Evaluation of NIR, FTIR and Raman Spectroscopies as Monitoring Techniques for Recombinant GFP Protein Production in Fed-Batch Cultures of *Escherichia coli*

by

Arnaldo José Rosario Rosario

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
CHEMICAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2008

Approved by:

_____	_____
Rodolfo Romañach, PhD Member, Graduate Committee	Date
_____	_____
Juan Carlos Sáez, PhD Member, Graduate Committee	Date
_____	_____
Lorenzo Saliceti Piazza, PhD President, Graduate Committee	Date
_____	_____
Lynette E. Orellana,PhD Representative of Graduate Studies	Date
_____	_____
David Suleiman, PhD Chairperson of the Department	Date

ABSTRACT

The expression of green fluorescent protein (GFP) was induced and monitored in fed-batch cultures of a genetically-engineered *Escherichia coli* strain. This recombinant *E. coli* is commercially available (Bio-Rad pGLO kit), and it synthesizes GFP when induced by addition of L-arabinose. The *E. coli* fed-batch culture considered in this work utilizes glycerol as carbon and energy source and ammonium ion as nitrogen source. Acetic acid (acetate) is produced as a metabolic by-product because of incomplete oxidation of glycerol (Hall et al., 1996). Given that higher concentrations of acetate, ammonium and glycerol can inhibit *E. coli* growth (can decrease the biomass bench scale concentration), all fed batch fermentations were inoculated to start at high initial cell mass concentrations in close analogy to an industrial setting. After inoculation, all fermentations were monitored and samples collected regularly to be analyzed at line using NIR, IR and Raman spectroscopies.

Acetate and glycerol concentrations were measured with an offline HPLC during each batch with very good accuracy (R^2 > 92%). To contribute to the monitoring and eventually to bioprocess automation and industrial applications improvements (such as PAT), Raman spectroscopy was used along with chemometric techniques to model and predict analyte concentrations. During the fermentation, samples were collected and analyzed with at-line Raman and the off-line HPLC to establish a correlation. Specifically, to accomplish the correlation, a design of experiments (DOE) was programmed with Minitab (software version

14). A mixture design was used to correlate the analytes concentrations and the Raman spectra using a chemometric techniques, partial least squares (PLS). The correlation between Raman spectroscopy and the analyte mixture showed excellent model results (R^2 > 99 % and Q^2 > 99 %) with three factors. The correlation between the HPLC and the Raman with fermentation data was determined and validated. Different modes were used, such as cross-validation (after model was obtained, it was tested by predicting the results with the set of data using in the model) and test set validation (reserve a representative data and predict it with the model). Both validation procedures presented a good model fit (linear y = x relation).

The biomass concentration is another critical variable to ensure rigorous fed-batch fermentation. Specifically, the biomass concentrations were measured using two methods: 1) an at-line absorbance method (ultraviolet spectrophotometry) and 2) dry cell mass concentration, or gravimetric determination. Both methods were then correlated with near infrared spectroscopy (NIR). PLS regression presented a good model fit (R^2 = 99%, two factors) as well as the validation analysis (y = x).

After obtaining a good correlation between different analytical methods used for estimating dry cell mass and by-products concentrations, a GFP protein fluorescence correlation was investigated using flourometry as the reference method and Fourier transform infrared (FT-IR) spectroscopy as the primary method. Analysis and PLS modeling resulted in a good fit for GFP concentration (protein concentration is proportional to the fluorescence). FT-IR was

used with chemometric techniques to model and predict the recombinant protein concentration. Following available literature, a partial least squares (PLS) regression analysis was used to establish the correlation between fluorescence and the FT-IR spectra (R^2= 99 %, two factors). Also a cross validation and test-set validation were achieved with good linear relationships (y = x).

Overall, a substantially better understanding on how we can measure the critical parameters of a recombinant protein fed-batch culture system using spectroscopic techniques and chemometrics was accomplished. Specifically, the ability to accurately determine glycerol, acetate, biomass and recombinant protein concentrations was achieved. More research is needed to test the techniques “in line” to determine if the bench scale experimental approach can be scaled-up to industrial and commercial scale working volumes.

RESUMEN

La expresión de la proteína de verde fluorescencia conocida como GFP, por sus siglas en inglés fue inducida y medida durante la fermentación de *Escherichia coli* en modo de semi-tanda. Una cepa de *E. coli* recombinante está comercialmente disponible (Bio-Rad pGLO kit), la cual es capaz de sintetizar la proteína de verde fluorescencia cuando es inducida mediante adición de L-arabinosa. La *E. coli* utilizada en este trabajo consume glicerol como su fuente de carbón y energía, y amoníaco como su fuente de nitrógeno. Ácido acético (acetato) es producido como un producto metabólico a consecuencia de la oxidación parcial del glicerol (Hall et al., 1996). Dado que las altas concentraciones de ácido acético, amoníaco y glicerol pueden inhibir el crecimiento de *E. coli* (puede disminuir la concentración de biomasa), el monitoreo y la automatización o ambas son necesarias para lograr una fermentación óptima y reproducible a escala industrial.

Las concentraciones de acetato y glicerol fueron medidas mediante cromatografía liquida de alta resolución (HPLC) durante cada fermentación para poder presentar un resultado lo suficientemente preciso ($R^2 > 92\%$). Como contribución a la automatización del bioproceso y mejorar las técnicas de monitoreo industriales, la técnica de espectroscopía analítica (Raman) fue utilizada junto a análisis de quimiometría para medir las concentraciones de los analitos. Durante las fermentaciones, las muestras fueron analizadas utilizando espectroscopía de Raman y HPLC para establecer una correlación entre ambas medidas. Específicamente, para

lograr la correlación entre ambas tecnologías, un diseño experimental fue diseñado usando Minitab (version 14) para modelar y predecir la concentración de los analitos y los espectros de Raman utilizando análisis de quimiometría. En nuestro caso, "partial least squares" (PLS) fue utilizado. La relación entre la espectroscopía de Raman y el diseño de experimento presentó excelentes resultados (R^2 > 99 % y Q^2 > 99 % con tres factores). La correlación entre HPLC y Raman en la fermentación fue completada y validada. Para esto, diferentes pruebas fueron usadas tales como "cross-validation" (el modelo es probado con los mismos datos que fueron utilizados para crear el modelo) y "test-set validation" (se reserva una porción representativa independiente de datos para probarla con el modelo creado). Ambos presentaron una buena validación y un buen modelo lineal (y = x).

La concentración de biomasa es una variable crítica para asegurar un nivel considerable del crecimiento de la bacteria para asegurar que la nuestra fermentación sería representativa y realizable a escala industrial. Específicamente, la concentración de biomasa fue medida por: 1) absorbancia, utilizando un equipo de espectrofotometría de ultravioleta y 2) como concentración de célula seca por el método conocido como gravimétrico. Luego, fue correlacionada con los espectros obtenidos mediante un equipo de cercano a infrarrojo (NIR por sus siglas en inglés). PLS presentó una buena precisión (R^2 = 99%, dos factores) así como también los análisis de validación (y = x).

Luego de obtener una correlación entre algunas de las variables mas críticas del proceso, la correlación de la producción de proteína recombinante fue establecida utilizando un fluorómetro. Esta tecnología presenta una manera fácil y precisa para la cuantificación de proteína recombinante (la concentración de proteína es proporcional a su fluorescencia). Infrarrojo mediano por transformadas de Fourier (FT-IR por sus siglas en inglés) fue utilizado junto con las técnicas de quimiometría para medir la concentración de la proteína. Tomando sugerencias de la literatura, el análisis de PLS fue realizado con muy buenos resultados (R^2= 99 %, dos factores). Además los análisis de validaciones como se habían expuesto anteriormente presentaron un modelo lineal (y = x).

En general, un mayor entendimiento fue obtenido sobre cómo medir las variables más críticas usando espectroscopía de infrarrojo y análisis quimiométrico en un bioreactor operado a modo de semi-tandas. Se pudo determinar con una buena precisión las concentraciones de proteína recombinante, analitos y biomasa. Mas investigación es necesaria, como por ejemplo, el utilizar estas técnicas en modo "in line" (sondas inmersas en el medio de cultivo) a escala industrial para determinar si la relación se asemeja a la de la escala pequeña.

DEDICATED TO MY PARENTS
JOSE ANTONIO ROSARIO BENITEZ AND ELMY ROSARIO GALARZA,
MY BROTHERS JONATAN ROSARIO, GISELA ROSARIO AND JOSE DAVID ROSARIO.

ACKNOWLEDGEMENTS

During the maturity of my graduate studies in the University of Puerto Rico several persons and institutions collaborated directly with my research. Without their support it would be impossible for me to finish my work. That is why I wish to dedicate this section to recognize their support.

I want to start expressing my gratitude to my advisor, Dr. Lorenzo Saliceti because he gave me the opportunity to perform research work under his guidance and I also thank him for his support in the laboratory. I also want to thank Dr. Rodolfo Romañach for his unconditional help and his interest in my work and also for treating me like one of his students. I aldo thank his support in every step of my graduate study, for trusting in my work and for the guidance I received from Dr. Carlos Rinaldi.

The Grant from INDUNIV provided the funding and the resources for the development of this research. I also would like to thank my family and friends, for their unconditional support and love.

TABLE OF CONTENTS

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

1 INTRODUCTION

Bioprocesses are used to produce a wide variety of metabolites: from chemicals such as ethanol, organic acids, amino acids, and antibiotics to therapeuticals such as monoclonal antibodies, and hormones. In those, bioreactors constitute the core components of their production processes. The ability to monitor on time the critical variables on time in the bioreactors is fundamental for both bioprocess supervision for quality control, and bioprocess development, where the microbial strain or cell line and the operating conditions always need consistency. For process supervision, at-line and in-situ analyses are needed, where the analyzer is within or inside the bioreactor, and sample collection is necessary for timely analysis and control of process parameters. For process development, the flexibility to analyze a wide range of substrates, biomass and products is critical and at-line measurements are useful for early identification of poor growth conditions, etc.

The inability to accurately measure concentrations of biomass, nutrients, products and by products in real time, obstruct effective and timely monitoring and controlling of batch and fed-batch fermentation bioprocesses [1]. Process control strategies must be evaluated during important instances; for example, during a shutdown of a process because of bacterial or viral contamination. In-process scheduling changes such as aeration rate, timing of inoculation, or delaying of harvest to increase titer concentration, have to be made while the process is running. For these reasons the traditional analytical off-line methods such as dry cell weight analysis and optical density measurements for biomass, enzymatic or chemical synthesis

mediated UV absorption reactions, liquid and gas chromatography are not adequate because of extensive sample preparation, prolonged analysis time and possible changes to samples after collection. At-line and in-situ analyses represent an improvement because of their rapid results, are analyzed rapidly represent an improvement, as opposed to conventional off-line analyses. In many production processes, off-line analyses are performed after the fermentation is finished or several hours have passed after the sample was collected, resulting in untimely analyses results [2]. For that reason, Federal agencies have promoted initiatives to improve the traditional methods and innovate novel methods to fully understand how products attributes and processes relate to final product performance [3]. Guidances like the Process Analytical Technologies (PAT) are intended to encourage the industry to voluntarily device and implement innovative pharmaceutical development, manufacturing, and quality assurance [3].

The ability to implement PAT in the biotechnology industry depends on the implementation of real time monitoring and control of bioprocesses [4]. The monitoring of the products, byproducts and nutrients at-line, preferably in-situ, where the analyzer is in direct contact with the process stream is an ideal approach of monitoring. In in-situ probes would help avoid contamination during the sample acquisition and would allow the process to continue forward without any interference and delay that may affect the process. The industry has already implemented the use of various in-situ probes (pH, dissolved oxygen, conductivity, and temperature). These technologies usually involve changes in potential, and although they provide useful data, they are insufficient for the understanding of critical variables such as

metabolic states of the culture, product and by-product synthesis, etc., that require other types of instrumentation.

Recent improvements in informatics, mathematical procedures, and multivariable analysis and technologies are established new tools that can be implemented as in-situ bioprocess monitoring tools [5]. Now a days the technologies available for in-situ determination of critical parameters such as biomass, are more an "experience than a science," they present difficulties and inconsistent pathways in fermentation processes during the growth phase [6]. For example, an inverse correlation exists between in-situ pH meters and biomass; as biomass increases, pH decreases due to acetate production. In the case of dissolved oxygen probes as biomass increases, oxygen demand also increases, therefore the dissolved oxygen concentration decreases. Infrared technology and Raman spectrometry show non-invasive measurement techniques with the possibility of multicomponent analysis, where the concentration of more than one culture broth constituent (substrate, biomass, products and by products) can be determined simultaneously [7]. There are also probes capable of withstanding the steam-sterilization at 120ºC without adverse effects on their calibrations [7].

