Universidade Católica Portuguesa

Faculdade de Engenharia

Infrared Spectroscopy: A groundbreaking tool for monitoring mammalian cells' processes

Filipa Orrico Pereira da Rosa

Dissertação para obtenção do Grau de Mestre em Engenharia Biomédica

Orientadora: Profª. Doutora Cecília Calado

Co-orientadora: Doutora Marta Belchior

Júri

Prof. Doutor Manuel Barata Marques (Presidente)

Prof. Doutor Luís Joaquim Pina da Fonseca (Vogal)

Doutora Carla Maria Cadete Martins Moita Brites (Vogal)

Profª. Doutora Cecília Calado (Orientadora)

Maio de 2014

To Grandfather and Grandmother

ACKNOWLEDGMENTS

First of all, I would like to express my truly gratefulness to Professor Cecília Calado, not only for allowing me to work with such an innovative technique, infrared spectroscopy, but also for all the guidance along this project. I would like as well to thank to my adviser, Marta Lopes, for all the patient and help during these last few months. It was truly an honor to work with both of you!

I also leave a special thanks to Professor Manuel Barata Marques, Director of the *Faculdade de Engenharia, Universidade Católica Portuguesa*, for all the understanding and supporting, since the first day of classes, through the last days of my Master's Thesis!

To Professor Lídia, from *Faculdade de Farmácia, Universidade de Lisboa*, thanks for providing me the AGS cells that allowed me to do this work. Also, to Professor Claúdia Lobato da Silva and Professor Joaquim Sampaio Cabral, from the Stem Cell BioEngineering and Regenerative Medicine Research Group, *Instituto Superior Técnico, Universidade de Lisboa,* and to Joana Carmelo, from *Instituto Superior Técnico, Universidade de Lisboa*, thanks for the collaboration and the supernatants from hMSCs expansion in bioreactors. To *Instituto Nacional de Saúde Dr. Ricardo Jorge*, and within, to Doctor Jorge Machado, Doctor João Brandão, Doctor Helena Rebelo and Doctor Raquel Rodrigues, thanks for all the collaboration and technical support.

I also appreciate to *Fundação para a Ciência e Tecnologia* and *Agência de Inovação* for supporting this work, through the project PTDC/BIO/69242/2006 and the project 03/293 (Cancer Light Assisted Receeding Oncologic), respectively.

To Kevin, my colleague, my friend, a giant "thank you" for helping with data interpretation and also for all the support and, of course, the nuts after lunch!

A heartily thanks to all of my friends that shared this academic journey with me, who supported me unconditionally and gave me the strength to continue. We started this journey together but, unfortunately, we had to finish it apart. However, I am very grateful for meeting you, you are the reason why the academic environment in our Faculty was so perfect! You mean the world to me! To my childhood friends, no words needed, the next dinner is on me!!

To my beloved boyfriend, Pedro Antunes, the greatest thanks for all the unconditionally support, I am truly sure that I could not made it without you!

Last but not least, an enormous thanks to my family. You supported me despite of anything, specially my bad temper! Mother, thank you for allowing me to have an education, I want to be half of the woman you are when I grow up! Little brother, the most special men, thank you for the never-ending cheering and for your wise words! Grandmother, you were always by my side, you are my inspiration and motivation. Grandfather, a true gentleman and a brilliant human being, thank you for always pushing me to be better, for being my example, for everything! We miss you and we love you “to the moon and back”!

ABSTRACT

Fourier Transform Infrared (FTIR) spectroscopy is a high sensitive technique, which is able to detect vibrational modes of biomolecules, with a consequent world of applications. In the present work, the potential of this technique, working in the mid-infrared region (MIR), was explored for studying three distinct mammalian cells' processes, working in a rapid, reagent free and in a high-throughput mode.

FT-MIR spectroscopy was applied to monitor the expansion of human mesenchymal stem cells (hMSCs) in microcarriers and cultured in spinner flasks. It was possible to develop partial least squares (PLS) regression models to quantify, directly from the spectral data, key analytes, e.g., glucose, lactate and ammonia. Also, information about the cellular growth stage was possible to be extracted by the development of principal component analysis (PCA) models.

Additionally, were developed PLS regression models for estimating the transfection efficiency in a cell population without the need for a reporter gene. The model is valid for two distinct cell lines, the adherent cell line AGS and the semi-adherent cell line HEK, both transfected with pVAX containing the GFP gene. Besides an accurate estimation of the transfection efficiency, it was also possible to extract some meaningful information about the biochemical cellular effect of the transfection reagent on cells and the transfection event itself, proving the sensitiveness of the technique.

Finally, AGS cells infected with ten different *Helicobacter pylori* strains were analyzed based on FT-MIR spectral data. The different *H. pylori* strains presented different CagA/VacA genotypes, and were isolated from patients with different gastric pathologies (non-ulcer dyspepsia, peptide ulcer disease and gastric cancer). It was possible to differentiate cell samples according to the strain causing the infection, through PCA and cluster analysis.

Key words: Infrared spectroscopy, high-throughput analysis, mammalian cells, chemometrics, transfection, mesenchymal stem cells, *Helicobacter pylori*

RESUMO

A espectroscopia de infravermelho é uma técnica extremamente sensível, com a capacidade de detetar modos vibracionais de biomoléculas, tendo, consequentemente, um mundo de aplicações. No presente trabalho esta técnica foi usada para estudar vários processos com recurso a células animais.

A espectroscopia FT-MIR (do inglês *Fourier transform mid-infrared region)* foi utilizada para estudar a expansão de células mesenquimais estaminais (hMSCs) em *microcarriers.* Foi possível estimar a concentração de metabolitos chave para o crescimento celular, como a glucose, o lactato e a amónia, diretamente a partir de dados espectrais. Foi ainda possível inferir informação relativa ao ciclo celular das células em crescimento, através do desenvolvimento de modelos PCA (do inglês *principal component analysis).*

Adicionalmente, desenvolveram-se modelos de regressão PLS (do inglês *partial least squares)* com o objetivo de estimar a eficiência de transfeção numa população de células. Trabalhou-se com duas linhas celulares distintas, uma linha aderente, AGS, e uma linha semiaderente, HEK, ambas transfetadas com o plasmídeo pVAX contendo o gene da GFP (do inglês *green fluorescent protein).* Para além de ter sido possível determinar, com elevada precisão, a eficiência de transfeção, independentemente do tipo de célula, foi ainda possível extrair informação extremamente relevante sobre o estado celular, nomeadamente o efeito da exposição ao reagente de transfeção e ainda alterações metabólicas resultantes do próprio evento de transfeção celular, provando a elevada sensibilidade da técnica.

Por fim, estudou-se o efeito da infeção por *Helicobacter pylori* em células gástricas humanas, AGS. As células foram infetadas com dez estirpes diferentes de *H. pylori*, incluindo estirpes com diferentes genótipos CagA/VacA e isoladas de pacientes com diferentes patologias gástricas, como gastrite, úlcera e cancro gástrico. Foi ainda possível distinguir as células infetadas de acordo com a estirpe responsável pela infeção, quer através de modelos PCA, e ainda com recurso a algoritmos de agrupamento.

Palavras-chave: Espectroscopia de infravermelhos, quimiometria, células animais, transfeção, células mesenquimais estaminais, *Helicobacter pylori*

TABLE OF CONTENTS

LIST OF SYMBOLS

General notation

$[\,]^{-1}$: Matrix inverse

$[\,]^{T}$: Transposed matrix or vector

^ : Estimated value

Notation for multiplicative scattering correction (MSC)

r_i: Values from the reference spectrum

x_i: Values from the initial spectral data

z_i: New values resulting from transformation

a: Intercept of a least-square regression of the x_n values on the r_n values

b: Slope of a least-square regression of the x_n values on the r_n values

i: Row index corresponding to each sample

k: Column index corresponding to each wavelength

p: Wavelengths

Notation for principal component analysis (PCA) and partial least squares (PLS) regression

E: Error matrix for the X-data matrix $(n \times p)$

F: Error matrix for the Y-data matrix $(n \times m)$

P: Loading matrix from X $(p \times g)$

T: Score matrix for X $(n \times g)$

W: PLS weights matrix $(p \times g)$

X: Descriptor data matrix (spectral matrix) $(n \times p)$

f: Error vector for the y-data matrix $(n \times 1)$

g: Number of chosen factors

m: Number of variables (wavelengths) in Y

n: Number of samples in X and Y (or y)

p: Number of variables in X

q: Loading matrix from y

y: Response data vector ($n \times 1$)

β: Vector containing the regression coefficients ($p \times 1$)

LIST OF ABBREVIATIONS

AGS: Humana gastric carcinoma

AP: Alkaline phosphatase

ATR: Attenuated Total Reflectance

BFS: Bovine fetal serum

CAT: Chloramphenicol acetyltransferase

CHO: Chinese hamster ovary

CV: Cross-validation

DAPI: 4,6-diamino-2-phenylinzon

DOC: dissolved oxygen concentration

ELISA: Enzyme-linked immunosorbent assay

ER%: Error as percentage of the range

FDA: Food and drugs administration

FTIR: Fourier Transform Infrared

GC: Gastric cancer

GFP: Green fluorescent protein

HEK: Human embryonic kidney

hMSCs: Human mesenchymal stem cells

HPLC: high-performance liquid chromatography

IMDM: Iscove’s Modified Dulbecco Medium

IR: Infrared

LOO: Leave-one-out

LV: Latent variables

MALT: Gastric mucosa associated lymphoid tissue

MIR: Mid-Infrared

MSC: Multiple Scattering Correction

MSCs: mesenchymal stem cells

NIR: Near-Infrared

NUD: Non-ulcer dyspepsia

ONPG: o-nitrophenyl β-D-galactopyranoside

PAT: Process analytical technologies

PC: Principal Component

PCA: Principal Component Analysis

PLS: Partial Least Squares

PUD: Peptide ulcer disease

R^2: Squared correlation coefficient

rhATIII: antithrombin III

RMPI: Roswell park memorial medium

RMSEC: Root mean squared error of calibration

RMSECV: Root mean squared error of cross validation

RMSEP: Root mean squared error of prediction

SEAP: Secreted alkaline phosphatase

SNR: Signal to Noise Ratio

TLC: Tin liquid chromatography

B-gal: β-galactosidase

LIST OF FIGURES

LIST OF TABLES

I. THESIS OVERVIEW

I.1 Objectives

The main goal of the present work was to evaluate the application of Fourier Transform Infrared (FTIR) spectroscopy, using the mid-infrared radiation (MIR) and working in a high-throughput mode, for studying several processes and mechanisms involving mammalian cells, namely, monitoring the expansion of human mesenchymal stem cells (hMSCs) in microcarriers, estimating and studying transfection events without using any reporter gene and studying infection by *Helicobacter pylori* in an adenocarcinoma gastric (AGS) cell line. The potential of the technique for detecting minor molecular alterations in cells, was also evaluated.

It was also aimed to optimize the procedures related to the acquisition and interpretation of infrared (IR) data and, by this way, to promote the use of the technique in a near future, not only for monitoring cellular processes, but also for other applications.

I.2 Thesis Outline

The present work is organized essentially in three sections. The first section includes a description of IR spectroscopy, where some theoretical explanations about the technique are presented. In a second section, chemometrics and spectral analysis are introduced, including the mathematical treatments applied to IR data in this work. The last section of this work concerns the experimental work realized for this thesis. Three different works using IR spectroscopy were conducted, and are separated in different subchapters, as they differ slightly from each other. For each work conducted, i.e., monitoring the expansion of hMSCs cells, estimating the transfection efficiency on a cell population, and studying AGS infection by *H. pylori*, four sections are presented: an introduction, an experimental section, a section dedicated to the results and discussion and, finally, the conclusions. At the end of this thesis a general conclusion is presented, encompassing the results achieved during all the work for this Master's thesis.

II. INFRARED SPECTROSCOPY

I.1 Theory of Infrared Spectroscopy

Originally spectroscopy was defined as the study of the interaction of electromagnetic radiation with matter, in function of wavelength. The concept was then extended to include any property that is function of wavelength or frequency of that radiation. Every spectroscopic technique is based on the same principle that, is given certain conditions, when the materials interact with radiation they emit or absorb energy. Some materials also reflect radiation and/or disperse/diffract radiation. Absorption occurs when the emitted radiation is attenuated by the sample and emission takes place when radiation is produced by the sample due to excitation by a light source. The reflection of radiation depends essentially of the material's surface and diffraction is mainly related to the composition, shape and microstructure of the sample (Nicolaï *et al.*, 2007)

IR spectroscopy uses the infrared region of the electromagnetic spectrum. The IR region is limited in the electromagnetic spectrum by the visible red light and the microwaves and ranges from 14000 to 4 cm^{-1} (0.7 to 250µm). There are three main regions of IR radiation: far-infrared (far-IR), mid-infrared (MIR) and near-infrared (NIR). Far-IR ranges from 400 to 4 cm^{-1} in the IR spectrum and will not be discussed in the present work. MIR represents the region of the IR spectrum between 4000 and 400 cm^{-1} and NIR the region between 14000 and 4000 cm^{-1} (Figure II.1). Both MIR and NIR radiation will be discussed later in this work since they represent the type of IR radiation that is most used in diverse applications of spectroscopy (Smith, 2011).

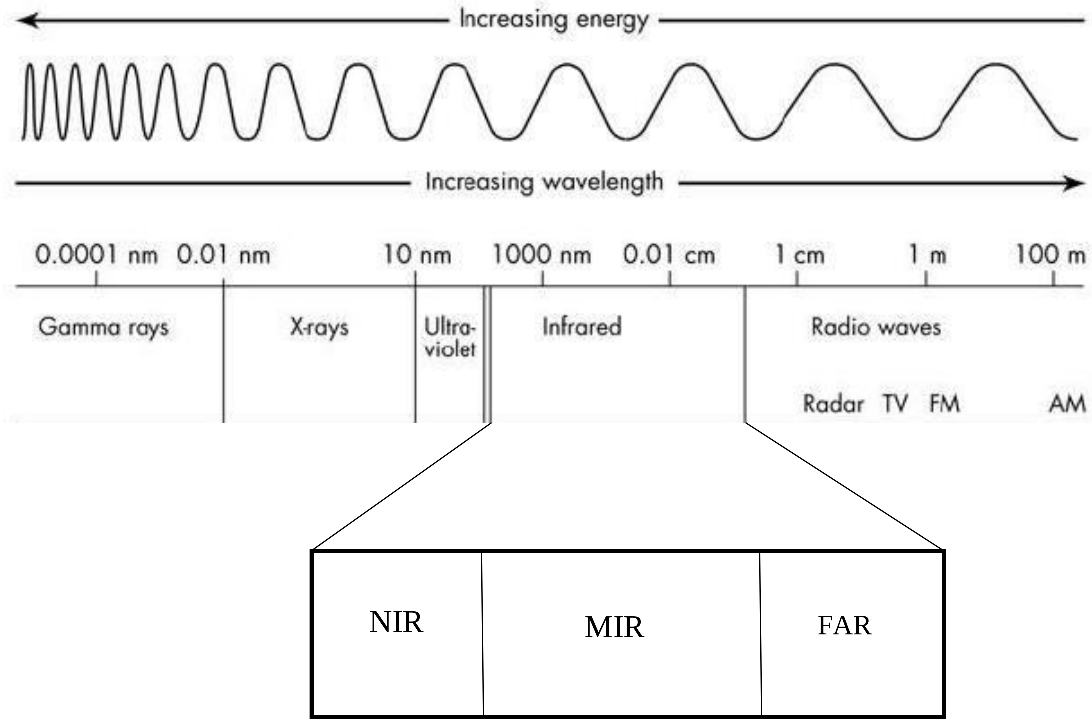

Figure II. 1 - Electromagnetic spectrum with IR radiations highlighted. (Adapted from: http://jasonbachand.blogspot.pt/2012/01/introduction-to-black-holes-via-black.html)

In IR spectroscopy the sample is irradiated with IR light and the absorption of this radiation by the molecules in the sample stimulates its molecular vibration. At temperatures above absolute zero the atoms in a molecule are in continuous vibration with respect to each other. When a molecule is exposed to IR radiation it only absorbs the frequencies corresponding to its own vibration frequency. These changes in the vibration mode of molecules, or in case of gases excitation of molecular rotational levels as well, due to interaction with radiation, produce the bands in the IR spectrum. Each band is characterized by a frequency and amplitude (Duygu, 2009).

Almost any molecule that possesses covalent bounds absorbs IR radiation, with exception of monoatomic molecules ($He\ or\ Ne$) and homopolar diatomic molecules (H_2, O_2 …). Monoatomic molecules are formed by only one atom so they don't have a dipolar moment. Homopolar diatomic molecules are formed by only one type of atoms so they don't have a dipolar moment as well, since the electronic filed of atoms is the same (Griffiths, 2002). All the other type of molecules will have a dipolar moment so they will absorb IR radiation and can be quantified or qualified by IR spectroscopy.

As mentioned before, the interaction of IR light, that possesses a specific wavelength or frequency, with a molecule with dipolar moment causes an alteration of its vibrational mode. The vibrational modes of a molecule that are infrared active are essentially stretching and bending (Figure II.2). The stretching can be symmetric or antisymmetric and results from an alteration of the molecular bound's length. The bending corresponds to a change of the angle between two atoms or two groups of atoms in a molecule (Babrah, 2009).

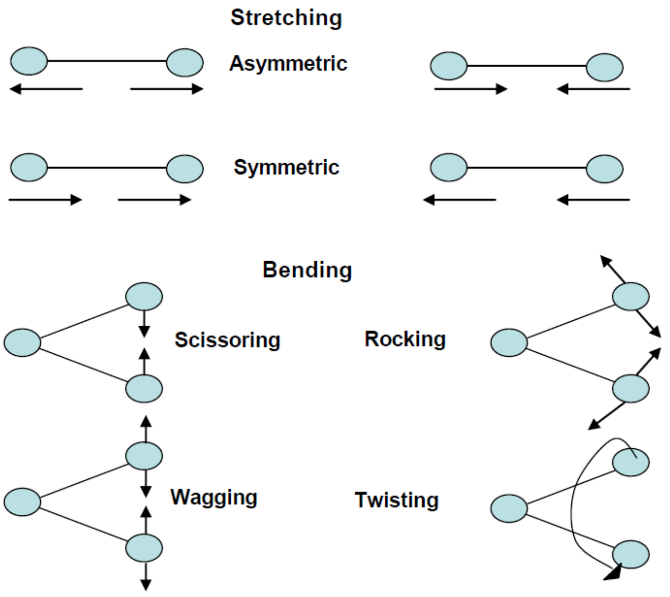

Figure II. 2 – Molecular vibrational modes (Babrah, 2009)

The IR spectrum is characteristic of each type of molecule, since it depends mainly of the mass of the atoms, their geometric arrangement and the bound forces between them. Given that, since there are no two different molecules that possess these three same characteristics, each molecule will have, in theory, a distinct spectrum. Extending that concept two different samples, with distinct molecular composition, they will also have different spectra. This principle is the foundation of IR spectroscopy and it makes possible to distinguish, qualify or quantify virtually any type of sample (Smith, 2011).

Since certain regions of the spectrum were already attributed to certain molecular bounds and combination of atoms, and we know the composition of every biomolecules, it is easy to extrapolate these results and to associate these biomolecules to certain regions of the IR absorption spectrum, especially in MIR region. This was the basis of application of spectroscopy in the biological field. The regions of MIR spectrum corresponding to each group of biomolecules are nowadays well characterized, namely regions associated with proteins, lipids and amino acids and even lactate, urea or glucose absorption (Figure II.3).

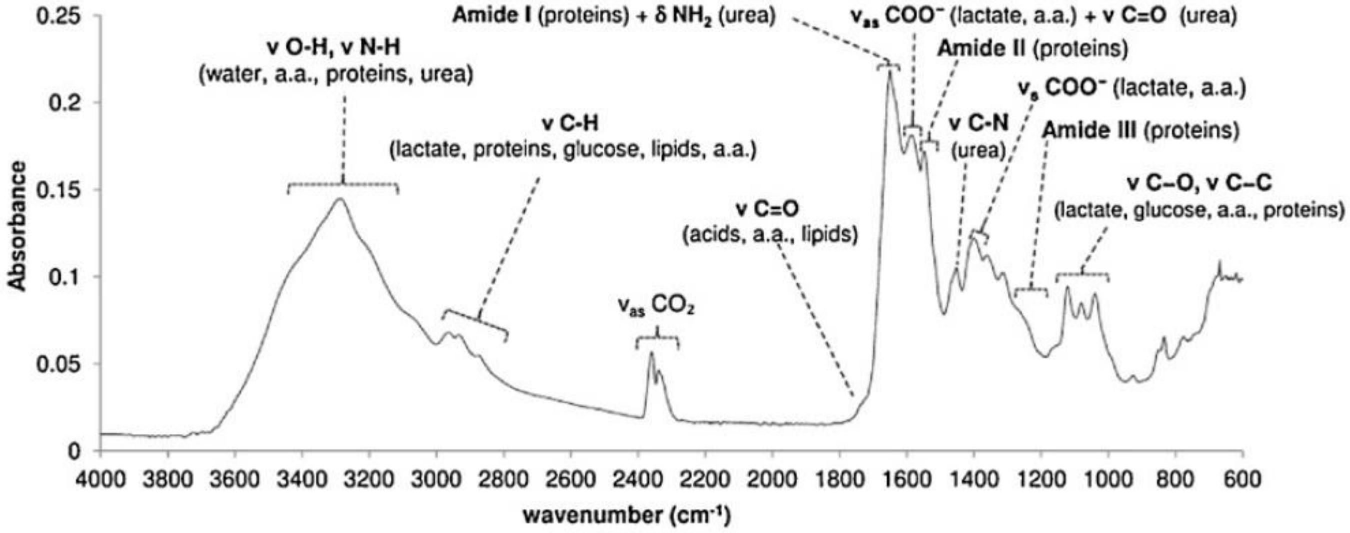

Figure II. 3 - MIR absorption spectrum with bands associated to molecular bounds of biomolecules highlighted (Graça *et al*. 2013).

Despite of the complex composition of biological samples, like cells and tissues, it can be observe that the most strong vibrational frequencies correspond to macro-biomolecules such as proteins, nuclei acids, lipids or glicids (Smith, 2011), due to its high concentration in the cell in relation to other biomolecules. The fact that different biological samples have different molecular compositions is reflected on their different absorption spectra.

II.1.1 Mid-Infrared (MIR) Spectroscopy

Mid-infrared spectroscopy uses the IR region of the electromagnetic spectrum that ranges from 4000 to $400\,cm^{-1}$. In MIR region the IR bands arise essentially from fundamental vibrational modes so they can be more easily attributed to specific molecular groups, which makes this technique more sensitive, allowing to extract more information of the spectra, comparing to NIR spectroscopy (Smith, 2011).

On the other hand, water absorbs much more radiation in the MIR region than in NIR region, which can be a problem when we are analyzing aqueous samples, being usually necessary an extra step by which the sample is dehydrated before spectral acquisition (Landgrebe *et al.*, 2010). In alternative, attenuated total reflectance (ATR) can be applied, as described in subchapter II.4.2.

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. Also MIR radiation is more difficult to transport so it is more difficult to achieve remote measures.

MIR spectroscopy, like NIR 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 (Lourenço *et al.*, 2012), which is extremely useful when we are dealing with samples we want to preserve. Moreover, spectra can be altered due to fluctuations in the equipment's environment and sometimes it is necessary to resort to chemometric methods due to the complexity of spectra, though rich in information, or when aiming quantitative analysis.

II.1.2 Near-Infrared (NIR) Spectroscopy

Near-infrared radiation ranges from 14000 to 4000 cm^{-1} in the electromagnetic spectrum and covers the transition from the visible light to the mid-infrared region (Smith, 2011).

The absorption spectrum that one can obtain applying NIR spectroscopy is usually very complex, mainly when this technique is used to analyze biological samples or monitoring bioreactors, since it results from a combination and overlap of vibrations from different chemical elements and functional groups. This is the main disadvantage of applying NIR spectroscopy and the main reason why NIR spectroscopy is less sensitive than MIR spectroscopy. Due to these complex spectra, usually NIR spectroscopy shows a great dependence on chemometric methods.

On the other hand generally the NIR radiation is less absorbed by samples than MIR radiation, which results in a great penetration power. Moreover, NIR radiation can be easily transported by optical fibers, then it is possible a remote acquisition of spectra (Lourenço *et al.*, 2012).

Sometimes NIR spectroscopy is preferred *versus* MIR spectroscopy since the water does not absorb so strongly NIR radiation, so the bands of water will not mask any information in the sample's spectrum. This is an important factor when we are dealing with biological samples where water is one of the main components.

Just like in MIR spectroscopy, NIR spectroscopy allows a fast acquisition of spectra, especially after the development of Fourier Transform Infrared (FTIR) spectroscopy, it is not necessary an extensive or any preparation of the sample, it is possible to measure several sample's properties at once, different kind of samples can be evaluated, it is a non-destructive technique and no reagents are necessary, which make each procedure less expensive (Smith, 2011).

Some typical applications of NIR spectroscopy are in the pharmaceutical industry as a tool for identification of compounds, test purity, structural investigation, quantitative measures, monitoring drug production or to verify drug's identity, as review by Kalinkova (1999). Other emerging applications include diagnosis (Kondepati *et al.*, 2008) or bioreactor monitoring essentially in biotechnology processes as reviewed by Lourenço *et al.* (2012).

II.2 Instrumentation

The instrument used in IR spectroscopy is called infrared spectrometer or 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 that will detect the radiation that is transmitted or reflect by the sample (Reich, 2005).

The beam source may consist on an inert solid thermally heated (Hsu, 1997) or in an incandescent filament like tungsten or quartz/halogen lamps, for NIR region, and carbon-silicon bars, for MIR radiation (Christian, 1994).

There are two types of detectors: thermal detectors and photon detectors. Thermal detectors measures the heat produced by IR radiation when in contact with the sample and photon detectors are based on the interaction of IR radiation with semiconductor materials where excitation of electron occurs and it is generated a small electrical current that can be quantified (Hsu, 1997).

The way that light is modulated, by a monochromator on by an interferometer, defines the type of spectrophotometer. There are essentially two types of spectrophotometer: Dispersive Infrared Spectrometers and Fourier-Transform Infrared Spectrometers. In both configurations the beam source, detectors and sample holders used are essentially the same.

Dispersive Infrared Spectrometers were the first kind of spectrophotometers to be developed. In this configuration a monochromator is used. A monochromator is a device used to separate a range of radiations in a certain range of wavelengths or frequencies and the most common kinds include prism and gratings coupled with systems of mirror and filters (Stuart, 2004).

Fourier-Transform Infrared (FTIR) spectrometers appeared later and basically the monochromator is substituted by an interferometer. The interferometer, the heart of every FTIR spectrometer, basically measures the interference pattern between two light beams. The IR radiation, after entering in the interferometer, is dived in two beams that will travel different paths (D1 and D2 in Figure II.4). After each beam travel its path, the two beams are reunited in a single beam again, leaving the interferometer. Michelson interferometer was perhaps the first spectrometer invented (Figure II.5), but the basically operation is common to all interferometers. Briefly, a collimating mirror receives the IR light from the source, and makes its beam parallel to each other, while directing them to the beam splitter. The beam splitter separates the radiation in two beams, redirecting one beam to the fixed mirror and the other to the moving mirror. These two beams, after traveling its path, return to the beam splitter, where they are combined in a single beam again, and send on to the sample (Figure II.5). The resulting spectrum is called an interferogram (a spectrum consisting on intensity *versus* acquisition time) that is later traduce to the final IR spectrum (intensity *versus* frequencies or wavelengths) by the mathematical operation called Fourier-Transform (Smith, 2011; Stuart, 2004).

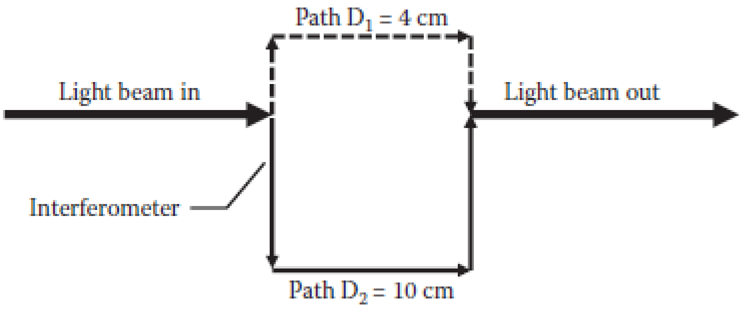

Figure II. 4 – Simplified illustration of an interferometer (Smith, 2011).

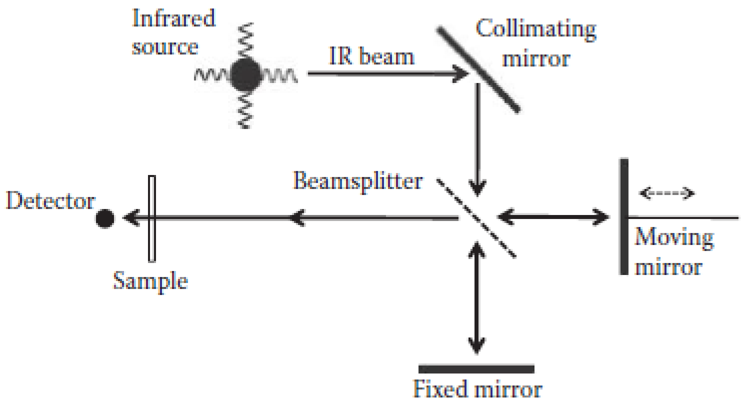

Figure II. 5 – Scheme of the Michelson interferometer (Smith, 2011).

FTIR spectrometers substituted the dispersive systems since they are faster, all the frequencies are examined at the same time, and essentially they have a higher signal-to-noise ratio (SNR) (Hsu, 1997; Pistorius, 1995). The amount of signal in a spectrum is highly dependent on the amount of light that reaches the detector. For the dispersive spectrometers, the beam need to travel through prisms, slits and gratings, before reaching the sample. Thus, the final beam that is detected have a much lower intensity, comparing to the beam that leaves the beam source and, consequently, the final spectrum has a lower SNR. Probably the big advantage of achieving a high SNR, is that it allows more sensitive measurements. For instance, in a less noisy spectrum even the smaller peaks became more evident. For FTIR spectrometers the SNR can be 10-100 higher than for the dispersive spectrometers (Smith, 2011).

II.3 Acquisition Modes

Depending on the samples properties the spectral data can be acquired essentially by two different modes: Transmission and Reflection.

II.3.1 Transmission

In transmission mode IR radiation passes through the sample and it is evaluated the decrease in the incident beam (Figure II.6). The obtain spectrum is the result of radiation that passes through the sample (that is proportional to the radiation absorbed by the sample) in function of wavelength, and depends of the radiation's pathlength (Hsu, 1997).

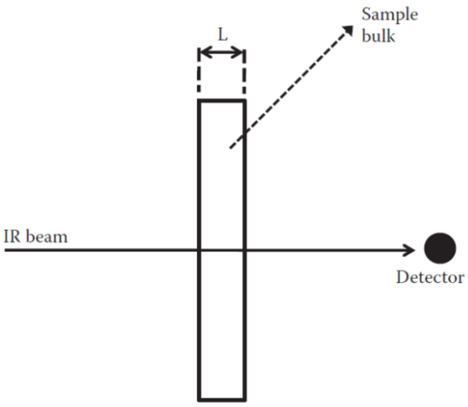

Figure II. 6 - Scheme of IR transmission mode (L represents the pathlength, by that means the thickness of the sample) (Smith, 2011)

Beer-Lambert law allows us to deduce concentrations of certain compounds in a sample through the IR radiation that is absorbed via the following equation:

$$A = \varepsilon \times L \times C = log_{10}(\frac{1}{T}), \qquad \text{(Equation II.1)}$$

where A is the absorbance, ε the coefficient of absorptivity, L is the pathlength, C the concentration and T the transmittance.

Absorption or transmission measures are universal, minimal preparation of the sample is necessary and usually it is possible to obtain spectra with good signal to noise ratio. In case of samples absorbing too much or too less radiation it is necessary to explore other modes, as reflectance (Smith, 2011).

II.3.2 Reflectance

In reflectance mode the IR radiation detected is the radiation reflected by the sample's surface, as is useful when the sample absorbs too much or too less energy or in case of samples that reflect the majority of the incident radiation (Figure II.7).

