needed to be induced to enable the acetate metabolizing is phosphotransacetylase (PTA) and acetate kinase (ACKA) genes, whose expression is induced by high acetate concentrations [Valgepea *et al.*, 2010]. In the feeding phase of culture B, there is an increasing of the mRNA, especially during the accumulation of glucose in the culture medium, which can be explained by the synthesis of proteins associated to the stress response, e.g. due to an overburden of the host cell metabolism [Dürrschmid *et al.*, 2008].

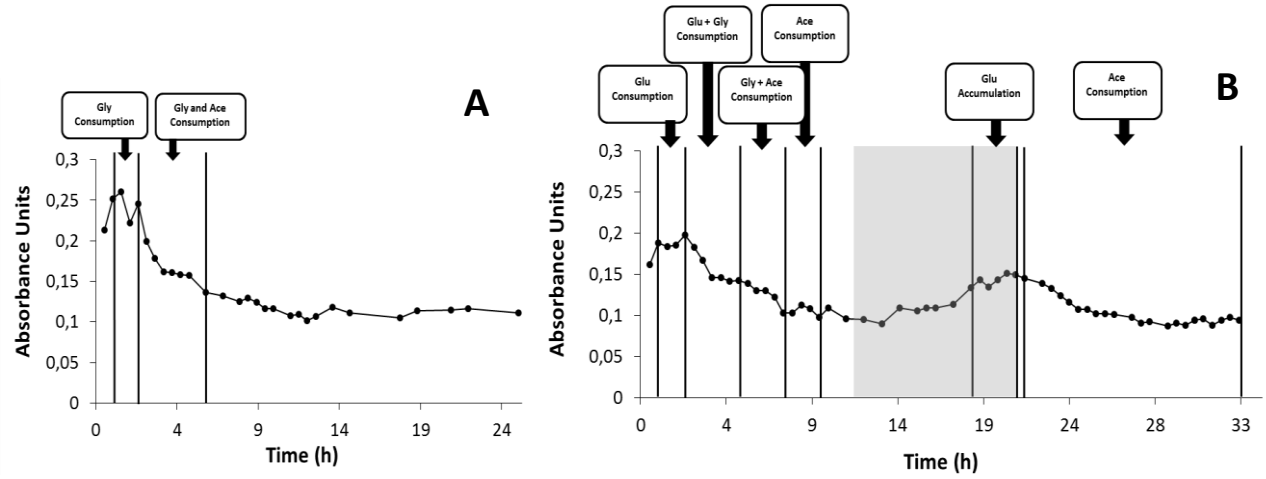


Figure IV.17: RNA concentration in the host cell, considering the 1111 cm^{-1} band, along the cultivation A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

It is well established in literature [Maziak *et al.*, 2007; Parker, 1971; Parker, 1983; Susi, 1969] that the peak maximum near 1650 cm^{-1} is correlated with the protein segments with α-helical structures. The component bands near 1688 and 1636 cm^{-1} are the amide I bands of the proteins segments with the β-sheet structure [Byler, 1986]. The changes in the relative intensities of the amide I bands described above have been widely used for monitoring the protein conformational changes in the cellular proteome [Baran *et al.*, 2013; Maziak *et al.*, 2007; Parker, 1971; Parker, 1983; Susi, 1969]. In this work general protein conformational changes along the bioprocess cultivation on both cultures could also be observed (**Figure IV.18**).

Besides nucleic acids and proteins, lipids are also biomolecules with major presence in the cell, being represented by spectral bands near 2850, 2920 and 2960 cm^{-1} [Baran *et al.*, 2013; Gaigneaux *et al.*, 2007; Wang *et al.*, 2010]. A higher lipid concentration in the phases either with high cell growth rates or at the end of the feeding phase was observed (**Figure IV.19**). According to Baran *et al.* (2013), this increase of the lipids levels, known as lipidation, is considered to be one of the general response of the cell to stress events.