Escherichia coli is perhaps the most commonly used industrial bacterium in biotechnology processes. It is the preferred microbial cell to express therapeutically useful protein molecules such as human insulin and Filgrastim (Neupogen). The high-cell-density *E. coli* fermentation considered in this study consumes glycerol as a carbon and energy source and synthesizes a recombinant protein called green fluorescent protein (GFP) in presence of

arabinose. Acetic acid (acetate) is produced as a metabolic by-product as a result of incomplete oxidation of glycerol [1]. In this study, we have used near-IR (NIR), IR and Raman spectroscopies to analyze bench-scale (1.5 L) fed-batch fermentations, which produce as time progresses different accumulation profiles for biomass, acetate and recombinant protein production.

1.1 Objectives

Our research objectives can be classified as general and specific.

1.1.1 General Objectives

- To monitor in a batch and fed-batch fermentation of *E. coli* the production of:
    - microbial cell mass and growth.
    - substrate concentration and utilization.
    - by-product concentration and production.
    - intracellular recombinant protein production and accumulation.

1.1.2 Specific Objectives

- To use multivariable analysis such as partial least squares (PLS) to create and validate a model for monitoring the *E. coli* growth using NIR spectroscopy via optical density and dry cell concentration (gravimetric) as reference methods.

- To Use multivariable analysis such as PLS to create and validate the substrate utilization and by-product production using Raman spectroscopy via HPLC as reference method.
- To use chemometrics for correlation of FT-IR spectra to characterize GFP expression via fluorescence spectrophotometry as reference method.

1.2 References

[1] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B.G.; Macaloney, G. Near-Infrared Spectroscopic Determination of Acetate, Ammonium, Biomass and Glycerol in an Industrial *Escherichia coli* Fermentation. *Applied Spectroscopy* **1996**, 50, 102-8.

[2] Macaloney, G. Near infrared sprectroscopy: the technology, its growing use in biotechnology and evaluation of its utility. *Society of Industrial Microbiology* **1996**, 46, 129-132.

[3] FPAC/PAT Summit with INDUNIV – Process Analytical Technology Pharmaceutical/ Biopharmaceutical Manufacturing, June 2007 San Juan, Puerto Rico, Essential Meeting in Puerto Rico for Process Analytical Technology, not accessed yet.

[4] Arnold, S.A.; Gaensakoo, R.; Harvey, L.M.; McNeil, B. Use of At-line and In-situ Near-Infrared Spectroscopy to Monitor Biomass in an Industrial Fed-Batch *Escherichia coli* Process. Wiley Periodicals, Inc. **2002**, 405-13.

[5] Raidyanathan, S.; Macaloney, G.; McNeil, B. Fundamental investigation on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *Analyst* **1999**, 124, 157-62.

[6] Vaidyanathan, S.; Macaloney, G.; McNeil, B. Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *Analyst*, **1999,** 124, 157-62.

[7] Lee, H.L.T.; Boccazzi, P.; Gorret, N.; Ram, R.J.; Sinskey, A.J. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, **2004,** 35, 131-7.

2 LITERATURE REVIEW

2.1 Historical Background

The need to understand fermentation processes dates back to the XIX century when Louis Pasteur brought microbiology to a new level of scientific excellence. In France in 1949, Monod demonstrated for *E. coli* that growth rate changed form first order to zero order kinetics as the initial substrate concentration was increased [1]. Later in Great Britain in 1965, Pirt developed a method for calculating the maintenance energy term using mass balance equations and demonstrating that the consumed energy was proportional to biomass yield and growth rate. [2].

In the pharmaceutical industry, the commercial fermentation processes started after the accidental discovery of penicillin by Alexander Fleming in 1941. The extreme demand (>100,000 patients per year) for penicillin during the Second World War was the reason for the design of the suspended cell culture reactor by Pfizer to achieve higher penicillin productivities. During this period, fermenters with rudimentary monitoring and feedback control were available in industry and academia [3, 4]. The need of technology which could improve the fermenters rudimentary control schemes was indispensable. In 1945, pH probes were able to withstand repeated sterilization. Later on, dissolved oxygen (DO) probes were also introduced [2]. After that, off gas analysis of oxygen and carbon dioxide using mass spectrometers became reliable and affordable, providing at-line data of the culture state in

several industrial fermentation plants [2, 5, and 6]. However, even with the advances in analytical technologies, from the mid-1970's to present the inability to measure important parameters in-situ has been a major task in developing mathematical models and control strategies [2, 7]. Even though these strategies are complex, the control and optimization of fermentation processes have become an "economic necessity" [8]. This conviction about the value of at-line sensors was the reason for a literature explosion that attempted in developing novel measurement applications during the decades of the 1980's and 1990's.

During the 1980's, the commercialization of several types of biosensors which provided in-situ measurements, increased the number of analyses obtainable under aseptic conditions and allowed a faster and less-labor intensive analysis methodology [9]. The advancement in computer technology has served to enhance the development of a multitude of novel measurement technologies [2]. Near infrared (NIR) techniques would be concerned with the rapid analysis, non-destructive sampling and possible in-situ analytical measurements. Also, this technique presents prominent expectations, reducing the use of tedious techniques such as HPLC technology or total solids weight determinations, to name a few. Using at-line sampling with an analysis time of less than a minute, NIR was implemented for an *E. coli* fermentation to determine acetate, ammonium, biomass and glycerol concentrations without pretreatment, and using either multiple linear least squares regression or partial least-squares regression [10]. Then at-line, flow-through, and in-situ IR probes were developed. The application of the multivariable analysis with the extensive calibration data and the lack of sophisticated mathematical techniques, limited early infrared applications in bioprocessing

[10, 11]. With computer advances and improvements, successful works in the collection of fermentation data have incremented [12]. For example, Novo Nordisk developed at-line data acquisition from 10-15 measurement devices with a sampling time ranging from 10 seconds to 30 min, over 100,000 data points accumulated when about 300 long fed-batch fermentations were conducted annually [13]. A tremendous amount of process data became archived without analysis for lack of correlations and the difficulty to analyze large amounts of data to assist in process decision-making.

Chemometrics is an innovative idea that was started in the 1970's. It was first reported in 1972 in a paper published by Wold, et al. who developed multivariable model analysis such as partial least squares [14]. Those model analyses were very useful in the 1980's to extract information from large sets of data. Chemometrics became a new chemical discipline that uses mathematical, statistical and logical information to provide maximal relevant chemical information by optimal experimental design. Chemometrics analysis ascertains the use of exploratory analysis with statistical property prediction such as regression modeling and factorial behavior to convey different spectra or patterns into a predictive model [14].

2.2 Difficulties of PAT Implementation in Biotechnology Processes

In a fermentation process, the necessity to implement monitoring and control techniques is fundamental for tight control of the life of organism's environment to optimize yield and

productivity [2]. The classical fermentation processes for milk and cheese have incorporated excellent applications for PAT. However, in the pharmaceutical industry PAT has not had the same response [2]. The frustrations have caused the fermentation scientists to establish that the protein-based biological production is more "an art than a science" [2, 15]. Fermentation has been considered by some to be "*the most complex step in the manufacturing process since the desired product typically is present at lower concentrations in the broth, at levels comparable to impurities*" [16]. Some of the difficulties consist in the inherent complexity of biopharmaceuticals, making it difficult to determine product characteristics critical to safety, efficacy and stability, and the different metabolic routes, which may produce different compounds depending on the raw feed materials [16]. Some of the complexities are dependent on the mass transfer limitations from high densities required for desired productivities, the inhibition of the bacteria and the strain causing low growth from the raw material, by-products excess concentration and the variability of the active seed culture. Despite these impediments, the PAT implementation is necessary to improve the quality without increasing costs and to reduce regulatory costs [2, 16]. The improvements of these techniques start low and increase over time [15, 17].

For the PAT implementation, the classical focus based on univariate experimentation and optimization is changing to a systematic multivariate analysis with robust modeling and control strategies [3]. Multivariate instead of univariate analysis permits us to see how the combination of experimental data could impact multiple process parameters and their interactions among the variables and observations to model and predict performance.

2.3 Infrared Spectroscopy Fundamentals

Spectroscopy in general is the study of the interactions between the radiation and matter by practically the entire spectrum (Figure 2.1). Infrared spectroscopy is a well-known characterization technique that uses simultaneous analysis of several components quickly, with no sample pretreatment and has the capability to be used in-situ and to provide data in real time. Spectroscopy can explain the chemical structural characteristic features of the molecule, if whether they are found in the backbone of the molecule or in the functional groups attached to the molecule.

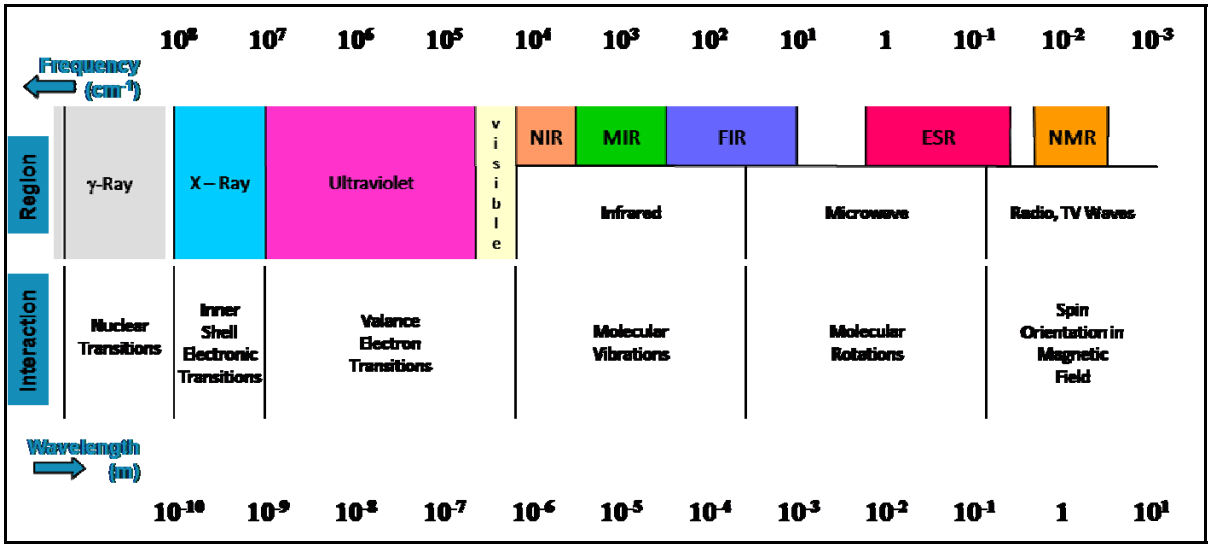


Figure 2.1 - Molecular frequency absorption through the spectra.

Previous studies have highlighted the use of infrared technology in many processes from simple (the transformation of progesterone at laboratory scale) to complex (high cell density *E. coli* fermentations) [18].

Coates et al. considered the infrared spectrum to be formed as a consequence of the absorption of electromagnetic radiation at frequencies that correlate to the vibration of specific sets of chemical bonds from within a molecule [19]. The distribution of the energy possessed by a molecule at any given moment is defined as the sum of the contributing energy terms (Eq. 2.1) [19].

$$E_{total} = E_{electronic} + E_{vibrational} + E_{rotational} + E_{translational} \quad \textbf{2.1}$$

The translational energy relates to the displacement of molecules in space as a function of the normal thermal motions of matter. Rotational energy is observed as the tumbling motion of a molecule, which is the result of the absorption of energy within the microwave region. The vibration energy component is a higher energy term and corresponds to the absorption of energy by a molecule as the component atoms vibrate about the mean center of their chemical bonds [19]. The electronic component is linked to the energy transitions of electrons. Planck established one of the most important general relativity-quantum mechanics equation (E = hv) that relates proportionally the energy with the frequency [19]. Ultraviolet and visible frequency absorption, which are shown in Figure 2.2 have the energy to make electron transitions, but in the infrared spectra the frequency absorption could only change the atom dipole making their vibration and rotation. For that reason this spectrum could tell us some physical as well as chemical properties.

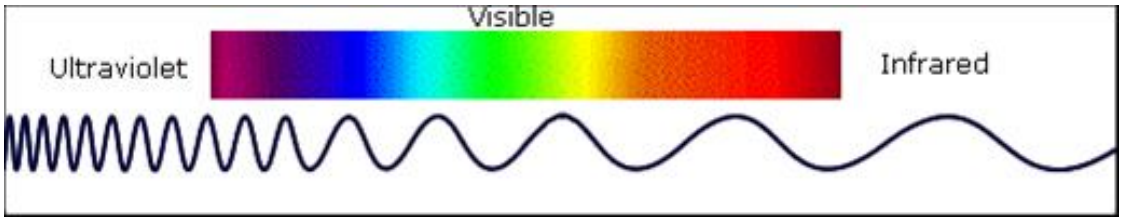


Figure 2.2 - UV-VIS and infrared frequencies are proportionally inverse to vibrational and rotational energies.

In IR spectroscopy, the vibration excitation is achieved by radiating the sample with a broad-band source of radiation within the 12820-200 cm^{-1} region [19]. Part of the spectrum is divided in the near infrared (NIR), 12820 -3959 cm^{-1} and mid infrared (MIR), 4000-200 cm^{-1}. Many NIR and MIR equipments are available in the market individually. One of the most significant explanations for that is because both NIR and MIR are part of the infrared spectrum and they provide us different information about the sample.