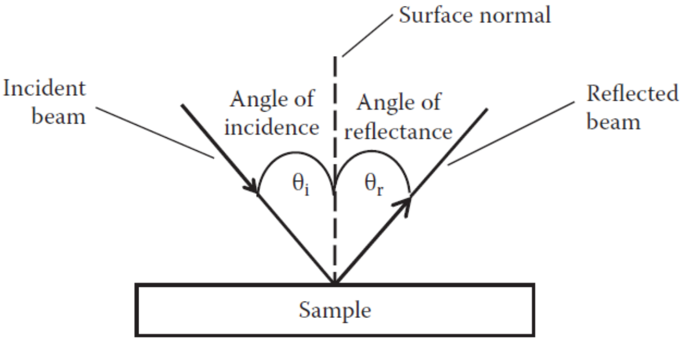

Figure II. 7 - Scheme of an IR beam reflected by the sample's surface. Θ_i is the incident angle of the incident beam and Θ_r is the angle of the reflected beam (Smith, 2011).

There are essentially three types of reflectance measures: Specular Reflectance, Diffuse Reflectance and Attenuated Total Reflectance (ATR). In Specular Reflectance the angle of the incident beam (Θ_i) is the same of the reflected beam (Θ_r) (Figure II.7). In Diffuse Reflectance the angle of the incidence beam and the angle of reflected beam are different (Figure II.8) and this usually happens in samples with rough surfaces (Smith, 2011). In Attenuated Total Reflectance (ATR) an IR beam travels through a crystal that is in close contact with the sample (Hsu, 1997).

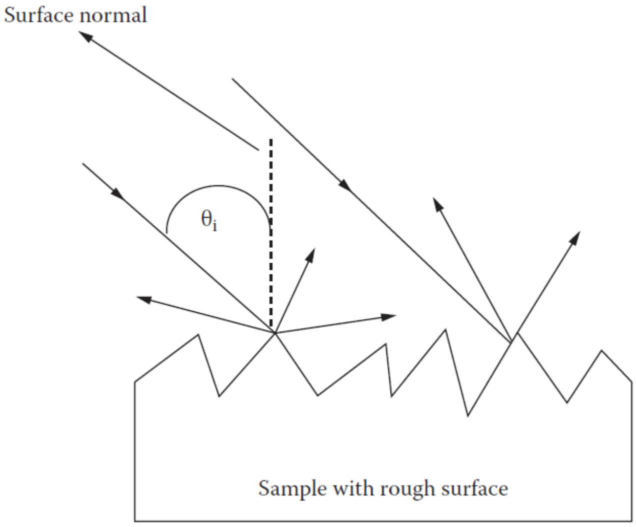

Figure II. 8 - Scheme of Diffuse Reflectance. Θ_i is the angle of the incident beam (Smith, 2011).

ATR was created by Harrick and Fahrenfort and is based on the transmission of IR radiation through a crystal which is in contact with the sample, Figure II.9 (Roychoudhury *et al.*, 2006). This method is becoming highly relevant for the study of living cells and biological tissues, since it is possible to reduce the water's interference in MIR spectroscopy.

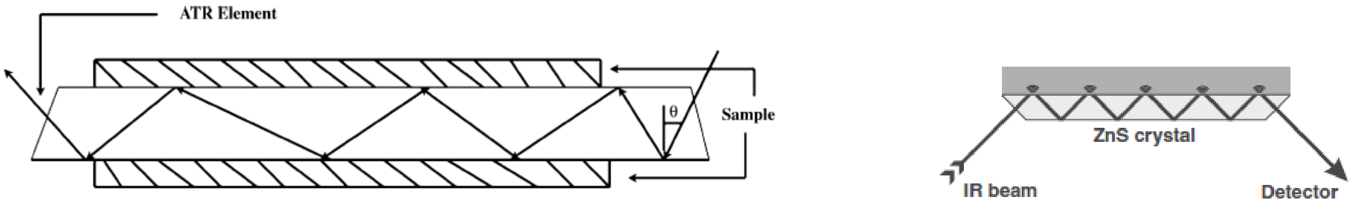

Figure II. 9 - Alternative configurations of ATR system (Roychoudhury *et al.*, 2006 and Adapted from Roth *et al.*, 2012).

An ATR accessory is a combination of IR radiation with reflection techniques and essentially operates by measuring the total reflection of an evanescent wave, which penetrates the sample in contact with a crystal. Briefly, an IR beam is reflected through the crystal and its intensity changes as the beam travels across it, if the sample absorbs MIR radiation, due to the interactions with sample. It is important that the sample is in close contact with ATR crystal (Khoshhesab, 2012). There are a few crystals that can be used in ATR, mainly zinc selenide (ZnS), silicon and diamond. It is important that the crystal have a high refraction index and do not absorb MIR radiation. ZnS is one of the most attractive materials, although it is damaged by acids and oxidant agents, so its performance decreases over time. Diamond is more resistant to chemical attacks but the C-C bounds absorb MIR radiation, so part of the sample's spectrum can be lost and also this is an expensive material (Landgrebe *et al.*, 2010).

ATR is a technique that makes possible to obtain rich spectra, since it works with MIR radiation, and it is also solved the problem of water absorption in the MIR region, as IR beam penetrates less in the sample the amount of water that is sampled is reduced. Another advantage of using an ATR system is that differences in sample's thickness, that causes alterations in conventional IR absorption spectra, do not affect ATR spectra because the penetration of the evanescent beam in the sample is determined by the sample's refraction index (Timlin *et al.*, 2009). Due to these characteristics ATR is an interesting technique to monitoring bioprocesses, to evaluate living cells or to analyze other biological samples, since the water is no longer a problem (Landgrebe *et al.*, 2010). Probably the main disadvantage of this technique is that it is not yet possible to perform high-throughput analysis, each sample needs to be evaluated independently.

III. THE BASICS OF CHEMOMETRICS

Chemometrics is the field that combines mathematical, statistics and computational methods in order to process data and to solve problems in chemistry, biochemistry or chemical engineering (Roggo *et al.*, 2007). It is therefore possible to extract relevant information from the data which otherwise would be very difficult. 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 producing broad and overlapping spectral bands, was only possible due to the development of chemometric methods. Even for MIR spectroscopy, where spectral bands are normally well defined, chemometrics may also play an important role, by making easier the interpretation and handling of large data sets, as well as by reducing the noise that is often present in spectra.

Chemometrics in spectroscopy can be divided into three main categories, mathematical pre-processing techniques, qualitative and quantitative methods. Mathematical pre-processing techniques include methods that work by eliminating spectral noise or effects of radiation scattering, i.e., the spectrum is "edited" so that only important information is kept. Qualitative methods on the other hand group samples according to their similarities, i.e., each member of a given group are more similar to the samples in its own group than to samples of other group. Quantitative methods usually resort to regression methods and are applied to predict samples' properties that can be quantified (Roggo *et al.*, 2007).

In the present work a few pre-processing techniques are review, including multiplicative scatter correction (MSC), baseline correction, normalization, smoothing and derivatives. Concerning quantitative and qualitative methods, only cluster analysis, principal component analysis (PCA) and partial least squares (PLS) regression will be discussed, since were the main techniques applied later for the data analysis.

III.1 Spectral pre-processing techniques

Spectral pre-treatments are essential for eliminating spectral alterations due to undesired variations, such as noise, differences along the sample thickness, differences in the number of cells across the sample and scattering events. The goal is that the final spectra possess the minimum irrelevant information as possible. The pre-processing techniques applied to IR data in the present work are described next, and that includes: multiplicative scatter correction (MSC), baseline correction, normalization smoothing, and derivatives.

III.1.1 Multiplicative Scatter Correction (MSC)

Multiplicative Scatter Correction (MSC) was first developed for NIR spectroscopy and it is used to eliminate changes in spectra due to radiation scattering. This transformation works on the influence of scattered radiation in a group of spectra from different samples. The goal is to find the “ideal” spectrum of the group. For this method a reference spectrum is necessary, which is usually the mean spectrum of all available spectra. MSC works by fitting each spectrum to the average spectrum, which is thought to be the ideal, performing a transformation where the spectral data (x_1, x_2, …, x_p) is converted into new values (z_1, z_2, …, z_p), where p corresponds to wavelengths (Fearn *et al.*, 2009). The following equation describes the transformation from x to z

$$z_i = \frac{x_i - a}{b}, \qquad \text{(Equation III.1)}$$

where a represents the intercept and b the slope of a least-squares regression of x_1, x_2, …, x_p on the values r_1, r_2, …, r_p coming from the reference spectra.

III.1.2 Baseline correction

Not always the obtained spectra are grounded at zero. First and second derivatives can be used to solve this problem, as well as other methods of baseline correction. The type of algorithm used depends on the baseline correction needed. Those spectra which are dislocated from zero by a constant value are the simpler cases, since subtracting the value in question from the spectrum is usually enough. Though there are more difficult cases, for instance cases in which the baseline presents a slope, or even spectra with curvatures. In these cases an algorithm generating a function, a linear or polynomial function, can bring the spectrum to zero (Otto, 1999; Smith, 2011).

There are probably two main disadvantages of using baseline correction algorithms. First, it is difficult to find a function that adjusts properly to the spectrum 'curvature, although there are already some good algorithms. Besides that, the curvature along the spectrum is not always equal, so a unique function will hardly adjusts correctly to the entire spectrum. In that way, sometimes it is preferred to apply derivatives for offset correction. The problem of applied derivatives 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).

III.1.3 Normalization

The goal of normalization is pre-processing the data in order to minimize 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. Of course 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 great review on this topic may be found at Randolph (2006). In the present work, all spectra were normalized using the Amide I band, at 1650 cm^{-1}. Basically, all spectra were divided by a previously determined constant, so all the spectra ended up having the same intensity at 1650 cm^{-1}.

III.1.4 Smoothing

Smoothing is used to treat spectra which have a low SNR, or before applying derivatives as they highlight the spectral noise. In these cases it is also important to try to solve this problem by other means before, such as using a higher number of scans or optimizing the sample preparation. If SNR is not improved in this way than pre-processing methods like smoothing should be applied. Smoothing makes possible to reduce the noise in the data so that bands can be better distinguished (Smith, 2011).

There are several kinds of filters or smoothing algorithms, being that the most commonly used are the ones based on average values. The method works by creating a "smoothing window" in the data. In this "window" the average of y's values (intensity or absorption) is calculated. This average is then associated to the p middle value of the smoothing window (wavenumber or frequency). The spectra are truncated at the end since the final frequencies will never represent the center of a smoothing window (Smith, 2011). The choice of the window size its close related to the noise reduction, the higher the window size greater the reduction of the noise. However, if smoothing is too strong, the window is too big, neighbor bands can be merged with each other and also important information in data can be lost and, so must be a tradeoff between loss of information and noise elimination

III.1.5 Spectral Derivatives

Spectral derivatives are normally employed to remove baseline offsets and for highlighting spectral information, through the resolution of overlapping bands. Applying the first derivate is extremely useful in cases where the offset is constant, since de first derivate of a constant is zero. The second derivative can also be applied and the result will be not only the removal of the baseline offset but also the resolution of overlapping bands. This is the reason why the second derivative is commonly applied to NIR spectra (Otto, 1999).

First and second derivatives can be also useful to highlight subtle differences between spectra, allowing extract more information from spectral data. However, before applying derivatives, it is important to have in mind that the derivative spectra will have more noise than the initial one, so first and second derivatives must be applied in spectral data with high SNR. If not the case, it is necessary to apply a smoothing algorithm first, otherwise derivatives will only enhance the existing noise. Savitzky-Golay is an algorithm commonly used with this purpose, since it applies a filter before derivation.

III.2 Cluster Analysis (CA)

Clustering is a non-supervised classification method that groups data in clusters according to their semblance, usually determined by pattern recognition algorithms that rely on distance measures. The shorter the distance between two objects or samples, the closer they are.

A cluster describes a group where the samples are more similar to each other than to those outside the group. Two types of clusters can be created, hierarchically and non-hierarchically (Otto, 1999).

In Hierarchical Cluster Analysis the objects or samples under study are combined according to their similarity or distance, as mentioned before. Commonly this method starts with a single sample and then other samples are added to create a cluster. Deciding the number of clusters can be a problem, although generally this number is known.

In Nonhierarchical Cluster Analysis samples are not grouped hierarchically. Generally a first division of the samples into clusters is first performed and then the cluster's centroid is calculated. If that the case samples are then allocated to another cluster wherewith they have more similarity or lower distances with the other samples in the new cluster (Otto, 1999).

Cluster analysis is also very useful in a way that it allows summarizing the information and gives an easier graphical output to analyze since the samples are grouped by classes. Cluster Analysis is less useful when the data is too similar, since only one cluster may be created, as well as when data are too heterogeneous allowing too many clusters to be created.

III.3 Principal Component Analysis (PCA)

Principle Component Analysis (PCA) is a data reduction algorithm that is very useful when we are dealing with high dimensional data as spectral data. It also allows performing qualitative analysis, while no information about the components of interest is necessary (unsupervised method). New variables retaining the maximum variance of the initial data are created through linear combinations of the spectral data, called principal components (PCs) (Jollife, 2002). Once the source of variance in data is identified, it is possible to visualize the major tendencies in data (Lourenço, 2012). PCs are ordered in terms of variance in the data set explained, 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). The initial data matrix is decomposed as following:

$$X = TP^T \ + E, \qquad \text{(Equation III.2)}$$

where n is the number of samples in X, p is the number of variables in X, g is the number of chosen factors, X $(n \times p)$ is the descriptor data matrix, T $(n \times g)$ is the score matrix, P $(p \times g)$ represents the loading matrix and E $(n \times p)$ is the error matrix for the X-data matrix. In that way a PC is described as a combination of loadings and scores, where loadings represent the contribution of each wavenumber to the PC and scores results from linear combinations of the initial data in X. Therefore each spectrum can be described as a combination of principle components (Kidder *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 the samples in a smaller dimensional space and consequently evaluate their semblance. Since 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.

III.4 Partial Least Squares (PLS) Regression

Partial Least Squares (PLS) is a data reduction and quantitative method developed by Herman Wold, in 1960s, with the goal of overcoming the tough decision of choosing the components of the model that are represent the best the variable(s) of interest, as for PCA. The PLS model find new components, factors or latent variables (LVs) and each component is obtained through the maximization of the covariance between the reference data (*y-data*) and all the linear combinations for the spectral data (*X-data*). So, in theory, the new variables obtained would be more related with the variance in the *y-data* than the components obtained through PCA. There are two variants of the PLS algorithm, PLS-1 and PLS-2. For PLS-1 a separate set of scores and loadings are calculated for each variable in the y-data and for PLS-2 a unique set of loadings and scores are calculated for all the y variables. PLS-1 is more accurate than PLS-2, since the scores and loadings are adjusted for each property of interest. There are a few methods for PLS development, being that the NIPALS and SIMPLS methods are the most commonly used. A more detailed information about the different PLS methods can be seen elsewhere (Hammond and Clarke, 2002; Yeniay and Göktş, 2002, Naes *et al.*, 2002).

The *X-data* and *y-data* are decomposed as following:

$$X = T.P^T + E \qquad \text{(Equation III.3)}$$

$$y = T.q + f, \qquad \text{(Equation III.4)}$$

where n is the number of samples in X and y, m is the number of variables in y, p is the number of variables in X, g is the number of chosen factors, X $(n \times p)$ is the descriptor data matrix, y $(n \times 1)$ is the vector of reference data (usually concentrations), T $(n \times g)$ is the latent variables score matrix, q $(m \times 1)$ and P $(p \times g)$ represent the loading matrix, from X and y respectively, and E $(n \times p)$ and f $(n \times 1)$ are the residual matrixes that contains irrelevant data from X and y, respectively. The scores in T will result from linear combinations of initial variables in X. In that way a LV is described as a combination of loadings and scores, where loadings represent the contribution of each wavenumber to the latent variable and scores results from linear combinations of the initial data in X, where the covariance with y is maximized. Therefore each spectrum can be described as a combination of these new variables that are created (Naes *et al.*, 2002).

The PLS regression coefficients $\hat{\beta}$ are given by:

$$\hat{\beta} = W(P^T W)^{-1}(T^T T)^{-1} T y, \qquad \text{(Equation III.5)}$$

where W $(p \times g)$ is the PLS weights matrix. Once the regression coefficients, $\hat{\beta}$, are determined they can be used to obtain the predictions:

$$\hat{y} = X\hat{\beta}, \qquad \text{(Equation III.6)}$$

IV. NEW APPROACHES TO STUDY MAMMALIAN CELLS' PROCESSES BY INFRARED SPECTROSCOPY

IV.1 Monitoring Mesenchymal Stem Cells expansion in xeno-free microcarrier-based reactor systems using FTIR spectroscopy

IV.1.1 Introduction

Stem cells are characterized by its self-renewal ability and the capability to give rise to at least one type of mature cells. Although mesenchymal stems cells (MSCs) do not fulfill these requirements, as they have limited *in vitro* proliferation, maybe related to culture and harvest conditions not yet optimized, they are often considered as truly stem cells (Le Blanc and Ringdén, 2005). MSCs can be found in the human adipose tissue, bone marrow, lung and umbilical cord, and are able to differentiate in different types of cells, such as osteoblasts, chondrocytes, adipocytes and stromal cells (Pittenger *et al*., 1999).

In the last years MSCs started to be seen as a very promising candidate to cell therapy, due to its ability to differentiate in several types of mature cells. Some clinical applications include delivery of anticancer agents (Dai *et al.* 2003), immunological regulators in patients with auto-immune diseases, treatment of the Graft-*versus*-Host (GVH) disease, preventing organ or cell rejection by patients who went through an organ or cell transplantation (Le Blanc and Ringdén, 2005; Friedman *et al*., 2007) and as a system for gene delivery system (Madeira *et al*., 2012).

Despite of the great potential of MSCs, they are considerably rare in human organism (0.001-0.01%), being the average of cells needed for therapy per patient $1-5\times 10^{6}\ cells/Kg$. Therefore, to make possible the therapeutic application of MSCs, the expansion of these cells represents a critical step. Furthermore, cells' expansion is dependent of several factors, like the age and condition of the donor, the isolation techniques chosen and the cells' source (Le Blanc and Ringdén, 2005). When it concerns to expand cells for future therapeutic applications, it is crucial to achieve a high cell number but also a reproducible process with the lower costs associated. For that, a rigorous real-time monitoring and control of the critical variables of the expansion processes is essential, as a small variation in the process can cause serious changes in the final product. Some of the variables in mammalian cells' cultures that are commonly monitored include pH, temperature, dissolved oxygen concentration (DOC) and key analytes such as glucose, glutamine lactate, ammonia or amino acids (Rhiel *et al.*, 2010). Sensors for on-line monitoring of temperature, pH and DOC are well established. However, the monitoring of the remaining critical variables of the bioprocess are usually based on time-consuming off-line analysis, such as enzymatic assays, high- performance liquid chromatography (HPLC) and immuno-assays (ELISA) (Harthun *et al.*, 1998).

Aiming at a rigorous monitoring of bioprocesses, the Food and drugs Administration (FDA) introduced the Process Analytical Technologies (PAT) Initiative with the goal of ensuring the quality of the final product, while achieving a high-knowledge of the process and preserving its reproducibility (Hakemeyer *et al.*, 2012). Fourier transform infrared (FTIR) spectroscopy, a technique that detects with a high sensitivity the vibrational modes of biomolecules, represents very promising candidate to achieve these goals. FTIR spectroscopy, combined with multivariate data analysis, allows the rapid quantification of several analytes from a single spectra, with no use of reagents, which make each measurement, besides highly sensitive and rapid, very economic. Additionally, it is a long stable method, which can perform high-throughput measurements, and high resolution chemical analysis even when the compounds of interest are present in very low concentrations (Scholz *et* al., 2012; Card *et al.*, 2008). A close follow up of the bioprocess using spectroscopic methods allows not only the quantification of key analytes, but also makes possible the detection of undesirable events, like cell death or contaminations. Therefore, FTIR spectroscopy meets the PAT's initiative goals and it has been used in several pharmaceutical applications (as reviewed by Roggo *et al.,* 2007) and for monitoring bacterial cultivations (as reviewed by Landgrebe *et al.*, 2007). The interest of using spectroscopic methods for monitoring mammalian cell's processes, however, started later and it is now on its early stages.

FTIR spectroscopy can be operated in the mid-infrared (MIR) and near-infrared (NIR) regions of the spectra. A MIR spectrum can be more informative than a NIR spectrum, since a MIR spectrum results from fundamental vibrations of molecules, so a higher sensibility it is expected for MIR spectroscopy. MIR spectroscopy also allows performing at-line high-throughput measurements. However, as water absorbs strongly in MIR region, it is necessary to use an ATR probe or to dehydrate the sample before spectral acquisition (Rhiel *et al*., 2010). On the other hand, NIR spectroscopy is preferable for in-situ analysis of bioprocesses, since this spectral region is less prone to water interference. However, NIR spectroscopy is not ideal for analyzing microencapsulated cells or cell grown in diluted media, due to its lower sensitivity, which can be a problem when monitoring mammalian cell's cultures, since one of the concerns is to maintain low levels of glucose and glutamine and to avoid accumulation of lactate and ammonia (Teixeira *et al*., 2009). For all these reasons the two techniques, *at-line* and *in-situ*, are complementary tools for bioprocess monitoring, with the goal of obtaining the largest amount of information about the bioprocess, for a correct and accurate process monitoring and optimization.

Few studies have been published on monitoring mammalian cell's cultures using IR spectroscopy. Harthun *et al*. (1998) used NIR spectroscopy to quantify a recombinant product, antithrombin III (rhATIII), as well as some key analytes such as glucose, lactate, glutamine, glutamate and ammonia, in a Chinese Hamster Ovary (CHO) culture. Partial Least Squares (PLS) models were built for predicting all components mentioned, with poor results since for the overall models a high number of latent variables were used, which have a risk of overfitting. Also Sandor *et al*. (2013) developed a comparative study using NIR and MIR-ATR spectroscopy for on-line monitoring of 8 CHO cell cultures, with different feeding regimes. PLS models were developed for several variables of the bioprocess, namely glucose, lactate, glutamine, glutamate, ammonia, cell viability and total cell concentration. The best PLS models were achieved using MIR-ATR spectroscopy especially for glucose, lactate, ammonia and cell viability, with a low number of latent variables (4/5) and high correlation coefficients between the real and predicted values. Card *et* al. (2008) had similar results using a HEK 293 cell line, proving that the technique is cell type independent. Hakemeyer *et al*. (2012) besides developing PLS models to quantify key analytes in a CHO cell culture, developed principal component analysis (PCA) models for monitoring bioprocesses with the purpose of ensuring a rigorous quality control. This is a very useful approach, as it allows detecting subtle undesirable events during the culture run in a rapid way, without quantifying any analytes. To date, no work was published using FTIR spectroscopy to monitor hMSCs expansion in microcarriers and only one study used FTIR spectroscopy for monitoring mammalian cells cultured in microcarriers. Petiot *et al.* (2012) used NIR spectroscopy for in-situ measurements of a Vero cell line (kidney epithelial cells extracted from an African green monkey) cultured in microcarriers, with the goal of predicting glucose and lactate concentrations.

The present work aims to apply MIR spectroscopy to monitor the expansion process of hMSCs, coming from living donors, cultured in different microcarriers and using different feeding regimes. MIR spectroscopy was preferred instead of NIR spectroscopy, in order not only to develop robust partial least squares (PLS) regression models for predicting key analytes as glucose, lactate and ammonia, but also aiming to acquire information concerning cellular events, as the follow-up of the cellular growth stages or the detection of undesirable events during the culture run.

All the experimental work associated with the expansion of hMSCs was conducted by Joana Carmelo, under the supervision of Professor Cláudia Lobato da Silva and Professor Joaquim Sampaio Cabral, from the Stem Cell BioEngineering and Regenerative Medicine Laboratory, *Instituto Superior Técnico, Universidade de Lisboa,* as described in Carmelo (2013).

IV.1.2 Materials and Methods

Samples

Human mesenchymal stem cells (hMSCs) derived from bone marrow samples were used in this study. BM samples were obtained from healthy donors after informed consent at *Instituto Português de Oncologia Francisco Gentil*, Lisbon, Portugal. hMSCs were recovered from cryopreservation and cultured in culture flasks with Iscove's Modified Dulbecco Medium (IMDM), xeno-free medium, supplemented with Penicillin (at a concentration of 0.025 U/mL) and Streptomycin (at a concentration of 0.025 µg/mL) (PenStrep, Gibco) and with GlutaMAX™-I CTS™ (Gibco) as a glutamine substitute. Cells were plated in T-flaks (BD Falcon™) at an initial cell density between 3000-6000 $cells/cm^2$ and incubated at 37°C, 5% CO_2, in a humidified atmosphere. The culture medium was renewed every 3 or 4 days. When cells reached about 70 or 80% confluence were detached with TrypLE™ Select (10X) (Gibco), and diluted in PBS. Cells were then seeded on a microcarrier culture and expanded on the microcarriers under dynamic conditions in spinner flask cultures, using Bellco® spinner flasks (Bellco Glass, Inc.) with a working volume of 80 mL, equipped with 90º paddles (normal paddles) and a magnetic stir bar.

For monitoring the consumption of nutrients and the production of metabolites, supernatant from the three culture runs were analyzed every day and glucose, lactate and ammonia concentrations were determined using an automatic analyzer YSI7100MBS (Yellow Springs Instruments). For cell counting the microcarrier cell cultures were harvested and incubated with TrypLE solution in the Thermomixter for 7 or 8 minutes at 37ºC and 750rpm. Then, IMDM with 10% of fetal bovine serum (FBS) in a proportion of 1:3 was added. The cells were separated from the microcarriers through filtration with a Cell Strainer (100µm) (BD Falcon™). Cell number and viability was determined by Trypan Blue exclusion method.

All steps of manipulation of the hMSCs were performed by Joana Carmelo, from IST, under her Master's Thesis, so for more detailed information on hMSCs culture's conditions see Carmelo (2013).

Three different hMSCs expansion processes were tested using FT-MIR spectroscopy:

Culture S - hMSCs derived from bone marrow cultured on plastic microcarriers (Solohill Engineering, Inc.) previously coated with CELLstart™CTS™ (Gibco). From the 3rd day on, 25% of the medium was renewed every day.

Culture A1 - hMSCs derived from bone marrow cultured on A microcarriers from the X Company. From the 3rd day on, 25% of the medium was renewed every day.

Culture A2 - hMSCs derived from bone marrow cultured on A microcarriers from the X Company. From the 3rd day on, 25% of the medium was renewed every 2 days.

For all the cultures, cells were culture in spinner flasks during 13 days and a medium sample was collected everyday (with an exception of day 2), with the samples being classified as *before* and *after*, if taken before or after the medium renewal, respectively. The conventional quantification of glucose, lactate and ammonia of these three cell cultures was performed, using an automatic analyzer YSI7100MBS (Yellow Springs Instruments). For detail information on the protocol see Carmelo (2013). At the time of the conventional measurements about 1 mL of the medium samples was preserved (-20°C), for further FTIR analysis.

Spectral acquisition

The supernatant samples, preserved at -80°C, were thawed at room temperature and then triplicates of 25 µL of each sample was transferred for a 96-wells KBr plate for the FT-MIR high-throughput measurements. Before spectral acquisition the samples were dehydrated for about 2 and a half hours in a desiccator under vacuum. The spectral data were collected using a FTIR spectrometer (Burker, HTS-XT) equipped with an HTS accessory. Forty scans, with a 4 cm^{-1} resolution, in transmission mode, were collected in the wavenumber region between 500 and 4000cm^{-1}. Each spectrum was baseline corrected with the OPUS software (Bruker, Germany) prior to data analysis.

Spectral data analysis

Data pre-processing, including multiplicative scatter correction (MSC), 1st and 2nd derivatives, and PCA and PLS regression models were carried out using Matlab R2012b (Matworks, Natick, MA, USA). Derivatives were computed using Savitzky-Golay algorithm, with a filter window of 15 data points and a 2nd order polynomial. Baseline correction were carried out using OPUS software (Bruker, Germany). The performance of the PLS models was accessed through the evaluation of root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV), root mean square error of prediction (RMSEP), the correlation coefficient (R^2) and the error as percentage of the concentration range (ER%), given by:

$$Error\ as\ \%\ of\ range = \frac{RMSEP}{range\ of\ concentration} \times 100\% \qquad \text{(Equation IV.1)}$$

The best pre-processing techniques were chosen based on the closeness of replicates in the PCA score plots and on lower RMSEC, RMSECV, RMSEP, ER% and higher R^2 in PLS models. The PLS regression vector and the percentage of variance explained *versus* the number of latent variables were also considered.

IV.1.3 Results and Discussion

Expansion of hMSCs requires not only the optimization of the culture's conditions but also the development of new methods that can offer real-time monitoring of the bioreactor. The main goals of the present work were to monitoring the cell's expansion and evaluate critical events concerning the consumption of key analytes, as glucose, the production of by-products that can be harmful to the cells (e.g., lactate or ammonia), while evaluating the reproducibility of the process in function of input perturbation, such as different microcarriers or feeding regimes. Furthermore, hMSCs culture conditions are not well optimized yet and the process of expansion and differentiation of these cells is not truly understood, FTIR spectroscopy can be a very useful tool to increase the knowledge about the bioprocess itself. For these purposes, three different expansion cultures, using different microcarriers and different feeding regimes, were analyzed. Several distinct analyses were performed based on the spectral data acquired for the medium samples of the three independent cultures. First, PCA models were able to find a relationship between spectral data and cellular events, like the cell's growth stages or toxic effects due to excessive concentrations of products such as ammonia and lactate, and also provided increased bioprocess knowledge. On the other hand, PLS regression models allowed the quantification of three principal components present in the media that are essential for cellular growth: glucose, one of the main energy sources to the cells, lactate and ammonia, metabolic waste by-products that, in excessive concentrations, may inhibit cell's growth, and therefore must be rigorously controlled (Rodrigues *et al.*, 2011)

Conventional analysis for metabolite determination

In order to monitor the consumption of nutrients and the production of metabolites, supernatant from the three culture runs were analyzed every day and glucose, lactate and ammonia concentrations were determined using an automatic analyzer YSI7100MBS (Yellow Springs Instruments) (Figure IV.1). In general, for the cultures in which the medium was renewed every day, i.e., cultures S and A1, (Figure IV.1) exhibited similar glucose, lactate and ammonia profiles, where slightly higher ammonia levels were reached for the culture A1 (Figure IV.1). On day 9, an ammonia concentration of 2.5 mM was reached, which can be considered as growth inhibitory. For culture A2 (Figure IV.1) the concentrations of the three metabolites showed slightly different profiles. As expected, by the less frequent medium renewal, culture A2 experiment reached lower concentrations for glucose and higher concentrations for lactate and ammonia.