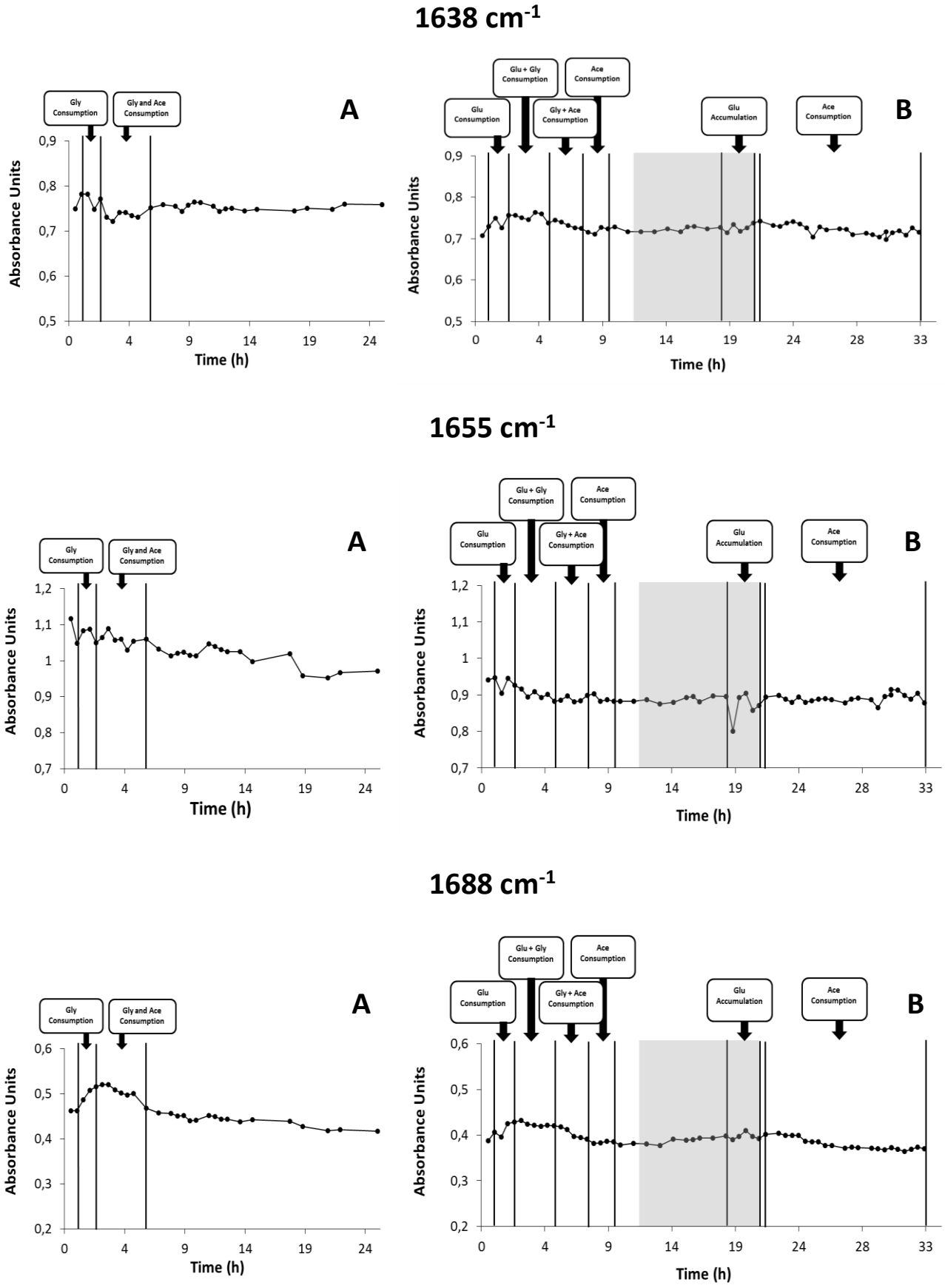


Figure IV.18: Intensities of the amide I bands (1638, 1655 and 1688 cm^{-1}) along the cultivations A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