In the case of near infrared (NIR) spectroscopy, because it has a lower broad energy, it is useful to determine physical interactions of the sample with the environment such as density, humidity and quantity of suspended particles. In our case, it was useful for determing biomass concentration in the fermentation broth.

The NIR absorption bands are typically broad, overlapping and 10-100 times weaker than their corresponding fundamental MIR absorption bands [20]. For that reason this technology could be useful for acquiring physical information. We agree that the low absorption permits large penetration depths and for that reason it could be useful for turbidity interpretation, such as in biomass in a fermentation process. It could also be an important tool for the determination of inclusion bodies within the bacteria. NIR could have an indirect relationship between the bacterial size and the over expressed recombinant protein concentrated as an inclusion body within the cell.

Fourier Transform Infrared (MIR and IR) spectroscopy is a powerful frequency absorbance energy that measures predominantly the bond strengths of molecules and the vibrations of bonds within functional groups [20]. These special characteristics can help us to determine the presence of particular bands within the spectrum, and to also notice the absence of other important bands [21].

2.4 Raman Spectroscopy Fundamentals

Raman spectroscopy is another important form of vibrational spectroscopy and also a complementary technique. The selection rules for Raman spectroscopy are different from the infrared spectroscopy [21]. The Raman scattering intensities are proportional to the change in molecular polarizability upon vibration excitation, whereas infrared absorption intensities are

proportional to the change in dipole moment of a molecule as it vibrates [19]. For these reasons, polar bonds or nonequal charges are more prominent to determine in IR spectra, while polarizable vibration modes are stronger in Raman spectra such as in diatomic molecules, aromatic ring and other vibrations involving multiple bonds or heavy atoms [22]. One of the benefits of Raman spectroscopy is the weak absorption of polar molecules such as water. The polar weakness on Raman spectroscopies could be useful for applications in aqueous solutions, because it possesses a weak OH stretch band and also could generate rich information from the sample. Raman spectroscopy might be an excellent technique during the fermentation process, because it presents a way to avoid the most significant source of interferences (water) at infrared frequencies. Also, this technique is useful in fermentation products as well as for the purification processes, where the expressed recombinant protein has a large molecular structure.

After explaining the different technologies explored in this work, it is necessary to establish that like other analytical equipments, the NIR, FT-IR and Raman spectroscopies will not fully replace the traditional wet technology. They are “old but new” tools which in some cases could be more convenient to apply than other techniques. Another important consideration is that these strong tools need a good preparation and a well devised, experimental design, statistical analysis and a rigorous validation of the process [23].

2.5 Application of NIR, IR and Raman Spectroscopies in Fermentation Monitoring

The scientific literature of simultaneous utilization of NIR, IR and Raman spectroscopies to monitor the concentration of biomass and recombinant proteins is very limited (Figure 2.3).

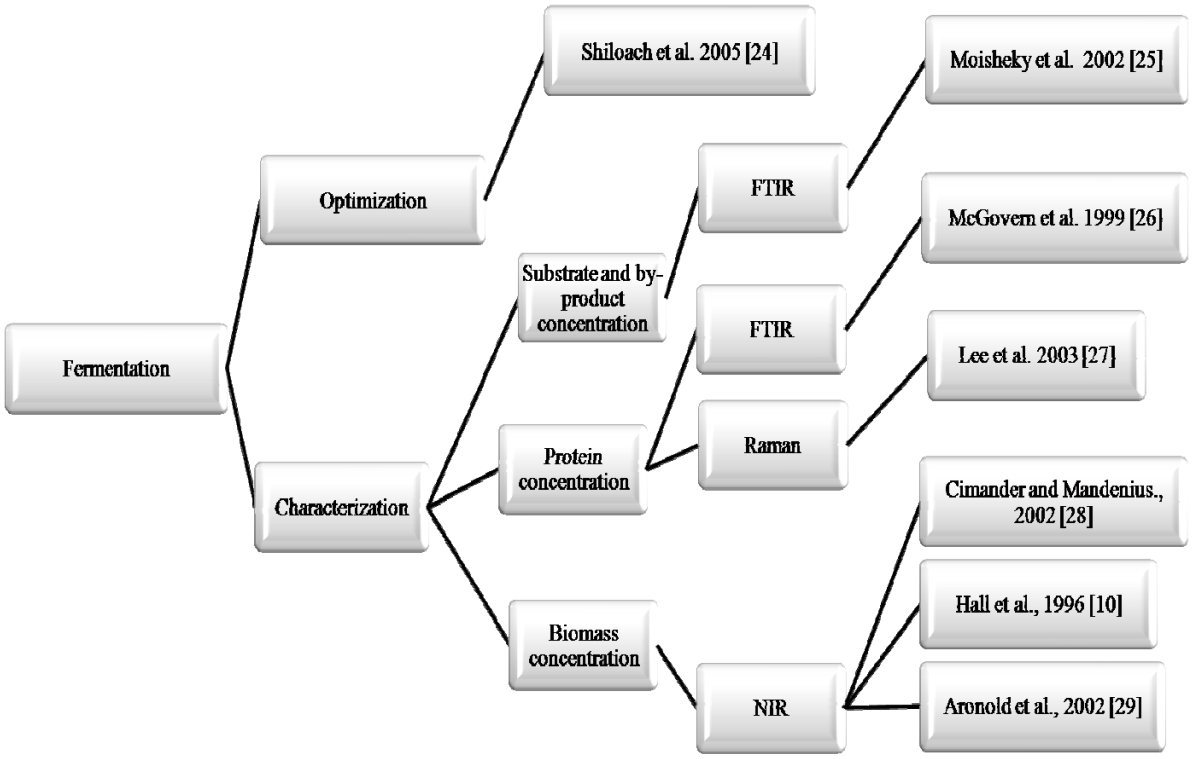


Figure 2.3 - Literature review on the use of NIR, IR and Raman spectroscopies to determine biomass, substrate, by-product and recombinant protein in an *E. coli* culture.

No matter the technology used or proposed by the authors, the purpose was to improve over the traditional methods to supply the necessary monitoring data from the fermentation

process. The extensive sample preparation and analysis time (1-24 hours, Appendix A) restricts the utility of the traditional assays for process monitoring, optimization and control [10]. Starting with NIR technology, its spectral information is secondary in nature, having its origin in the mid-IR region. This non-invasive technique possesses attractive features in particular: the possibility of multicomponent analysis, where the concentration of more than one culture broth constituent (e.g., biomass, substrates, and metabolic products) can be determined simultaneously by means of non-invasive sample analysis and real-time measurements. These features make of this technique an attractive option for real-time bioprocess monitoring [27]. Raman spectroscopy, like NIR, can be used to estimate the concentration of chemical components and biomolecules in bioprocess media. The main difference between Raman and NIR resides on the low interference from water for Raman spectroscopy [22]. This technique validates the demonstrated simultaneous concentration estimation of fed-batch culture substrate (glycerol) and by-products (acetate, formate, lactate and phenylalanine) reported by Lee et al. in 2004 [22]. Although a good correlated spectrum is available in the literature, Lee and coworkers agreed in the necessary additional experiments and analyses to identify and correct the model errors and to obtain a confidence interval around concentration estimates [22].

The last technique is the Fourier transform infrared spectroscopy (FTIR), which is a physic-chemical method that measures predominantly the bond strengths of molecules and the vibrations of bonds within functional groups, respectively [21]. Infrared spectra of proteins exhibit strong amide absorption bands at 1650 cm^{-1} associated with the characteristic

stretching of C=O and C—N and the bending of the N—H bond [21]. The problem is that this technique, like NIR, provides water interferences and for that reason it becomes necessary to explore various chemometric approaches which use information from the entire spectrum.

2.6 Statistics Approach in the Application of Chemometrics

Statistics can be used for the extraction of maximal and relevant information from raw data. Its uses the experimental design to conduct experiments and determine how and where to find additional information of the processes.

In the pharmaceutical and biotechnology industry, hundreds of variables are desirable to fully understand the product and the process, and to optimize production. In view of the facts above, process monitoring and control need to improve continuously by installing probes at-line or in-situ to obtain data for the acquisition of the critical variables within the process. Multivariable methods such as principal components analysis (PCA) or partial least squares (PLS) have demonstrated to be ideal in such cases were huge amounts of data are collected. This is would be the case with the kinetic data from a heterogeneous system in which the metabolites vary or the metabolic routes could change with respect to their environments. The non-linearity of the system studied plus the numerous amounts of data could avoid a full characterization of the system. The multivariate analysis does not pretend to substitute univariate analyses, but in some cases whereas the lack of information is not the problem and

the model is complex, multivariate methods should provide more information than univariate methods.

PCA is an algorithm that recognizes the raw data (usually a spectrum) as variation patterns (intrinsic correlation), performs a mathematical decomposition of the patterns, and reconstructs the patterns using orthogonal vectors (uncorrelated to each other) [30]. PCA could be useful for finding the individual relationship between the analytical observation such as experimental tests, batches, chemical compounds, etc., to the spectroscopic quality or quantity aspect of the product [31]. Some analysis advantage consists of the loading dependence of the variable that permits an "a priori" control of the loading, if the researcher knows which variable will be more characteristic of the process. If physical interferences are known to affect the process modeling, pretreatment of the data is required (smooth normalized variance, log, etc.). One example of the pretreatment would be the usage of normalized pretreatment to avoid fluorescence effects (due to lower frequency interference) on a synthesis of a fluorescent protein in a fermentation run [32]. In fermentation, the microbial growth and the recombinant protein synthesis depend on achieving their optimal parameters (decreasing the inhibitory effect in the bacteria, metabolic routes, among others). The nutrient quantities are a critical variable in the model established in this research.

PCA analysis that explains the variability using vectors or factors is essential in the use of other analyses to establish the relation with the quality or productivity data. This analysis is called partial least squares (PLS) and its objective is to explain the correlation and variability

of the observations and variables (X, usually spectra) to the response of the process (Y, usually absorbances or other assays such as concentrations determined using HPLC)

PLS regression estimates the X projection and maximizes its correlation with Y. One of the advantages of this analysis is that each process variable is treated as a vector behavior which provides an easy and effective way to determine and establish a good model where the loading is sought in the vector direction and sign, and not in the magnitude. There techniques are useful because they reduce the noise and compress the long raw data in a better view, within vector and effect plots.

In general, PCA and PLS are good statistical methods to build vigorous models with an accurate prediction and adjustment of the process.

2.7 Geometrical Chemometrics Interpretation

The univariate least squares measures a variable response for each known independent variable, plots variable responses vs. properties, and finds the slope and intercept which minimize the residual sum of squares of the response variable. This technique is useful to analyze small clusters of data or where a linear relation is obvious because it presents a rapid and simple model analysis. To be sure the model is significant, different statistical tools such as ANOVA, null hypothesis test, factorial, repeatability and reproducibility (R&R), split plot, are available to determine the relationship between the observation (X) and the response (Y).

If the data contains interferences, the linearity of the model is not obvious and one response (Y) for observations is not sufficient to explain and obtain a good reproducibility and repeatability of the model. If the collected data is extremely large, the limits of this process are notable, because the model has to measure one variable response for each known independent variable and also without adding the effect of responses in the model.

The factor-based regression replaces original X by X_{aprox}, which has uncorrelated columns. Two of them are principal components analysis (PCA) and partial least squares (PLS). PCA considers a matrix X with **N** observations and **K** variables. Each observation (each row) of the X- matrix is placed in the K-dimensional variables space. Consequently, the rows in the data table form a swarm of points in the space. Once we have the space point, the mean centering involves the subtraction of the variable averages from the data. After mean centering and scaling to unit variance, the data set is ready for the computation of the first principal component where is the line in the K-dimensional space that best approximates the data in the least squares sense. The second component is also represented in the K-dimensional variable space, which is orthogonal to the first component. When two principal components have been derived they together define a plane, a window into the K-dimensional variables space [33]. The coordinate values of the observations on this plane are called scores of the vector for the first ($\mathbf{t_1}$) and the second ($\mathbf{t_2}$) components.

PLS is a method for relating two data matrices, X and Y, to each other by a linear multivariate model. For parameters related to the observations (samples, compound, object,

items), the precision of a PLS model improves with the increasing number of relevant X-variables. This corresponds to one of the most important concepts in different areas such as statistics, engineering and chemistry, when many variables provide more information about the observations than just a few variables do.

Historically, the PLS methodology has proven to be successful in different application areas such as quantitative structure-activity relationship (QSAR), multivariable calibration and process monitoring and optimization [33]. Our objective was in the process modeling, were it is used to find the relationship between spectroscopic variables measured on the process (X) at N time points and corresponding observations (Y) such as product properties and quantity. For this process, there were several responses needed to explain a system and others that with a single response would be sufficient to obtain the model.

Like PCA, every observation in a data set may be understood as one point in the X-space and another point in the Y-space. The first PLS component is a line in the X-space and another line in the Y-space. These components are calculated such that they could well approximate the point-swarms in X and Y and provide a good correlation between the positions of points along these lines in X and Y.