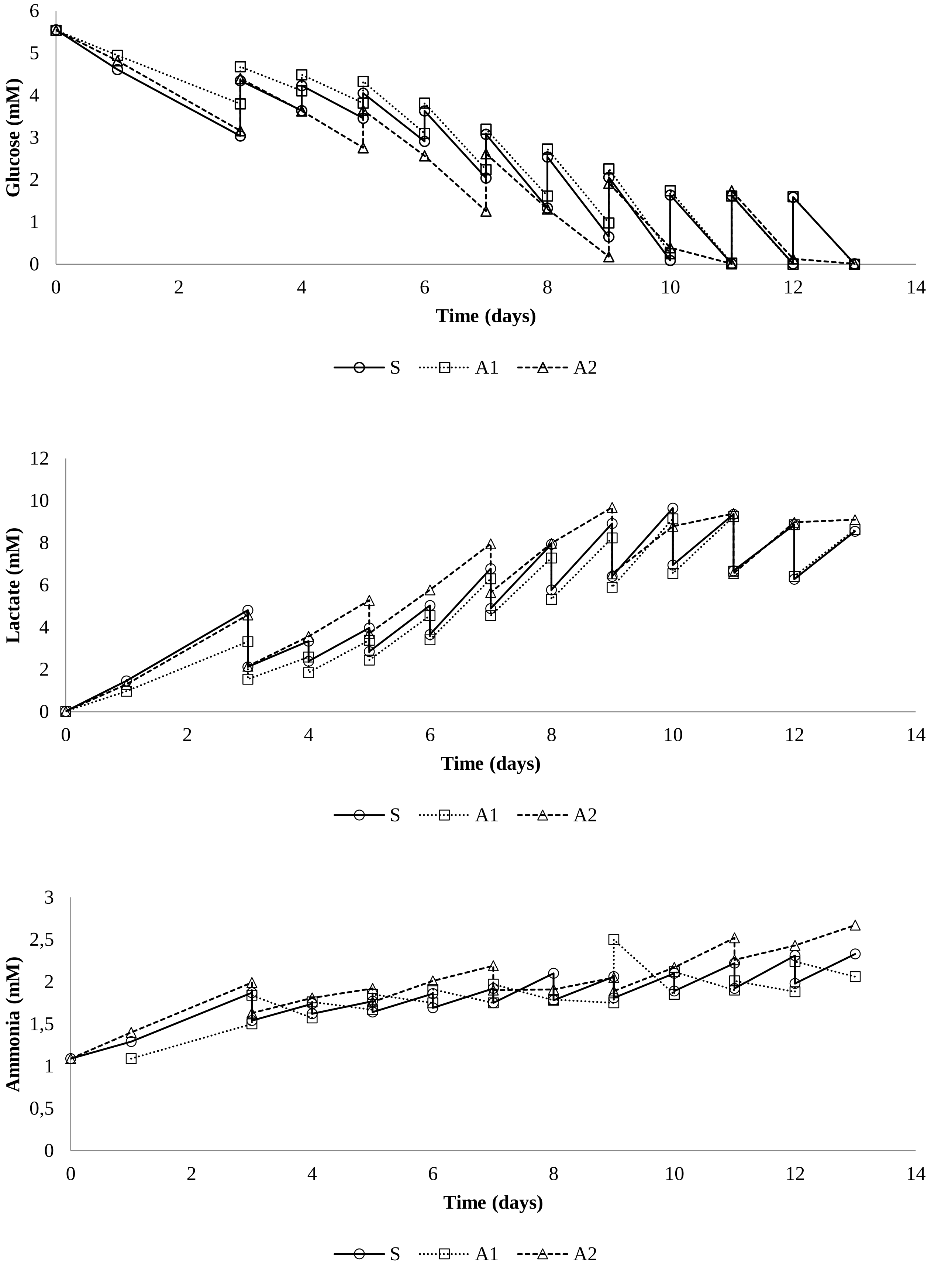

Figure IV. 1 - Concentrations profiles for glucose, lactate and ammonia during hMSCs expansion in Solohill microcarriers (culture S), A microcarriers with the media renewed every day (culture A1) and A microcarriers with the media renewed every 2 days (culture A2).

Conventional analysis for the evaluation of hMSCs proliferation

The three independent cultures presented a similar lag growth phase, which last until day 5 of the culture run, and a similar exponential growth phase, since day 5 to day 8 (Figure IV.2). Cultures S and A1, with a daily medium renewal, showed more similar profiles and by the end of day 13 the same cell density was reached, i.e. 2.5×10^5 cells/mL. However, a slower expansion was observed for the culture A1, even with a similar final cell density (Figure IV.2). For culture A2, the cell density along the culture and at the end of the culture was lower, which may be due to the accumulation of by-products or lower levels of glucose. Also, since day 9, cell-microcarriers aggregates could be seen and becoming larger with time, even if no cell detachment from the microcarriers was observed. However, the cell-microcarrier aggregates can difficult the assessment of cell viability and the sampling process (For more detailed information about the cultures see Carmelo (2013)).

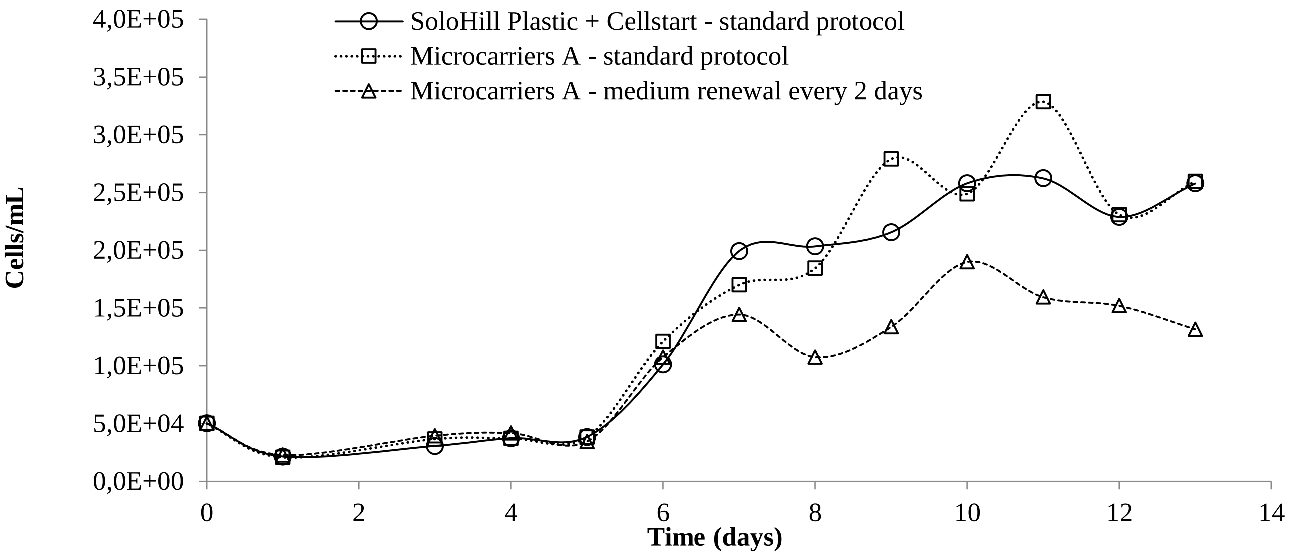

Figure IV. 2 – Cells' concentration for hMSC cultured on Solohill microcarriers (culture S), A microcarriers with the medium renewed every day (culture A1) and every 2 days (culture A2).

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a data reduction method often used in spectral data analysis that decompose the spectral data into new variables, called principal components (PCs), that capture the most variance in data (Jollife, 2002). PCA models were developed with the goal of finding meaningful relationships between the spectral data and cellular events, such as different phases of the cell’s cycle, and to evaluate differences due to the medium composition After building an appropriate and robust PCA model able to monitoring hMSCs cultures, it will be possible to apply such model for new coming cultures and evaluate if the culture conditions are the desired ones for hMSCs expansion and, more important, if none of the key metabolites under study have reached critical values. Moreover, PCA provides increased understanding about the culture itself and allows detecting undesirable events during the culture run, such as contaminations, only currently detected at the end of the culture run with conventional methods.

Different pre-processing techniques were applied to the spectral data of each culture before developing the PCA models. The best PCA results were obtained using the data pre-processed with baseline correction, MSC, 1st and 2nd derivatives, allowing the replicates to be aggregated and showing tendencies in data that were not visible before pre-processing.

Figure IV.3 shows the PCA score plot of the A1 experiment, with PC1 and PC2 representing 94.9% and 1.9% of the variance explained in the data, respectively, after applying baseline correction and MSC to the data. Grouping of the samples according to the cellular stage (Figure IV.4) can be clearly observed: A first group consisting of samples from the early culture days (from day 1 to day 3), representing a period of adaptation of the cells to the culture conditions - the lag phase of cell’s growth (according to Figure IV.4); a second group consisting of samples reflecting the exponential phase of cellular growth (from day 3 to day 8), where an exponential proliferation happens (according to Figure IV.4); a third group, consisting of samples representing the plateau/decline phase (from day 8 to day 13), where there is no increase in the cell number and cellular death start to occur (according to Figure IV.4).

Interestingly, samples from day 8 seem to be isolated from the other groups. The reason why this happen is not clear and more experiments shall be carried out to explain this behavior.

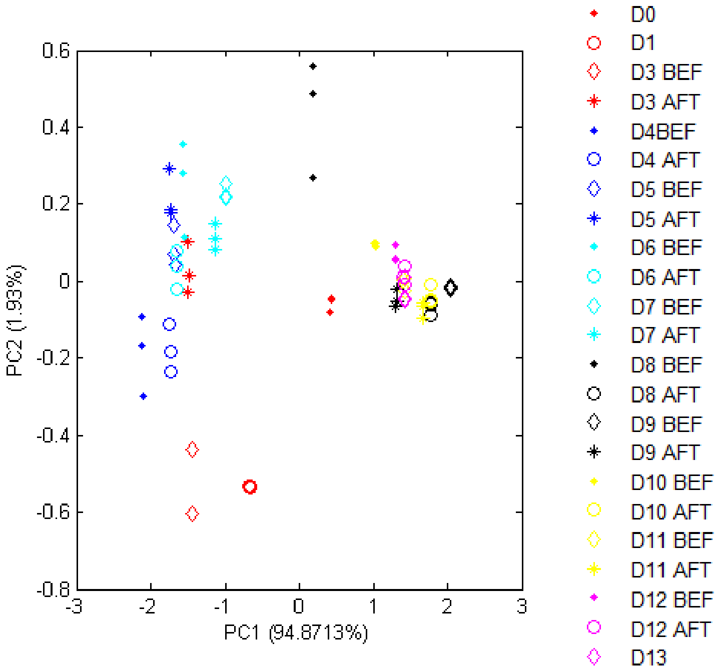

Figure IV. 3 - PCA analysis (after applying MSC to the replicates) of the hMSCs cultured microcarriers A from Company X, with a medium renewal every day (Culture A1). Three groups can be observed: samples from day 1 (D1) and day 3 (D3) before the medium renewal (bef); samples from day 3 (D3) to day 7 (D7) after the medium renewal (aft); samples from day 8 after the medium renewal (D8 aft) to day 13 (D13)

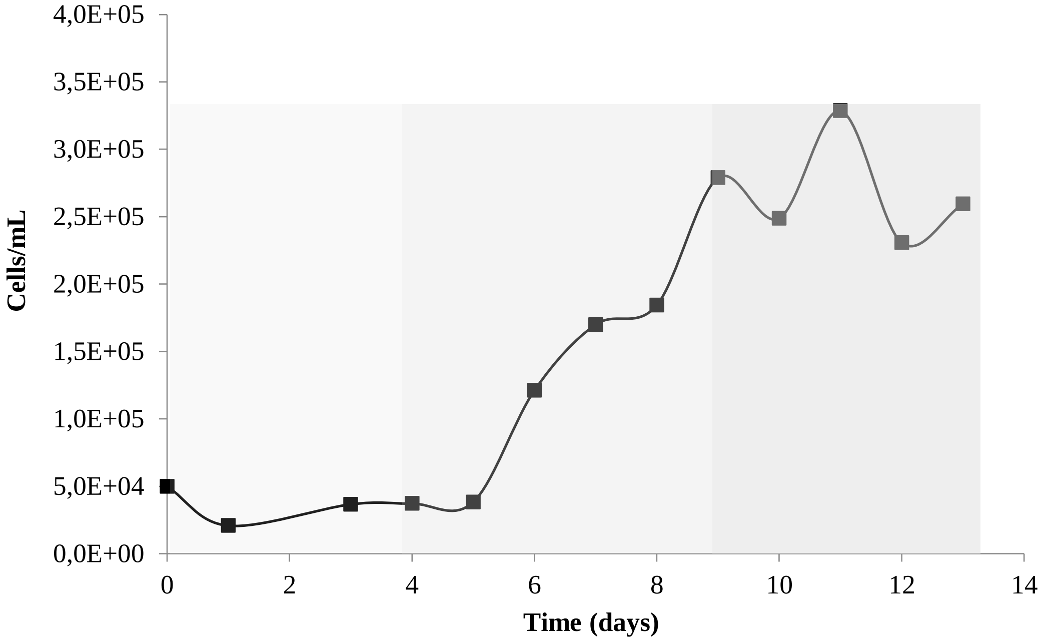

Figure IV. 4 – Cells' concentration for hMSC cultured on A microcarriers, with the medium renewed every day (Culture A1).

The same approach was conducted for culture A2 (Figure IV.5). In the PCA scores plot, with PC1 and PC3 representing 70.9% and 5.9% of the variance explained in the in data, respectively, the grouping of samples was not as clear as for culture A1, however there was a separation of the samples before day 7 from the samples after day 7. According to the data of the cellular growth, cells stopped growing after day 7 (Figure IV.6), which may explain the separation of the samples in the PCA score plot.

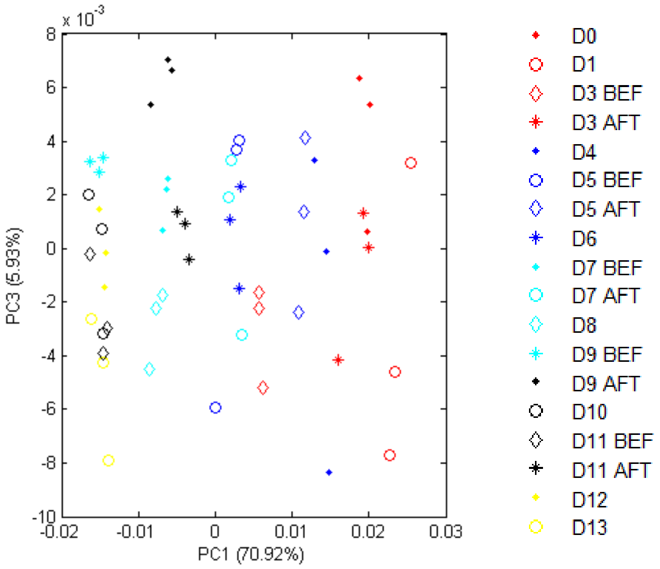

Figure IV. 5 - PCA analysis (after applying MSC to the replicates and first derivative) of the hMSCs cultured microcarriers A from Company X, with the medium renewed every 2 days (culture A2). BEF – before the media renewal; AFT – After the media renewal; D – day; the numbers correspond to the day of the culture run, from day 1 to day 13.

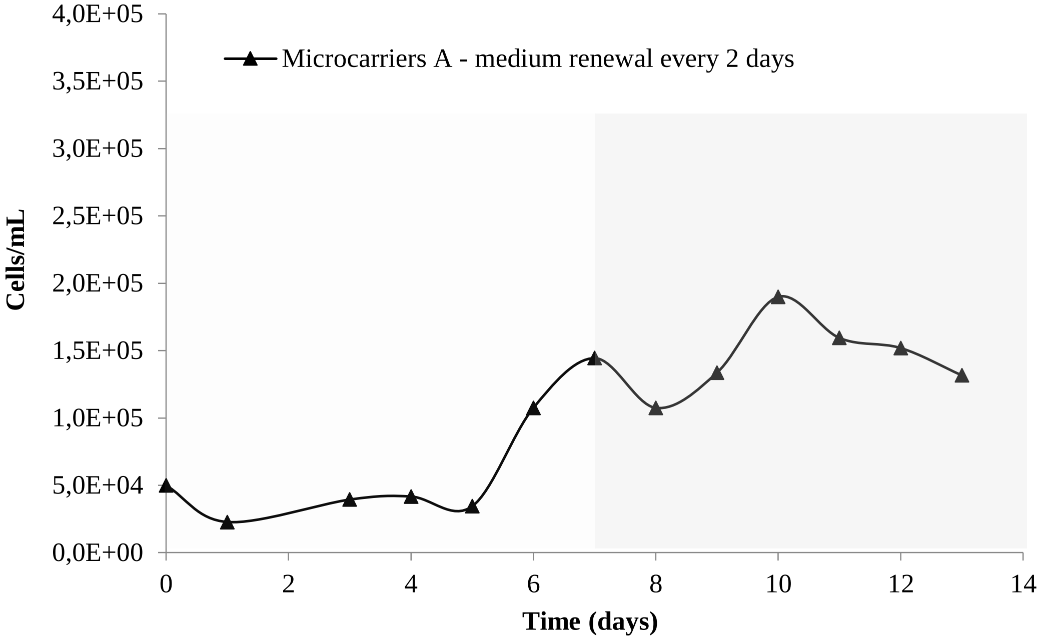

Figure IV. 6 – Cells' concentration for hMSC cultured on A microcarriers, with the medium renewed every 2 days (Culture A2).

Interestingly, it was only possible to visualize grouping of the samples according to the cellular growth stage in the PCA score plots for the culture where hMSCs were cultured in microcarriers A from the X Company (cultures A1 and A2). For the hMSCs cultured on the Solohill microcarriers (culture S) there was no grouping according to the cellular stage, even after trying different pre-processing techniques (Figure IV.7).

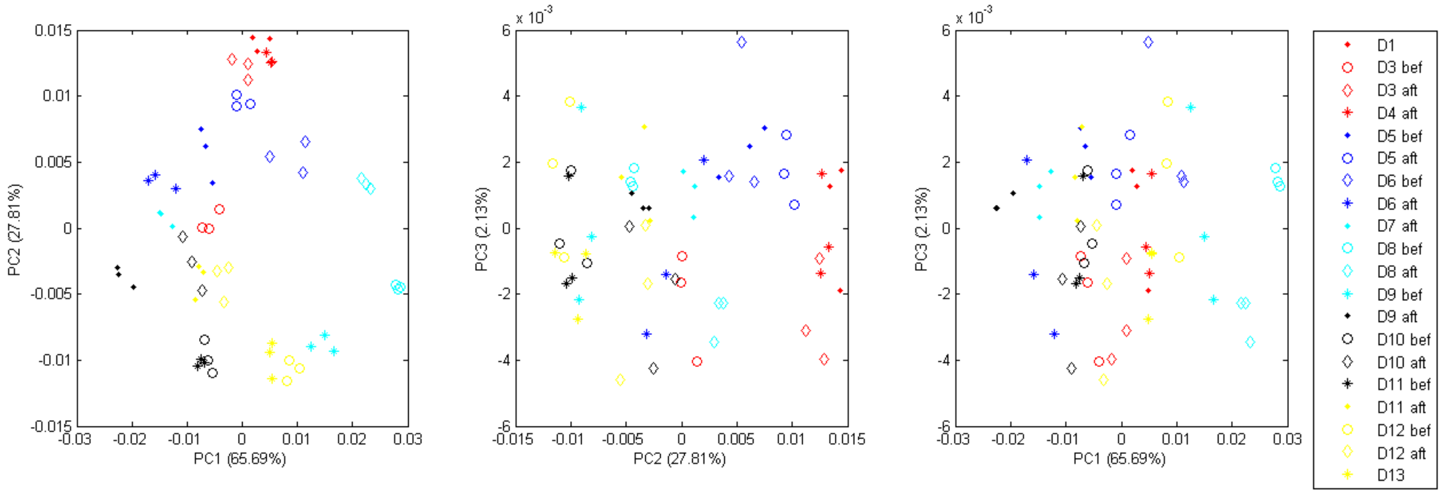

Figure IV. 7 - PCA analysis (after applying MSC to the replicates and first derivative) of the hMSCs cultured microcarriers Solohill microcarriers (Culture S), with the medium renewed every day (culture S). BEF – before the media renewal; AFT – After the media renewal; D – day; the numbers correspond to the day of the culture run, from day 1 to day 13.

Sandor *et al*. (2013) did a similar approach, they used MIR and NIR spectral data acquired along several CHO culture runs, for the development of PCA models. The PCA score plot obtained by these authors showed the same progression of samples along all the culture runs. They suggested that the trajectory in the 1st PC was related to the cell growth, while the 2nd PC was related with lactate, glutamate and ammonia concentrations. In spite of grouping of the scores according to the cell grown, the above trajectories were not observed in this work, probably due to the fact that a different cell type was used. Also, on the contrary of the work conducted by Sandor *et al*. (2013), hMSCs were grown in microcarriers.

The PCA scores plots for the cultures studied provided valuable information, indicating that FTIR spectroscopy could be used to monitoring hMSCs expansion. More culture runs should be followed using FTIR spectroscopy, not only to retain as much information as possible through this technology, but also to develop robust models than can later predict the behavior of a given culture, thus providing a valuable help in the optimization of the conditions of bioreactors.

Partial Least Squares (PLS) regression models

For the development of the PLS models two distinct approaches were carried out, using a leaving-one-out (LOO) cross-validation (CV) methodology and using a calibration and a test set validation approach. PLS models based on LOO cross-validation were developed for each culture independently, since there were fewer samples available (up to 68 for each culture, including the three replicates for each sample). For the construction of the PLS models using all the 176 samples (including the replicates), from the three cultivations, samples were divided in two sets: a calibration set for the model development, 120 samples randomly chosen from the three cultures, and a test set used for external validation, consisting on 56 samples randomly chosen from the three culture runs. The models were built for the three components under study: glucose, lactate and ammonia.

A very important factor to have into consideration when building a PLS model, especially when dealing with mammalian cell cultures that are very complex, is that as much variance as possible should be introduced during the model construction, i.e., the model must be based on experiments with different culture conditions (Petiot *et al*., 2010):

different feeding regimes and different types of microcarriers can improve PLS model performance.

To enhance the predictive ability of the PLS models, different pre-processing techniques were applied to the spectral data, namely MSC and derivatives. MSC was applied to the data in order to eliminate the effect of physical phenomena like the light scattering effect of particles of different sizes and shapes (Helland *et al*., 1995), highly relevant in case of dehydrated films. During the dehydration process an irregular surface may be formed, which introduces a certain degree of error when we are analyzing replicates, due to effects of light scattering. The main goal of applying derivatives to the spectral data is to enhance the information in data, while eliminating physical interferences that can compromise the relationship between the data and the biological samples.

The number of latent variables (LVs) must be carefully chosen when building a PLS model, as too many variables can lead to overfitting and a very specific model, which means that the prediction ability outside these conditions is reduced. On the other hand, a very low number of latent variables can lead to underfitting. The ideal is to select the number of new variables that allows covering the complexity of the data, while avoiding overfitting. The number of latent variables always depends on the model, but in general models with a lower number of LVs are more robust. However, they may fail if not enough variance is covered in the construction of the model (Haaland *et al*., 1988; Teixeira *et al*., 2009).

The best models, with an optimum number of LVs, and the highest predictive ability were chosen based on: the highest R^2, the lowest RMSEP and/or RMSECV and a high percentage of variance explained in data (~ 98%). Based on the above measures of performance, the best models were obtained using the data prior pre-processed with MSC and first derivative (Table IV.1).

Table IV. 1 – PLS regression models, with data pre-processed differently, for glucose prediction of hMSCs cultured on Solohill microcarriers with the media renewed every day (Culture S). MSC – multiplicative scatter correction

	Without pre-processing		MSC		MSC + 1st derivative		MSC + 2nd derivative	
LV	R^2	RMSECV	R^2	RMSECV	R^2	RMSECV	R^2	RMSECV
2	0,79	0,70	0,81	0,66	0,95	0,34	0,95	0,34
3	0,92	0,43	0,94	0,37	0,97	0,26	0,96	0,31
4	0,944	0,36	0,95	0,32	0,98	0,22	0,98	0,23
5	0,96	0,32	0,96	0,32	0,99	0,18	0,98	0,19
6	0,96	0,29	0,97	0,25	0,99	0,18	0,99	0,18
7	0,97	0,28	0,98	0,23	0,99	0,17	0,99	0,15

The best PLS models obtained for culture S (Table IV.2), using LOO CV, were based on 4 LVs for glucose and lactate and 5 LVs for ammonia prediction. The PLS model for glucose explained about 98% of the variance in data, with a R^2 of 0.98 and a RMSECV of 0.22 (mM), corresponding to an error as percentage of range of 4.7%. About 98% of the variance in the data was explained by the lactate model, with a R^2 of 0.98, a RMSECV of 0.36 (mM) (error as percentage of the range of 4.4%). The RMSECV obtained was higher for the lactate than for the glucose model, with an error as percentage of the range similar to the glucose model, given the concentration range of the lactate was higher. For ammonia the best model yielded a R^2 of 0.95, a RMSECV of 0.06 (error as percentage of the range of 5.7%) and explaining about 95% of the variance in data.

Very similar models for glucose, lactate and ammonia were obtained for culture A1, where the hMSCs were cultured on A microcarriers from the X Company (Table IV.2).

The PLS models developed for the culture A2, where hMSCs were cultured in A microcarriers from X Company with a medium renewal every 2 days (Table IV.2), provided the best results for glucose and lactate predictions, using less LVs than the previous models. The lower error as percentage of the concentration range achieved for glucose and lactate are the result of the lower RMSECV and also the larger concentration range, since the medium was changed every 2 days so these cellular products achieved higher concentrations when comparing with the first two cultures, where the medium was renewed every day since day 3. A slight decrease in the predictive ability for ammonia was observed, however, yielding a R^2 0.94 of and a RMSECV of 0.09 (representing 5.7% of the concentration range).

Table IV. 2 – PLS models for glucose, lactate and ammonia prediction for hMSCs cultured on Solohill microcarriers (culture S), A microcarriers with daily media renewal (culture A1) and hMSCs cultured on A microcarriers with the medium renewed every 2 days (culture A2).

Culture	Analytes	LV	R²	RMSECV (mM)	Concentration range (mM)	Error as % of the concentration range
S	Glucose	4	0.98	0.22	0.01 – 4.61	4.7
	Lactate	4	0.98	0.36	1.39 – 9.65	4.4
	Ammonia	5	0.95	0.06	1.29 – 2.33	5.7
A1	Glucose	4	0.98	0.23	0.00 – 5.54	4.2
	Lactate	5	0.98	0.39	0.02 – 9.24	4.3
	Ammonia	5	0.94	0.06	1.09 – 2.27	5.4
A2	Glucose	3	0.99	0.17	0.01 – 5.54	3.16
	Lactate	3	0.99	0.25	0.02 – 9.68	2.54
	Ammonia	4	0.94	0.09	1.09 – 2.67	5.68

In general, the ammonia models yielded higher errors, compared to glucose and lactate predictions, probably because the range of concentrations tested was lower and because it was present in the samples in a lower concentration that the other analytes evaluated.

Similar results were obtained by Card *et al*. (2008) for glucose, lactate and ammonia prediction for HEK cells cultured in bioreactors, using on-line NIR spectroscopy for the development of PLS models using selected spectral regions. Also Harthburn *et al*. (1998) performed on-line NIR monitoring of a CHO cell line and, regarding glucose, lactate and ammonia concentrations, using the entire spectral window. Poor results were obtained by these authors, compared to our study, especially for ammonia for which a correlation coefficient between measured and predicted values of 0.76 was obtained. Petiot *et al*. (2010) used on-line NIR spectroscopy for the quantification of glucose and lactate for Vero cells grown in microcarriers. The results were significantly worse than the ones obtained here and by other authors, namely a correlation coefficient of 0.86 and 0.88 for glucose and lactate, respectively. Actually, Petiot *et al*. observed that the presence of the microcarriers caused an increase of intensity of the spectra, due to an increase of the pathlength. It is probably necessary to optimize the culture conditions when using NIR spectroscopy for on-line measurements, such as the size and concentration of the microcarriers, or alternatively, to consider if MIR at-line analysis as the best solution for these cases. Sandor *et al*. (2013) did a comparative study between NIR and MIR spectroscopy for monitoring mammalian cell cultures. In generally NIR spectra performed better for the development of PCA models and MIR spectra yielded more accurate PLS models. It would be interesting to see more comparative studies between the two techniques, essentially as they can be complementary.

One factor that must be taken into account when evaluating PLS models for different parameters while studying a cell culture, is that some of these parameters are highly correlated with each other (for instance lactate is a product of glucose consumption, as mentioned earlier). The different feeding conditions applied to the culture runs can be very useful to break these correlations between the compounds under study (Sandor *et al*., 2013). Also some useful information can be obtained through the regression vector, as it provides information about the spectral regions that were used for the model construction, allowing to ensure that one component is not being predicted based on others.

The regression vectors obtained for glucose and lactate (Figures IV.8 to IV.10) are more similar as expected, since the two have similar molecular composition, so they absorb in similar regions and those regions will contribute more for the PLS regression model. Nevertheless, within those regions, both analytes show distinct intensities which, along the different feeding regimens, allows discarding the hypotheses of predicting one component from the other

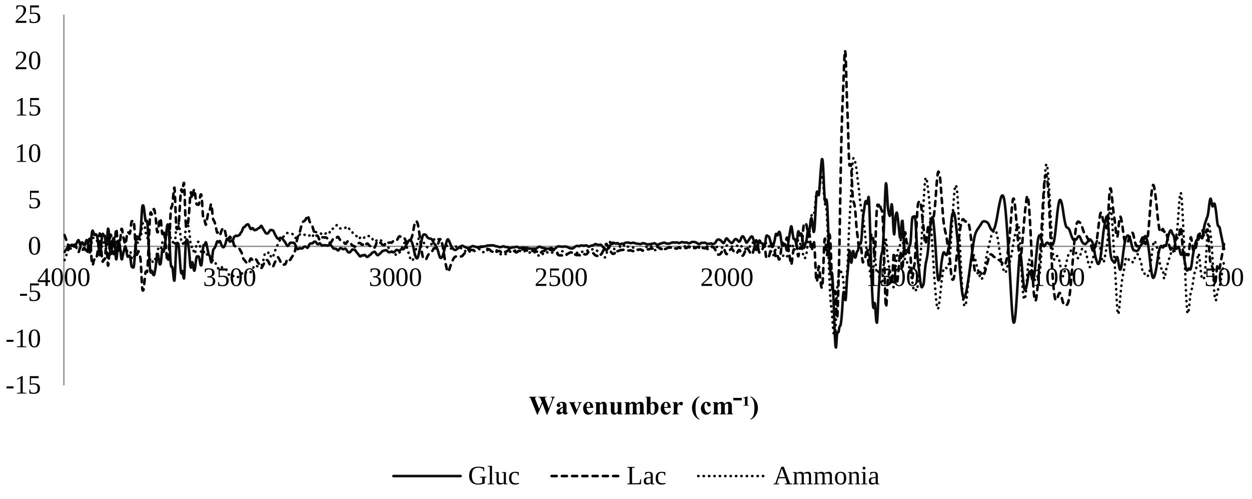

Figure IV. 8 - Regression vector from the best PLS models build for glucose, lactate and ammonia, with medium samples from hMSCs cultured in Plastic Solohill microcarriers coated with CELLstart (Culture S).

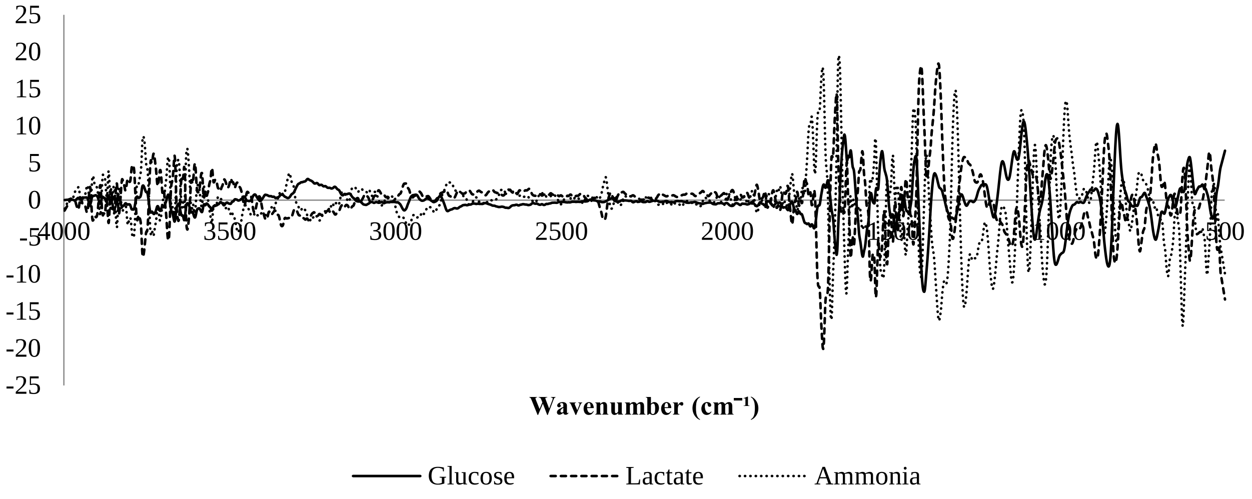

Figure IV. 9 - Regression vector from the best PLS models build for glucose, lactate and ammonia, with medium samples from hMSCs cultured in A microcarriers from X Company, with a daily medium renewal (Culture A1).