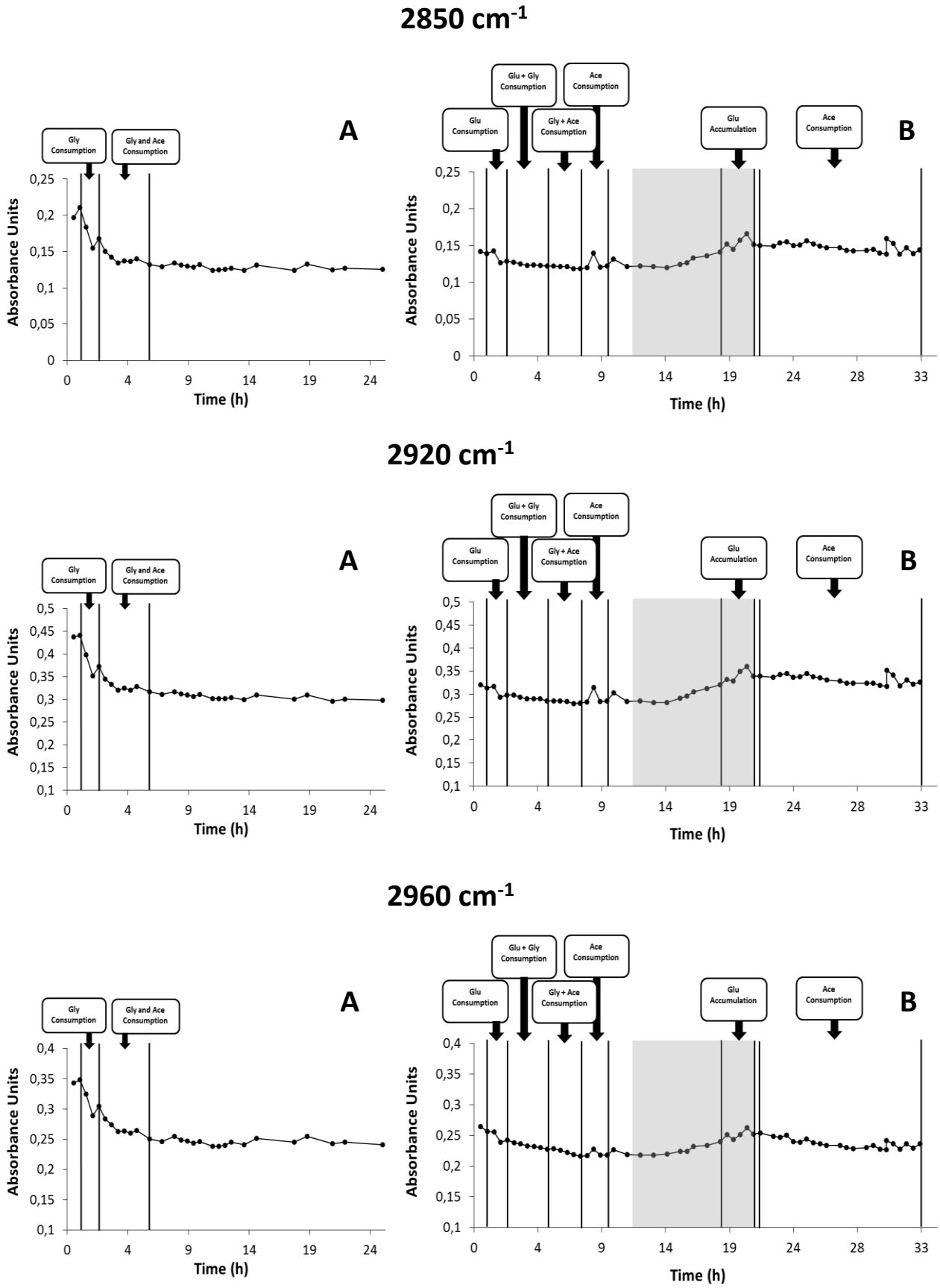


Figure IV.19: Intensities of the lipids bands (2850, 2920 and 2960 cm^{-1}) along the cultivations A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

The intensity ratio between spectral bands at 1111 cm^{-1} (RNA band) and 1550 cm^{-1} (amide II band) was also considered in order to understand the transcriptional status of the host cell, as presented by Baran *et al.* (2013) in their carcinogenic studies. During the batch phase, an increase of this intensity ratio in both cultures immediately before the beginning of the C-source consumption (**Figure IV.20**) was observed, meaning that the bacteria was transcribing the necessary gene to the carbon source metabolism. This trend is corroborated by the trends observed in the RNA concentrations along both cultivations.

After the beginning of the feeding phase of culture B, an increase of the intensity ratio happened, probably due to the transcription of genes that encode the proteins associated to the stress response [Dürrschmid *et al.*, 2008].

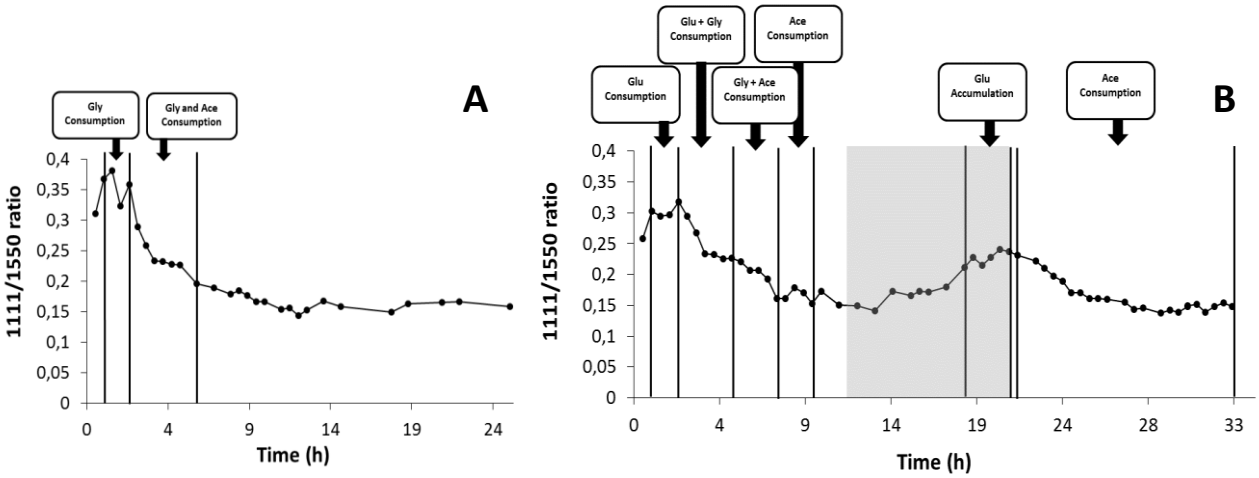


Figure IV.20: Intensity ratio of the 1111 cm^{-1} and amide II along the cultivations A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

Complementary to the transcription levels, the protein expression status of the host cell, represented by the intensity ratio between 1550 cm^{-1} and 1082 cm^{-1} [Baran *et al.*, 2013], was also studied. As expected, there was an increase of the intensity ratio during the C-source consumption, which is related to the need of the host cell to synthetize the proteins involved in metabolism of the carbon sources (**Figure IV.21**). After the feeding phase of culture B, the intensity ratio increased again, that is once more related to the increase of the protein expression, namely due to synthesis of proteins associated to the stress response [Dürrschmid *et al.*, 2008].