After the average of all the vectors is established, the two score vectors $\mathbf{t_1}$ and $\mathbf{u_1}$, may be displayed in a scatter plot and then are connected through the inner relation $\mathbf{ui_1} = \mathbf{ti_1} + \mathbf{hi}$, where hi is a residual. This relation tells us the visualization of the correlation structure

between X and Y such as presences of outliers in the X-data, Y-data and/or in the relation between X and Y. Furthermore, when there are non-linearities between the predictors and the response (Figure 2.4). This projection is also used in the ANOVA analysis were the plot of Y predicted by the model and the Y variable to make sure the analysis satisfy with the presumption of the model.

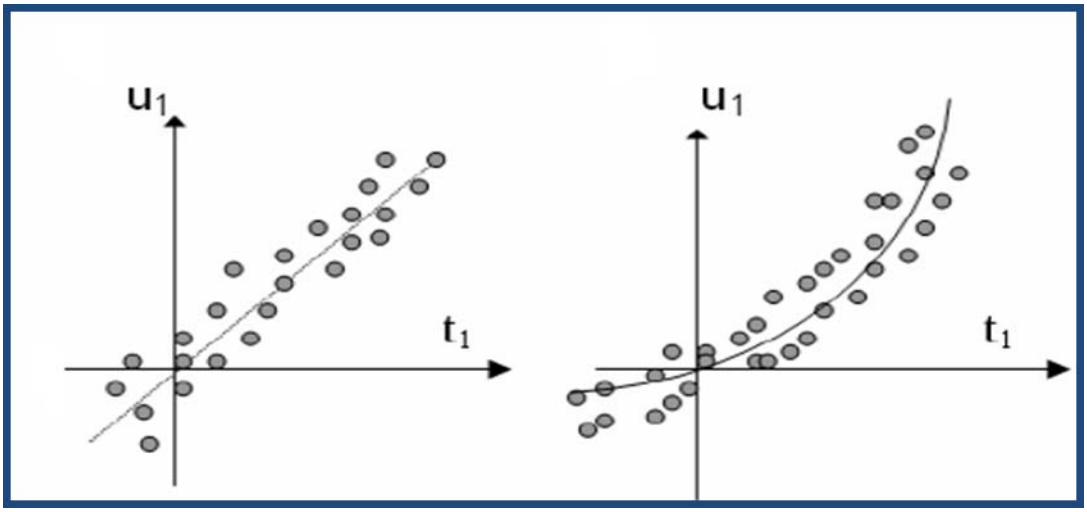


Figure 2.4 - (Left) Linear projection t_1 and u_1, in the two spaces, X and Y, were connected and correlated through the inner relation $u_{i1} = t_{i1} + h_i$. (Right) PLS score plot t_1/u_1 is useful for identifying curved (non-linear) relationships between the predictors and responses. Adapted from Eriksonn, et al. [33].

2.8 Culture of Genetically-Engineered *Escherichia coli* at High Cell Densities

High cell density fermentations require an extensive knowledge of the process. These growth strategies, together with optimization of the media composition and the application of

molecular biology methods, make it possible to grow *E. coli* to cell densities of up to 190 g/L (dry weight), while avoiding media precipitation and preventing acetate accumulation (waste by product that inhibits the microbial growth) [24] . The growth optimization depends on the dissolved oxygen, medium composition, acetate inhibition and growth techniques.

Results published in 1965 by Herbert et al. [24], indicate that the yield-constant for oxygen of an *E. coli* strain growing on glucose as a carbon source is close to 1 (1 g of oxygen is needed for the production of 1.06 g of dry *E .coli* biomass). The need of strong aeration is required for a high cell density growth, however, the major obstacle to achieve high-density growth of bacteria is in the initial substrate concentration in the liquid medium [24].

To accomplish high cell density cultures, Luria Bertani (LB) medium does not contain phosphorus, sulfur and trace elements which are fundamental to the bacterial growth [24]. For these reasons, is it more convenient to formulate the broth with this in consideration. Instead, the bacterium needs all of these nutrients, high quantities of medium ingredients become inhibitors to *E. coli* when added. In the literature, the specific recipe for high cell density does not exist, because it depends on the bacterium metabolism. It is recommendable to determine experimentally which recipe is more convenient for microbial growth. However, literature reviews such as in Riesenberg, et al. in 1991 [37] established that nutrients such as glucose at concentrations of 50 g/L, ammonium at 3 g/L iron at 1.15 g/L, magnesium at 8.7 g/L, and phosphorus at 10g/L and zinc at 0.038 g/L inhibit *E. coli* growth. These concentrations are just a guide to follow, and not a mandatory law. In 1996 Lee [27] established that *E. coli* growth that contained the maximum non-inhibitive concentration of

nutrients allowed growth to a cell density of about 15 g/L dry cell weight (dcw) compared to LB medium which is 1 g/L dcw. The anion concentration which the bacterium does not use instantly could precipitate and affect bacterial growth. Precipitates occur when non-soluble complexes of divalent metal-ammonium phosphates, magnesium phosphates and other phosphates are formed [24, 38]. The precipitation could increase the osmotic pressure and conductivity that may affect membrane potential and may activate different stress mechanisms that induce a decrease in growth rate or termination of the growth cycle [39]. To avoid the nutrient precipitation it is recommended by same authors [37] that a minimal nutrient medium containing only simple inorganic salts and a defined carbon source should be used Matsui et al. (1989) [38] improved Neighardt's method by adding solid glucose powder to a growing culture, achieving a cell density of 134 g/L dcw. Korz and coworkers improved the Risenberg method by substituting the glucose carbon source with glycerol because of its higher solubility to achieve a cell density of 148 g/L, this was the method implemented in this work.

Acetate at high concentrations might be another reason for bacterium inhibition. For this reason, the acetate metabolism and its effect on *E. coli* have been studied extensively. In 1982, some authors described acetate excretion under aerobic conditions, resulting from excess carbon source, especially glucose [37]. Others cited overloading of the TCA cycle by fast oxidation through glycolysis as the main reason for acetate accumulation [38]. During the 1990's, it became a general consensus that acetate accumulation above 2 g/L in the growth media slow down *E. coli* growth, may stop biomass build-up and may inhibit

recombinant protein biosynthesis [24]. It is known that one way to lower acetate accumulation is to reduce the growth rate by supplying the carbon source slowly. Fed batch techniques would become necessary for achievement of this goal. In the industry, fed batch technology, also known as semi-continuous or variable-volume continuous culture, is common for the following reasons: overcomes substrate inhibition or catabolite repression, increases productivity and lowers downtime cycles.

2.9 References

[1] Monod, J. The growth of bacterial cultures. *Ann. Rev. Microbiol.* **1949**, 3, 371-94.

[2] Junker, B.H.; Wang, H.Y. Bioprocess Monitoring and Computer Control: Key Roots of the Current PAT Initiative. *Wiley InterScience*, **2006**, 226-61.

[3] Fuld, G.J.; Dunn, C.G. A 50-gallon pilot plant fermenter for classroom instruction. *Appl. Microbiol.*, **1958**, 6, 15-23.

[4] Shichiji, S.; Futai, N. Control element for the process automation control of continuous fermentation. *J. Ferment. Technol.*, **1962**, 40(3), 131-9.

[5] Buckland, B.C.; Brix, T.; Fastert, H.; Gbewonyo, K.; Hunt, G.; Jain, D. Fermentation exhaust gas analysis using mass spectrometry. *Bio/Technol.*, **1985**, 3(11), 982, 985, 987-8.

[6] Neves, A.A.; Vieiria, L.M.; Menezes, J.C. Effects of preculture variability on clavulcanic acid fermentation. *Biotechnol. Bioeng.*, **2001**, 72(6), 628-33.

[7] Armiger, W.B.; Humphrey, A.E. Computer Applications in Fermentation Technology. In: Microbial Technology. Peppler, H.J.; Ed; New York: Academic Press, 2nd ed., Vol 2, Ch. 15, 1979, 375-401.

[8] Ryu, D.D.Y.; Humphrey, A.E. Examples of computer-aided fermentation systems. *J. Appl. Chem. Biotecnol*., **1973**, 23, 283-95.

[9] Schugerl, K. Progress in monitoring, modeling and control of bioprocess during the last 20 years. *J. Biotechnol.*, **2001**, 85, 149-73.

[10] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B.G.; Macaloney, G. Near- Infrared Spectroscopic Determination of Acetate, Ammonium, Biomass and Glycerol in an Industrial *Escherichia coli* Fermentation. *Applied Spectroscopy,* **1996**, 50, 102-8.

[11] Yu, K.; Phillips, J.A. The use of infrared spectroscopic techniques in monitoring and controlling bioreactors. In: Karim MN. Stephanopoulos, G.; Ed; Modeling and Control of Biotechnical Processes, 2nd IFAC Symposium. New York: Pergamon Press. 1992, 7-14.

[12] Locher, G.; Sonnleitner, B.; Fiechter, A. Automatic bioprocess control. 2. Implementations and practical experiences. *J. Biotechnol.*, **1991**, 19, 127-44.
[13] Gram, A. Mini-review: Biochemical engineering and industry. *J. Biotechnol.*, **1997**, 59, 19-23.
[14] Massart, D.L. In *Handbook of Chemometrics and Qualimetrics: Part A;* Ed.; Elsevier; **1997**; pp 1-12.
[15] DePalma, A. Moving forward with FDA's PAT initiative. *Gen. Eng. News*, 25(15), 2005, 56-8.
[16] Webber, K. FDA update: Process analytical technology for biotechnology products. *PAT*, 2(4), 12-4.
[17] Ellis, S.; Davies, B. The PAT front end using FT-NIR. *Pharm. Formulation Qual.*, **2005**, 10, 56-63.
[18] Wang, F.; Wachter, J.A.; Antosz, F.J.; Berglund, K.A. An investigation of solvent-mediated polymorphic transformation of progesterone using Raman spectroscopy. *Organic Process Research & Development,* **2000,** 4, 391-5.
[19] Coates, J. Interpretation of Infrared Spectra, A Practical Approach. *Encyclopedia of Analytical Chemistry*, Meyers, R.A.; Ed.; Wiley: Chichester, 2000; 10815-37.
[20] Reich, G. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Advanced Drug Delivery Review*, **2005**, 57, 1110-5.
[21] Raidyanathan, S.; Macaloney, G.; McNeil, B. Fundamental investigation on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring. *Analyst* **1999**, 124, 157-62.
[22] McClain, B.L.; Clark, S.M.; Gabriel, R.L.; Ben-Amotz, D. Educational Applications of IR and Raman Spectroscopy: A Comparison of Experiment and Theory. *J. Chemical Education,* **2000**, 77, 654-60.
[23] Ramos, S.; Rohrback, B. FDA Update-QbD, Quality Assessment and PAT. *IFPAC/PAT Summit with INDUNIV*, June 14, 2007, 1-23.
[24] Shiloach, J.; Fass, R. Growing *E. coli* to high cell density – A historical perspective on method development. *Biotechnology Advances*, **2005**, 23, 345-57.
[25] Moisheky, Z.; Melling, P. J.; Thomson, M.A. In situ real-time monitoring of a fermentation reaction using a fiber optic FT-IR probe. *ADVANSTAR*, **2001**, 1-5.
[26] McGovern, A.C.; Ernill, R.; Kara, B.V.; Kell, D.B.; Goodacre, R. Rapid analysis of the expression of heterologous protein in *Escherichia coli* using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production. *J. Biotechnology.*, **1999**, 72, 157-67.
[27] Lee, H.L.T.; Boccazzi, P.; Gorret, N.; Ram, R.J.; Sinskey, A.J. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, **2004,** 35, 131-7.
[28] Cimander, C.; Mandenius, C.F. Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. *J. Chem. Technol. Biotechnol.,* **2002**, 77, 1157-68.

[29] Arnold, S.A.; Gaensakoo, R.; Harvey, L.M.; McNeil, B. Use of At-line and In-situ Near-Infrared Spectroscopy to Monitor Biomass in an Industrial Fed-Batch *Escherichia coli* Process. *Wiley Periodicals*, **2002**, 405-13.

[30] Kokot, S.; King, G.; Keller, H.R.; Massart, D.L. Application of Chemometrics for the Selection of Microwave Digestion Procedures; *Analytica Chimica Acta*, **1992,** 268, 81 – 94.

[31] Ferrer, A. Técnicas de análisis multivariante para el análisis, monitorización, detección de fallos, predicción y optimización de procesos altamente automatizados. *Advanced Hands-on PAT Seminar Series: 2.1 Real – Time Process Monitoring Using MVA*, April 26, 2006.

[32] McClain, B.L.; Clark, S.M.; Gabriel, R.L.; Ben-Amotz, D. Educational applications of IR and Raman Spectroscopy: A comparison of experiment and theory. *Journal of Chemical Education*, **2000,** 77, 654-60.

[33] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Tryg, J.; Wikstrom, C.; Wold, S. *Multi- and Megavariate Data Analysis Part 1*. Umetrics AB, Ch. 16, 2006, 337-60.