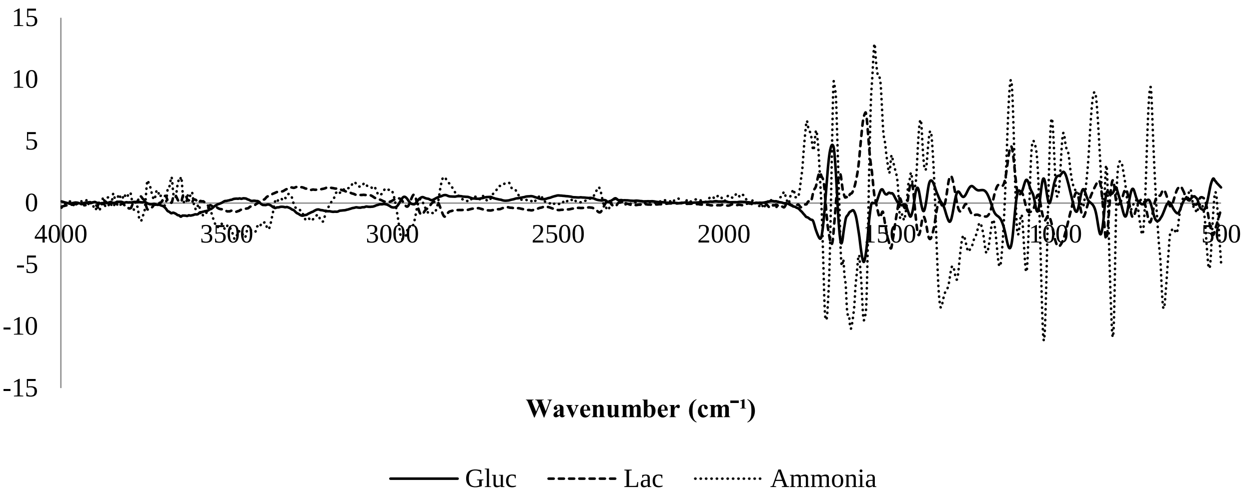

Figure IV. 10 - Regression vector from the best PLS models build for glucose, lactate and ammonia, with medium samples from hMSCs cultured in A microcarriers from X Company, with a medium renewal every 2 days (Culture A2).

An additional PLS model was developed using all samples available from the three independent cultures (176 samples with replicates), divided in a calibration and a test set. All the cultures were performed in different conditions, as concerning the microcarriers where the cells were expanded (Solohill or X microcarriers) and the frequency of the medium renewal (every day or every two days). When choosing the calibration and the test sets two important "rules" were followed: First, the calibration and test sets consisted of samples randomly chosen from the three cultures with the range of the variables measured for the validation samples being represented in the calibration set; Second the calibration set must contain about 2/3 of the available samples and the test set about 1/3. The calibration set contained 120 samples and the test set 56 samples.

When comparing the PLS models using all samples available with the models developed using only one culture, a slightly decrease in the predictive ability was observed, as expected, since the three cultures were performed in different conditions and the number of the total samples available was considerably low. Nevertheless, the models for glucose and lactate estimation showed very good prediction ability.

A glucose model based on 5 LVs was chosen (Figure IV.11), as for 6 LVs an increase of the RMSEC was observed, which may indicate overfitting. A RMSEC of 0.33

and a RMSEP of 0.32, with a R^2 of calibration of 0.95 and a $R^2$of validation of 0.97 were obtained. Compared to the previous models, a considerable increase in the RMSEC (about 58% more than the mean of the previous RMSECV) and a RMSEP slightly higher than the RMSECV obtained earlier. The RMSEP obtained correspond to an error of 5.7% as percentage of the concentration range. The correlation coefficients obtained also indicate a god fit.

For the lactate prediction (Figure IV.12), 5 LVs produced an increase of about 69.5% for RMSEC, compared to the mean RMSECV obtained for the previous models. Again the RMSEP here is also higher than the mean RMSECV previously obtained, representing an error of 5.5% as percentage of the concentration range which is again not very high. Also a correlation coefficient of 0.97 for the validation samples indicates a good predictive performance of the model.

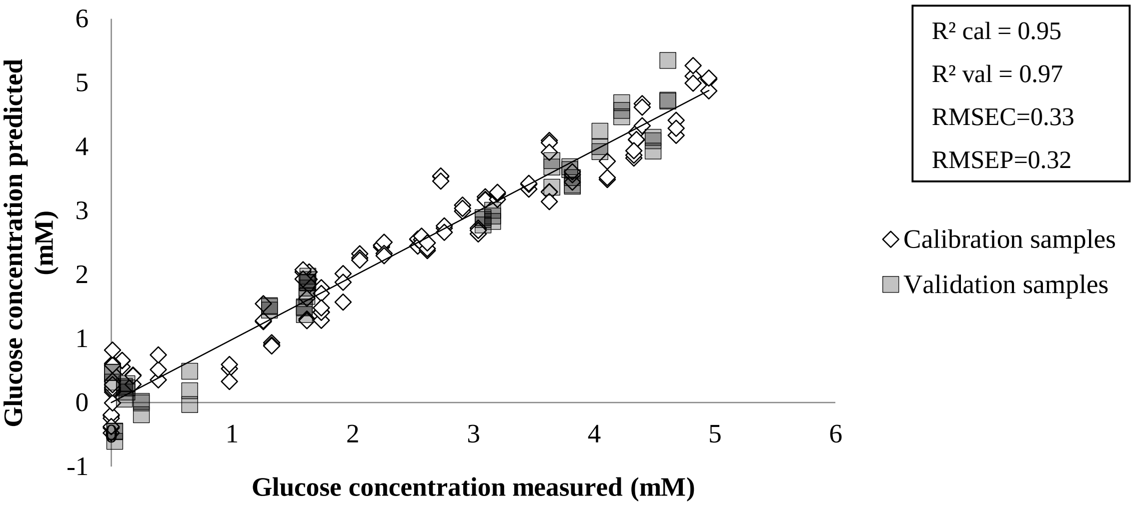

Figure IV. 11 - Glucose concentration measured and predicted (5 LVs), using the entire spectral region and the spectral data from the three independent cultures, divided in a calibration and test sets.

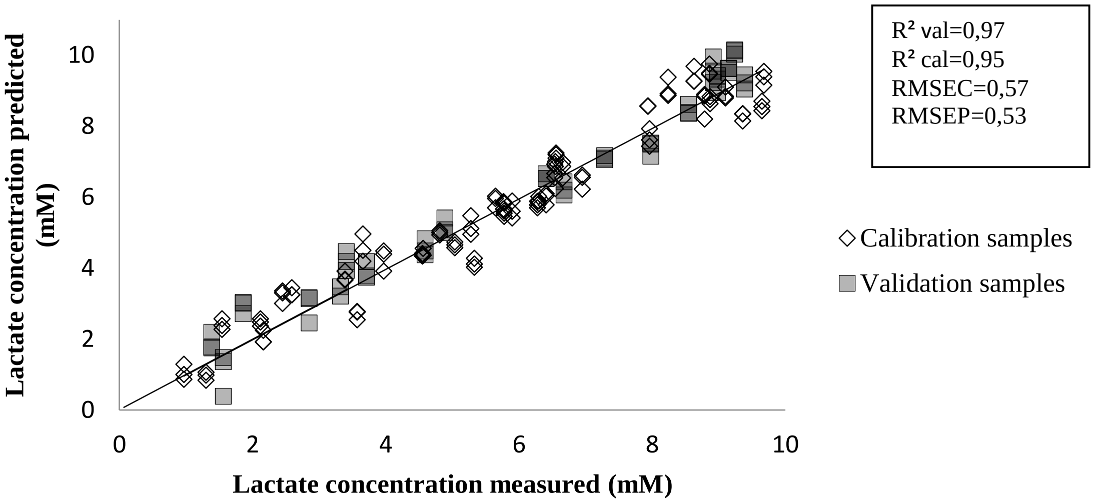

Figure IV. 12 - Lactate concentration measured and predicted (5 LVs), using the entire spectral regions and the spectral data from the three independent cultures, divided in a calibration and test sets.

The PLS model obtained for ammonia, using all samples, divided in two sets as described earlier, yielded considerably worse predictions when compared to the PLS models for each individually cultures. As an attempt to improve the model's predictive performance, two models with selected spectral regions were built: First, using the spectral region between 500 and 1900 cm^{-1}, as by the regression vector this region showed a high contribution for model development (darker gray - Figure IV.13); second, eliminating the spectral regions 3500-4000cm^{-1} and 1872-2700cm^{-1}, showing to have a lower contribution for the model development (lighter gray - Figure IV.13).

The model using the spectral region between 500 and 1900 cm^{-1} showed no improvements, which may indicate that even showing a lower contribution to the model, these regions still provide meaningful information. For the second model, for which two spectral regions were eliminated, the prediction results slightly improved. For the ammonia prediction (Figure IV.14), 5 LVs were selected and a RMSEC of 0.09 (mM) was obtained, representing an increase of about 31% when comparing with the mean of the results obtained for the previous models, and a RMSEP of 0.14 (mM), representing an error of 11.6% as percentage of the concentration range, twice the errors previously obtained. The correlation coefficient also decreased, with a R^2 of calibration of 0.86 and a R^2 of validation of 0.79. Again, ammonia is present in very low concentrations and also the concentration range is very limited, which can explain the poorer predictions compared to the other analytes.

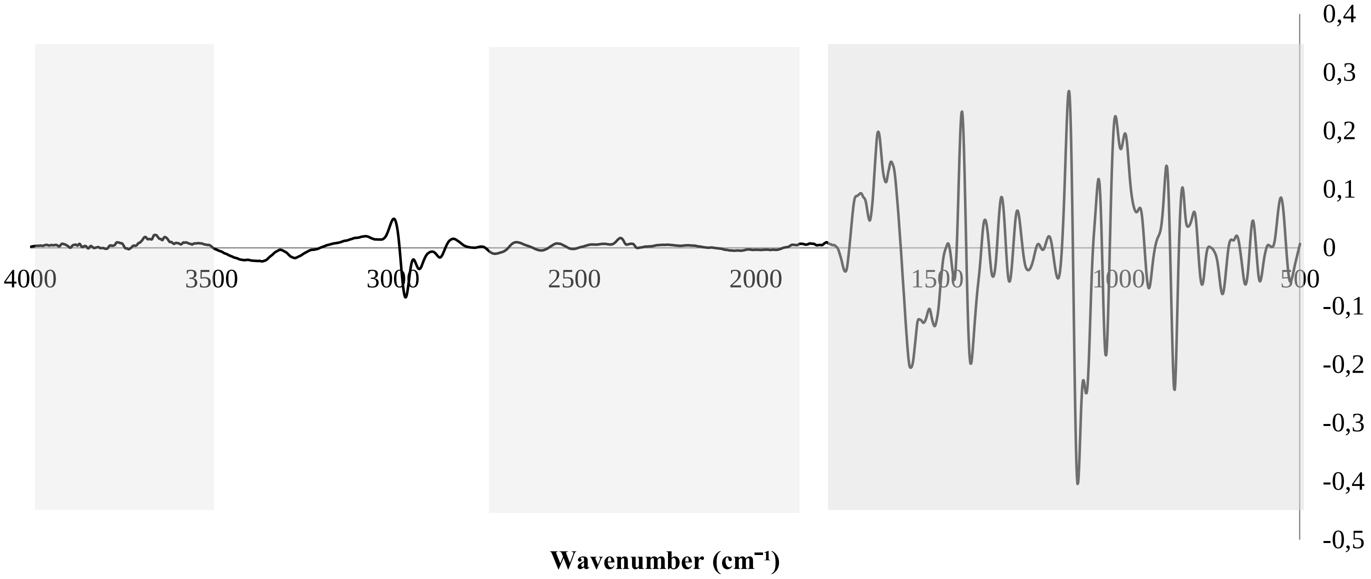

Figure IV. 13 – Regression vector from the ammonia PLS model (5LVs), using the entire spectral region and the spectral data from the three cultures, divided in a calibration and a test sets.

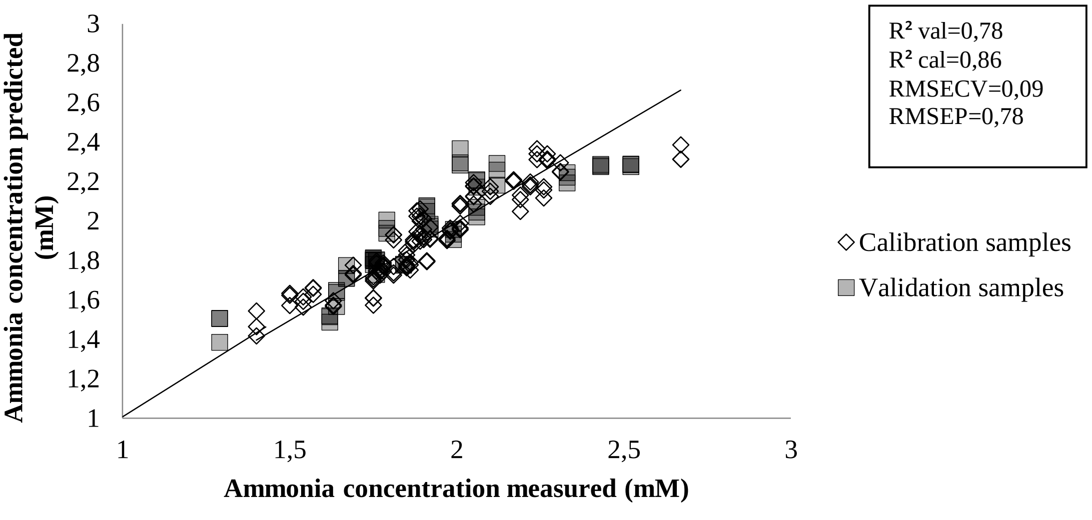

Figure IV. 14 – Ammonia concentration measured and predicted (5 LVs), eliminating selected spectral regions and using the spectral data from the three independent cultures, divided in a calibration and test sets.

IV.1.4 Conclusions

Human mesenchymal stem cells have received considerable attention in the past years, as they are very promising candidates for cellular therapy, due to its high differentiation and immunomodulatory abilities. Being present in the human body at a very low number, its expansion in bioreactors is still a critical step. It is necessary not only to obtain the high number of hMSCs needed for therapy, but at the same time to ensure safety and reproducibility of the process.

Parameters that are normally monitored on-line during bioprocesses include pH, dissolved oxygen concentration and dissolved carbon dioxide concentration. Glucose, lactate and ammonia are also currently monitored, however, through off-line methods. Normally HPLC or an automatic analyzer as the one applied here for the conventional analysis are used. These conventional methods are usually time-consuming, do not allow to quantify all the desired components, require reagents or standard solutions and a high sample’s volume is usually necessary. On the other hand, FTIR spectroscopy allows to rapidly evaluate several components at a time, from a single spectrum, the sample’s volume necessary is very low (25µL or lower) and with no need of reagents.

The potential of MIR spectroscopy to monitoring hMSCs’ expansion in bioreactors was studied in the present work, testing three independent hMSCs experiments, where the cells were grown in microcarriers. Qualitative and quantitative analysis were carried out, PCA and PLS regression models, respectively, in order to characterize the cultures. While PCA allowed to understand the processes, media composition and the cell’s growth, PLS allowed estimating glucose, lactate and ammonia concentrations.

Through PCA models were possible to observe the grouping of the samples according to the cellular growth. This approach can be extremely useful when analyzing hMSCs expansion. An interesting experiment would be contaminating purposely the cells during the culture run and trying to detect that contamination through PCA analysis. Also, very accurate results were obtained through the PLS models development, concerning all variables, specially glucose and lactate. Were achieved correlation coefficients between measured and predicted values of 0.99 for both glucose and lactate. For ammonia the results were slightly poorest, probably as the concentrations of ammonia in the media were very low and also the concentration range was much reduced, when comparing with the other metabolites studied.

In the present work FTIR spectroscopy, combined with multivariate data analysis, proved to be an ideal tool to monitoring cell bioreactors, since it allows increasing the knowledge of the process itself in a rapid and simple way, and to estimate the concentration of several parameters in one step, instead of performing a high number of expensive and time-consuming analyses. In that way, the parameters can also be optimized to ensure the quality of the process and that the necessary cell's number is achieved.

IV.2 Estimating the transfection efficiency on a cell population by FTIR spectroscopy without the need for reporter genes

IV.2.1 Introduction

Reporter gene is a common term for describing genes with measurable characteristics, which allow them to be easily distinguished from other endogenous proteins (Naylor, 1999). Reporter genes are routinely used in every molecular and cellular biology laboratory for numerous applications, such as cytotoxic assays (Parekh et al., 2012), drug discovery (Stadel et al., 1997), for studying gene expression as the roll of particular promoters (Jeyaseelan et al., 2001), for monitoring transcriptional activities (Yang et al., 1997) or the production of recombinant proteins (Durocher et al., 2001), understanding cell communication, cellular development, regulation of cell's growth and proliferation, protein-protein interactions, protein sub-cellular location or even studying transfection events (Jiang, 2008).

Examples of reporter genes commonly used include chloramphenicol acetyltransferase (CAT), alkaline phosphatase (AP), β-galactosidase (β-gal), green fluorescent protein (GFP) and luciferase. The choice of the ideal reporter gene is motivated by the type of the cell used, as it is important to ensure that no endogenous activity exists, and the adaptability of the experiment to the detection assays.

CAT is a bacterial enzyme that catalyzes the transference of acetyl groups from acetyl CoA to chloramphenicol, prior labeled with a radioisotope. In a tin liquid chromatography (TLC) the acetylated forms of chloramphenicol will migrate faster, as compared with the non-acetylated forms. The amount of acetylation is proportional to the expression of the CAT reporter gene. CAT has the great advantage of not detecting endogenous activity when dealing with mammalian cells, however the assay involves cells disruption and the need for radioisotopes, which limits its *in vivo* application (Jiang *et al*., 2008). Furthermore it was observed by Zhang *et al.* (2013) a decrease in gene expression, when using CAT for promoter activity qualification and quantification analysis in HepG2 transfected cells, which suggests that this reporter gene has a silencer activity. Two explanations for CAT silencer activity were suggested by the authors. Promoters or regulatory elements, both crucial elements for gene expression, exist in the context of the overall gene structure and chromatin conformation. Adding extra nucleotide sequences without specific structural elements can compromise this harmony, the promoter regulation, and, consequently, change the gene expression. Another hypotheses, according to the authors, is that CAT reporter gene can actually possess structural elements that negatively affect gene expression.

AP assays involve a substrate that is hydrolyzed by the AP, with, the resulting absorbance changes being detected by a chemiluminescent or fluorescent assay. Probably the main disadvantage of using AP, apart from requiring the use of substrates, is the background activity, as mammalian cells also express endogenous AP, which will limit the sensitivity of the technique. Secreted alkaline phosphatase (SEAP) are also available, with the AP being secreted by the cells, thus facilitating the sampling procedures as cell disruption is no longer necessary. The background activity is also reduced for SEAP (Jiang *et al*., 2008; Yang *et al*., 1997).

Similar to CAT, the *E. coli* β-gal, is a bacterial enzyme, which can be detected by hydrolysis of a substrate, usually o-nitrophenyl β-D-galactopyranoside (ONPG). β-gal also presents no endogenous activity, however, besides requiring substrate and cell lysis the assay has a narrow dynamic range and a very poor sensitivity.

GFP gene, isolated from the bioluminescent jellyfish *Aequorea*, is probably the most frequently used reporter gene for studying biological systems. The great advantages of GFP include the fact that it is an autofluorescent protein, thus not dependent of any substrate, making possible the study of intracellular events without disrupting the cells, and no endogenous activity is detected in mammalian or even bacterial cells. Also, several mutants of GFP gene are available nowadays, with improved fluorescence. However, GFP gene has a considerably size, so precious plasmid space can be lost, and GFP protein is very sensitive to changes in temperature and oxygen concentrations (Jiang et al., 2008).

Luciferase is a generic term for describing bioluminescent proteins that catalyzes the oxidation of a substrate, emitting light. No endogenous activity is detected in mammalian cells and luciferase assays are sensitive. Additionally, luciferases have a large emission broad, and emits light at wavelengths capable of penetrating cells and tissues, so the detection does not involve cell lysis. However, bacterial luciferases, often used as reporter genes, have limited applications as they are unable to continuously produce light, it has limited cellular permeability and limited contrast. Belancio et al. (2011) reported artifacts in the luciferase activity, when comparing a wild type protein and the same protein with a point mutation, and obtained the same luciferase activity. Since luciferase activity is an indirect method to evaluate protein expression, it may happen that mutations in the target gene may be overlooked.

As pointed, reporter genes represent a generalized tool in cell and molecular biology based work, with a high diversity of applications. However, reporter genes present serious limitations, most of the times not considered, and that may impair the final work's goals. The main drawbacks are the possible background activity, usually the assays involve cell disruption and have low sensitivity. Also the silencer effect described for CAT and the misleading results provided by luciferase represent serious disadvantages. Therefore, the ideal reporter gene is still to be discovered as it should have none or reduced background activity, meaning that the gene should not be naturally expressed by the host cell, the detection should not involve cell disruption and the assay must be sensitive, reproducible and not time-consuming (Jiang et al., 2008). The development of alternative techniques that can bypass the need for reporter genes is thus highly desired.

Fourier transform infrared (FTIR) spectroscopy can be a promising candidate to bypass the need for reporter genes. Besides allowing quantitative analysis, FTIR spectroscopy also enables the extraction of meaningful information about biochemical cellular events. Particularly in the mid infrared region (MIR), the fundamental vibration modes of biomolecules can be measured, with the resulting FT-MIR spectra of a cell reflecting its specific metabolic status, such as the general gene expression. For example, FTIR spectroscopy has been used to detect cell cycle events (Pacifico et al., 2003), stem cells differentiation (Ami et al., 2008) and carcinogenic processes (Gazi et al., 2006). FTIR spectroscopy has also been used for monitoring cell cultures, allowing the estimation, from a single spectrum, of several variables such as the concentration of biomass (cells), carbon sources, by-products (e.g., ethanol and acetate), plasmid (Lopes *et al*., 2013) and recombinant proteins (Ami *et al*., 1999; Arnold *et al*., 2003; Harthun *et al*., 1998, McGovern *et al*., 1999; Rhiel *et al*., 2010; Sandor *et al*., 2013).

The present work aims to evaluate the application of FTIR spectroscopy to detect, in a mammalian cell populations, the percentage of cells expressing the target gene, without the need for detecting the reporter gene phenotype, in a simple, rapid and cheap way. Also, the measurements were conducted in a high-throughput mode, without needing for cell disruption or using any other reagents. FT-MIR spectra were collected from dehydrated cell pellets and using a 96-microwells plate. Partial least squares (PLS) models were developed for quantitative analysis, based on the spectral data acquired. Also, FTIR spectra was analyzed directly, were studied the 1st derivatives spectra and some bands' ratios between some crucial bands, e.g., DNA, RNA, protein and lipid bands, with the goal of extracting biochemical cellular information related not only with the transfection event itself, but also related with the exposure to the transfection reagent, highlighting the sensitivity of the technique. For these purposes, a semi-adherent cell line (Human Embryonic Kidney 293 cells, HEK) and adherent cell line (Human gastric carcinoma, AGS), and the plasmid pVAX-GFP were used as model systems.

IV.2.2 Materials and Methods

Cell culture

HEK and AGS cells were preserved at -80°C. The cells were grown in an CO_2 incubator (Blinder CB150, Tuttlingen, Germany) at 37˚C, 5% CO_2 and 99% of humidity. HEK cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Thermo Scientific) and AGS cells were cultured in Roswell Park Memorial Institute Medium (RPMI) (Thermo Scientific), both supplemented with 10% (v/v) of bovine fetal serum (BFS) (Thermo Scientific), previously inactivated by heat (30 minutes at 56˚C), and with 1% (v/v) of penicillin/streptomycin and 15% (v/v) of L-glutamine (Thermo Scientific), in 75 cm^2 T-flasks. Cells were cultured until they had reached about 80-90% of confluence. For releasing the AGS cells between passages trypsin was used (for 15 minutes at 37˚C). HEK cells are semi-adherent cells, no trypsin was necessary to release the cells between passages, as well as for further spectral analysis. For the transfection procedures, about 500000 cells were transferred to each well of a 6-wells plate.

Plasmid

Both AGS and HEK cells were transfected with pVAX plasmid (Invitrogen), containing the GFP gene.

Transfection

Several mixes were prepared for transfection, according to the manufacture's recommendations. A first mix, Mix A, consisting on 15µL of DNA (pVAX with GFP gene), 900 µL of DMEM0 (DMEM medium without FBS) (Thermo Scientific) and 9µL of Plus reagent (Invitrogen) was incubated for 10 minutes at room temperature. A second mix, Mix B, similar to Mix A but with no DNA, was also incubated for 10 minutes at room temperature. Two falcons containing Mix C, consisting on 900µL of DMEM0 (Thermo Scientific) and 10µL of Lipofectamine (Invitrogen). Mix A was added to Mix C, called from now on transfection reagent, and incubated for 30 minutes at room temperature. Mix B was also added to the other falcon containing Mix C for the control cells, called from now on transfection reagent without DNA.

Three types of controls were carried out: one control consisting exclusively of HEK or AGS cells not exposed to any reagent; HEK or AGS cells exposed to 10µL and 200µL of the transfection reagent without DNA. The two last controls were chosen since the volumes represented the minimum and maximum of transfection reagent volume used for transfection.

Different volumes of the transfection reagent were prepared, with the purpose of producing cell populations with growing transfection efficiencies: 10µL, 15µL, 25µL, 50µL, 75µL, 100µL, 120µL, 160µL and 200 µL.

Just before the transfection experiments, the medium was removed and 500µL of fresh DMEM0 or RPMI0 (DMEM or RMPI without FBS) was added to each well. Then, the different volumes of the transfection reagent, with and without DNA, were added and 4 hours later 2mL of DMEM10 was also added to each plate. The fluorescent imaging and FTIR analysis were carried out the day after.

Fluorescent imaging and conventional quantification

The GFP fluorescence in HEK and AGS cells was observed, after 24h of cells being transfected, using a Axiovert 40CFL microscope (Zeiss, Germany) and the images were recorded using a digital high resolution camera (Axiocam Mrc5 (Zeiss)) and the Axiovision Rel. 4.6.3. software (Zeiss). The transfection rate for each sample, corresponding to different volumes of the transfection reagent added, was determined as follow:

$$\frac{Number\ of\ green\ fluorescent\ cells}{Total\ cell\ number} \times 100\%, \qquad \text{(Equation IV.2)}$$

where the number of green fluorescent cells as well as the total number of cells in each well was assumed as being the mean of several fields.

FTIR measurements

Twenty-four hours after the transfection procedures, samples were collected from the wells, centrifuged (5 minutes at 1500 rpm) and resuspended in 500µl of PBS. From each sample, 25µl were placed on IR-transparent Zn-Se-microliter plates (Bruker, Germany) and dehydrated for about 2h50 hours, in a desiccator under vacuum. The spectra were collected using a FTIR spectrometer (Burker, Germany) equipped with an HTS-XT accessory (Bruker, Germany). In order to achieve a high signal-to-noise-ratio (SNR), 64 scans, with a 2 cm^{-1} resolution, in the wavenumber region between 400 and 4000 cm^{-1} , were collected. The FTIR spectra were acquired in triplicate and quintuplicate, for the HEK and AGS cells, respectively.

Data Analysis

Spectral pre-processing, namely multiple scattering correction (MSC) and derivatives, was carried out using Matlab R2012b (Matworks, Natick, MA, USA) and the OPUS software (Bruker, Germany), for baseline correction. Partial Least Squares (PLS) regression were carried out also using Matlab R2012b (Matworks, Natick, MA, USA), for the quantitative analysis. The performance of the PLS models was evaluated based on the root mean square error of cross-validation (RMSECV), the correlation coefficient (R^2) and the percentage of error according to the transfection efficiency range, as follows:

$$Error\ as\ \%\ of\ range = \frac{RMSEP}{range\ transfection\ efficiency} \times 100\% \qquad \text{(Equation IV.3)}$$

The percentage of variance in data explained *versus* the number of latent variables was also considered for model evaluation. The best pre-processing method leading to the best PLS models was selecting as the one providing the lowest RMSECV, a low error as percentage of the transfection efficiency range and a high R^2, while explaining a high percentage of variance in data and using a reduced number of latent variables.

IV.2.3 Results and Discussion

Transfection analysis based on GFP expression

The conventional method chosen to estimate the percentage of transfected cells in a cell population, or sample, was based on the microscope observation of the expression of the reporter gene GFP (Figure IV.15), one of the reporter genes most used for non-invasive monitoring. Two different cell lines were used, HEK and AGS cell lines, representing two types of cells, semi-adherent and adherent cells, respectively. The transfection of the cell lines, analyzed by microscopic observation of the GFP fluorescence, implied counting different fields of observation. The transfection efficiency is known to be cell-dependent. Indeed, for the same transfection protocol, AGS had a lower transfection efficiency in relation to HEK cells (Figure IV.16).

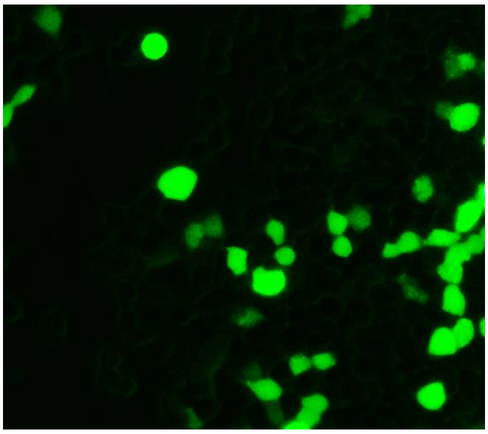

Figure IV. 15 – Microscopic observation of the green fluorescence emitted by the expression of GFP in HEK cells.

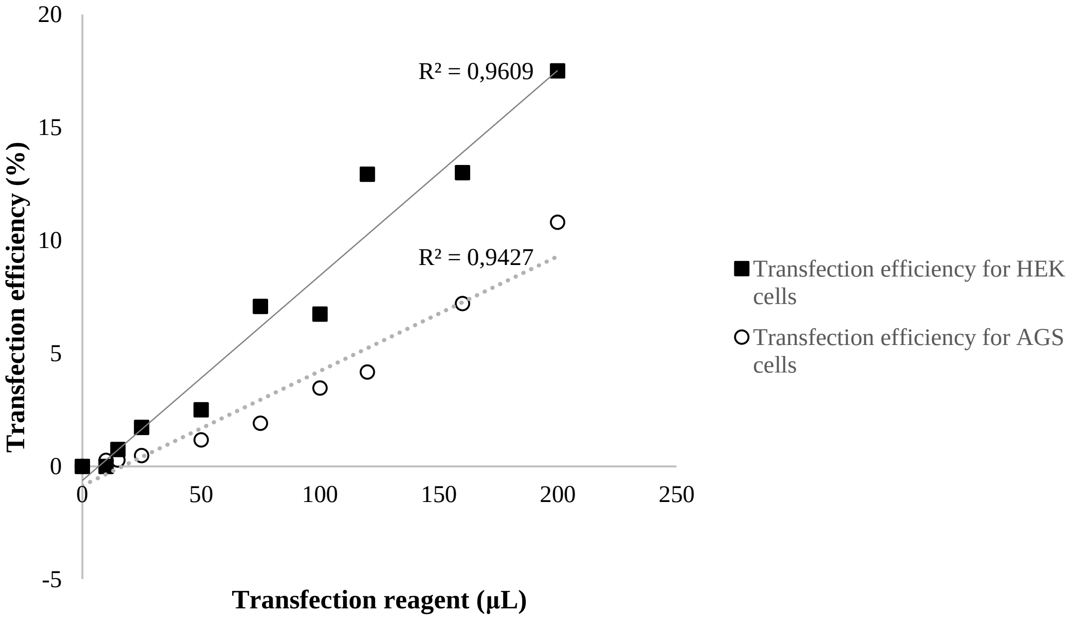

Figure IV. 16 – Transfection efficiency for HEK and AGS cells, as determined by fluorescence detection of the reporter gene expression, GFP.

Spectral pre-processing and data analysis

In order to evaluate if FT-MIR spectroscopy could be used for estimating the transfection efficiency, 3 or 5 replicates of dehydrated cell pellets, HEK and AGS cells, respectively, were analyzed. While possessing a huge amount of information concerning the sample being measured, a FT-MIR spectrum also has some undesirable noise associated, especially when dealing with dehydrated samples, which are subjected to physical interferences, such as light scattering resulting from irregularities on the samples' surface or particles with different sizes and shapes. It is thus necessary to pre-process the spectral data in order to minimize unwanted spectral interferences, while highlighting important information about the sample (Geladi, 2003). Baseline correction, multiplicative scatter correction (MSC), normalization and derivatives are probably the most commonly pre-processing techniques applied to spectral data. Baseline correction is used for removing spectral offsets (Figure IV.17 - A). MSC is often used to eliminate changes in spectra due to radiation scattering. MSC minimizes the differences between replicates, by the elimination of those undesirable physical phenomena (Figure IV.17 - B). Normalizing the spectra is also very important, since sometimes there are differences in the spectra's intensities that are not related to the property of interested, but are a result of differences in the cells' number instead. In the present work spectra were pre-processed using the Amide I band, at 1650 cm^{-1} (Figure IV.17- C) (Naes *et al*., 2002).