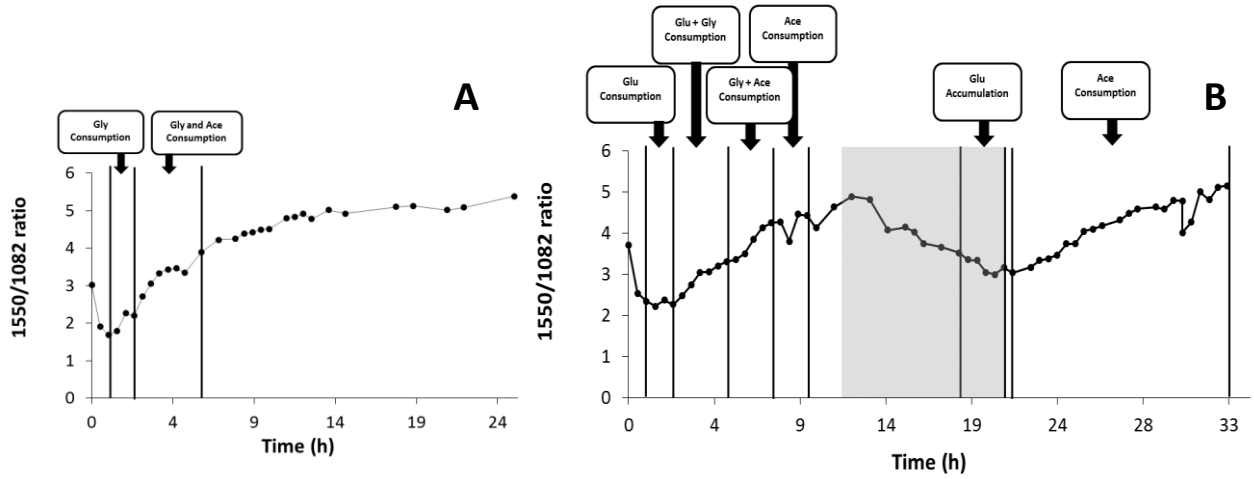


Figure IV.21: Intensity ratio of the amide II and 1080 cm^{-1} along the cultivations A (**A**) and B (**B**). The feeding phase is represented in the graph by the grey area.

IV.4. Conclusions

The present work shows the ability of FT-IR spectroscopy to extract metabolic information about the host cell, namely the identification of the general metabolic switches along the different phases of C-source consumption, by PCA, and the biomolecules' concentrations or metabolic status like translational levels along the cell culture, by direct spectral analysis. Regarding specific biomolecules' concentrations present in the cell, glycogen levels trended to decrease due to high cellular growth rates, namely during the carbon sources consumption. The RNA concentrations increased mainly before the beginning of the carbon sources consumption, due to the need of the bacteria to transcribe the genes the start new C-source metabolism. Protein structural changes in the cell proteome were also identified by FT-IR spectral analysis, considering the amide bands. The RNA/amide II ratio enabled to monitor the transcriptional status of the host cell, being higher immediately before the consumption of the carbon sources. A higher protein expression by the amide II/nucleic acids total ratio was observed immediately after the increase of the transcriptional level, i.e. during the carbon sources consumption. Therefore, FT-IR spectroscopy proved to be a highly promising tool for monitoring the structural and functional changes in host cell during the biopharmaceuticals production.

Chapter V

General Conclusions

Due to the relevance of *Escherichia coli* as a recombinant cell host to produce heterologous products, it is important to develop new techniques that enable in a fast, sensitive and in an *in-situ* or high-throughput mode to estimate critical variables of the process culture and the metabolic characteristics of the cell in response to different cultivation conditions. The present work shows the potential of FT-IR spectroscopy to achieve that purposes. The application of chemometrics methods to the IR spectral data also showed to be very important, since they highly influence the output of data analysis, allowing extracting more detailed information that is often hidden in the raw IR spectra. Therefore, the strategies to be used for each dataset must be carefully chosen.

The present thesis shows FT-IR spectroscopy combined with PLS regression as a powerful tool to quantify of critical variables of the bioprocess (as biomass growth, plasmid production, carbon source consumption and acetate production and consumption), either by *in-situ* NIR or *at-line* high-throughput MIR spectroscopy. Furthermore, this work also shows how FT-IR spectroscopy can be used to monitor the metabolism of the bacteria cell host, considering proteins, nucleic acids, lipids and others biomolecules present in the cell, during a biopharmaceutical's production, both by a direct spectral analysis and by PCA.

In a future work, it will be interesting to:

- Develop a deconvolution algorithm in a programming language, e.g. Matlab, based on the deconvolution methods described in scientific articles, like in Kauppinen *et al.* (1981) and Kochev *et al.* (2001). This need is related to the limitations of the software used in this work, since it works as a "black box";
- Conduct a more detailed characterization of the *E. coli* cultivation, by conventional methods, e.g., analyzing total nucleic acids, mRNA, genomic DNA, glycogen, total proteins and lipids, and other metabolites that would enable for example to characterize stress response metabolism;
- Find the biochemical meaning of some spectral bands that present a specific profile along the bacteria cell cultivation process, based for example on a complement metabolic characterization of the *E. coli* cultivation process.

In sum, FT-IR spectroscopy was presented as a highly promising tool for bioprocess monitoring, as it enables the quantification of critical variables and the biochemical and metabolic characterization of the cell host. The present results may certainly contribute to the design of more economic and robust processes ensuring reproducibility and quality of the final product in accordance to the PAT initiative.