[34] Riesenberg, D. High-cell density cultivation of *E coli*. *Curr. Opin. Biotechnol.*, **1991**, 2, 380-4.

[35] Dean, J.A. Lang's Handbook of Chemistry. McGraw-Hill; Ed; NY, 1979, 5-12.

[36] Winzer, K.; Hardie, K.R.; Williams, P. Bacterial cell to cell communication: sorry, can't talk now- gone to lunch. *Curr. Opin. Microbiol*., **2002**, 5, 216-22.

[37] Neidhardt, F.C.; Bloch, P.L.; Smith, D.F. Culture medium for enterobacteria. *J. Bacteriol*., **1974**, 119, 736-47.

[38] Matsui, T.; Yokota, H.; Sato, S.; Mukataka, S.; Takahashi, J. Pressurized culture of *Escherichia coli* for high concentration. *Agric. Biol. Chem.*, **1989**, 53, 2115-20.

[39] Korz, D.J.; Rinas, U.; Hellmuth, K.; Sanders, E.A.; Deckwer, W.D. Simple fed-batch technique for high cell density cultivation of *Escherichia coli*. *J Biotechnol*, **1995**, 39:59-65.

3 MATERIALS AND METHODS

3.1 Microorganism

The strain used in this research work was *Escherichia coli* K-12, which is a genetically engineered bacterium that expresses the green fluorescent protein (GFP) [1]. GFP is a recombinant protein that consists of 238 amino acids and has a molecular weight of 23 kDa [1]. This protein is found naturally in the jellyfish *Aequorea victoria* and fluoresces green when exposed to ultraviolet (UV) light [1]. The *E. coli* strain and the plasmid that expresses GFP are commercially available (Bio-Rad pGlo kit). The plasmid contains the P_{BAD} promoter and the *ara*C repressor gene. However, the genes which code for arabinose catabolism, araB, araA and araD, have been replaced by the simple gene which codes for the GFP. Therefore, in the presence of the arabinose in the cell culture, araC promotes the binding of RNA polymerase and GFP is produced [1].

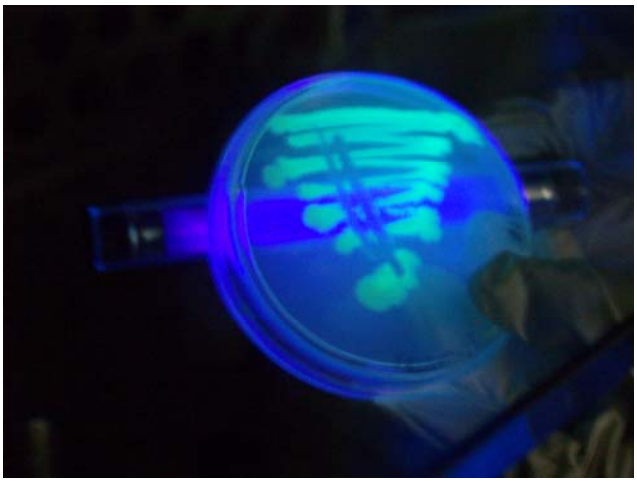


Figure 3.1 - Petri dish cultured with *E. coli* K-12. Green fluorescence can be observed upon application of UV light to the system [1].

3.2 Inoculum Preparation

Inocula were prepared using a two-stage process as described in Hall et al. [2]. Stock mutant culture from a petri dish with LB/ampiciline and arabinose were first inoculated into 280 mL of filter-sterilized nutrient medium containing the following components (amount in g/L): glycerol, 3.53; ammonium sulfate, 5.35; yeast extract, 8.57; KH_2PO_4, 4.62; K_2HPO_4, 25; citric acid, 1.7; EDTA, 0.0084 and ampiciline, 0.1. The medium composition was followed according to Hall et al. [2] and Korz et al. [3]. The cultures were incubated for 15 hours at 37°C and 300 rpm in a temperature controller orbital Innova 4000 shaker (New Brunswick, NJ) (Figure 3.2).

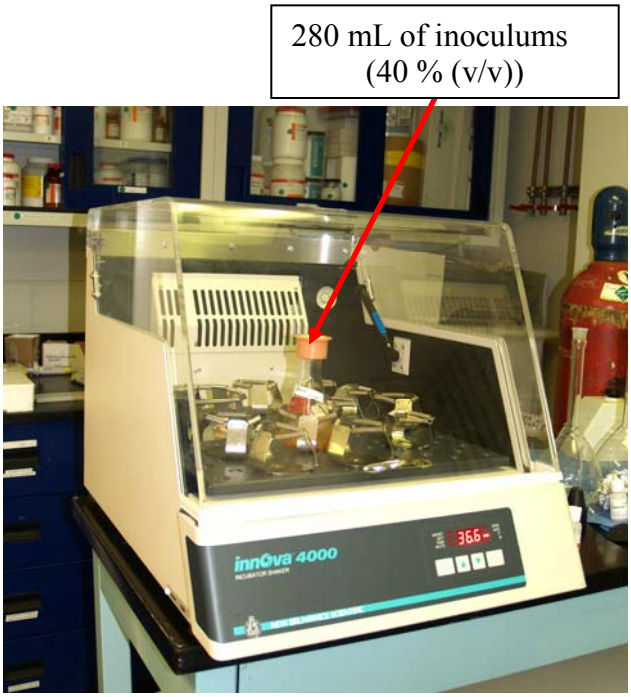


Figure 3.2 - Inoculum preparation in an Innova 4000 shaker incubator.

3.3 Batch and Fed-Batch Fermentation and Setting-Up

Inocula cultured as described above were transferred at a 40% (v/v) level to a 2-L sterilized-New Brunswick Bioflow 3000 bioreactor (Figure 3.3) containing the following components (amount in g/L): glycerol, 3.53; ammonium sulfate, 2.67; tryptone, 8.57; yeast extract, 17.1; KH_2PO_4, 4.62; K_2HPO_4, 25; citric acid, 1.7; EDTA, 0.0084 and ampiciline, 0.1. In all runs, the pH was monitored using a Mettler Toledo pH probe and was controlled with proportional-integral-derivative (PID) control system with a pH 7.0 set point using 2N sulfuric acid and 4N sodium hydroxide solutions. Dissolved oxygen was measured using a Mettler Toledo DO membrane probe and was controlled with a PID controller at a set point of 30% oxygen saturation by oxygen-air enrichment with pure oxygen gas addition. Temperature was monitored using an resistant temperature detector (RTD) and was controlled at a set point of 37°C through addition of hot water flow through a concave bottom jacket. Bioreactor contents were mixed at 350 rpm using Rushton turbine impellers.

After the inoculation, the fermentation process was run for the first four hours in a batch mode and then it was continued in a fed-batch mode by adding feed at a rate of 75 mL/h containing the following components (amount in g/L): glycerol, 200; yeast extract, 360; tryptone, 180; and ampiciline, 0.1. At the moment the feed was started, bacterial growth should be in its exponential phase, and for this reason, the carbon source is required for bacterial growth [4]. Since the beginning of the batch fermentation, samples were collected every hour for fluorescence (SpectraMax Gemini EM fluorometer), HPLC (Waters), NIR (Buker 300), FT-IR (Varian 800) and Raman (Kaiser Optical System Inc.) assays.

Fermentation samples were collected periodically for optical density (OD_{600}) measurement at 600 nm (Genesys 6 spectrophotometer).

After ten hours, an arabinose stock solution was added to induce recombinant protein (GFP) expression; this activated the plasmid arabinose operon and induced GFP expression [1]. In this process of protein production, since GFP fluoresces, samples were collected every hour and a SpectraMax Gemini EM fluorometer was used to measure protein in the sample.

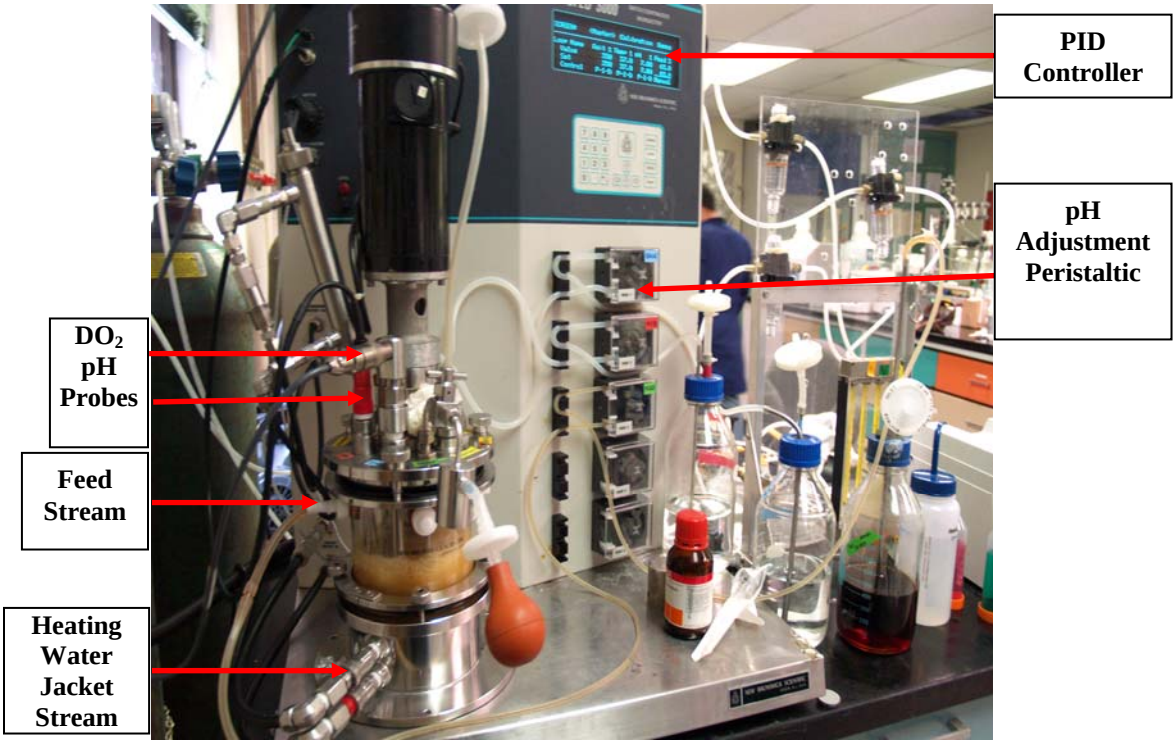


Figure 3.3 - Fermentation in a fed-batch reactor (BioFlow 3000).

3.4 Determination of Glycerol and Acetate Concentrations

Glycerol and acetate concentrations were determined using high performance liquid chromatography (HPLC). Fermentor samples were collected and centrifuged using an Eppendorf centrifuge system (5415 D). The supernant was filtered using a 0.22 μm Millipore membrane and was then analyzed using high performance liquid chromatography with a BioRad HPX-87H, 300 mm 7.8 mm organic acids column (BioRad Labs, CA) operating at 55ºC. A refractive index detector was used for compound detection. Dilute sulfuric acid (0.1 N) at a flow rate of 0.6 mL/min was the mobile phase. Mixed component concentration verification standards containing glycerol and acetate were periodically injected to the HPLC to verify calibration accuracy. A similar procedure was already reported by Sáez Miranda et al., [5].

3.5 Determination of Cell Mass Concentration and Cell Mass Growth

Cell mass concentration was determined by a gravimetric method. One mL of whole culture broth samples were filtered using a PTFE Cole Parmer manifold filtration unit. A pre-weighed cellulose acetate membrane filter (25 mm, 0.22 μm, Cole Parmer Instruments) was used to retain the cell mass, was washed three times with deonized water and dried in an oven at 50°C until constant weight was attained. Triplicate analyses of a single sample were performed at each sampling time to obtain reliable results.

Cell mass was also alternatively determined using a Genesys 6 spectrophotometer at a wavelength of 600 nm. The absorbance or optical density (OD_{600}) measurement is linearly related to the cell dry weight. To maintain Beer's Law linearity, samples measuring over 0.6 units of absorbance were diluted for corrected OD determination. The dilution range depends on the spectrophotometer linear range, which for our spectrophotometer is 0.01 - 0.6 at 600 nm. For this method, sterile medium was used as the "blank" to zero the spectrophotometer and was also used as sample diluent, when necessary.

Like Saez Miranda et al., [5] established in a recent article, a linear relation between the dry cell mass concentration (g/L dcm) and optical density (absorbance units) exists for a gram-negative bacteria, which is in agreement with the *E. coli* results shown in Figure 3.4. The following equation was obtained as the slope of the linear correlation between the dry cell mass concentration and the absorbance at 600 nm using whole broth samples collected in several runs (Equation 3.1).