The main goal of applying derivatives to the spectral data is to enhance information in data, as it is possible to resolve overlapping bands and eliminating physical differences, as differences in cell number, which can compromise the relationship between the data and the biological sample. First derivative is helpful for resolution enhancement of overlapping bands and for offset correction, as the 1st derivative of a constant is zero (Fearn *et* al., 2009; Otto, 1999; Smith, 2011). Fist derivative spectra were obtained using the Savitzky-Golay algorithm, that applies a filter/smoothing before derivatives to avoid noise amplification. When applying smoothing to the data, there is a risk of losing information while eliminating noise, so it is important to choose the appropriate number of data points of the window and also the appropriate polynomial order. In the present work a window of 15 points and a 2nd order polynomial was used, since it eliminates the noise, while maintaining the peaks in the spectra, meaning no information regarding the samples is lost (data not shown). First derivative spectra and the information that was possible to extract, concerning cells exposed to the transfection reagent and transfected cells, are discussed later in this work.

By pre-processing the data with the methods mentioned above, it is possible to highlight relevant information in data, while the undesirable noise is removed.

In Figures IV.17 (C) and IV.18 are represented the normalized spectra, using Amide I band, from AGS and HEK cells, respectively. By normalizing the spectra it is possible to see interesting differences. Interestingly, the FTIR spectra capture in a highly sensitive way not only the effect of the transfection itself, but also the effect of the transfection reagent on the cell's metabolism. The spectra from AGS cell exposed to the transfection reagent without DNA are apart from all the other cell's spectra. It is not clear the reason why the control cells (cells not exposed to the transfection reagent) are so close to the transfected cells, representing a transfection efficiency of 10.8%. However, even being the spectra from the control cells closer to the spectra from the transfected cells, they are different from each other, reflecting, in this way, the transfection event (Figure

IV.17-C). For HEK cells, interestingly, can be clearly observed the impact of the transfection reagent used (Figure IV.18). The spectra from HEK cells not exposed to the transfection reagent are apart from the spectra of HEK cells exposed to the transfection reagent, with our without DNA. The transfection reagent used, Lipofectamine, had a great impact on cells, as expected, and this was reflected in the spectral data. The results showed to be dependent on the cell-type, as expected.

Even sometimes the spectra can be informative after pre-processed, as shown, it is always important to develop other analysis based on spectral data, especially when the goal is to estimate any parameter, as it will be discussed along this work. To better evaluate the application of FT-MIR spectroscopy for the estimation of transfection efficiency, the subsequent spectra analysis was conducted:

- Quantitative model for the estimation of the transfection efficiency, based on partial least squares (PLS) regression
- Direct biochemical interpretation, exploring the first derivative spectra and a some statistical significant bands’ ratios, associated with the absorption of key cellular biomolecules

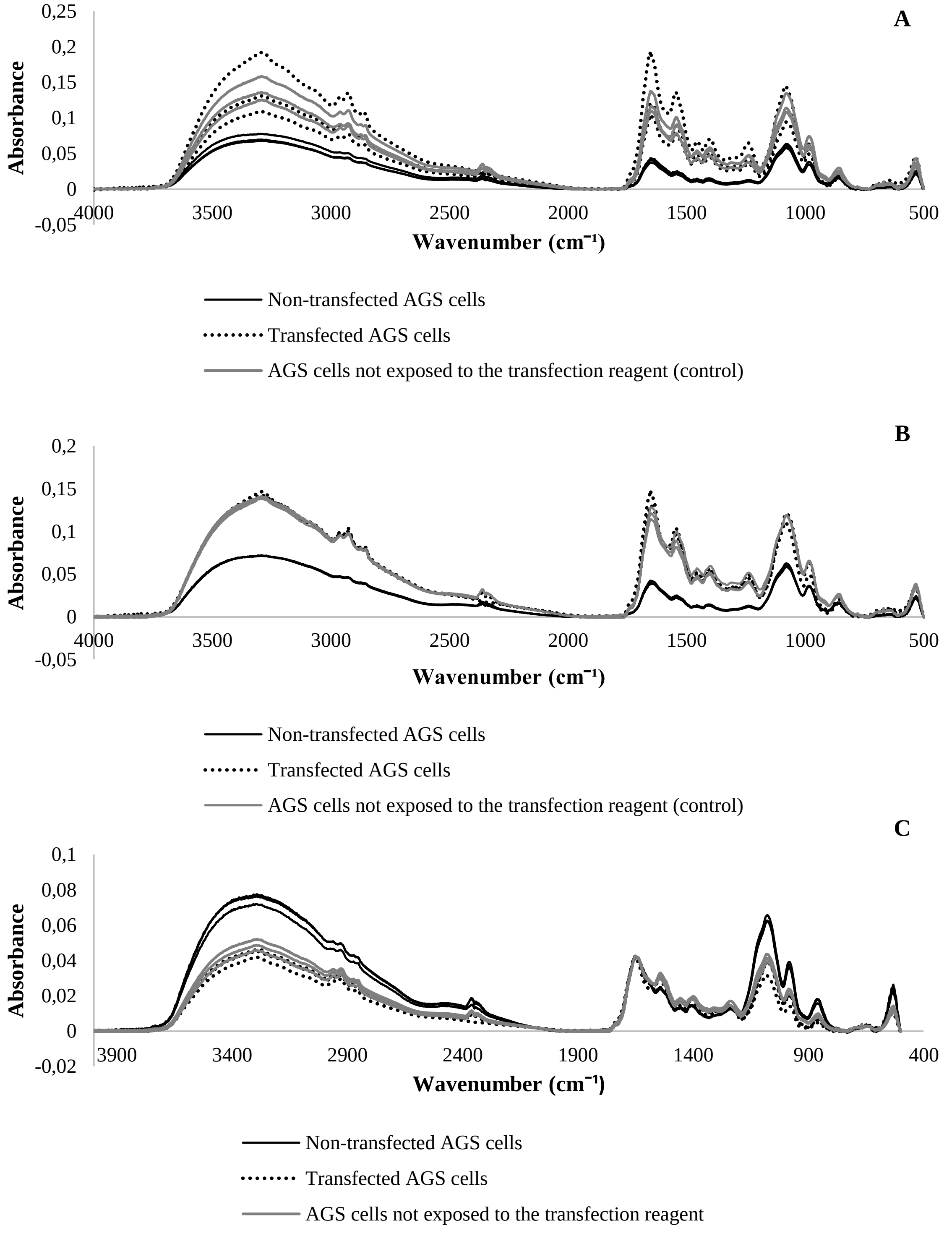

Figure IV. 17 - FT-MIR spectra, with triplicates, pre-processed with baseline correction (A), pre-processed with MSC (B), and normalized with Amide I band (C) from transfected AGS cells (black dashed lines), representing a transfection efficiency of 10.8%, non-transfected AGS cells (black bold lines), meaning cells that were exposed to the transfection reagent without DNA, and AGS cells not exposed to the transfection reagent (gray lines). The spectra were acquired in the 500-4000 cm^{-1} spectral region, with 64 co-added scans and a resolution of 2 cm^{-1}.

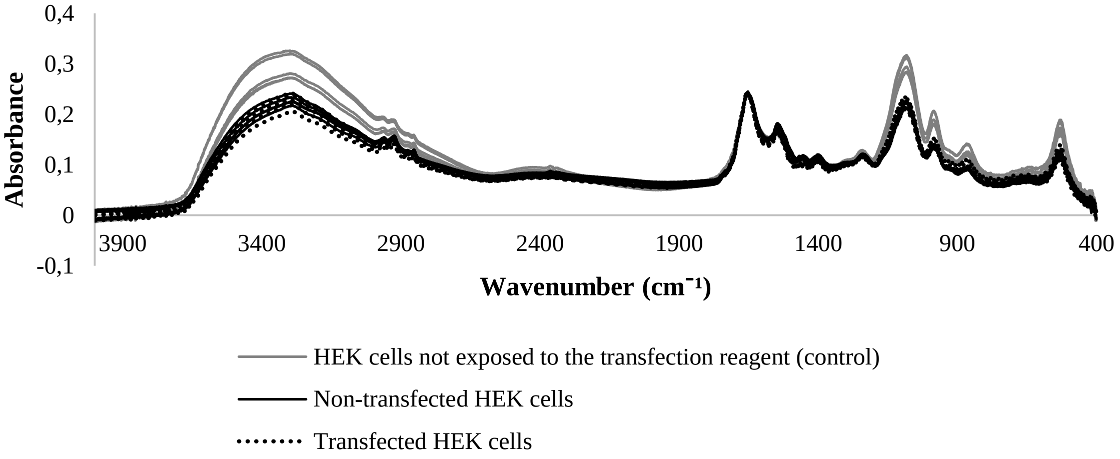

Figure IV. 18 - FT-MIR spectra of HEK cells not exposed to the transfection reagent (gray lines), non-transfected HEK cells, meaning cells that were exposed to the transfection reagent without DNA (black bold line), and transfected HEK cells, representing a transfection efficiency of 17.5% (black dashed line). The spectra were normalized with Amide I band and acquired in the 500-4000 cm^{-1} spectral region, with 64 co-added scans and a resolution of 2 cm^{-1}.

Partial least squares (PLS) regression models

PLS regression models were developed with the goal of predicting the transfection efficiency for HEK and AGS cells from FTIR spectral data

Different pre-processing techniques were applied to data, in order to obtain the best predictive performance. The best PLS models concerning the RMSECV, R^2 and error as percentage of the transfection efficiency range were obtained for data pre-processed with MSC and 1st or 2nd derivatives (Table IV.3). As an example, Table IV.3 presents the correlation coefficient (R^2) and the RMSECV of the PLS regression models from the transfection experiment using the HEK cell line, were is observed that different pre-processing techniques significantly affect the model's performance

Table IV. 3 - Correlation coefficients (R²) and RMSECV for the PLS models (5 latent variables) were the data is pre-processed with different techniques (data from the transfection experiment with the HEK cell)

	R^2	RMSECV	Error as % of the range
Baseline correction	0.63	3.64	20.8
Baseline correction + MSC	0.69	3.34	19.1
1st derivative	0.91	1.77	10.11
MSC + 1st derivative	0.93	1.57	8.9
MSC + 2nd derivative	0.92	1.67	9.5

For the experiment using the HEK cell line the best PLS model for quantifying the transfection efficiency was built on data pre-processed with MSC and 1st derivative and using 5 LVs explaining about 92% of the variance in data, yielding a correlation coefficient between real and predicted values of 0.93, which indicates a good fit (Figure IV.19). The RMSECV was 1.57 (%), representing 8.34% of the range of transfection efficiency.

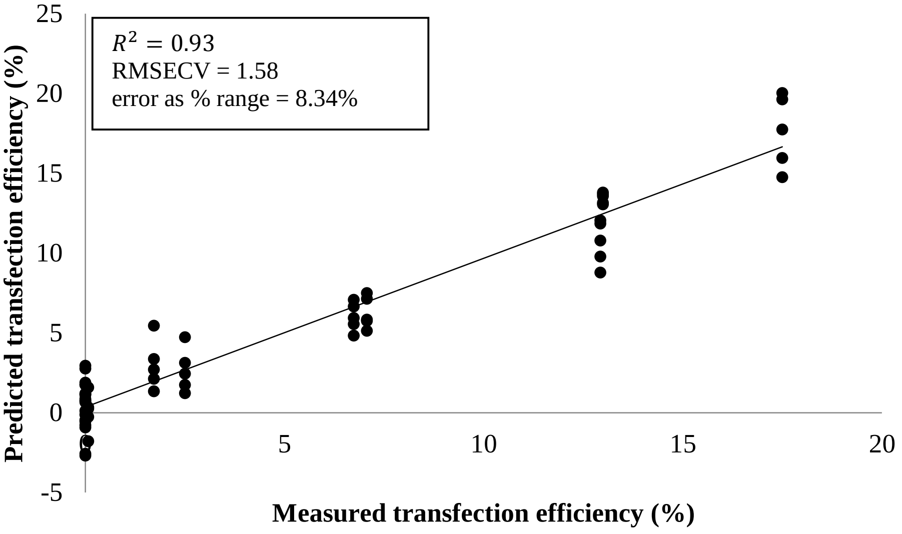

Figure IV. 19 - Predicted (5LVs) and measured values for the transfection efficiency of the HEK cells, using the spectral region between 400 and 4000 cm^{-1} and data pre-processed with MSC and 1st derivative.

The best PLS model for quantifying the transfection efficiency using the AGS cell line was built on data pre-processed with MSC and 2nd derivative and using 5 LVs explaining about 95% of the variance in data, achieving a R^2 between the true and the predicted transfection efficiency of 0.95 and a RMSECV of 0.71 (%), representing 6.64% of the transfection range (Figure IV.20). By pre-treating the data with MSC and 1st derivative, as for the model above using HEK cells, the results were significantly worse, using 5LVs the model yielded a R^2 of 0.78 and a RMSECV of 1.55, representing about 14.4% of the transfection efficiency range.

Both models performed very well, proving that PLS models based on spectral data are cell line-independent and can be used as an accurate tool for estimating the transfection rate or monitoring the production of heterologous proteins without the need of using any reporter gene.

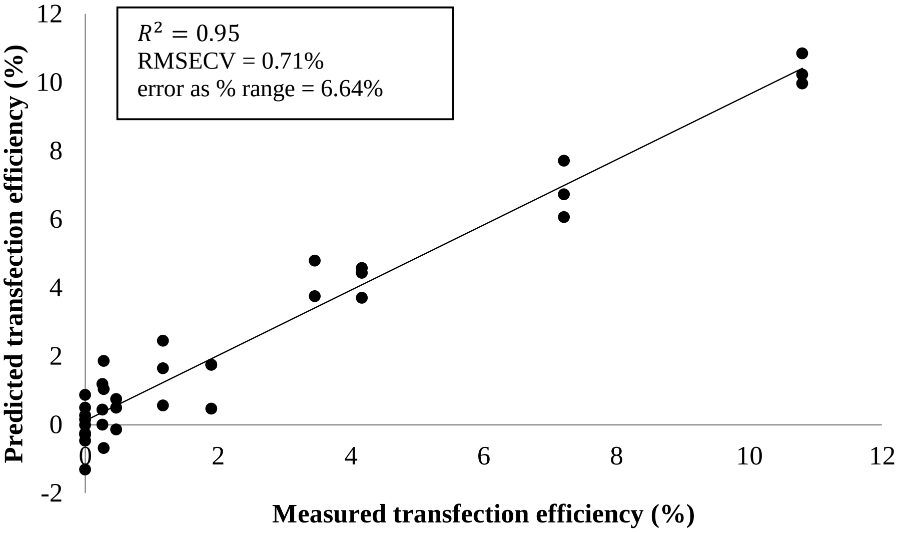

Figure IV. 20 – Predicted (5LVs) and measured values for the transfection efficiency of the AGS cells, using the spectral region between 400 and 4000 cm^{-1} and data pre-processed with MSC and 2nd derivative.

Besides trying different pre-processing techniques to improve the performance of the PLS models, for both HEK and AGS transfection experiments, the selection of different spectral regions, according to the regression vector obtained from the PLS model built on the entire spectral region (Figures IV.21 and 22), was also evaluated. In terms of comparison, it was considered for both AGS and HEK cells the regression vector of the models build based on the 1st derivative spectra and after pre-processing the data with

MSC, as the regression model build from the 2nd derivative spectra presents, as expected, a higher noise. Interestingly, for both cell lines, it seems that the spectral regions contributing more for the PLS models are very similar (Figures IV.21 and IV.22). The following regions, contributing to the PLS models with the highest coefficients were selected, after applying MSC and 1st derivative to the spectral data of HEK and AGS transfection experiments:

- Protein region: 1800 - 1480 cm^{-1};
- DNA and RNA region: 1425 - 900 cm^{-1};
- Protein and DNA region: 1000 - 1700cm^{-1};
- Region between 3400 $cm^{-1}$and 700 cm^{-1}, in an attempt to eliminate the noise.

Curiously, none of the models built on specific spectral regions performed better than the ones using the entire spectral window (Table IV.4). Although some spectral regions seem to poorly contribute for the PLS model, they also contain important information that improves the model's predictive performance.

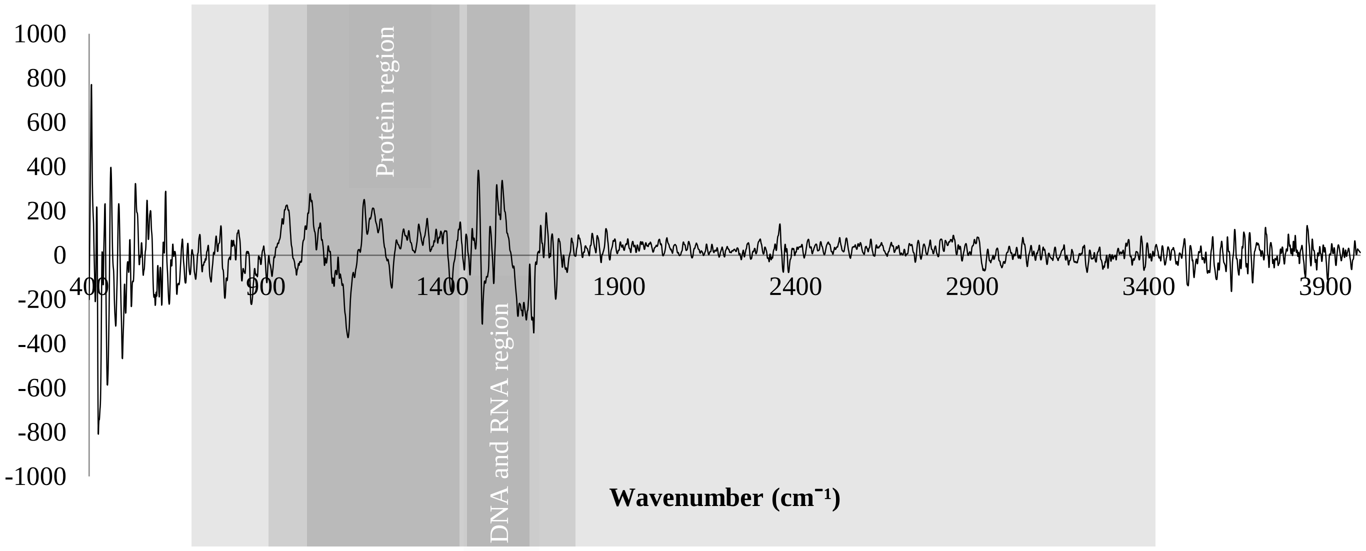

Figure IV. 21 - Regression vector from the PLS model for the HEK transfection experiment using 5 LVs with data pre-processed with MSC and 1st derivative and considering the entire spectral window.

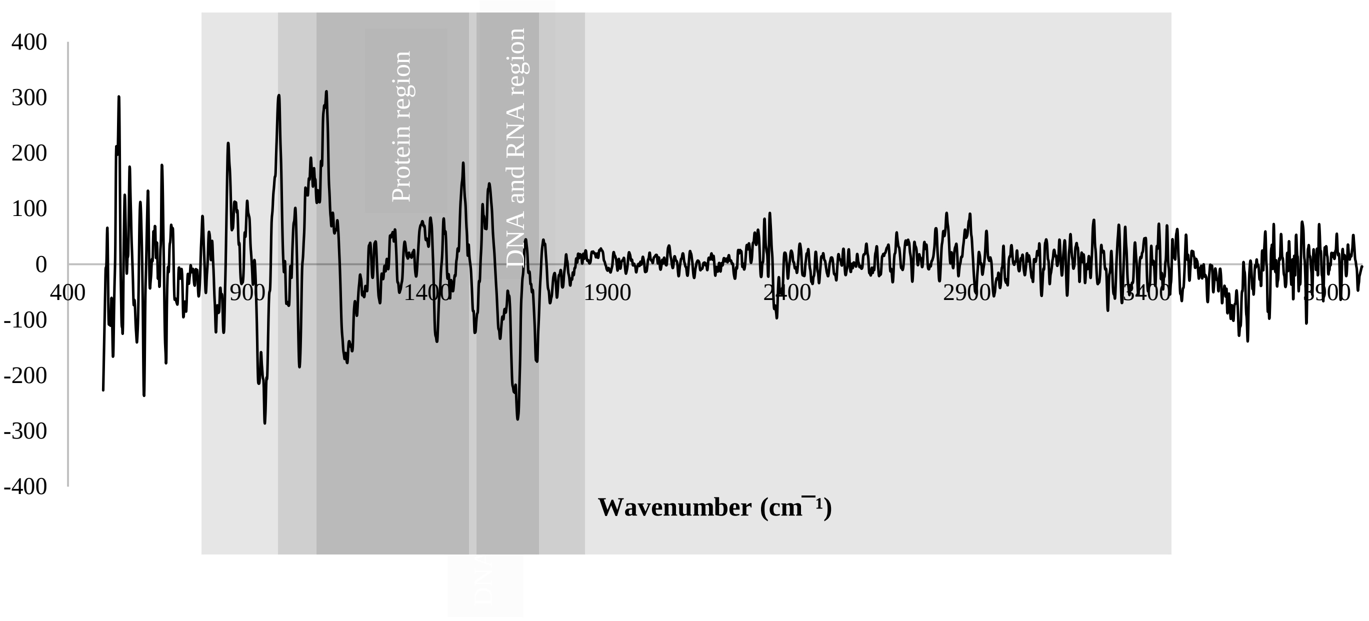

Figure IV. 22 - Regression vector from the PLS model for the AGS transfection experiment, using 5 LVs, with data pre-processed with MSC and 1st derivative, considering the entire spectral window.

Table IV. 4 - PLS regression models (5 LVs) using different spectral regions and data pre-processed with MSC and 1st derivative for the HEK cells and MSC and 2nd derivative for the AGS cells

		All spectral region	**Protein Region (1800-1480 cm^{-1})**	**DNA and RNA region (1425-900 cm^{-1})**	**DNA and Protein region (1000-700 cm^{-1})**	**Region between 3400-700 cm^{-1}**
HEK cells	R^2	0.93	0.87	0.91	0.89	0.91
	RMSECV (%)	1.57	2.16	1.78	2.04	1.83
	Error as % of the range	8.34	12.36	10.17	11.64	10.48
AGS cells	R^2	0.95	0.71	0.94	0.82	0.91
	RMSECV (%)	0.72	1.76	0.84	1.42	1.02
	Error as % of the range	6.64	16.30	7.80	13.15	9.44

An additional PLS regression model based on LOO cross-validation was built, considering simultaneously data from the transfection of AGS and HEL cell lines and after data being pre-processed with MSC and 2nd derivative (Figure IV.23). This model, obtained using 5LVs, yielded a R² between measured and predicted values of 0.93 and a RMSEV of 1.45 (%), representing about 8.3% of the transfection efficiency range. Even when introducing more variability in the model, namely different cell lines grown in media with different compositions, it is still possible to achieved good estimates of the transfection efficiency. These results are very promising, and they suggest that it is possible to develop good PLS models for the prediction of the transfection efficiency, independently of the cell line or the cells' growing medium.

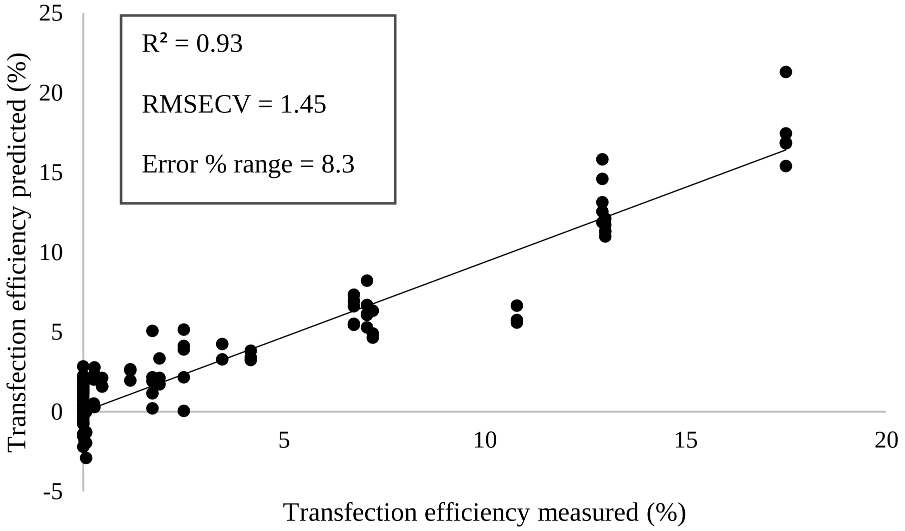

Figure IV. 23 – Predicted (5LVs) and measured values for the transfection efficiency of the AGS and HEK cells, using the spectral region between 400 and 4000 cm^{-1} and data pre-processed with MSC and 2nd derivative.

Analysis of the first derivative spectra

First derivative spectra can highlight crucial biochemical information about the samples that sometimes is hidden in the initial spectra. In the present work, differences in the 1st derivative spectra that could be after used as biomarkers for identifying transfected cells or a reflection of the exposure to the transfection reagent were selected. Aiming at finding differences between cell samples due to the maintenance of heterologous genetic material and the production of an heterologous protein, 1st derivative spectra of HEK and AGS cells with the maximum transfection efficiency (17.5% and 10.8%, respectively), were compared to the 1st derivative spectra of the cells exposed to the same volume of the transfection reagent without DNA, assumed as non-transfected cells. For identifying changed bands associated with the exposition to the transfection reagent, 1st derivative spectra of HEK and AGS cells not exposed to the transfection reagent, were compared with the 1st derivative spectra of cells exposed to the maximum volume of the transfection reagent used, 200 mL, called before non-transfected cells.

For the HEK cells the main differences between transfected and non-transfected cells were found in the region between 2700 and 3800 cm^{-1}, attributed to proteins with some influence of lipids (Lamberti *et al.*, 2010), proving that FTIR spectroscopy has the ability to detect the production of new proteins by the cell, even when they are in much lower concentrations when comparing to the host proteins. The first derivative bands that showed the greater differences were at 3645 cm^{-1}, 3629 cm^{-1}, 3625 cm^{-1}, 3609 cm^{-1}, 3584 cm^{-1} and 3563 cm^{-1} (A-F in Figure IV.24). These bands, mainly attributed to proteins and amino acids (Graça *et al.*, 2013), have a very low intensity in non-transfected cells and greatly increase their intensity in transfected cells. Also, the region between 1500 and 1600 cm^{-1}, Amide II band region (Lamberti *et al.*, 2010), is markedly different for transfected cells (G in Figure IV.24). Additionally, bands corresponding to DNA absorption appear slightly different, namely a peak displacement of the band at 1000cm^{-1} for transfected cells, attributed mainly to PO_2^- symmetric stretching (Liu $et\ al.$, 2005) (marked as H in Figure IV.24).

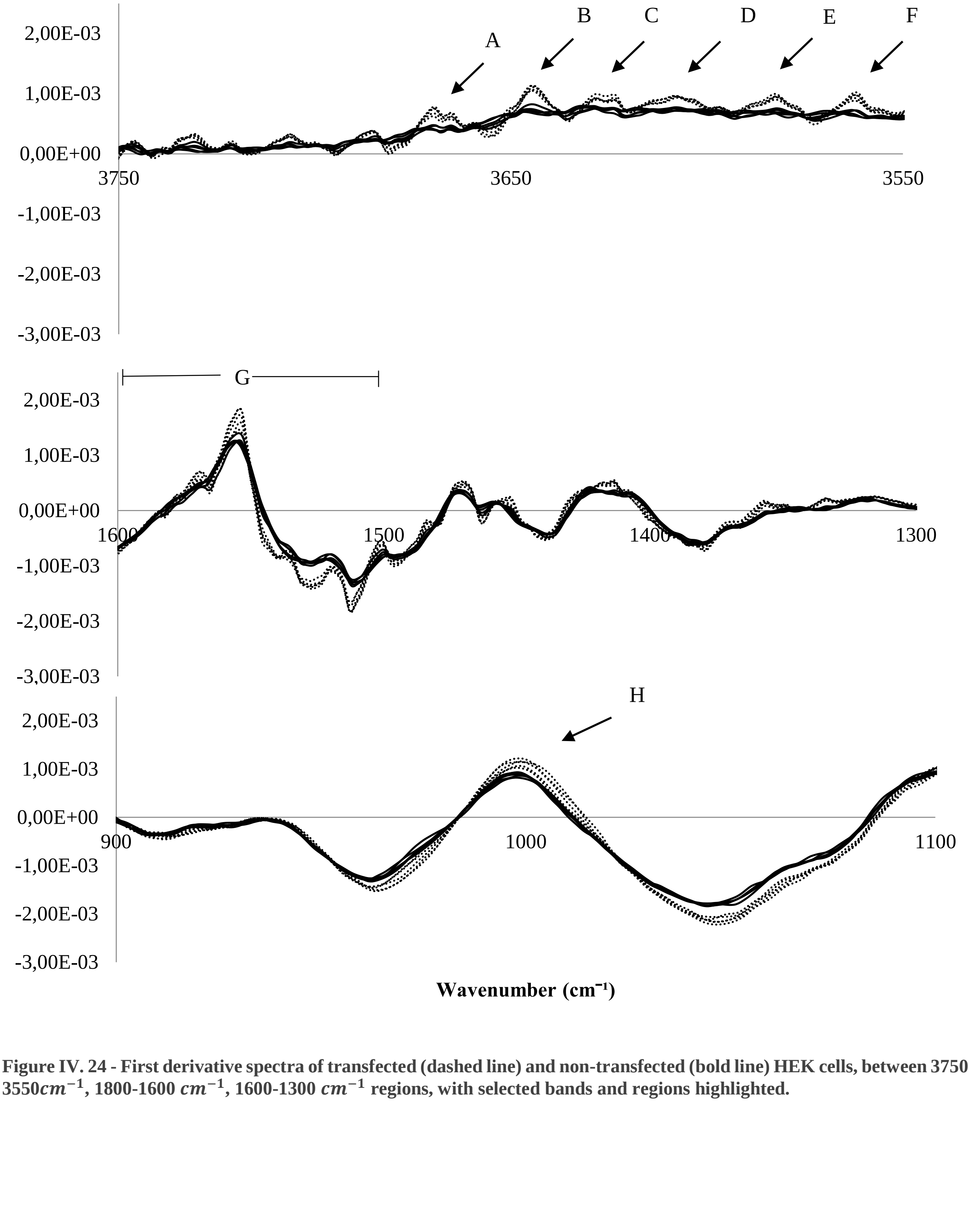

Figure IV. 24 - First derivative spectra of transfected (dashed line) and non-transfected (bold line) HEK cells, between 3750-3550cm^{-1}, 1800-1600 cm^{-1}, 1600-1300 cm^{-1} regions, with selected bands and regions highlighted.

By analyzing the first derivative spectra of HEK cells exposed and not exposed to the transfection reagent (without DNA), the main differences could be observed in bands associated with lipid absorption (Figure IV.25), in agreement with the reagent chosen for the transfection experiments, Lipofectamine. As the transfection event was based on the fusion of liposomes with the cellular membrane. Therefore, the cellular membrane structure will be affected, in accordance with the 1st derivative spectra.