References

Arnold SA, Gaensakoo R, Harvey LM and McNeil B (2002). Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch *Escherichia coli* process. Biotechnology and Bioengineering 80: 405-413.

Babrah J (2009). A study of FT-IR spectroscopy for the identification and classification of hematological malignancies. PhD Thesis, Cranfield University, United Kingdom.

Baran Y, Ceylan C and Camgoz A (2013). The roles of macromolecules in imatinib resistance of chronic myeloid leukemia cells by Fourier transform infrared spectroscopy. Biomedicine and Pharmacotherapy 67 (3): 221-227.

Byler, D.M. and Susi, H. (1986). Examination of the Secondary Structure of Proteins by Deconvolved FT-IR Spectra. Biopolymers 25: 469-487.

Card C, Hunsaker B, Smith T and Hirsch J. (2008). Near-infrared spectroscopy for rapid, simultaneous monitoring. BioProcess International 6: 59-67.

Carnes AE (2005). Fermentation design for the manufacture of therapeutic plasmid DNA. Bioprocess Technical 3: 36-44.

Cash P (2014). Proteomic analysis of uropathogenic *Escherichia coli*. Expert Review of Proteomics 11 (1): 43-58.

Christian GD (1994). Analytical Chemistry. WILEY: United States.

Ciccollini LAS, Shamlou PA and Titchener-Hooker N (2002). A mass balance study to assess the extent of contaminant removal achieved in the operations for the primary recovery of plasmid DNA from *Escherichia coli* cells. Biotechnology and Bioengineering 77(7): 796-805.

Cimander C and Mandenius C-F (2002). Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modeling. Journal of Chemical Technolology and Biotechnology 77: 1157-1157.

Coban C, Kobiyama K, Aoshi T, Takeshita F, Horii T, Akira S and Ishii KJ (2011). Novel strategies to improve DNA vaccine immunogenicity. Current Gene Therapy 11(6): 479-484.

Coban C, Kobiyama K, Jounai N, Tozuka M and Ishii KJ (2013). DNA vaccines: A simple DNA sensing matter? Human Vaccines & Immunotherapeutics 9 (10).

Danquah MK and Forde GM (2007). Growth medium selection and its economic impact on plasmid production. Journal of Bioscience and Bioengineering 104: 490-497.

Dhanoa MS, Lister SJ, Sanderson R and Barnes RJ (1994). The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. Journal of Near Infrared Spectroscopy 2(1): 43-47.

Di Egidio V, Sinelli N, Giovanelli G, Moles A and Casiraghi E (2010). NIR and MIR spectroscopy as rapid methods to monitor red wine fermentation. European Food Research and Technology 230: 947-955.

Durland RH and Eastman EM (1998). Manufacturing and quality control of plasmid-based gene expression systems. Advanced Drug Delivery Reviews 30: 33-48.

Dürrschmid K, Reischer H, Schmidt-Heck W, Hrebicek T, Guthke R, Rizzi A and Bayer K (2008). Monitoring of transcriptome and proteome profiles to investigate the cellular response of *E. coli* towards recombinant proteins expression under defined chemostat conditions. Journal of Biotechnology 135: 34-44.

Duygu D, Baykal T, Açikgöz I and Yildiz K (2009). Fourier Transform Infrared (FT-IR) Spectroscopy for Biological Studies. Journal of Science 22: 117-121.

El-Mansi EMT and Holms WH (1989). Control of carbon flux to acetate excretion during growth of *Escherichia coli* in batch and continuous cultures. Journal of General Microbiology 135: 2875-2883.

Fearn T, Riccioli C, Garrido-Varo A and Guerrero-Ginel JE (2009). On the geometry of SNV and MSC. Chemometrics and Intelligent Laboratory Systems 96: 22-26.

Fox DK, Meadow ND and Roseman S (1986). Phosphate transfer between acetate kinase and enzyme I of the bacterial phosphotransferase system. Journal of Biological Chemistry 261: 13498-13503.

Gaigneaux A, Decaestecker C, Camby I, Mijatovic T, Kiss R, Ruysschaert JM and Goormaghtigh E (2004). Experimental Cell Research 297: 294-301.

Garip S, Gozen AC and Severcan F (2009). Use of Fourier transform infrared spectroscopy for rapid comparative analysis of *Bacillus* and *Micrococcus* isolates. Food Chemistry 113: 1301-1307.

Gasper R, Dewelle J, Kiss R, Mijatovic T and Goormaghtigh E (2009). Biochimica et Biophysica Acta 1788: 1263-1270.

Gazi E, Baker M, Dwyer J, Lockyer NP, Gardner P, Shanks JH, Reeve RS, Hart CA, Clarke NW and Brown MD (2006). A correlation of FTIR spectra derived from prostate cancer biopsies with Gleason grade and tumor stage. European Urology 50: 750-761.

Gazi E, Dwyer J, Gardner P, Ghanbari-Siahkali A, Wade AP, Miyan J, Lockyer NP, Vickerman JC, Clarke NW, Shanks JH, Scott LJ, Hart CA and Brown M (2003). Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. A pilot study. Journal of Pathology 201: 99-108.