3.6 Recombinant Protein Concentration

In this research work, the transformed *E. coli* has the betalactamase (*bla*) plasmid, which provides physiological advantages such as ampicillin resistance and recombinant protein expression in the presence of arabinose [1]. The process where the arabinose is related to the operon activation and the protein expression is called induction. In our fermentations, the induction was activated during the exponential phase of the culture to achieve a high protein expression. Once the bacteria cells start to produce the protein, a fluorometry

spectrophotometer was used to quantify the protein expression by fluorescence. GFP has two excitation peaks, a major one at 395 nm and a minor one at 475 nm [6]. Its emission peak is at 509 nm in the lower green portion of the visible spectrum. We used this information to determine the GFP synthesis quantification on the fermentation sample for each hour after induction (Figure 3.5). Figure 3.6 shows an image from confocal microscopy that displays the inclusion body in the bacteria cytoplasm.

$$DCM = -0.217 + 0.579 \cdot OD\,at\,600nm \qquad \textbf{3.1}$$

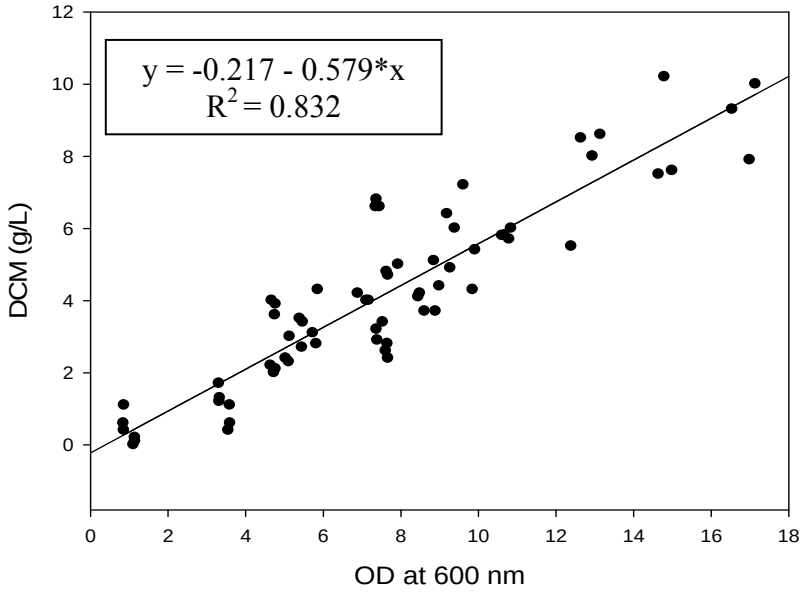


Figure 3.4 - Linear correlation between dry cell mass concentration (g/L dcm) and OD at 600 nm for *E. coli* K-12 strain. Data shown includes different fermentation experiments.

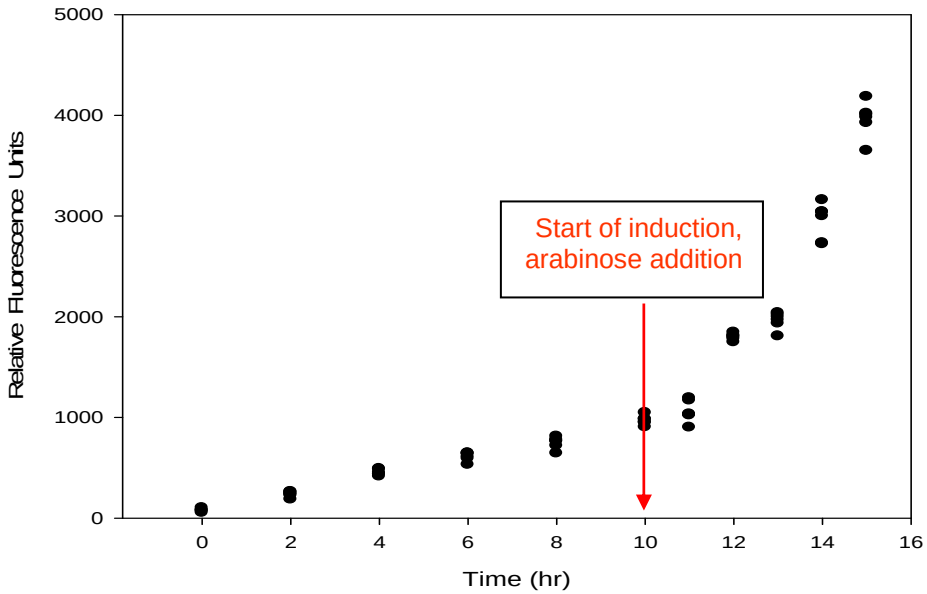


Figure 3.5 - Variation of fluorescence after protein synthesis is induced by addition of arabinose.

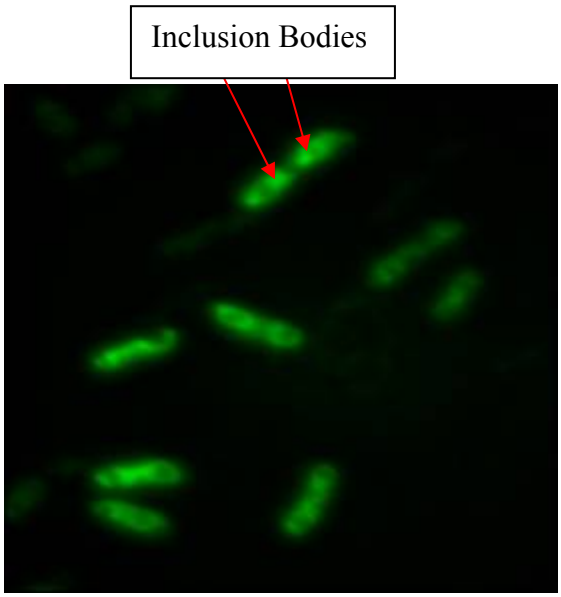


Figure 3.6 - Image of *E. coli* K-12 expressing GFP. Taken with a confocal microscope at Microscopy Laboratory, Department of Biology (UPR-Mayagüez).

3.7 Fermentation Process Simulation

In this work, we used Super Pro Designer software, which is the leading process simulator in the biotechnology industry to simulate the whole fed-batch process to produce the GFP. This software is used throughout the life cycle of product development and commercialization to facilitate process optimization, cycle time reduction, improve team collaboration, and shorten the time to market.

Figure 3.7 shows the sketch of simulation of all the process in this research starting with the inoculum, batch process, to the fed-batch fermentation, and finally to the purification procedure using a hydrophobic interaction column (HIC). Figure 3.8 can help us to visualize the time schedule schematic as a Gantt chart.

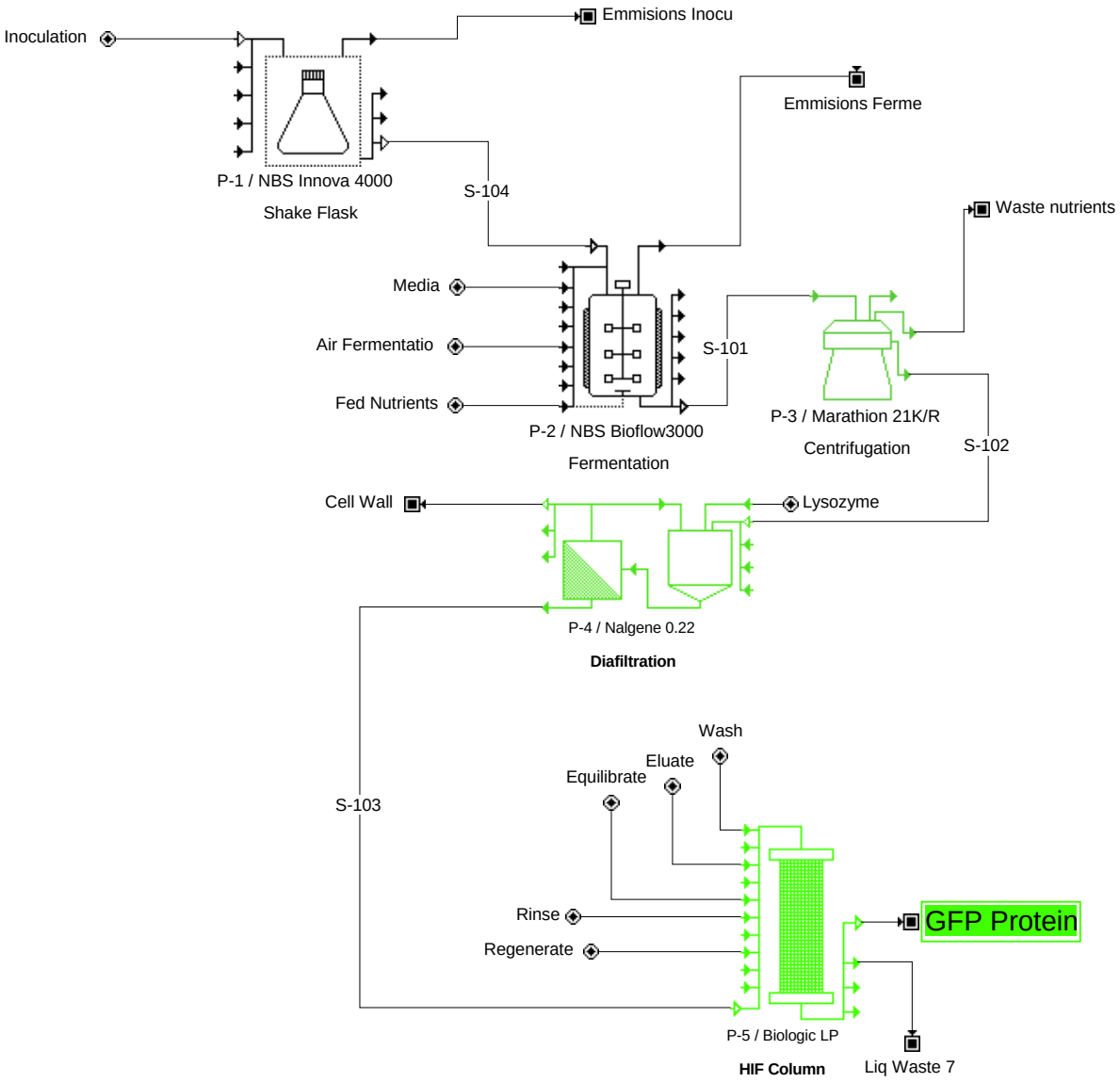


Figure 3.7 - Green fluorescent protein (GFP) production flowsheet using Super Pro Design software.

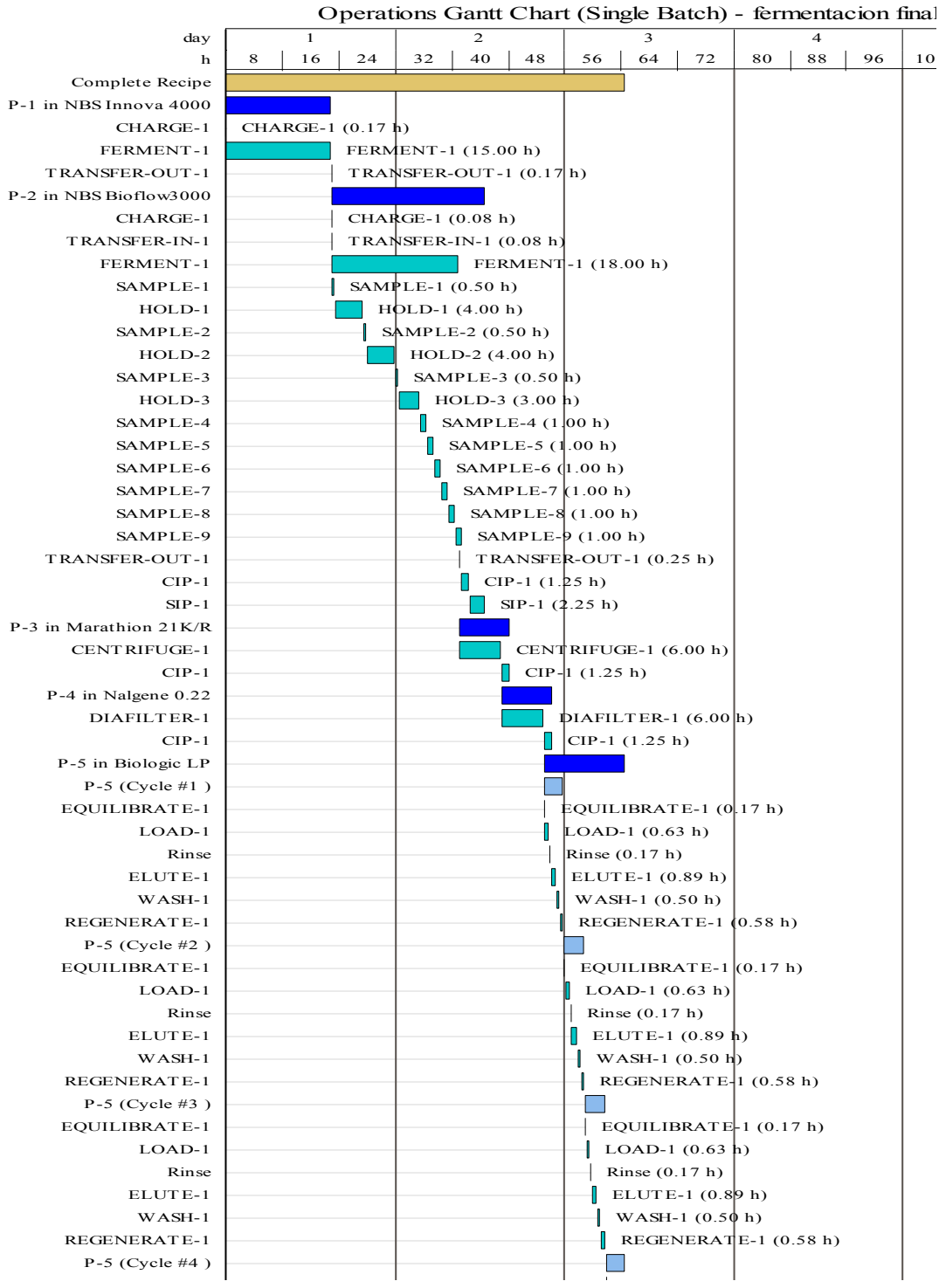


Figure 3.8 - Gantt chart schedule for GFP production and purification, as generated by Super Pro Designer software.