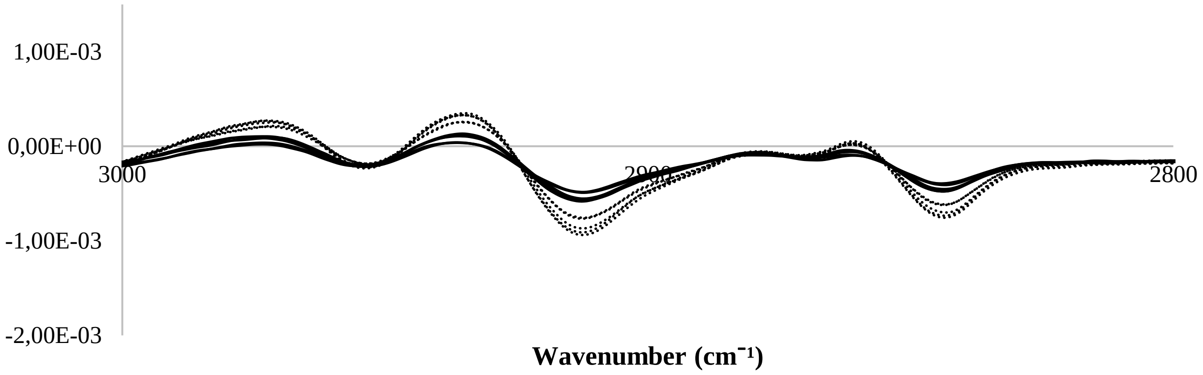

Figure IV. 25 - First derivative spectra of HEK cells not exposed to the transfection reagent (bold line) and HEK cells exposed to the transfection reagent without DNA (dashed line), between 3000-2800 cm^{-1}.

The same analysis based on the 1st derivative spectra was performed for the transfection experiment using the AGS cell line. As for HEK cells, the spectral regions with the major differences between transfected and non-transfected cells are located at 1800 - 810 cm^{-1} (Figure IV.26), a region mainly attributed to protein absorption (Walsh *et al.*, 2009). In this region the changes between transfected and non-transfected cells were concerning some strong bands corresponding to amide I, II and III, located at 1750 – 1600 cm^{-1}, 1600 – 1480 cm^{-1}, and 1300 – 1180 cm^{-1}, respectively (Lamberti *et al.*, 2010). For both transfection experiments, using AGS and HEK cell lines, the major differences regarding the protein spectral region show the ability of MIR spectroscopy for detecting the production of a heterologous protein, even when the protein is present in a very lower concentration compared to the other host proteins. As for the HEK cell line, a displacement of the band located at 1000cm^{-1} and attributed to DNA absorption (Liu *et al.*, 2005) (Figure IV.26 – B), could also be observed.

When comparing AGS cells exposed and not exposed to the transfection reagent (without DNA) the greater changes were also in bands related to lipid absorption (Figure

IV.27), as observed for HEK cells and in agreement with the transfection reagent used, Lipofectamine.

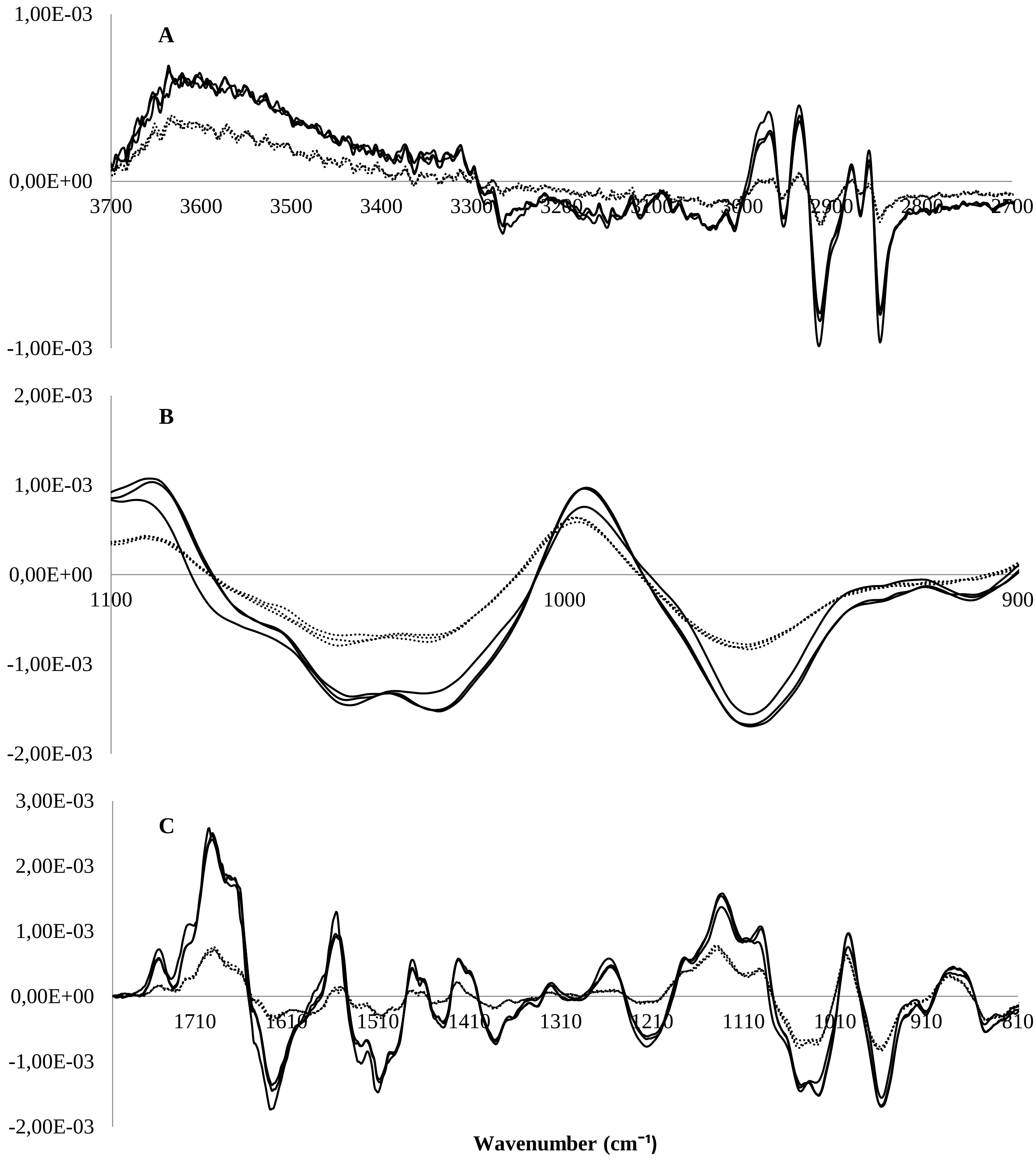

Figure IV. 26 - First derivative spectra of transfected (dashed line) and non-transfected (bold line) AGS cells, 3 replicates each, in the 3700-2700cm^{-1} (A), 1100-900cm^{-1} (B) and 810-1800cm^{-1} (C) spectral regions.

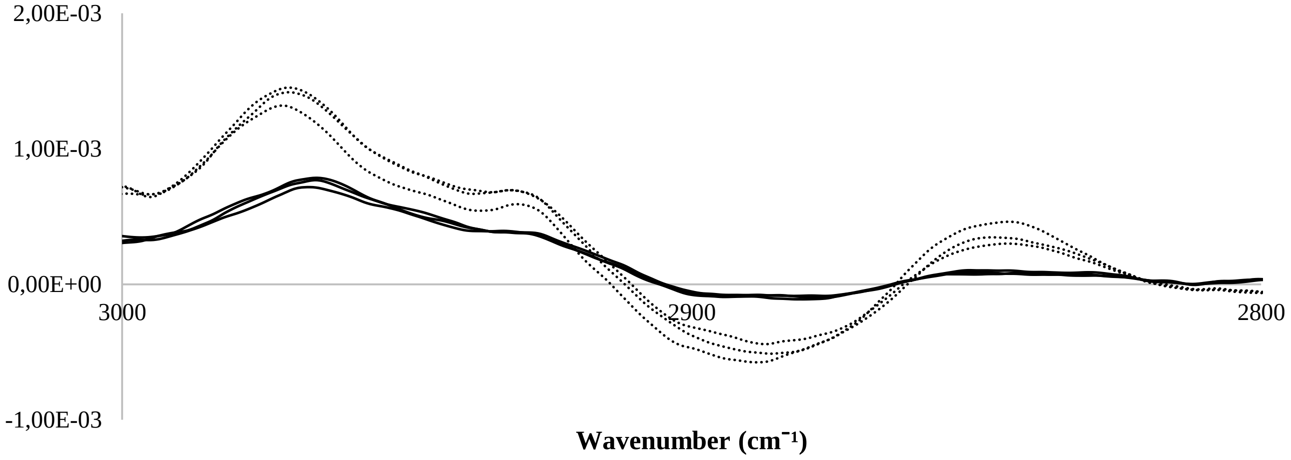

Figure IV. 27 - First derivative spectra of AGS cells not exposed to the transfection reagent (bold line) and AGS cells exposed to the transfection reagent without DNA (dashed line), between 3000-2800cm^{-1}.

By comparing the 1st derivative spectra from the different cell lines, common differences were found, with bands' changes due the transfection event and even changes related to the exposure to the transfection reagent. It was observed that were specific regions that were changed for both cell lines, namely the 3000-2800 cm^{-1} region, which changed as a result of the exposition to the transfection reagent, and the 810-1800 cm^{-1} region, that changed for both cell lines as a result of the transfection event.

Bands' ratios

With the goal of extracting additional information on transfected cells and on the effect of the transfection reagent, the ratio between some relevant bands in the spectra, e.g. bands attributed to DNA, proteins and lipids absorptions, were also evaluated. IR normalization, has conducted at the beginning of the chapter, can minimize the effect of differences in the cells' number. However, to normalize the spectra it is necessary to choose a spectra region as constant, as the Amide I band, that however can be different in case of the expression of an heterologous gene or due to stress response.

FTIR spectroscopy is based on the absorption of radiation by a sample, therefore the pathlength is a critical variable to take into account. However, the pathlength of the sample is irreproducible, as it is impossible to have two different samples with the same thickness or the same number of cells. Calculating the ratios between important spectral bands can thus be very useful, as it allows to eliminating these physical interferences while comparing different samples.

The band's ratios of some important spectral bands were determined for three groups of samples: transfected cells with the higher transfection efficiency (17.5% for the HEK cells and 10.8% for the AGS); cells exposed to the transfection reagent without DNA, called non-transfected cells; and cells that were never exposed to the transfection reagent (control). The spectral data were pre-processed with baseline correction and MSC and the differences in the bands' ratios were considered statistically significant for a p-value lower than 0.05.

The effect of the transfection reagent was studied based on the ratios between bands resulting from lipid absorption, for both HEK and AGS cells exposed and not exposed to the transfection reagent. For both cell lines the ratio between the spectral bands $2925cm^{-1}$ and $2960cm^{-1}$, associated to the asymmetric stretching of CH_2 and CH_3 end-groups of membrane lipids, respectively (Liu *et al.*, 2005), increased for cells exposed to the transfection reagent (Figure IV.28). These results are in agreement with the ones obtained through the analysis of the 1st derivative spectra (shown before), as well as with the transfection reagent used, Lipofectamine.

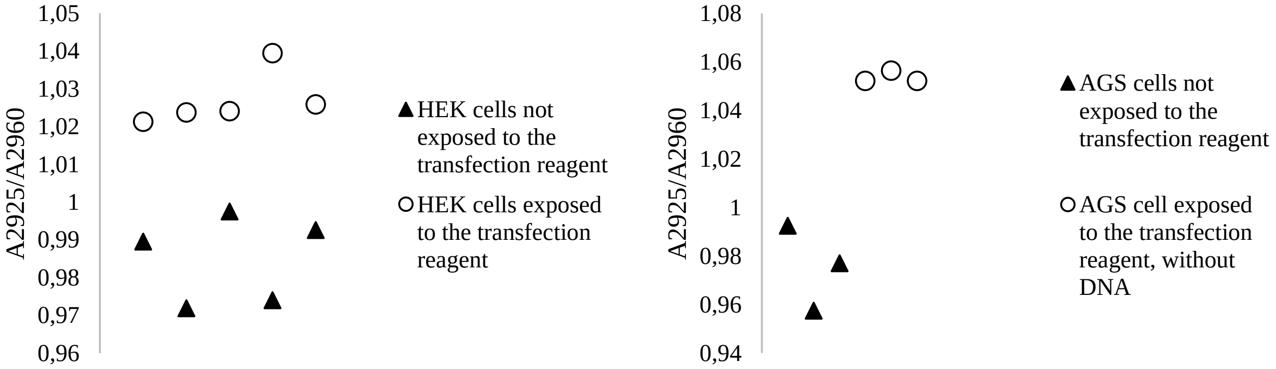

Figure IV. 28 - Ratio between the bands at $2925cm^{-1}$ and 2960 $cm^{-1}$for HEK and AGS cells lines (p value < 0.05). The triangles represent cells that were never exposed to the transfection reagent and the circles cells that were exposed to the reagent.

By comparing the ratios of spectral bands from transfected and non-transfected HEK cells the following differences were found:

1) The A_{1121}/A_{1020} ratio, associated with the RNA and DNA absorption (Walsh *et al.*, 2009), respectively, increased for transfected cells, indicating a higher transcriptional activity (Figure IV.29);
2) The A_{1121}/A_{2852} ratio, associated with the RNA and lipid absorption (Walsh *et al.*, 2009; Liu *et al.*, 2005), respectively increased for transfected HEK cells, also indicating a higher transcriptional activity (Figure IV.29).

Both of these bands' ratios shown for HEK cells shown a higher transcriptional activity for transfected cells. So these results are in accordance with each other and with the transfection event, proving again the sensitivity of the technique.

For AGS cells, differences in bands' ratios were also observed, that might be attributed to the transfection event, however changes were not consistent with the ones observed for HEK cells and also revealed to be less informative. For instance the A_{1087}/A_{1570} ratio, associated to the DNA and Amide II absorptions, respectively, is often used to detect changes in the DNA and protein content (Liu *et al.*, 2005; Walsh *et al.*, 2009). This ratio decreased for transfected AGS cells and may be related with the increase in the protein content. However, when relying on ratios using DNA absorption bands, some caution must be taken, as those bands are highly influenced by the cell's stage due to different chromatin conformations. Depending on the chromatin conformation the absorption of infrared radiation can be different (Figure IV.30).

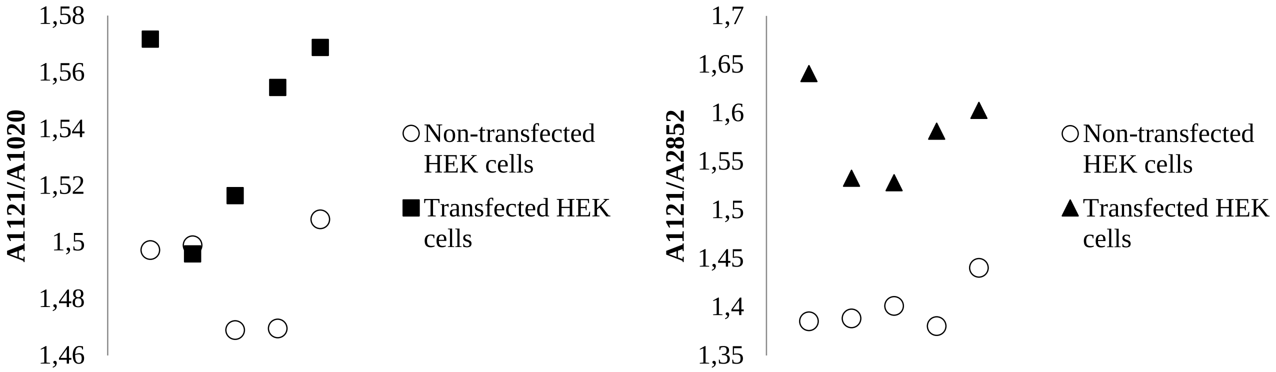

Figure IV. 29 - Ratio between the bands at 1121cm^{-1} and 1020 cm^{-1} (associated with RNA and lipids absorption, respectively), and between 1121 cm^{-1} and 2852 cm^{-1} (associated with RNA and lipid absorption, respectively) for HEK cell lines (p value < 0.05). The triangles represent transfected HEK cells and the circles the non-transfected HEK cells.

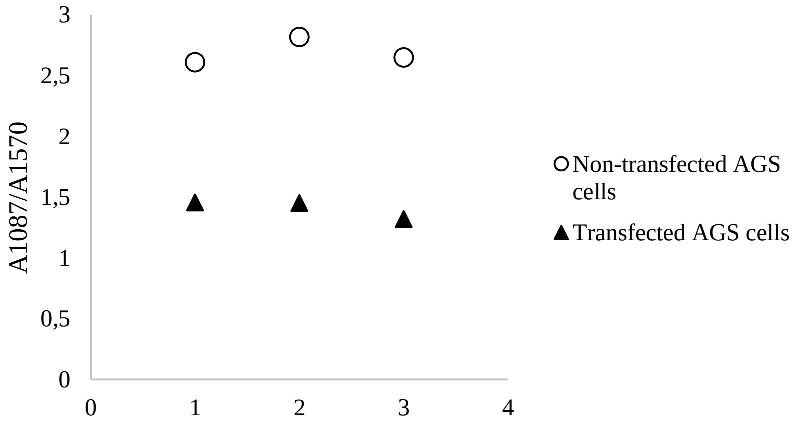

Figure IV. 30 - Ratio between the bands at 1087cm^{-1} and 1570 cm^{-1} (associated with DNA and Amide II absorptions, respectively) for AGS cell lines (p value < 0.05). The triangles represent transfected AGS cells and the circles the non-transfected AGS cells.

The ratios determined earlier, although cell-type dependent, can be very informative and can be used as biomarkers for the analysis of transfected cells, once again proving the potential of FTIR spectroscopy to detect transfected cells without using any gene probes. Moreover, important information about the effect of the transfection reagent could be extracted.

IV.2.4 Conclusions

Reporter genes, are routinely used in molecular and cellular biology laboratories for several applications, with fully characterized and broadly implemented procedures. However, reporter genes present serious disadvantages, as many of them have background activity, the assays involve very time-consuming procedures and have poor sensitivity, and cell disruption is sometimes needed for the detection assays. Other drawbacks have also been reported, such as a silencer effect as described for CAT (Zhang *et al*., 2013) or misleading results as described for luciferase (Belancio *et al*., 2011).

In the present work FT-MIR spectroscopy was evaluated as a substitute for reporter genes for estimating the transfection efficiency in a cell population, as it is a very sensitive technique, it requires minimal sample preparation and no reagents are necessary, and it can be operated in a high-throughput mode. Two different cell lines, HEK and AGS, transfected with pVAX plasmid containing the GFP gene, were studied as model system in an attempt of evaluating the suitability of this approach.

Accurate PLS regression models were built to predict the transfection efficiency, in HEK and AGS cell lines. For HEK cells, a PLS model built on data pre-processed with MSC and 1st derivative, yielded a R^2 between the real and the predicted values of 0.93 and a RMSECV of 1.57 (%), representing about 8% of the transfection efficiency rate range. For the AGS cells, a slightly better model was obtained for spectral data pre-processed with MSC and 2nd derivative, yielded slightly better results, with data being prior pre-processed with MSC and 2nd derivative, yielding a R^2 of 0.95 and a RMSECV of 0.72 (%), representing about 6.6% of the transfection efficiency range. It was also possible to build a robust PLS regression model using both cell lines (HEK and AGS cell lines), a R^2 between measured and predicted values of 0.93 and a RMSECV of 1.45 (%), representing 8.3% of the transfection efficiency range.

Besides allowing to estimate the transfection efficiency, FTIR spectroscopy shown the ability to simultaneously provide information about the cells' biochemical status. Through the study of the first derivative spectral and a few relevant bands' ratios, was possible to detect statistically significant changes in cells resulting from the transfection event as well as the exposition to the transfection reagent.

For both HEK and AGS cell lines, it was observed that the spectral regions that reflected the greater differences between transfected and non-transfected cells were associated with proteins, as expected. Additionally, bands associated with lipid absorption reflected the cells' exposition to the transfection reagent, also as expected.

Very interesting results were also achieved when studying some relevant bands' ratios. The RNA/DNA (A_{1121}/A_{1020}) and RNA/Lipids (A_{1121}/A_{2852}) ratios, increased for transfected HEK cells, probably indicating a higher transcription activity, in accordance with the transfection event. For both HEK and AGS cells, changes were detected in ratios of bands associated with lipid absorption, A_{3010}/A_{2960} and A_{2925}/A_{2960}, due to the cell exposition to the transfection reagent, again, in accordance with the results provided by the first derivative spectra and in accordance to the transfection reagent used, Lipofectamine.

In resume, all the above approaches for post-transfection analysis yielded very good results, proving the potential of the technique not only to accurately estimating the transfection efficiency but also to provide valuable information about cells' biochemical events, related not only with the transfection event, but with the effect of the transfection reagent in cells, particular the cell membrane.

IV.3 Study of *Helicobacter pilory* infection of AGS cells using FTIR spectroscopy

IV.3.1 Introduction

Helicobacter pylori is a gram negative bacterium that has the ability of infecting the human stomach, being estimated that approximately 50% of the world population is infected. The infection by *H. pylori* can lead to several pathologies, namely chronic gastritis, asymptomatic or a late non-ulcer dyspepsia (NUD), peptide ulcer disease (PUD), gastric or duodenal, or even gastric cancer (GC), adenocarcinoma or gastric mucosa associated lymphoid tissue (MALT) lymphoma (Kusters *et al*., 2006). It is estimated that 10 to 20% of the individuals infected with *H. pylori* develop PUD and 1 to 2% develop GC, the two most aggressive pathologies caused by *H. pylori* infection. Moreover, *H. pylori* infection is responsible for 75% of the GC cases worldwide (Azevedo *et al*., 2007).

Infection by *H. pylori* triggers an acute immunity response, though not efficient in eradicating the bacterium. The growing resistance to antibiotics, the lack of a mechanism to avoid *H. pylori* infection and an efficient treatment to irradiate the bacterium, makes *H. pylori* infection a serious health problem worldwide.

The different pathologies mentioned above are a result of different factors, such as the virulence of *H. pylori* strains and their interaction with the host's immune system, as well as environmental factors, such as smoking, alcoholism, wrong eating habits or the use of anti-inflammatory drugs. The contribution of each factor mentioned above for the development of the disease (NUD, PUD or GC) is still poorly understood (Kusters *et al*., 2006; Amieva e El-Omar, 2008). It is thought that the virulence factors modulate the interaction between the bacteria and the gastric epithelial cells and also the interaction with the immunity system (Akhter *et al*., 2007).

Cag A (cytotoxin-associated antigen A) and VacA (vacuolating cytotoxin A) are probably the most relevant virulence factors of *H. pylori*. Cag A protein, delivered from the bacteria towards the human cells, is a cytotoxin that interact with several cellular proteins, causing alterations in the host's signaling pathways and, consequently, leads to modifications in the cellular morphology, can compromises cell-cell adhesion, cell polarity and promotes cell proliferation (Hatakeyama and Higashi. 2005). Not all *H.*

pylori express CagA, but in general the most virulent strains, usually associated to the most severe gastric diseases, express this cytotoxin. VacA is also a cytotoxin that can induce vacuolation in the host's cells when internalized by endocytosis. Additionally, it has been suggested that VacA can lead to cellular apoptosis, it creates pores in the cellular membrane and has an immunomodulatory effect (Isomoto *et al*., 2010). All identified *H. pylori* strains possesses the VacA gene, however, there are diverse VacA genotypes, classified in function of the gene variability present at the signal sequence, at the mid-region and at the intermediate region of the gene, designated by s-, m- and i-regions respectively. Most of *H. pylori* associated with serious gastric diseases present a s1 genotype, in relation to the two distinct genotypes of the s-region. The exact way how CagA, VacA and other virulence factors are related to the development of each pathology is still not well understood. However, it looks that these two virulence factors act as functional antagonists, as for examples, it was observed that VacA, that promotes cellular apoptosis, may inhibit the morphologic alterations and mitogenic effect induced by CagA. In the same way, it has being suggested that CagA can inhibit apoptosis induced by VacA. Interestingly, the most virulent strains usually express the CagA protein as well as have the most severe VacA genotype, s2. One possible explanation for the presence of these antagonist virulence factors on the most virulent strains is to able the bacteria to take control of the host cell, but without causing gross cellular damage. The final effect of VacA and CagA is dependent of several factors, such as the quantity of each cytotoxin expressed (Palframan *et al*., 2012).

From the, exposed above, increasing the knowledge about the effect of *H. pylori* infection on human gastric cells is of paramount importance, and especially how the main virulence factors affect the infection process. The main goal of the present work was to evaluate if Fourier transform infrared (FTIR) spectroscopy could be used for analyzing *H. pylori* infection *in vitro,* using AGS (human gastric carcinoma) cell lines. FTIR spectroscopy in the middle region of the spectra (MIR) represents the fundamental vibrations modes of chemical bonds, and, therefore, can theoretically capture the general cellular metabolic status as well the specific biomolecular composition and conformation of a cell at a specific state. Furthermore, the FTIR spectra could be acquired from a cellular population, after a simple dehydration step, and in a high-throughput mode. If FTIR-spectroscopy reveals as a very useful technique to analyze and monitor the infection process, it could be applied in high-throughput mode, to further promoting the evaluation of interrelationships, for example, between different virulent factors on the infection process.

The present work aims to evaluate, based on spectral data, the effect of different *H. pylori* strains, with different CagA/VacA genotypes, and associated with different pathologies (NUD, PUD and GC), on the general metabolism of AGS cells. With that goal, FT-MIR spectra of AGS cells infected with different *H. pylori* strains were acquired, and the following spectral analysis were conducted:

i) Principal component analysis (PCA) to visualize tendencies in the samples' distribution in the space of the principal components explaining the most variance in data, that might be related to the *H. pylori* strain causing the infection, namely its virulence and disease;

ii) Clustering analysis, using the k-means algorithm to classify samples into groups based on spectral data, that can be related, again, to the infection, *H. pylori* virulence and related disease;

iii) Analysis of samples' first derivative spectra with the goal of not only finding spectral changes associated to infection, to be used as biomarkers, but also spectral changes that can be related with CagA/VacA genotypes

IV.3.2 Materials and Methods

Cells and culture conditions

AGS cells were grown in an CO_2 incubator (Blinder CB150, Tuttlingen, Germany) at 37˚C, 5% CO_2 and 99% of humidity. Cells were cultured in Roswell Park Memorial Institute Medium (RPMI) (Thermo Scientific), supplemented with 10% (v/v) of bovine fetal serum (BFS) (Thermo Scientific), previously inactivated by heat (30 minutes at 56˚C), 1% (v/v) of penicillin/streptomycin and 15% (v/v) of L-glutamine (Thermo Scientific), in 75 cm^2 T-flasks. Cells were cultured until they had reached about 80-90% of confluence. After 1 passage, cells were released with trypsin (incubated for 15 minutes at 37˚C) and about 1×10^6 cells were transferred to each well of a 6-wells plate for the infection procedures, which took place the day after the cells were plated into the wells.

For the infection experiments 10 strains of *Helicobacter pylori* were used (Table IV.5). *H. pylori* was grown in Columbia agar (Thermo Scientific), supplemented with horse blood (Thermo Scientific), for 36h, under microaerophilic conditions.

Table IV. 5 – Strains used for the infection experiments, characterized according the pathology associated and the expression (+) or not (-) of the CagA gene and the presence of the genotype s1 (+) or s2 (-) of the VacA gene, respectively. All *H. pylori* were isolated from Portuguese patients, with the exception of J99, isolated from an American patient. GC – Gastric cancer, PUD – Peptide ulcer disease, NUD – Non-ulcer dyspepsia.

Strains	CagA/VacA	Pathology
JP1	-/-	GC
P3/92	-/-	GC
JP26	-/+	GC
JP22	+/+	GC
J99	+/+	PUD
1152	+/+	PUD
93/00	+/+	PUD
147	-/+	PUD
173	-/-	NUD
228	-/-	NUD

Infection

Before infection, the AGS cells were washed 2 times with PBS and the media was replaced by RMPI, supplemented with FBS (10% v/v) and L-glutamine (15% v/v), but without antibiotics. Bacteria were harvested with this same media and the estimation of the desired number of bacteria was obtained by optical density, at 600nm, using the following relation (Zhang *et al*., 2007):

$$1OD_{600nm} = 10^6 \ bacteria/mL \qquad \text{(Equation IV.4)}$$

AGS cells were infected with different *H. pylori* strains with a MOI of 100 (100 bacteria for each AGS cell). The plates were kept in an CO_2 incubator (Blinder CB150, Tuttlingen, Germany) at 37˚C, 5% CO_2 and 99% of humidity, to ensure the viability of gastric cells. After 24h FTIR analysis were performed.

Spectral acquisition

For FTIR spectral measurements the sample cells were released with trypsin (infected AGS cells and control), centrifuged 15 minutes at 1500 RPM and resuspended in 200µL of PBS. Then, 25 µL of each sample was transferred for a 96-wells KBr plate for the FT-MIR high-throughput measurements. The samples were dehydrated for about 2 and a half hours, in a desiccator under vacuum, before spectral acquisition. The spectral data were collected using a FTIR spectrometer (Burker, HTS-XT) equipped with an HTS accessory. Each spectra represent sixty-four scans, with a 2 cm^{-1} resolution, and were collected in transmission mode, in the wavenumber region between 400 and 4000cm^{-1}. Five replicates were conducted to ensure the reproducibility of spectral information.

Spectral data analysis

Data pre-processing, including multiplicative scatter correction (MSC), 1st and 2nd derivatives, principal component analysis (PCA) and cluster analysis were carried out using Matlab R2012b (Matworks, Natick, MA, USA). Derivatives were computing using Savitzky-Golay algorithm, with a filter window of 15 data points and a 2nd order polynomial. Baseline correction was carried out using OPUS software (Bruker, Germany). The PCA models where the replicates are closer together and it was possible for extracting meaningful information about the infection event, were considered as being the ones with the best performance. For clustering, the performance of the models built was evaluated based on the mean silhouette value. Silhouette value is a measure of how similar a sample is to the rest of the samples in its own cluster, providing, in this way, information about the strength of the clusters (Lopes and Wolff, 2009).

IV.3.3 Results and Discussion

In order to evaluate if FTIR spectroscopy could be applied for studying the effect of *H. pylori* infection in AGS cell lines, the following spectral analysis were conducted

- Principal Component Analysis (PCA) models;
- Clustering analysis using the k-means algorithm;
- Analysis of the first derivative spectra.

These analyses were performed with the goal of correlating the spectral data with the virulence factors, VacA and CagA, as well as the gastric pathology from which the strains were isolated.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a data reduction method often used in spectral data analysis that decompose the spectral data into new variables, called principal components (PCs), that capture most variance in the data (Jollife, 2002).

PCA models were performed with the goal of finding meaningful relationships between the spectral data and the effect of *H. pylori* infection in AGS cells, e.g., the effect of the strains' virulence or the pathology associated to each strain. Prior to PCA, data were pre-processed with different methods, namely, baseline correction, multiplicative scattering correction (MSC), normalization and 1st and 2nd derivatives. MSC is often applied when working with dehydrated samples, as it helps eliminating spectral alterations that are not related with the property of interest but with undesirable light scattering events. Normalizing the spectra is also very useful, as it is virtually impossible to ensure that all samples have the same number of cells. Differences in the cell number can change the spectral intensity and lead to misleading results. First and second derivatives resolve overlapped bands and eliminate spectral offsets, providing valuable information that sometimes cannot be observed in the raw initial spectra (Otto, 1999). Also, PCA performed on selected spectral regions, as regions associated with protein and DNA absorptions, was evaluated, with the goal of capturing the most relevant information associated with *H. pylori* infection.

Figure IV.31 shows the PCA scores plot from data pre-processed with MSC, applied to the replicates using as the reference spectrum the mean spectrum of all replicates, and 2nd derivative. Four groups of samples can be observed, that are related to the day/week at which the experiments were conducted. The infection experiments were developed in four different sets, being that AGS cells were infected with different *H. pylori* strains each set: in the first set, cells were infected with J99 and JP1 strains; 1152, P3/92 and 93/00 strains were used in the second set; 147, JP26 and JP22 strains were used in the third set; and 228 and 173 strains were used in the last set. This grouping of the samples according to the date of the experiments is probably related to the metabolic status of the AGS cells, accounting for most variability in data. In fact, it has been reported that the cell state can be reflected in spectral data (Boydston-White *et al.*, 2009) and it is not possible to ensure that the cells are in the same cell stage in all experiments, even when the procedures are the same. For this reason, control cells are not close together in the space of principal components, as expected, but rather grouped with the corresponding infected cells in the same experiment. For evaluating the effect of the day of the experiment, it would be interesting to perform an additional infection experiments, in two different days, but using the same *H. pylori* strain. However, within the same group corresponding to each experiment set, control cells are still apart from the corresponding infected cells. In the second PC, in the PC1 *versus* PC2 scores plot, (Figure IV.31), two groups of samples can also be seen, one composed of cells infected with 173 and 228 strains, which are CagA-/VacA- associated to non-ulcer dyspepsia (NUD), and the other composed by the remaining samples.