Geladi P (2003). Chemometrics in spectroscopy. Part 1. Classical chemometrics. Spectrochimica Acta Part B 58: 767-782.

Guernec A, Robichaud-Rincon P and Saucier L (2013). Whole-genome transcriptional analysis of *Escherichia coli* during heat inactivation processes related to industrial cooking. Applied and Environmental Microbiology 79 (16): 4940-4950.

Guillen M and Cabo N (1997). Infrared spectroscopy in the study of edible oils and fats. Journal of the Science of Food and Agriculture 75 (1): 1-11.

Hall JW, Valentine KG, Lefrant S,Mevellec JY and Mulazzi E (1996). Spectroscopic properties of polyacetylenes synthesized via three modifications of Ziegler-Natta catalytic system. Synthetic Metals 79:183–88.

Hansen R and Eriksen NT (2007). Activity of recombinant gst in *Escherichia coli* grown on glucose and glycerol. Process Biochemistry 42(8): 1259-1263.

Helland IS, Naes T and Isaksson T (1995). Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. Chemometrics and Intelligent Laboratory Systems 29(2): 233-241.

Hsu C-PS (1997). Infrared Spectroscopy, in Settle F. (Ed.). Instrumentation Techniques for Analytical Chemistry. Prentice Hall PTR: New Jersey, USA.

Huang WE, Hopper D, Goodacre R, Beckmann M, Singer A and Draper J (2006). Rapid characterization of microbial biodegradation pathways by FT-IR spectroscopy. Journal of Microbiological Methods 67: 273-280.

Johnston W, Stewart M, P. Lee and Cooney M (2003). Tracking the acetate threshold using DO-transient control during medium and high cell density cultivation of recombinant *Escherichia coli* in complex media. Biotechnology and Bioengineering 84: 314-323.

Jolliffe IT (2002). Principal Component Analysis. Springer: New York, USA.

Kalams SA, Parker SD, Elizaga M, Metch B, Edupuganti S, Hural J, De Rosa S, Carter DK, Rybczyk K, Frank I, Fuch J, Koblin B, Kim DH, Joseph P, Keefer MC, Baden LR, Eldridge J, Boyer J, Sherwat A, Cardinali M, Allen M, Pensiero M, Butler C, Khan AS, Yan J, Sardesai NY, Kublin JG and Weiner DB (2013). The Journal of Infectious Diseases 208(5): 818-829.

Kansiz M, Gapes JR, McNaughton D, Lendl B and Shuster KC (2001). Mid-infrared spectroscopy coupled to sequential injection analysis for the on-line monitoring of the acetone-butanol fermentation process. Analytica Chimica Acta 438: 175-186.

Kauppinen JK, Moffat DJ, Mantsch HH and Cameron DG (1981). Fourier Self-Deconvolution: A Method for Resolving Intrinsically Overlapped Bands. Applied Spectroscopy 35 (3): 271-276.

Kochev NT, Rogojerov MI and Andreev GN (2001). A new graphical approach for improved user control of Fourier self-deconvolution of infrared spectra. Vibrational Spectroscopy 25: 177-183.

Korz DJ, Rinas U, Hellmuth K, Sanders EA and Deckwer WD (1995). Simple fed-batch technique for high cell density cultivation of *Escherichia coli*. Journal of Biotechnology 39(1): 59-65.

Landgrebe D, Haake C, Höpfner T, Beutel S, Hitzmann B, Beutel S, Hitzmann B, Scheper T, Rhiel M and Reardon KF (2010). On-line infrared spectroscopy for bioprocess monitoring. Applied Microbiology and Biotechnology 88: 11-22.

Lee SY, Yoon K-A, Jang SH, Ganbold EO, Uuriintuya D, Shin S-M, Ryu PD and Joo S-W (2009). Infrared spectroscopy characterization of normal and lung cancer cells originated from epithelium. Journal of Veterinary Science 10 (4): 299-304.

Lenahan M, Sheridan A, Morris D, Duffy G, Fanning S and Burgess CM (2013). Trancriptomic Analysis of Triclosan-Susceptible and –Tolerant *Escherichia coli* O157:H19 in Response to Triclosan Exposure. Mirobial Drug Resistance. October 8, *in press*.

Lewis PD, Lewis KE, Ghosal R, Bayliss S, Lloyd AJ, Wills J, Godfrey R, Kloer P and Mur LAJ (2010). Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using sputum. BioMed Central Cancer 10: 640.

Lopes MB, Sales KC, Lopes VV and Calado CRC. Real-Time Plasmid monitoring of batch and fed-batch *Escherichia coli* cultures by NIR spectroscopy. 3rd Portuguese BioEngineering Meeting. IEEE-EMB Portuguese Chapter, IEEE, 20- 23th February 2013, Minho University.

Lourenço ND, Lopes JA, Almeida CF, Sarraguça MC and Pinheiro HM (2012). Bioreactor monitoring with spectroscopy and chemometrics: a review. Analytical and Bioanalytical Chemistry 404: 1211-1237.

Luli GW and Strohl WR (1990). Comparison of growth, acetate production, and acetate inhibition of Escherichia coli strains in batch and fed-batch fermentations. Applied and Environmental Microbiology 56: 1004-1011.