3.8 NIR, IR and Raman Equipment Procedure

Each sample collected from the fed-batch experiments was analyzed at-line using NIR, IR and Raman spectrometry.

The NIR uses an immersion (transmission) probe in 15 mL of sample with a pathlength of 1 mm. Each sample was not only referenced against the internal reference fiber, it was adjusted to give specific illumination energy on installation. The probe was also referenced with air (external reference), which must be collected prior to probe insertion in the sample. After equilibrating the sample to room temperature, samples were scanned in triplicate. The parameters on the software were: 128 scans, 4 cm^{-1} resolutions and the method used Zaineett fiber. The OPUS software was used for the data acquisition and transformation into GRAMS data. Figure 3.9 shows a representative NIR spectrum for a sample collected during a fermentation run, where the high-density growth was proportional at 5580 cm^{-1} [2].

Raman spectra were collected using a Raman probe immersed in 15 mL of sample. After sample analysis, calibration was performed with a cyclohexane standard. Samples were analyzed with ~ 10 exposures and ~ 5 accumulations. GRAMS software was used for the data acquisition. Figure 3.10 shows representative Raman spectra of samples collected from a fermentation run.

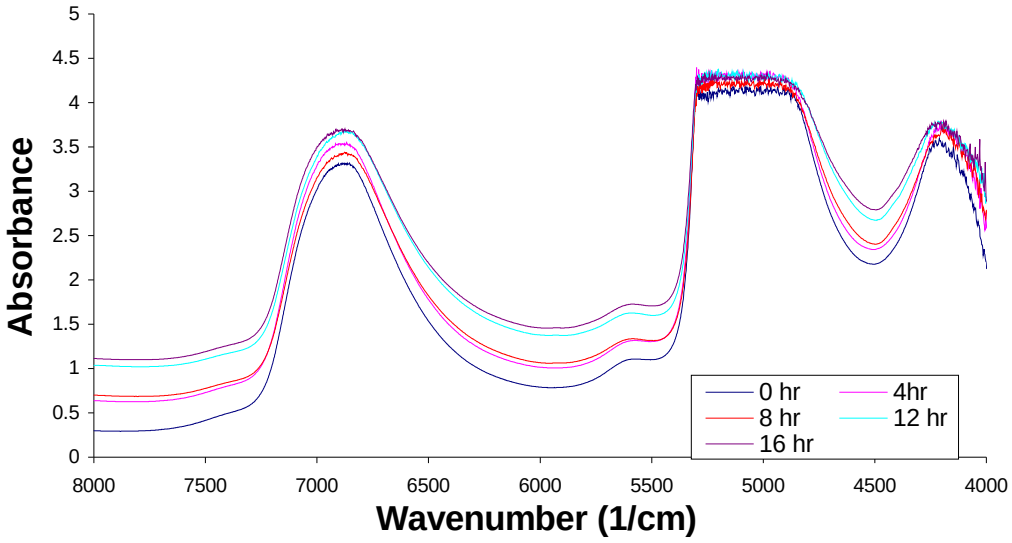


Figure 3.9 - Sample collected from a NIR spectrum of *E. coli* fermentation producing recombinant GFP protein. Induction was started at the 10th hour.

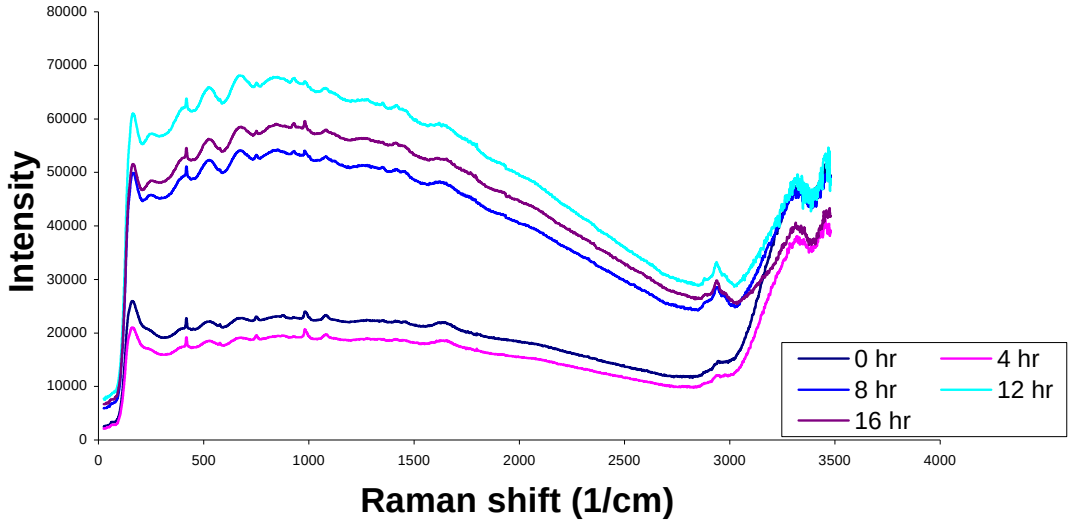


Figure 3.10 - Raman spectroscopy of fermentation samples with induction started between samples 3 and 4.

FTIR uses an attenuated total reflectance probe (ATR) to obtain spectra during the course of fermentation. Three spectra were collected from the culture broth after background was collected, which adjusted the equipment to its electric potential and also with an external reference (air and CO_2). The collection of each spectrum was at the rate of 4 cm^{-1} using 100 scans per spectrum. Figure 3.11 shows a result of the measurements of the nutrients and by-products using FT-IR vibration energy stretch for fermentation run.

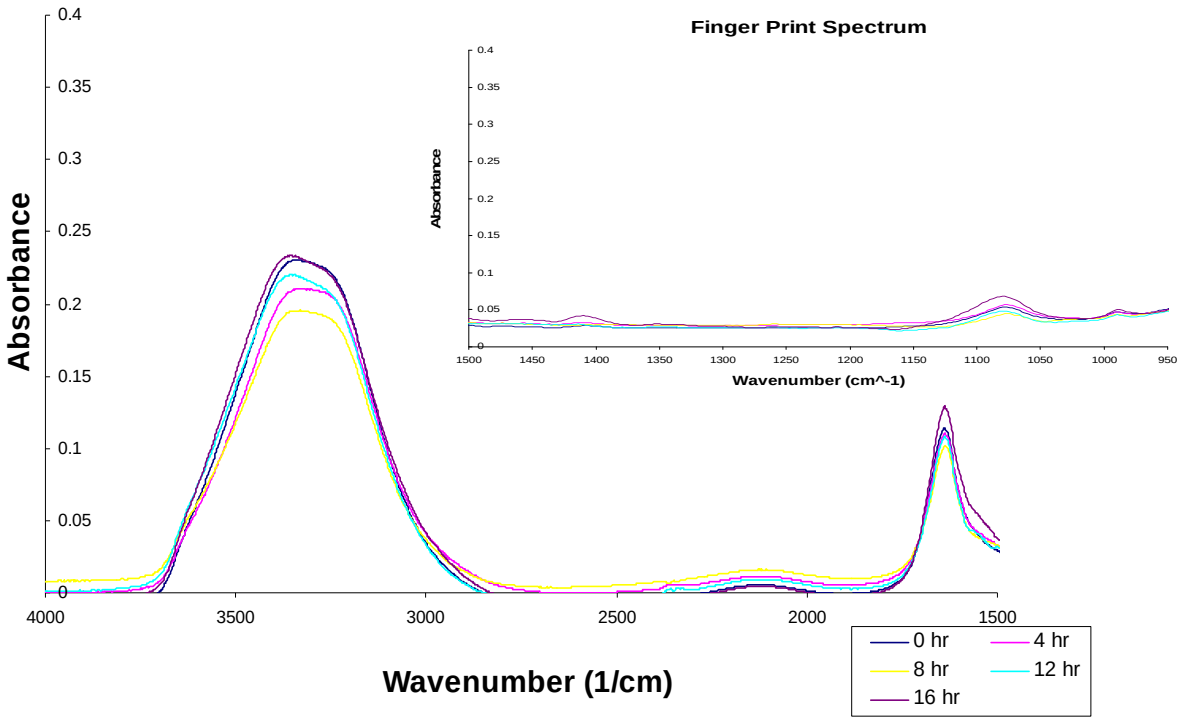


Figure 3.11 – FT-IR spectra of fermentation samples. Calibration was performed to correlate the characteristic peaks of glycerol, acetate and ammonium sulfate.

3.9 Analysis of NIR, Raman and FTIR Spectra

In the project, different types of components were monitored using different reference techniques such as HPLC, UV, fluorescent spectrometry, and gravimetric analysis. Some of these techniques are laborious and time consuming. To speed up the sample analysis, NIR, IR and Raman, along with multivariable data analysis were performed using SIMCA-P+11 (Umetrics) and Pirouette software. The steps followed in SIMCA-P+11 and/or Pirouette are:

1. Definition of project- Import the raw primary data set. The raw material consists of the NIR, IR and Raman spectra, which will be on GRAM files and the nutrients such as glycerol using HPLC, by products such as acetate using HPLC. The products such as recombinant protein concentration using fluorescence spectrophotometry, and microbial concentration using optical density and the dry cell concentration (gravimetric) method.
2. Preparation of data- Specify which variables are process variables (X) and which are responses (Y). Transform the data and group the observations if necessary. The spectrum wavenumber would be the X primary variable and the experimental analysis data would be the Y response [8].
3. Fitting the model, a PCAX of all the data- The PCAX analysis will be useful to determine the scores or loading of each factor, find out which variable has the same information and the variation and the predictive ability explanation [8]. In the experiment the PCAX is determined, which spectroscopies explain better the recombinant protein, nutrients and by-product concentration. These results will be

compared with the theorical hypothesis, using the energy vibration spectra, Raman spectroscopy may explain better the nutrient and by-product concentrations [9], NIR may explain the biomass concentration [2] and IR may explain the protein production [10].

4. Detection of possible outliers- After the PCAX is set, two different analyzes are need to detect the presence of outliers. Two different kinds of outliers may exist: the outliers that are out of the hyperplane and the outliers that are out of the interval of confidence. To determine the hyperplane outliers a DMoxd plot analysis will be useful. This plot displays the residual standard deviation (RSD) of the observations in the X space, after all the computed dimensions. The RSD is proportional to the distance of the observation to the model hyper plane. Larger DModX than the critical limit indicates that the observation is an outlier in the X space. The 95 % interval confidence outliers can be detected with the Score plot. The plots show the possible presence of outliers, groups, similarities and other patterns in the data [8].
5. Fitting of model by PLS of all the data- After the determination of the representative factor, a PLS analysis is required to determine the relation between the X and Y. This model shows which factor could explain and predict better. Also, the analysis will tell us the difference among each variable in the group and which variables and which factors better predict and explain that group. It is very important to know that the PLS like PCAX could locate the outlier's presence [8].

6. Examination of results- After running the model, a careful examination of the model is required. For that it is important to compare with the theorical hypothesis and develop new perspectives on the underlying meaning in the data [11].
7. Validation of the model- To determine the reliability of a model it is necessary to include a validation step. This means to determine where the model breaks off the critical points of the process. The two different tests used are: test set analysis, where the properties of samples (which had not been included in the calibration models) were predicted using the calibration models and the leave-one-out analysis were the properties of the samples (which had been included in the calibration models) is used to determine if there are outliers [12].
8. Prediction and classification of PLS model- Compare the results with previous expectations. Determine the variance of the prediction of the new experimental set and determine if it could be applicable to another fermentation setup [11].

3.10 References

[1] Mardigian, R. Biotechnology Explorer™ pGLO™ Bacterial Transfomation Kit. Bio-Rad Laboratories. 166-0003EDU, catalog number, 1-62.

[2] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B.G.; Macaloney, G. Near-Infrared Spectroscopic Determination of Acetate, Ammonium, Biomass and Glycerol in an Industrial *Escherichia coli* Fermentation. *Applied Spectroscopy,* **1996**, 50, 102-8.