In order to be able to visualize tendencies in the PCA scores plots related with *H. pylori* infection, the data were pre-processed in a different way. MSC was applied to the replicates of each strain, as for the above PCA, and applied to all the control samples, being the reference spectrum the one resulting from the mean of all spectra from control cell samples, with the goal of eliminating differences between cells that can be related with the cellular stage, as consequence of the experiments being conducted in different days. Clear grouping of samples according to the infection and the *H. pylori* strains responsible for the infection could then be observed (Figure IV.32). The control samples became all aligned in the first PC, explaining 78.6% of the variance in data, separated from the infected cells. Also, the samples infected with different strains were more clearly separated from each other. Interestingly, the cells infected with *H. pylori* strains associated with non-ulcer dyspepsia (NUD) and CagA-/VacA-, 228 and 173 strains, were closer to the controls than the samples infected with other strains associated with peptide ulcer disease (PUD) and gastric cancer (GC). Additionally, all the associated with PUD and having a CagA+/VacA+ genotype (1152, 93/00 and J99), were closer together, with all the strains with a CagA-/VacA- genotype, associated to GC (P3/92 and JP1). The remaining samples in the right side of the PC1 axis, accounted for more variability with respect to the associated disease (GC or PUD) or the genotypes concerning CagA and VacA, and were strains from GC with VacA+ and a strain from PUD, with CagA-.

Although the grouping of sample in the PCA scores plot revealed interesting tendencies, it is not possible to affirm that the grouping of the infected AGS cells observed was only related with the CagA/VacA genotype or the pathology associated. Indeed, the development of a pathology associated with *H. pylori* infection is dependent on several factors, such as, the host immune response and the host environment (Kusters *et al*., 2006; Amieva e El-Omar, 2008). To further explore the tendencies in data would be necessary to use a higher number of *H. pylori* strains.

Although some important factors were not mimicked in the present study, the results presented are very promising, as shown by the ability of FTIR spectroscopy to identify infected cells and infection according to the *H. pylori* strain responsible.

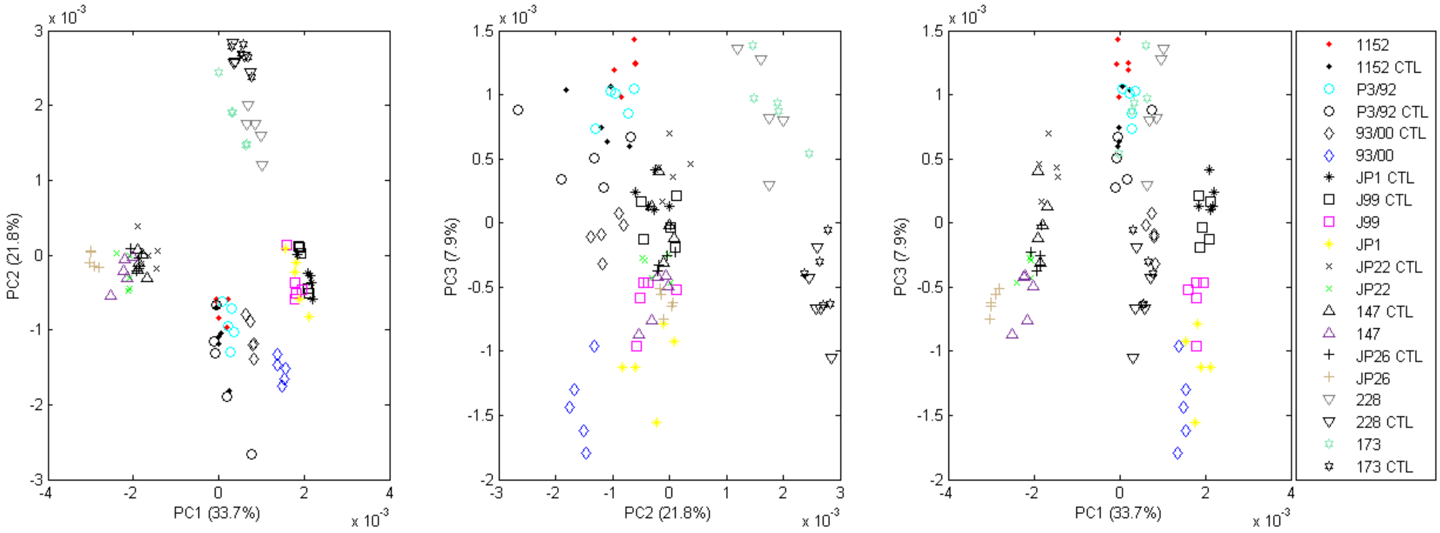

Figure IV. 31 - PCA of data from the infection experiment. AGS cells were infected with different *H. pylori* strains (MOI=100): 1152, P3/92, 93/00, JP1, J99, JP22, 147, JP26, 228 and 173. For each infection experiment using a single strain, a control was used (CTL), corresponding to AGS cells not infected. The data was pre-processed with MSC and 2nd derivative and the entire spectral window was considered.

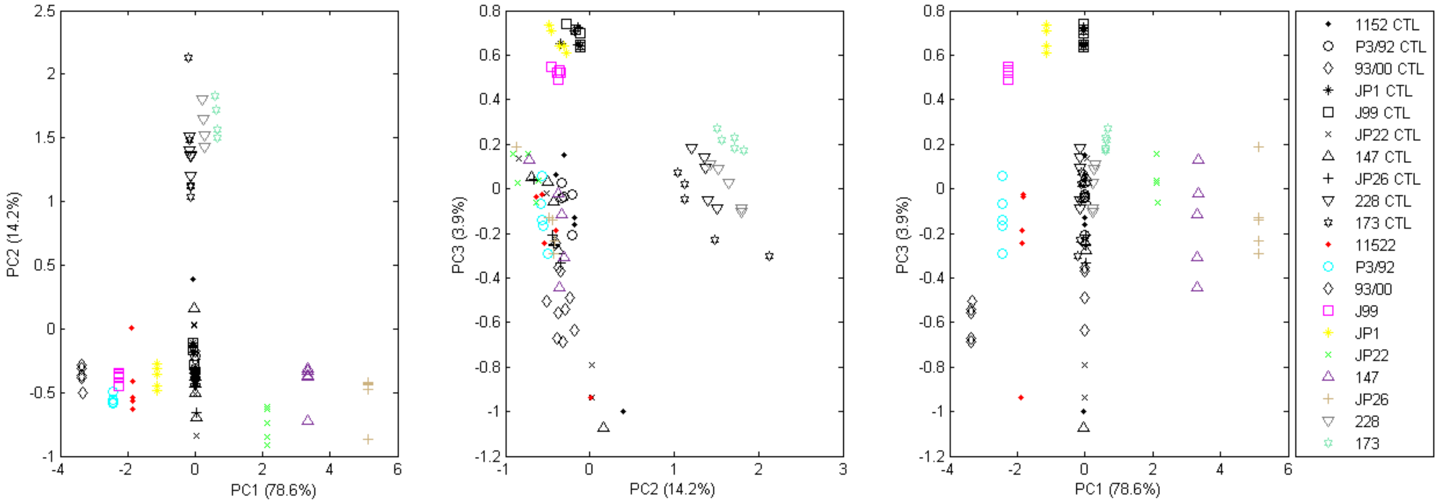

Figure IV. 32 - PCA of data from the infection experiment. AGS cells were infected with different *H. pylori* strains (MOI=100): 1152, P3/92, 93/00, JP1, J99, JP22, 147, JP26, 228 and 173. For each infection experiment using a single strain, a control was considered (CTL). MSC was applied to all the control samples and to the replicates of each strain.

Cluster analysis

Clustering analysis was evaluated in the present work in an attempt of classifying AGS cells, according to the strain used. Although, PCA allowed a visualization of some grouping in data, PCA is not a truly a classification method, as it does not provide a clear cut-structure that allow a real classification of the samples, therefore being somehow subjective.

Clustering is a non-supervised classification method and it is a tool that groups data into clusters according to their semblance, usually determined by pattern recognition algorithms that rely on distance measures. The shorter the distance between two objects or samples, the closer they are. A cluster describes a group where the samples are more similar to each other than to those outside the group (Otto, 1999). In this work, cluster analysis was performed using the k-means algorithm. With this algorithm, each cluster is defined by its elements and its centroid, the point for which the sum of squares of the distances of all cluster's elements is minimum. Thus, the objects are arranged between the groups or clusters in a way that the distance to the cluster's centroid is minimized (Seber, 1984).

To date, some authors have already applied clustering with classification purposes. Some examples include the discrimination of counterfeit drugs (Lopes and Wolff, 2009), study of bacteria isolated from sputum samples from cystic fibrosis patients (Bosch *et al*., 2008) and pre-screening of childhood acute leukemia (Zelig *et al*., 2011).

The results of cluster analysis are represented in Figure IV.33 with data pre-processed with MSC, applied to all the control cells and to each replicate of the infected cells. The optimal number of clusters was assessed by the combinations of a high mean silhouette value, and a non-partition of control samples in different groups, meaning that the variability within the control cells (not infected AGS cells), a result of different cellular stages, was not being considered for grouping purposes. The data was fractionated in four groups, as it was found that for a higher number of groups the control group became fragmented. A high mean silhouette was obtained (0.712), indicating that a strong grouping structure was found (Lopes and Wolff, 2009). It was found that the 228 and 173 strains were in the same group as the control cells (Cluster 1 – Figure IV.32). Actually, for the PCA results (shown before) the cells infected with these strains, associated with NUD, were also closer to the control than to the other infected populations. A strong second cluster can also be seen in Figure IV.31, with all strains with a CagA+/VacA+ genotype, and associated to PUD, and all the strains CagA-/VacA-, and associated with GC, grouped in this cluster (Cluster 4 – Figure IV.33). Although closer to the neighbor samples in group 4, cells infected with JP1 are the most different samples in group 4. The only difference observed during the experimental work was that JP1 was the strain that always presented the faster grown *in vitro*. Also, other factors related with *H. pylori* infection and the variability among the different strains were not considered.

As mentioned for PCA results, it is dangerous to associate these results exclusively with the CagA/VacA genotype and with the pathology associated to each strain. However, the groups produced by cluster analysis are not yet subjective, so if strains are grouped together they must have a similar effect on AGS cells. Overall, the PCA and the clustering analysis yielded similar results and an interesting grouping of some *H. pylori* strains, essentially those with a CagA+/VacA+ genotype associated to PUD and a CagA-/VacA- associated with GC.

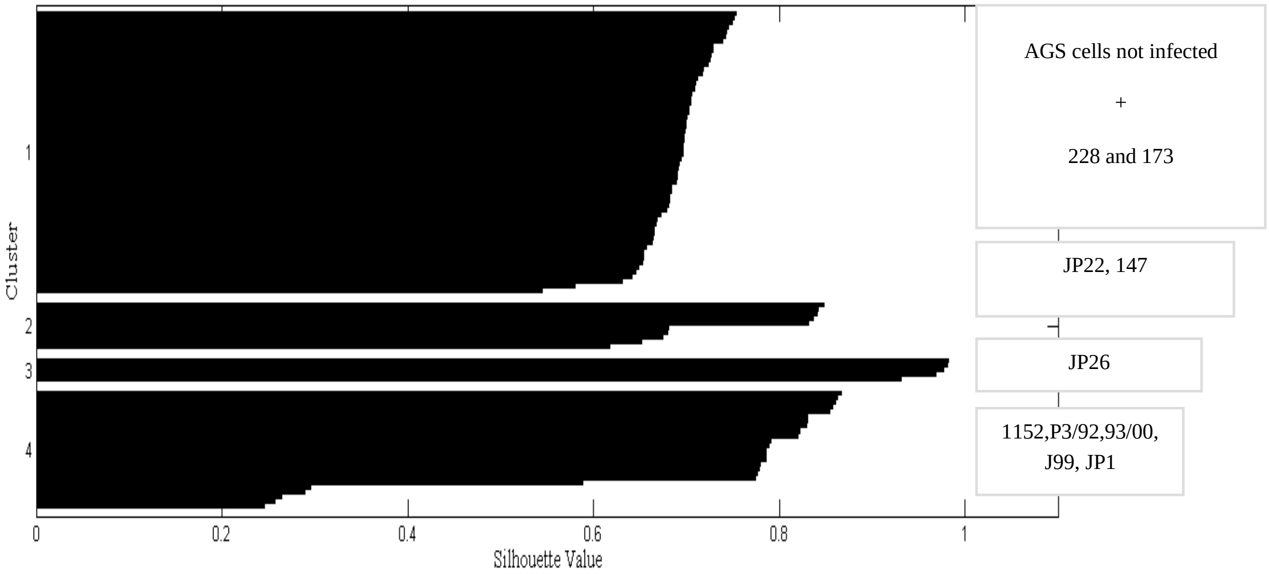

Figure IV. 33 – Silhouette plot of the 4 groups found with k-means analysis on all control cells and cells infected with different *H. pylori* strains, with 5 replicates each. Data were pre-processed with MSC, applied to all the control cells and applied to the replicates of each infected cell

Analysis of the first derivative spectra

First derivative spectra were also considered for the analysis of the impact of the infection according with the *H. pylori* strain. First derivative spectra can offer valuable information that sometimes can't be accessed through the analysis of the initial spectra, by increasing the spectra resolution over overlapping bands (Otto, 1999). The goal was to find differences in the 1st derivative spectra related to the general infection and with the strain causing the infection. The spectra of control cells, AGS cells not infected with *H. pylori*, were compared with the spectra of AGS cells infected with different *H. pylori* strains, namely: 1152, P3/92, 93/00, JP1, J99, JP22, 147, JP26, 228 and 173.

Consistent differences between infected and non-infected cells were found. The changes in the 1st derivative spectra caused by the infection with *H. pylori*, were mainly in Amide I band at 1687 cm^{-1} (Walsh *et al.*, 2009), a protein absorption band, and also a band associated with glicids absorption, at 1025 cm^{-1} (Wong *et al.*, 1991). Alterations in bands associated with protein absorption, can be a result of the metabolic stress due to infection and/or due to the injection into the cell, by the bacteria, of CagA and VacA proteins. Due to *H. pylori* infection, cell will change its metabolic status, consuming more energy, depleting its energy reserves, as glicids. In Figure IV.34 is the first derivative spectra of control cells and cells infected with *H. pylori* 1152 (Figure 34 – A) and 228

(Figure 34 – B), with the bands mentioned above highlighted. These bands changed as a result of infection for all the 10 strains used (Figures IV.34 to IV.41).

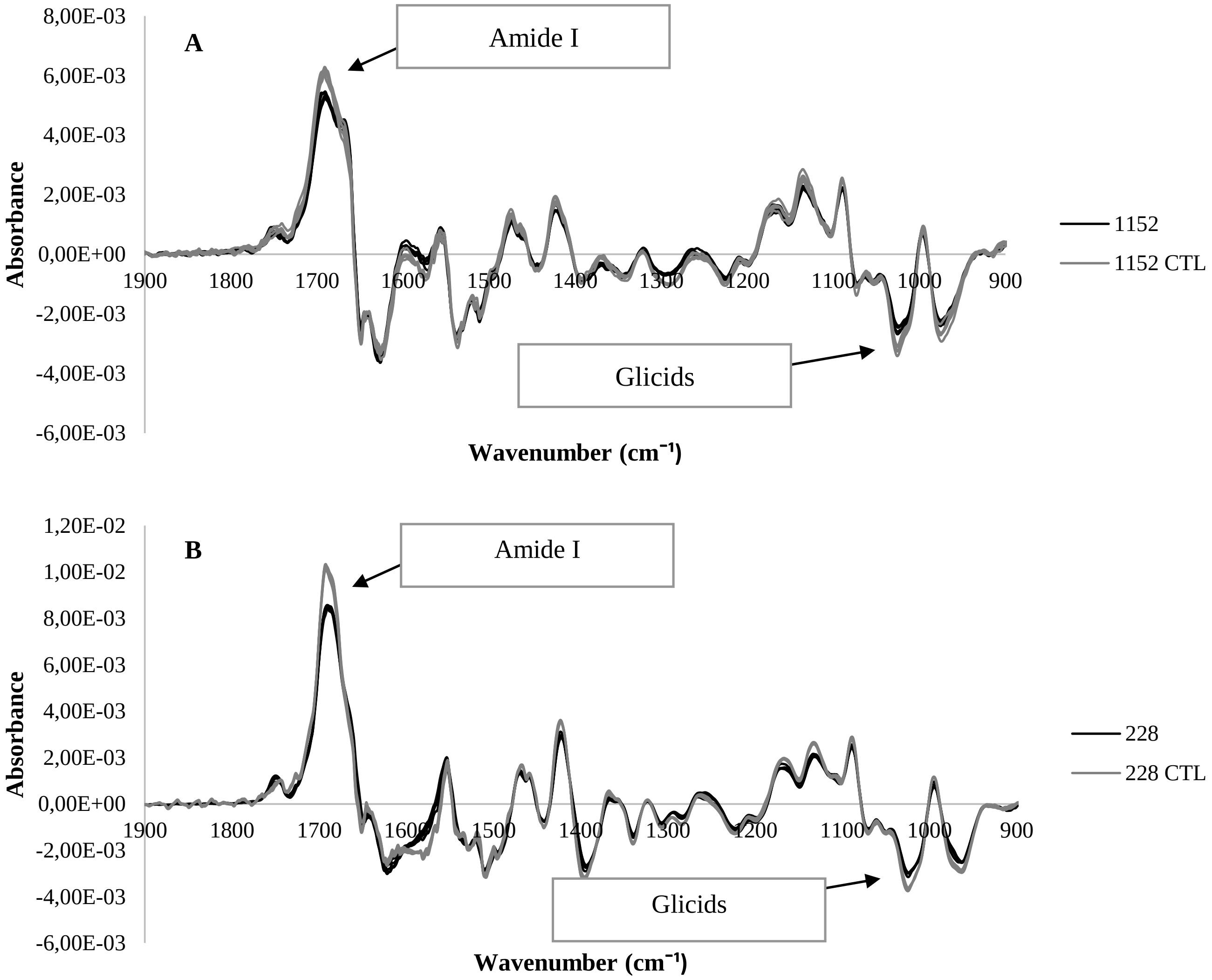

Figure IV. 34 - First derivative spectra of non-infected cells and cells infected with *H. pylori* 1152 (A) and 228 (B), with bands associated with protein and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

As expected, some differences, as the ones mentioned above, appeared in all the 1st derivative spectra of cells infected with the different strains. However, specific regions of the 1st derivative spectra, related to each strain, were also found.

H. pylori 93/00, presenting a CagA+/VacA+ genotype and isolated from a PUD patient, showed to be the strain with the strong impact on AGS cells, according to the 1st derivative spectra (Figure IV.35). Greater differences were observed in bands at

1648 $cm^{-1}$and 1624 cm^{-11}, associated to Amide I, and 1537 cm^{-1}associated to Amide II, all bands associated with protein absorption (Walsh *et al.*, 2009). Additional differences in the first derivative spectra can be seen in bands associated with DNA absorption at 1169 cm^{-1}, 1080 cm^{-1} and 970 cm^{-1} (Walsh *et al.*, 2009; Wong *et al.*, 1991; Liu *et al.*, 2005). Another band associated with glicids, at 1169 cm^{-1} (Mordechai *et al.*, 2003) also changed for cells infected with *H. pylori* 93/00. The reason of the changes observed for the cell's proteins and glicids content had already been proposed here.

For infection with *H. pylori* P3/92, presenting a CagA-/VacA- genotype and isolated from a GC patient, the main differences were observed for bands associated with glicids and DNA absorptions, with no so dramatic alterations in the protein absorption regions (Figure IV.36). For the strains 173, JP1 and JP26 (Figures IV.37, IV.38 and IV.39, respectively), form patients with gastritis and gastric cancer, respectively, the main differences in the 1st derivative spectra were in bands associated with protein absorption. For cells infected with *H. pylori* J99 smaller differences were observed in the 1st derivative spectra and all associated with protein absorption (Figure IV.40), that could be a result from the fact that this strain presents a higher number of passages through laboratories after its isolation from an american PUD patient. The other strains used in the present work were isolated recently from Portuguese patients. Interestingly, for cells infected with *H. pylori* 147 (isolated from a GC patient and presenting a CagA-/VacA+ genotype) and JP22 (isolated from a PUD patient and presenting a CagA+/VacA+ genotype, there are practically no changes in the first derivative spectra (Figures IV.41 and IV.42, respectively). Probably 24h were not enough for seeing the effect of the infection.

Also, some differences in the 1st derivative spectra of the control cells can also be observed. Cells were treated equality, however, as the experiments were developed in different days, the cells may had different metabolic status at the time and these differences eventually were reflected by the spectra.

Overall, the results provided by the analysis of the 1st derivative spectra were very interesting. It was possible to see differences related with infection, independently of the strain used, mainly in bands associated with protein and glicids absorption. Additionally, differences in the 1st derivative spectra related with the strains were also found, in regions associated to proteins and DNA absorption. This approach can be very useful for studying the effect of *H. pylori* infection. It would be also interesting to evaluate infection in different time points and also evaluating different strains.

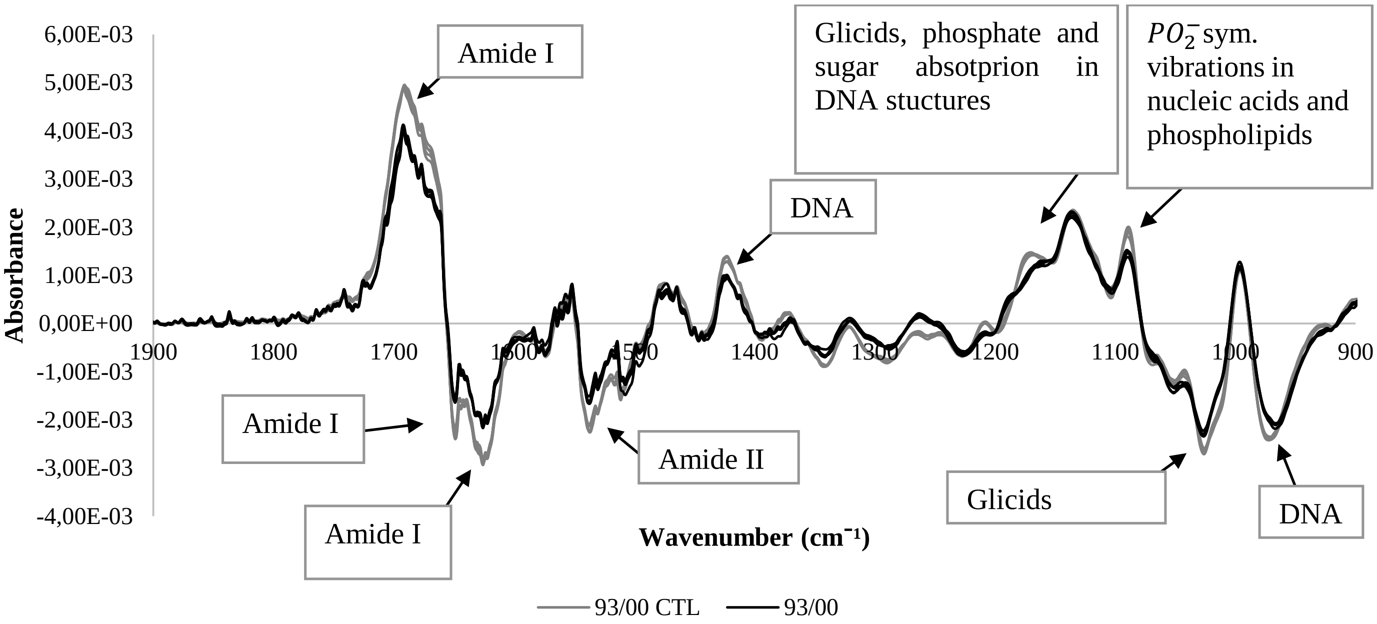

Figure IV. 35 - First derivative spectra of non-infected cells and cells infected with *H. pylori* 93/00, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

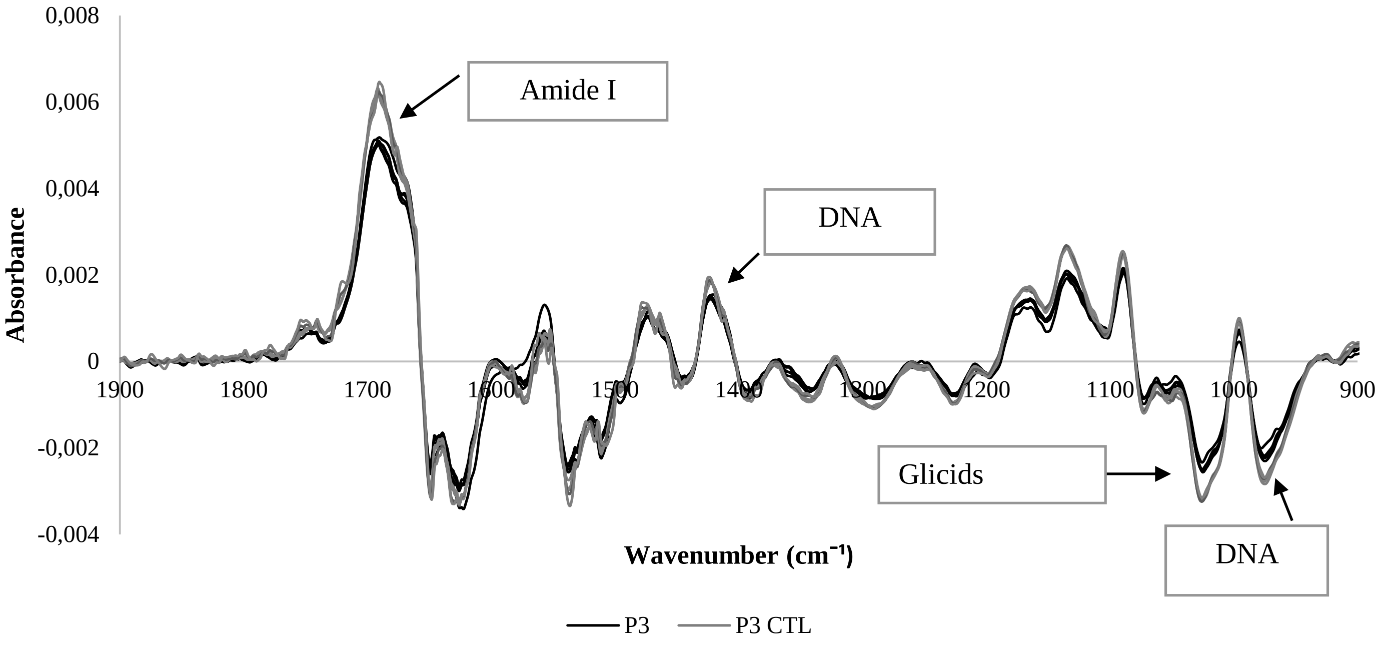

Figure IV. 36 - First derivative spectra of non-infected cells and cells infected with *H. pylori* P3/92, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

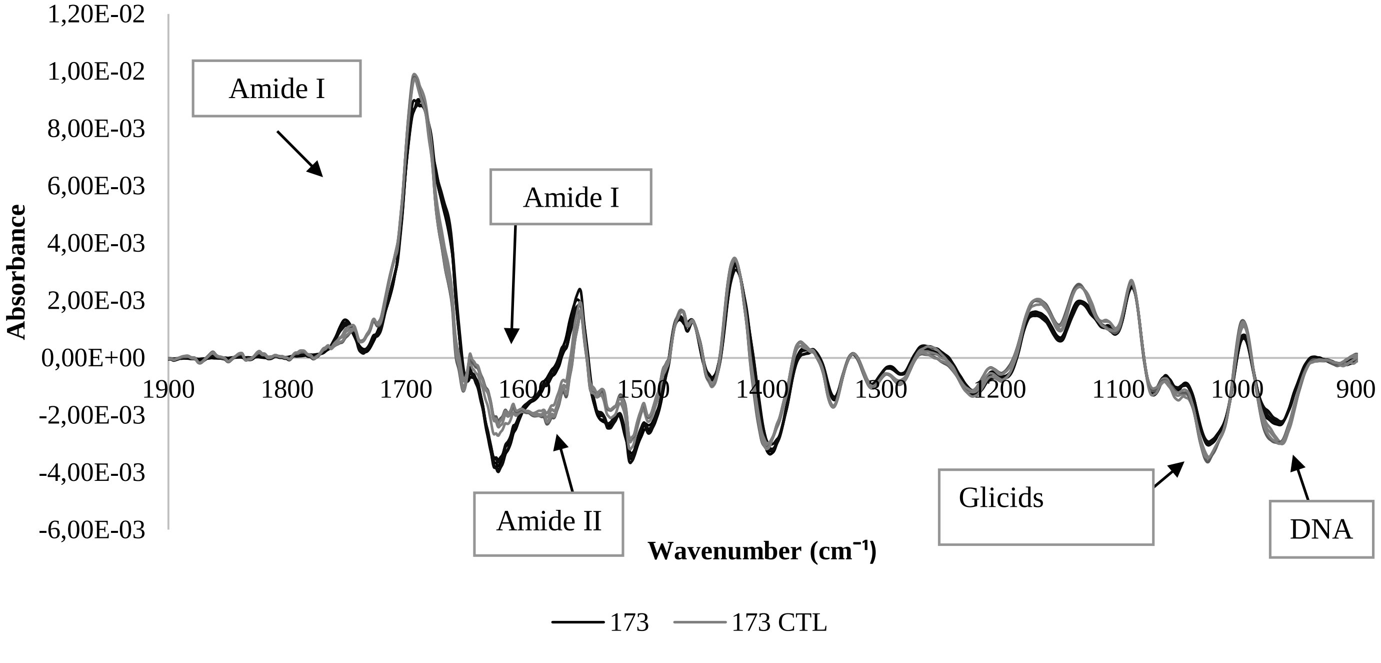

Figure IV. 37 - First derivative spectra of non-infected cells and cells infected with *H. pylori* 173, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

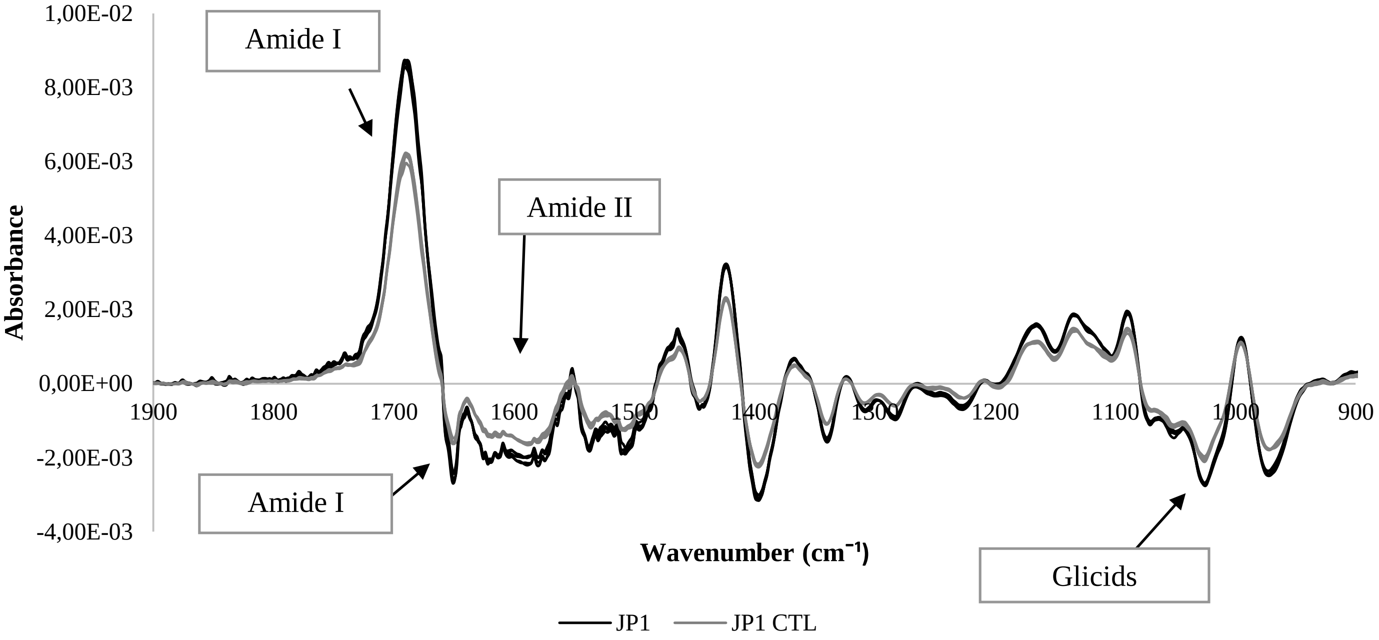

Figure IV. 38 - First derivative spectra of non-infected cells and cells infected with *H. pylori* JP1, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