MacDonald HL and Neway JO (1990). Effects of medium quality on the expression of human interleukin-2 at high cell density in fermentor cultures of *Escherichia coli* K-12. Applied and Environmental Microbiology 56: 640-645.

Majewski RA and Domach MM (1990). Simple constrained optimization view of acetate overflow in E. coli. Biotechnology and Bioengineering 35: 732-738.

Maquelin K, Kirschner C, Choo-Smith L-P, van den Braak N, Endtz HPh, Naumann D and Puppels GJ (2002). Identification of medically relevant microorganism by vibrational spectroscopy. Journal of Microbiological Methods 51: 255-271.

Maziak DE, Do MT, Shamji FM, Sundaresan SR, Perkins G and Wong PTT (2007). Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: An exploratory study. Cancer Detection and Prevention 31: 244-253.

McGovern AC, Broadhurst D, Taylor J, Kaderbhai N, Winson MK, Small DA, Rowland JJ, Kell DB and Goodacre R (2002). Monitoring of Complex Industrial Bioprocesses for Metabolite Concentrations Using Modern Spectroscopies and Machine Learning: application to Gibberellic Acid Production. Biotechnology and Bioengineering 78(5): 527-538.

McNeil B and Harvey LM (1990). Fermentation a practical approach. 1st ed., Oxford University, England.

McQuillan JS and Shaw AM (2014). Differential gene regulation in the Ag nanoparticle and Ag+ - induced silver stress response in *Escherichia coli*: A full transcriptomic profile. Nanotoxicology. January 6, *in press*.

Moen B, Janbu AO, Langsrud S, Langsrud Ø, Hobman JL, Constantinidou C, Kohler A and Rudi K (2009). Global responses of *Escherichia coli* to adverse conditions determined by microarrays and FT-IR spectroscopy. Canadian Journal of Microbiology 55(6): 714-728.

Movasaghi Z, Rehman S and Rehman I (2008). Fourier Transform Infrared (FTIR) spectroscopy of Biological Tissues, Applied Spectroscopy Reviews 43 (2): 134-179.

Naes T, Isaksson T, Fearn T and Davies T (2002). A User Friendly Guide to Multivariate Calibration and Classification. NIR Publications, UK.

Navrátil M, Norberg A, Lembrén L and Mandenius C-F (2005). On-line multi-analyzer monitoring of biomass, glucose and acetate for growth rate control of *Vibrio cholera* fed-batch cultivation. Journal of Biotechnology 115: 67-79.

Nicolaï BM, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron KI and Lammertyn K (2007). Nondestructive measurement of fruit and vegetables quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology 46: 99-118.

O'Kennedy RD, Ward JM and Keshavarz-Moore E (2003). Effects of fermentation strategy on the characteristics of plasmid DNA production. Biotechnology and Applied Biochemistry 37: 83-90.

Orsini F, Ami D, Villa AM, Sala G, Bellotti MG and Doglia SM (2000). FT-IR microspectroscopy for microbiological studies. Journal of Microbiological Methods 42: 17-27.

Otto M (1999). Chemometrics: Statistics and Computer Application in Analytical Chemistry. WILEY-VCH: Freiberg, Germany.

Ow S-W, Lee M-Y, Nissom PM and Philp R (2007). Inactivating FruR global regulator in plasmid-bearing *Escherichia coli* alters metabolic gene expression and improves growth rate. Journal of Biotechnology 131: 261-269.

Ow S-W, Yap G-S and Oh K-W (2009). Enhancement of plasmid DNA yields during fed-batch culture of a fruR- knockout *Escherichia coli* strain. Biotechnology and Applied Biochemistry 52: 53-59.

Parker FS (1971). Application of infrared spectroscopy in biochemistry, biology and medicine. New York: Plenum.

Parker FS (1983). Application of infrared and Raman and resonance Raman spectroscopy in biochemistry. New York: Plenum.

Pistorius AMA (1995). Biomedical applications of FT-IR spectroscopy. Spectroscopy Europe 7: 8-15.

Prather KJ, Sagar S, Murphy J and Chartrain M (2003). Industrial scale production of plasmid DNA for vaccine and gene therapy: plasmid design, production, and purification. Enzyme and Microbial Technology 33: 865-883.

Randolph TW (2006). Scaled-based normalization of spectral data. Cancer Biomarkers 2: 135-144.

Reich G (2005). Near-Infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. Advanced Drug Delivery Reviews 57: 1109-1143.

Rinnan A, van den Berg F and Engelsen SB (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28(10):1201-1222.

Roychoudhury P, Harvey LM and McNeil B (2006). At-line monitoring of ammonium, glucose, methyl oleate and biomass in a complex antibiotic fermentation process using attenuated total reflectance-mid-infrared (ATR-MIR) spectroscopy. Analytica Chimica Acta 561: 218-224.

Sandor M, Rüdinger F, Bienert R, Grimm C, Solle D and Scheper T (2013). Comparative study of non-invasive monitoring via infrared spectroscopy for mammalian study cultivations. Journal of Biotechnology 168(4): 636-645.