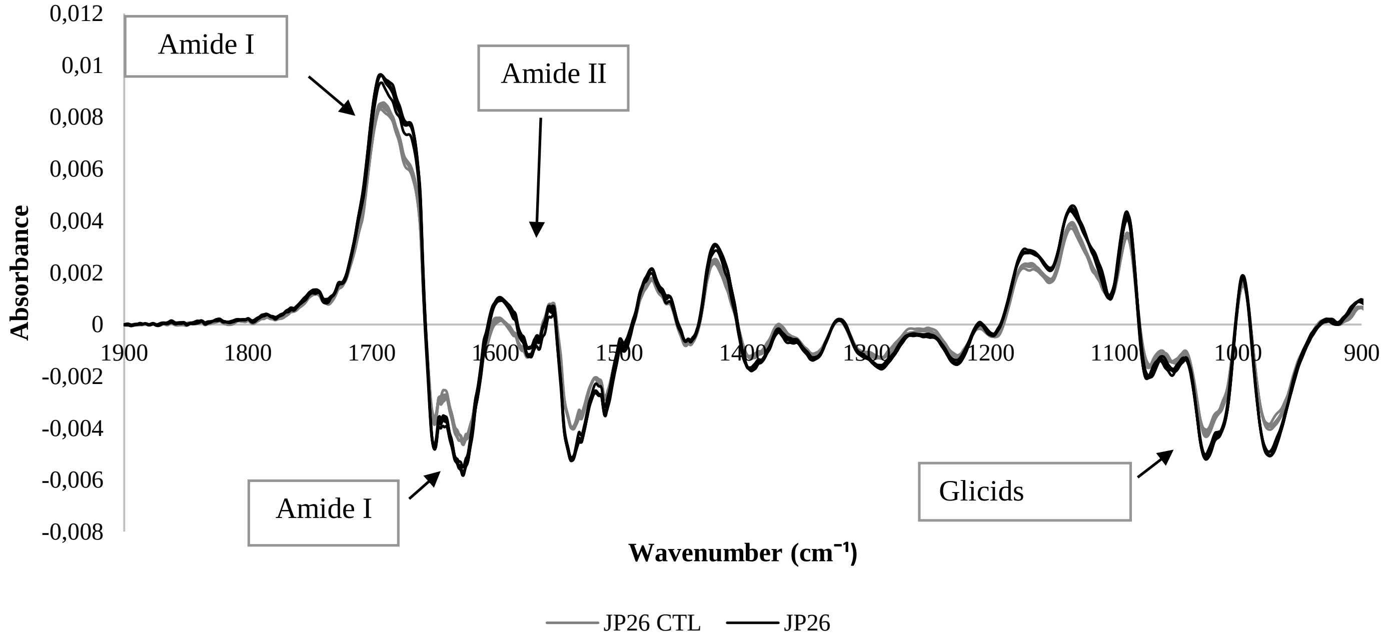

Figure IV. 39 - First derivative spectra of non-infected cells and cells infected with *H. pylori* JP26, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

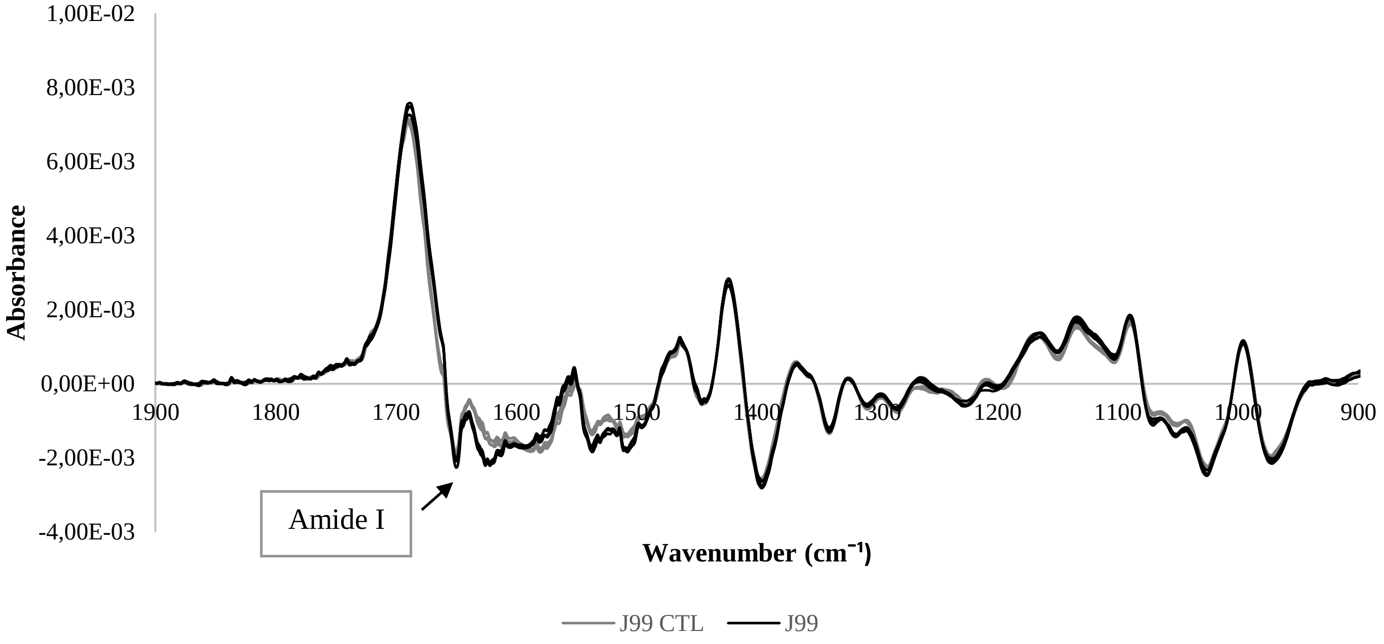

Figure IV. 40 - First derivative spectra of non-infected cells and cells infected with *H. pylori* J99, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

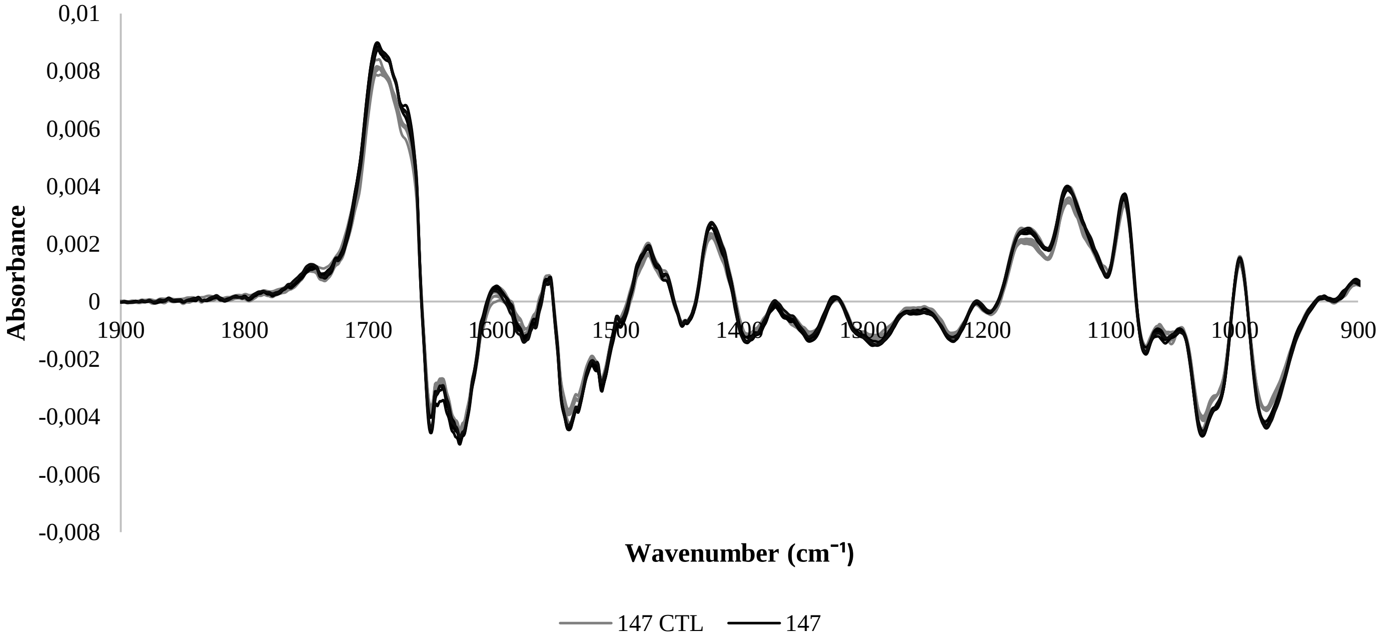

Figure IV. 41 - First derivative spectra of non-infected cells and cells infected with *H. pylori* 147, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

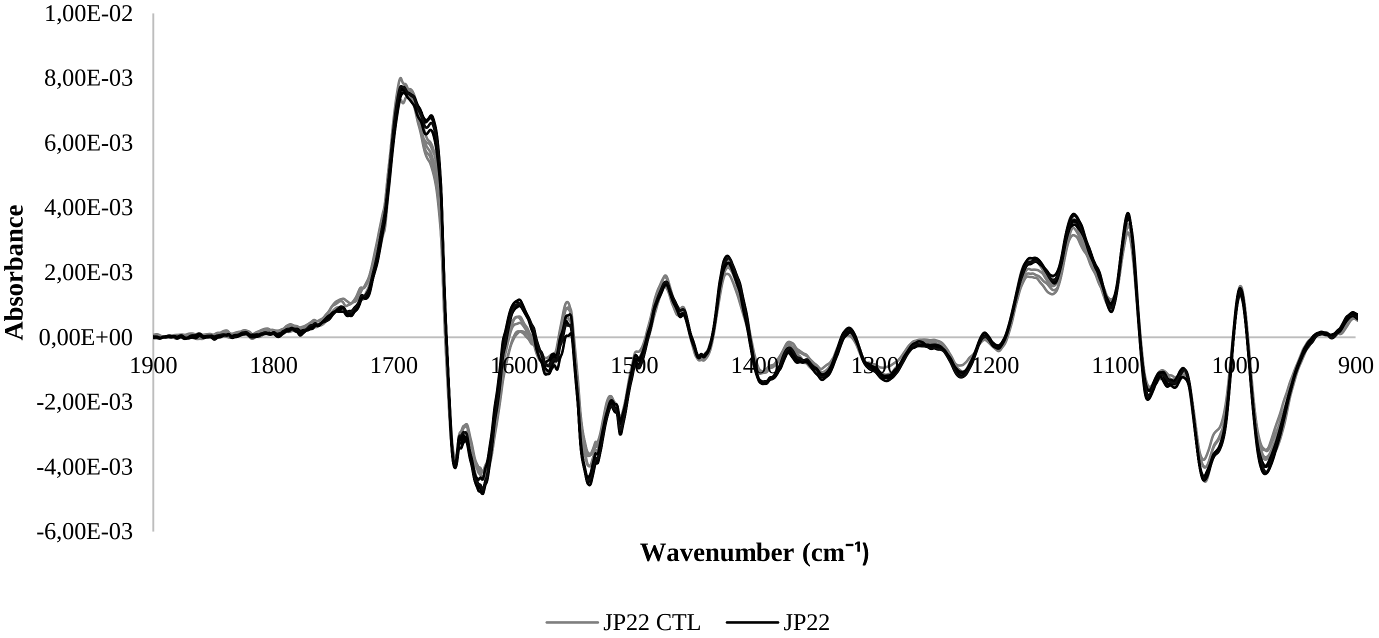

Figure IV. 42 - First derivative spectra of non-infected cells and cells infected with *H. pylori* JP22, with bands associated with protein, DNA and glicids absorptions highlighted. MSC was applied to the replicates and 1st derivative spectra was determined using the Savitzky-Golay algorithm, using a 15 points window and a 2nd order polynomial.

IV.3.4 Conclusions

The development of the different pathologies associated with *H. pylori* infection is dependent of several factors, such as the virulence of the strains, the immune response of the host and of environmental factors. The way the bacteria infects the human gastrointestinal tract and, more important, the mechanism by which it interacts with the host and leads to development of a variety of pathologies is not well understood yet. Additionally, there is no effective approach for eliminating *H. pylori* and to prevent the development of the disease (Kusters *et al.*, 2006). Thus, it is opportune to study, using different approaches, not only the mechanism by which the infection leads to the disease, but also the impact of the infection in the host's cells.

In the present work the potential of FTIR spectroscopy for studying the impact of *H. pylori* infection on gastrointestinal cells, AGS cells, was evaluated. Different approaches were conducted, yielding very interesting results. Principal component analysis (PCA) was performed for trying to differentiate not only infected and non-infected cells, but also the infection caused by different *H. pylori* strains. Cluster analysis was also conducted with the same purpose of PCA, however clustering is a true classification tool that offers a measure of how strong the groups created are. Again, an interesting grouping of the samples was achieved, namely for the *H. pylori* strains with a CagA+/VacA+ genotype associated with the development of peptide ulcer disease (PUD) and with a CagA-/VacA- genotype associate with the development of gastric cancer (GC). Additionally, it was observed from the analysis of the 1st derivative spectra differences related with the infection by *H. pylori* and other specific differences related with the strains infecting the AGS cells.

Overall, the results provided by the present work strongly suggest the high potential of FTIR spectroscopy for studying the infection by *H. pylori*. In the future it would be interesting to test the ability of FTIR spectroscopy to evaluate the infection by others *H. pylori* strains. Also, additional analysis on the effectiveness of *H. pylori* infection, such as the adherence of the bacteria to the AGS cells, and different infection times should also be evaluated.

V. GENERAL CONCLUSIONS

In the last decade infrared (IR) spectroscopy started to be seen as a very powerful tool for several applications. The main reasons for that are the fact that this technique gives information about vibrational states of molecules in a highly sensitive and fast way, is reagent free, it allows high-throughput measurements, it measures several analytes at once from a single spectra with a lower human error compared to the conventional techniques, and it allows extracting relevant biochemical and physiological information (Shaw and Mantsch, 1999; Ellis and Goodacre, 2006).

IR spectroscopy may use near-infrared (NIR) and mid-infrared (MIR) radiations. MIR spectroscopy provides more informative spectra, since the bands essentially arise from fundamental molecular vibrational states, making this method more sensitive, while allowing an easier interpretation of the spectral data. Since water strongly absorbs IR radiation in the MIR gion, it is usually necessary an extra step by which the sample is dehydrated or, alternatively, using an attenuated total reflection (ATR) accessory. On the other hand, MIR spectroscopy has the great advantage of working in a high-throughput mode, by using microplates with multi-wells. NIR-spectroscopy, is less informative as represents overtones and combinations of the fundamental vibrations, in spite of being less prone to water interference. Therefore, to achieve a high quality information concerning biological processes it is usually preferable to use MIR spectroscopy. One exception to that would be the *in-situ* monitoring of cultures of living cells inside bioreactors vessels, using fiber-optic probes working on the NIR region of the spectrum. But even on that case, NIR spectroscopy can have a low sensitivity, especially when monitoring mammalian cell cultures, where the concentration of key analytes, such as glucose or lactate, are usually very low (Smith, 2011; Lourenço *et al*, 2012).

The importance of chemometrics was also a subject of discussion in the present work. Chemometrics is a field that combines mathematical, statistics and computational methods, that make possible to extract relevant information from the data, which otherwise would be very difficult. The pre-processing techniques applied to the spectral data acquired for this work are briefly reviewed, namely baseline correction, multiple scatter correction (MSC), normalization, smoothing and derivatives. These methods allow to reduce the noise in data and undesirable effects like differences in sample´s thickness, differences in cell´s number or radiation scattering, while highlighting relevant information in data. Other chemometric methods applied to the data are also described, namely principal component analysis (PCA), clustering and partial least squares (PLS) regression.

The second part of the present Thesis describes the experimental work conducted. The applications described in this work involved the evaluation of Fourier Transform Infrared (FTIR) spectroscopy for studying several mammalian cell associated processes, namely:

- Monitoring the expansion process of human mesenchymal stem cells directly obtained from human donors, and conducted in spinner flasks with different microcarriers and feeding regimes, for estimating key analytes in cell culture, e.g., glucose, lactate and ammonia;
- Estimating the transfection efficiency in a cell population using two distinct cell lines: an adherent cell line (AGS) and an semi-adherent cell line (HEK), and grown on two distinct media compositions;
- Studying *Helicobacter pylori* infection, *in vitro*, using AGS cell lines, using different *H. pylori* strains, with different CagA and VacA genotypes and isolated from patients with different gastric pathologies.

In general, analysis based on spectral data provided not only accurate quantitative models, classification models and qualitative biochemical information, as discussed in the conclusions sub-chapter of each main section, which will not be replicate in this Final Conclusions chapter.

Very interesting results were achieved in the present work, showing the ability of the technique for monitoring mammalian cells’ processes, which therefore could strong promote the future use of the FTIR spectroscopy to evaluate biological processes.

VI. BIBLIOGRAPHY

- Akhter Y., Ahmed I., Devi S.M., Ahmed N. (2007). The co-evolved *Helicobacter pylori* and gastric cancer: trinity of bacterial virulence, host susceptibility and lifestyle. *Infectious Agents and Cancer* 2:2
- Ami D., Bonecchi L., Cali S., Orsini G., Tonon G., Doglia S.M. (2003). FT-IR study of heterologous protein expression in recombinant *Escherichia Coli* strains. *Biochimica et Biophysica Acta* 1624:6-10
- Ami D., Neri T., Natalello A., Mereghetti P., Doglia S.M., Zanoni M., Zuccoti M., Garagna S., Redi C.A. (2008). Embryonic stem cell differentiation studied by FT-IR spectroscopy. *Biochimica et Biophyisica Acta* 1783:98-106
- Amieva M.R., El-Omar E.M. (2008). Host-bacterial interactions in *Helicobacter pylori* infection. *Gastroenterology* 134:306-323
- Arnold S.A., Crowley J., Woods N., Harvey L.M., McNeil B. (2003). In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation. *Biological and Bioengineering* 84:13-19
- Azevedo N.F., Almeida C., Cerqueira L., Dias S., Keevil C.W., Vieira M.J. (2007). Coccoid form of *Helicobacter pylori* as a morphological manifestation of cell adaptation to the environment. *Applied Environmental Microbiology* 73:3423-3427
- Babrah, J. (2009). A study of FT-IR spectroscopy for the identification and classification of hematological malignancies. PhD Thesis, Cranfield University, United Kingdom
- Belancio V.P. (2011). Importance of RNA analysis in interpretation of reporter gene expression data. *Analytical Biochemistry* 417:159-161
- Bosch A., Miñán A., Vescina C., Degrossi J., Gatti B., Montanaro P., Messina M., Franco M., Vay C., Schmitt J., Naumann D., Yantorno O. (2008). Fourier Transform Infrared Spectroscopy for Rapid Identification of Nonfermenting Gram-Negative Bacteria Isolated from Sputum Samples from Cystic Fibrosis Pateints. *Journal of Clinical Microbiology* 46:2535-2546
- Card C. Hunsaker B., Smith T., Hirsch J. (2008). Near-infrared spectroscopy for rapid, simultaneous monitoring. *BioProcess International* 6:59-67

- Carmelo J.I.G. (2013). Optimizing the production of human mesenchymal stem/stromal cells in xeno-free microcarrier-based reactor systems. Master Thesis, Instituto Superior Técnico, Portugal
- Christian G.D. (1994). *Analytical Chemistry*. WILEY: United States
- Durocher Y., Perret S., Kamen A. (2002). High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-RBNA1 cells. *Nucleic Acids Research* 30:2-9
- Duygu D., Baykal T., Açikgöz I., Yildiz K. (2009). Fourier Transform Infrared (FT-IR) Spectroscopy for Biological Studies. *Journal of Science* 22:117-121
- Fearn T., Riccioli C., Garrido-Varo A., Guerrero-Ginel J.E. (2009). On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems* 96:22-26
- Gaudenzi S., Pozzi D., Toro P., Silvestri I., Morrone S., Castellano A.C. (2004). Cell apoptosis specific marker found by Fourier Transform Infrared Spectroscopy. *Spectroscopy* 18:415-422
- Gazi E., Baker M., Dwyer J., Lockyer N.P., Gardner P., Shanks J.H., Reeve R.S., Hart C.A., Clarke N.W., Brown M.D. (2006). A correlation of FTIR spectra derived from prostate cancer biopsies with Gleason grade and tumor stage. *European Urology* 50:750-761
- Geladi P. (2003). Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B* 58:767-782
- Giambattista L.D., Pozzi D., Grimaldi P., Gaudenzi S., Morrone S., Castellano A.C. (2011). New marker of tumor cell death revealed by ATR-FTIR spectroscopy. *Analytical and Bioanalytical Chemistry* 399:2771-2778
- Graça G., Moreira A.S., Correia A.J.V., Goodfellow B.J., Barros A.S., Duarte I.F., Carreira I.M., Galhano E., Pita C., Almeida M.C., Gil A.M. (In press). Mid-infrared (MIR) metabolic fingerprinting of amniotic fluid: A possible avenue for early diagnosis of prenatal disorders? *Analytica Chimica Acta*
- Griffiths P.R. (2002). Introduction to Vibrational Spectroscopy, in Chalmers J.M. and Griffiths P. (Eds.). *Handbook of Vibrational Spectroscopy*. Wiley: United States

- Haaland D.M., Thomas E.V. (1988). Partial least-squares Methods for spectral analysis. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* 60:1193-1202
- Hakemeyera C., Straussa U., Werza S., JosebG.E., Folque F., Menezes J.C. (2012). At-line NIR spectroscopy as effective PAT monitoring technique in Mab cultivations during process development and manufacturing. *Talanta* 90:12-21
- Hammond S.V. and Clarke F.C. (2002). Near-infrared Microscopy, in Chalmers J.M. and Griffiths P. (Eds.). *Handbook of Vibrational Spectroscopy*. Wiley: United States
- Harthun S., Matischak K., Friedl P. (1998). Simultaneous prediction of human antithrombin III and main metabolites in animal cell culture processes by near-infrared spectroscopy. *Biotechnology Techniques* 23:393-398
- Helland S., Naes T., Isaksson T. (1995). Related version of multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* 29:233-241
- Hsu C.-P.S. (1997). Infrared Spectroscopy, in Settle F. (Ed.). *Instrumentation Techniques for Analytical Chemistry*. Prentice Hall PTR: New Jersey, USA
- Jeyaseelan K., Ma D., Armugam A. (2001). Real-time detection of gene promoter activity: quantification of toxin gene transcription. *Nucleic Acids Research* 29:11-15
- Isomoto H., Moss J., Hirayama T. (2001). Pleiotropic actions of *Helicobacter pylori* vacuolating cytotoxin, VacA. *Tohoku Journal of Experimental Medicine* 220:3-14
- Jiang T., Xing B., Rao J. (2008). Recent developments of biological reporter technology for detecting gene expression. *Biotechnology and Genetic Engineering Reviews* 25:41-76
- Jolliffe I.T. (2002). *Principal Component Analysis*. Springer: New York, USA
- Kalinkova G.N. (1999). Infrared spectroscopy in pharmacy. *Vibrational Spectroscopy* 19:307-320
- Khoshhesab Z.M. (2012). Reflectance IR Spectroscopy, in Theophile T. (Ed.). *Infrared Spectroscopy - Materials Science, Engineering and Technology*. InTech: Croatia

- Kidder L.H., Haka A.S., Lewis E.N. (2002). Instrumentation for FT-IR Imaging, in Chalmers J.M. and Griffiths P. (Eds.). *Handbook of Vibrational Spectroscopy*. Wiley: United States
- Kondepati V.R., Heise H.M., Backhaus J. (2008). Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy. *Analytical and Bioanalytical Chemistry* 390:125-139
- Kusters J.G.,van Vliet A.H., Kuipers E.J. (2006). Pathogenesis of *Helicobacter pylori* infection. *Clinical Microbiology Reviews* 19:449-490
- Lamberti A., Sanges C., Arcari P. (2010). FT-IR spectromicroscopy of mammalian cell culture during necrosis and apoptosis induced by drugs. *Spectroscopy* 24:535-546
- Landgrebe D., Haake C., Höpfner T., Beutel S., Hitzmann B., Beutel S., Hitzmann B., Scheper T., Rhiel M., Reardon K.F. (2010). On-line infrared spectroscopy for bioprocess monitoring. *Applied Microbiology and Biotechnology* 88:11-22
- Le Blanc K., Rindgén O. (2005). Immunobiology of human mesenchymal stem cells and future use in hematopoietic stem cell transplantation. *Biology of Blood and Marrow Transplantation* 11:321-334
- Liu K., Shi M., Mantsch H.H. (2005). Molecular and chemical characterization of blood cells by infrared spectroscopy. A new optical tool in hematology. *Blood Cells, Molecules and Diseases* 35:404-412
- Lopes M.B., Sales K.C., Lopes V.V., Calado C.R.C. (2013). Real-time plasmid monitoring of batch and fed-batch *Escherichia coli* cultures by NIR spectroscopy. *Proceedings of the Third IEEE-EMBS Portuguese Bioengineering Meeting* (ENEBERG, Braga, Portugal)
- Lopes M.B., Wolff J.-C. (2009). Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging. *Analytica Chimica Acta* 633:149-155
- Lourenço N.D., Lopes J.A., Alemida C.F., Sarraguça M.C., Pinheiro H.M. (2012). Bioreactor monitoring with spectroscopy and chemometrics: a review. *Analytical and Bioanalytical Chemistry* 404:1211-1237
- Madeira C., Santos F., Andrade P.Z., Silva C.L., Cabral J.M.S. (2012). Mesenchymal stem cells for cellular therapies. *Stem cell and Cancer Stem Cells*

- McGovern A.C., Ernill R., Kara B.V., Kell D.B., Goodacre R. (1999). Rapid analysis of the expression of heterologous proteins in *Escherichia Coli* using pyrolysis mass spectroscopy and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production. *Journal of Biotechnology* 72:157-167
- Mordechai S., Sahu R.K., Hammody Z., Mark S., Kantarovich K., Guterman H., Podshyvalov A., Goldstein J., Argov S. (2004). Possible common biomarkers from FTIR microspectroscopy of cervical cancer and melanoma. *Journal of Microscopy* 215:86-91
- Naes T., Isaksson T., Fearn T., Davies T. (2002). *Multivariate Calibration and Classification*. NIR Publications: Chichester, United Kingdom
- Naylor L.H. (1999). Reporter genes technology: The future looks bright. *Biochemical Pharmacology* 58:749-757
- Nicolaï B.M., Beullens K., Bobelyn E., Peirs A., Saeys W., Theron K.I., Lammertyn K. (2007). Nondestructive measurement of fruit and vegetables quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology* 46:99-118
- Otto M. (1999). *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. WILEY-VCH: Freiberg, Germany
- Pacifico A., Chiriboga L.A., Lasch P., Diem M. (2003). Infrared Spectroscopy of cultured cells II. Spectra of exponentially growing, serum-deprived and confluent cells. *Vibrational Spectroscopy* 32:107-115

- Parekh B.S., Berger E., Sibley S., Cahya S., Xiao L., LaCerte M.A., Vaillancourt P., Wooden S., Gately D (2012). Development and validation of an antibody-dependent cell-mediated cytotoxic-reporter gene assay. *Landes Bioscience* 4:310-318
- Petiot E., Bernard-Moulin P., Magadoux T., Gény C., Pinton H., Marc A. (2010). In-situ quantification of microcarrier animal cultures using near –infrared spectroscopy. *Process Biochemistry:* 45:1832-1836
- Pistorius A.M.A. (1995). Biomedical applications of FT-IR spectroscopy. *Spectroscopy Europe* 7:8-15

- Pittenger M.F., Mackay A.M., Beck S.C. (1999). Multilineage potential of adult human mesenchymal stem cells. *Science* 284:143-147
- Randolph T.W. (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
- Rhiel M., Ducommum P., Bolzonella I., Marison I., Stockar U. (2010). Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. *Biotechnology and Bioengineering* 77:174-185
- Rodrigues C.A.V., Fernandes T.G., Diogo M.M., Silva C.L., Cabral J.M.S. (2011). Stem cell cultivation in bioreactors. *Biotechnology Advances* 29:815-829
- Roggo Y., Chalus P., Maurer L., Lema-Martinez C., Edmond A., Jent N. (2007). A review of near infrared spectrosc-opy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis* 44:683-700
- Roychoudhury P., Harvey L.M. 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
- Palframan S.L., Kwok T., Gabriel K (2012). Vacuolating cytotoxin A (VacA), a key toxin for *Helicobacter pylori* pathogenesis. *Frontiers in Cellular and Infection Microbiology* 2:92
- Sandor M., Rüdinger F., Bienert R., Grimm C., Solle D., Scheper T. (in press). Comparative non-invasive monitoring via infrared spectroscopy for mammalian cell cultivations. *Journal of Biotechnology*
- Scholz T., Lopes V.V., Calado C.R.C. 2012). High-throughput analysis of the plasmid bioproduction process in *Escherichia coli* by FTIR spectroscopy. *Biotechnology and Bioengineering* 109:2279-2285
- Seber G.A.F. (1984). *Multivariate Observations*. John Wiley and Sons: Mew York, United States
- Smith B.C. (2011). *Fourier Transform Infrared Spectroscopy*. New York: CRC Press

- Stadel J.M., Wilson S., Bergsma D.J. (1997). Orphan G protein-coupled receptors: a neglected opportunity for pioneer drug discovery. *TiPS* 18:430-437
- Steele D. (2002). Infrared Spectroscopy: Theory, in Chalmers J.M. and Griffiths P.R. (Eds.). *Handbook of Vibrational Spectroscopy*. Wiley: United States
- Stuart B. (2004). *Infrared Spectroscopy: Fundamentals and Applications*. WILEY: United States
- Teixeira A.P., Oliveira R., Alves P.M., Carrondo M.J.T. (2009). Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnology Advances* 27:726-732
- Timlin J.A., Martin L.E., Lyons C.R., Hjelle B., Alam M.K. (2009). Dynamics of cellular activation as revealed by attenuated total reflectance infrared spectroscopy. *Vibrational Spectroscopy* 50:78-85
- Walsh M.J., Hammiche A., Fellous T.G., Nicholson J.M., Cotte M., Susini J., Fullwood N.J., Martin-Hirsch P.L., Alison M.R., Martin F.L. (2009). Tracking the cell hierarchy in the human intestine using biochemical signatures derived by mid-infrared microscopy. *Stem Cell Research* 3:15-27
- Wong P.T.T., Wong R.K., Caputo T.A., Godwin T.A., Rigas B. (1991). Infrared spectroscopy of exfoliated human cervical cells: Evidence of extensive structural changes during carcinogenesis. *Proc. Natl. Acad. Sci.* 88:10988-10992
- Yang T.-T., Parisa S., Kitts P.A., Kain S.R. (1997). Quantification of gene expression with a secreted alkaline phosphatase reporter system. *BioTechniques* 23:1110-1114
- Yeniay Ö., Götkaş A. (2002). A comparison of partial least saquares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics* 31:99-111
- Zelig U., Mordechai S., Shubinsky G., Sahu R.K., Huleihel M., Leibovitz E., Nathan I., Kapelushnik J. (2011). Pre-screening and follow-up of childhood acute leukemia using biochemical infrared analysis of peripheral blood mononuclear cells. *Biochimica et Biophysica* 1810:827-835
- Zhang G., Kurachi S., Kotoku (2013). Limitations in use heterologous reporter genes for gene promoter analysis: Silencer activity associated with the chloramphenicol acetyltransferase reporter gene. *The Journal of Biological Chemistry* 278:4825-4830

- Zhang M.J., Meng F.L., Ji X.Y., He L.H., Zhang J.Z. (2007). Adherence and invasion of mouse-adapted H pylori in different epithelial cell lines. *World Journal of Gastroenterology* 13:845-850