Schenk J, Marison IW and von Stockar U (2006). A simple method to monitor and control methanol feeding of *Pichia pastoris* fermentations using mid-IR spectroscopy. Journal of Biotechnology 128: 344-353.

Scholz T, Lopes VV and Calado CC (2012). High-Throughput Analysis of the Plasmid Bioproduction Process in *Escherichia coli* by FTIR Spectroscopy. Biotechnology and Bioengineering 109(9): 2279-2285.

Sharaf MA, Illman DL and Kowalski BR (1986). Chemometrics. John Wiley & Sons.

Shenk JS, Workman JJ and Weterhans MO (2001). Application of NIRS to agricultural products. In Handbook of Near-Infrared Analysis. Ed. DA Burns, EW Ciurczak, 16: 419–74. New York: Marcel Dekker.

Shibui A, Nakae S, Watanbe J, Sato Y, Tolba ME, Doi J, Shiibashi T, Nogami S, Sugano S and Hozumi N (2013). Screening of novel malarria DNA vaccine candidates using full-length cDNA library. Experimental Parasitology 135(3): 546-550.

Silva T, Lima P, Roxo-Rosa M, Hageman S, Fonseca LP and Calado CRC (2009). Prediction of dynamic plasmid production by recombinant Escherichia coli fed-batch cultivations with a generalized regression neural network. Chemical and Biochemical Engineering Quarterly 3: 419-427.

Sivakesava S, Irudayaraj J and Ali D (2001). Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR and FT-Raman spectroscopic techniques. Process Biochemistry 37: 371-378.

Smirnova G and Oktyabrskii O (1985). Influence of acetate on the growth of escherichia coli under aerobic and anaerobic conditions. Microbiology (USSR) 54: 205-209.

Smith BC (2011). Fourier Transform Infrared Spectroscopy. New York: CRC Press.

Stuart B (2004). Infrared Spectroscopy: Fundamentals and Applications. WILEY: United States.

Susi H (1969). Infrared spectra of biological macromolecules and related systems. In: Timashell SN, Fasman CD, eds. Structure and stability of biological macromolecules. New York: Marcel Dekker: 575-663.

Tamburini E, Vaccari G, Tosi S and Trilli A (2003). Near-infrared spectroscopy: a tool for monitoring submerged fermentation process using an immersion optical-fiber probe. Appllied Spectroscopy 57: 132-138.

Tosi S, Rossi M, Tamburini E, Vaccari G, Amaretti A and Matteuzzi D (2003). Assessment of in-line near-infrared spectroscopy for continuous monitoring of fermentation processes. Biotechnology Progress 19: 1816-1821.

Trauchessec M, Jaquinod M, Bonvalot A, Brun V, Bruley C, Ropers D, De Jong H, Garin J, Bestel-Corre G and Ferro M (2014). Mass spectrometry-based workflow for accurate quantification of *E. coli* enzymes: how proteomics can play a key role in metabolic engineering. Molecular & Cellular Proteomics, *in press*.

Triadaphillou S, Martin E, Montague G, Norden A, Jeffkins P and Stimpson S (2007). Fermentation process tracking through enhanced spectral calibration modelling. Biotechnology and Bioengineering 97 (3): 554-567.

U.S. Department of Health and Human Services, Food and Drug Administration (2004). Guidance for industry: PAT—a framework for innovative pharmaceutical development, manufacturing and quality assurance. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf Accessed 11 Nov 2013.

Valgepea K, Adamberg K, Nahku R, Lahtvee P-J, Arike L and Vilu R (2010). Systems biology approach reveals that overflow metabolism of acetate in *Escherichia coli* is triggered by carbon catabolite repression of acetyl-CoA synthetase. BMC Systems Biology 4: 166.

Voss C, Schmidt T, Schleef M, Friehs K and Flaschel E (2003). Effect of ammonium chloride on plasmid DNA production in high cell density batch culture for biopharmaceutical use. Journal of Chemical Technology and Biotechnology 79: 57-62.

Wang J, Zhang J, Wu W, Duan X, Wang S, Zhang M, Zhou S, Mo F, Xu Y, Shi J and Wu J (2010) Evaluation of Gallbladder Lipid Level During Carcinogenesis by an Infrared Spectroscopic Method. Digestive Diseases and Sciences 55: 2670-2675.

Xiong Z-Q, Guo M-J, Guo Y-X, Chu J, Zhuang Y-P and Zhang S-L (2008). Realtime viable-cell mass monitoring in high-cell-density fed-batch glutathione fermentation by *Saccharomyces cerevisiae* T65 in industrial complex medium. Journal of Bioscience and Bioengineering 105 (4): 409-413.

Xu Z, Shen W, Chen H and Cen P (2005). Effects of medium composition on the production of plasmid DNA vector potentially for human gene therapy. Journal of Zhejiang University-Science B 6: 396-400.

Yang Y-T (1999). Metabolic Flux Analysis of *Escherichia coli* Deficiente in the Acetate Production Pathway and Expressing the *Bacillus subtilis* Acetolactate Synthase. Metabolic Engineering 1: 26-34.

Yano K, Ohoshima S, Shimizu Y, Moriguchi T and Katayama H (1996). Evaluation of glycogen levels in human lung carcinoma tissues by an infrared spectroscopic method. Cancer Letters 110: 29-34.