[3] Korz, D.J.; Rinas, U.; Hellmuth, K.; Sanders, E.A.; Deckwer, W.D. Simple fed-batch technique for high cell density cultivation of *Escherichia coli*. *J. Biotechnol.*, **1995**, 39, 59-65.

[4] Shiloach, J.; Fass, R. Growing *E. coli* to high cell density – A historical perspective on method development. *Biotechnology Advances*, **2005**, 23, 345-57.

[5] Sáez-Miranda, J.C.; Saliceti-Piazza, L.; McMillan, J.D. (2005). Measurement and Analysis of Intracellular ATP Levels in Metabolically Engineered *Zymomonas mobilis* Fermenting Glucose and Xylose Mixtures. *Biotechnol. Prog.,* **22**(2): 359-368.

[6] Title of Site. http://www.biotek.de/products/tech_res_detail.php?id=54 (February 20, 2001).

[7] Lun-Vien, D.; Colthup N.B.; Fateley, W.G.; Grasselli, J.G. The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecule. New York: Academic Press, Ch. 10-15, 1991, 155-261.

[8] FPAC/PAT Summit with INDUNIV – Process Analytical Technology Pharmaceutical/ Biopharmaceutical Manufacturing, June 2007 San Juan, Puerto Rico, Essential Meeting in Puerto Rico for Process Analytical Technology, not accessed yet.

[9] Lee, H.L.T.; Boccazzi, P.; Gorret, N.; Ram, R.J.; Sinskey, A.J. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, **2004,** 35, 131-7.

[10] McGovern, A.C.; Ernill, R.; Kara, B.V.; Kell, D.B.; Goodacre, R. Rapid analysis of the expression of heterologous protein in *Escherichia coli* using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production. *J. Biotechnology,* **1999**, 72, 157-67.

[11] Advanced Hands-on PAT Seminar Series: 2.1 Real – Time Process Monitoring Using MVA, April 2006 San Juan, Puerto Rico, IBS Caribe, Inc., not accessed yet.

[12] Cooper, J.B. Chemometrics analysis of Raman Spectroscopic data for process control application. *Chemometric and Intelligent Laboratory System,* **1999**.46, 231-47.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Fermentation Profile

In this study, intracellular GFP-producing *E. coli* fed-batch cultivations were monitored with three different spectroscopies: FT-IR, NIR and Raman. Figure 4.1 shows values from quantitative analysis for biomass, intracellular GFP, glycerol and acetate during a representative run. The ±2 standard deviation error bars represent the average of nine runs performed as described earlier in Chapter 3. There is good run-to-run reproducibility in the concentration profiles for all components.

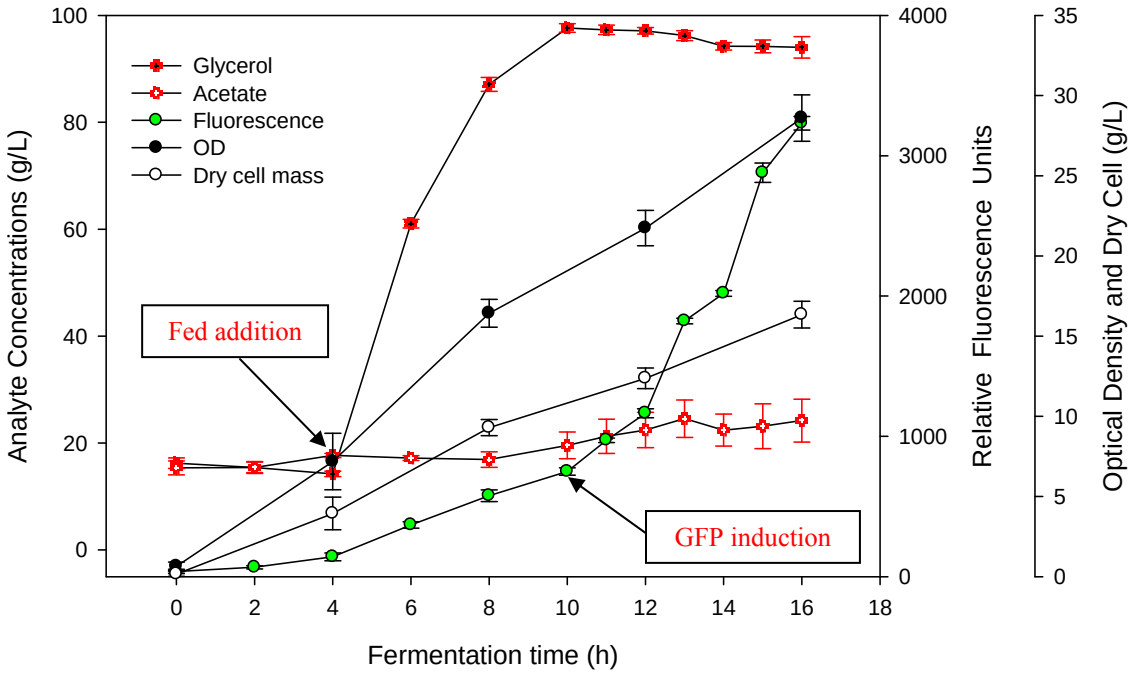


Figure 4.1 – Accumulation profiles for high-density *E. coli* fed-batch fermentation bioprocess. The plot includes: quantitative analysis of biomass, glycerol, acetate and intracellular GFP.

4.2 Determination of Acetate and Glycerol Concentrations using Raman Spectroscopy

The following data analysis was performed to generate a model for the prediction of acetate and glycerol concentrations in a fed-batch reactor. Raman spectroscopy is ideally suited for the analysis of acetate and glycerol due to the strong absorption of Raman energy by organic molecules and the weak absorption for -OH- band, which enables direct spectra acquisition. Conversely, we believe that the monitoring of acetate and glycerol during fermentations using Raman is not abundant in the literature. Hall et al. [1] used multiple linear regression (MLR) to derive a linear relationship between the absorption of NIR energy at discrete wavelengths and the analyte concentrations. Their group basically determined bands, where differences at the spectrum by the pure analytes and by regression analysis were subtracted at analyte concentration during fermentation. MLR is a classical method that assumes statistical independence of the X-variables and Y-variables. In the Hall et al. case, the relationship between the bands and the responses were not included, and the X-variables were exact and completely (100 %) relevant (each variable and noise were be incorporate in the model). Based on their assumptions, truncate the intrinsic relationship by the organism pathways by analyte consumption and production.

To distinguish the difference between the MLR analyses and PLS, Moisheky et al. [2] used FT-IR spectroscopy and PLS to quantitate the critical analytes in an alcoholic fermentation. They used a factorial design that consisted in different combinations of maximum and minimum levels of concentration of analytes (n=98 standards) to derive a calibration model

and to later predict several fermentations. Their method consisted in a factorial design as a calibration model, which was not in agreement with the microorganism's pathway. Chemical engineering fermentations are analogous to a chemical reaction among reactants converting to products and by-products. In each reaction there is a limiting reactant that in fermentation vocabulary is called a "substrate," and unwanted products are called by-products. One of the most significant differences between chemical reactions and bioreactions is that substrates can inhibit microorganism growth, for that reason the monitoring of these concentrations is fundamental for a proper microorganism growth. Because factorial design assumes that each factor (analyte) is independent, and because we already explained there is an indirect relationship between the substrates and by-products, a full factorial design might not be a good method to use in such cases for different reasons. First, it would require a lot of samples to accomplish the calibration method (n= 98), were fifty percent of these experiments would physically be impossible. As an example, having a combination of simultaneous high concentrations of substrates and byproducts is not real. Another problem is that a lot of unuseful combinations promote noise that would affect the model. For instance, to have a good model, Moisheky had to use approximately 13 factors to explain each analyte in the fermentation, using a full factor design as a calibration method.

A better suited experimental design was required to obtain the acetate, ammonium and glycerol calibrations to determine the proportion of the components along the fermentation knowing that the analytes are dependently related. Also, the substrate and byproduct

concentrations were analyzed using high performance liquid chromatography (HPLC) as the reference method.

The calibration of the culture media with and without the microorganism was needed to correlate the change of concentrations of the substrate and by-products using Raman. The calibration was performed using an extreme vertices mixture design programmed in Minitab software [3]. The mixture design consisted of nine different samples in triplicate of the substrates and by-products: glycerol, acetate and ammonium. Preliminary experimental studies induced to use the intervals (scaled concentration): $0.05 < x / \% < 0.90$ for glycerol, $0 < x / \% < 0.85$ for acetate and $0.1 < x / \% < 0.3$ for ammonium. The mixtures are related one to another as shown in Figure 4.2.

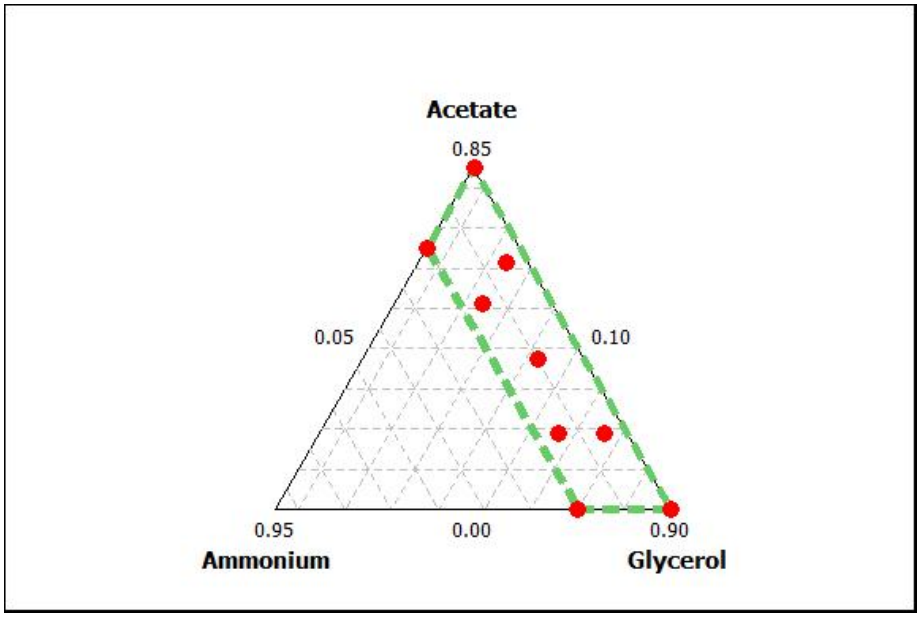


Figure 4.2 - Experimental design: extreme vertices of substrate mixture test analysis.

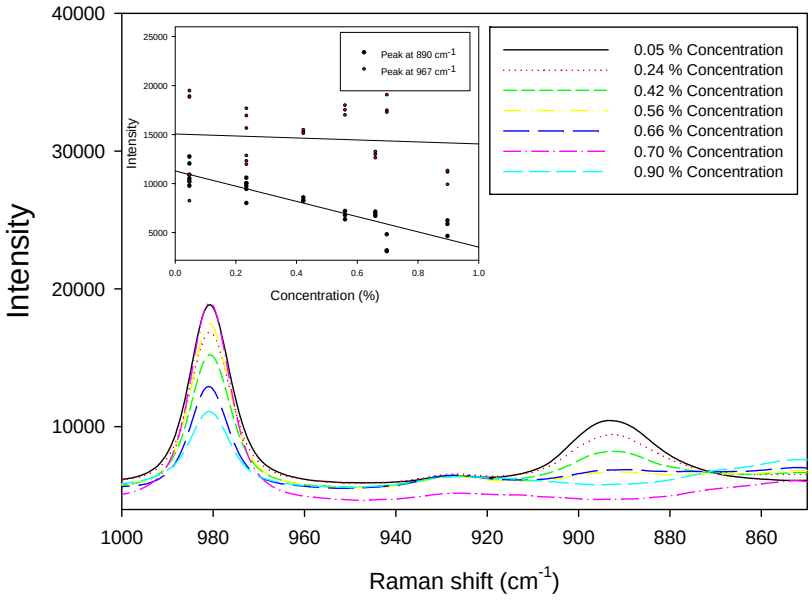
4.2.1 Exploratory Spectral Analysis

The selection of the spectral region of interest was the primary step in the chemometric exploratory data analysis. Patterns of association exist in the whole spectral data set, but the relationship between the sample and the variable of interest can be difficult to predict if uncorrelated spectra associated to noise may produce possible outliers. Figures 4.3 and 4.4 present the spectra of mixture spectra with a variation at the concentration of glycerol and acetate and a plot between the peaks that better fit the analyte concentration profile. The plots fit very well between the concentration profiles of the analytes and the intensity of Raman except for the 920-860 cm^{-1} range.

4.2.2 Description of the Primary Data Set

The primary data consisted of 27 samples of acetate, ammonium and glycerol mixtures, prepared and analyzed at-line, with HPLC as reference method. Raman spectra were acquired from each sample of mixtures at triplicate. The spectra were obtained in the 3000 - 100 cm^{-1} range but it was analyzed between the 3000 - 920 cm^{-1} range, giving a total of 3332 X-variables (absorbance), FT-IR spectral wavenumbers, and three are Y variables (mixture proportions). The data set has 27 observations (samples, row vectors).

a)



b)

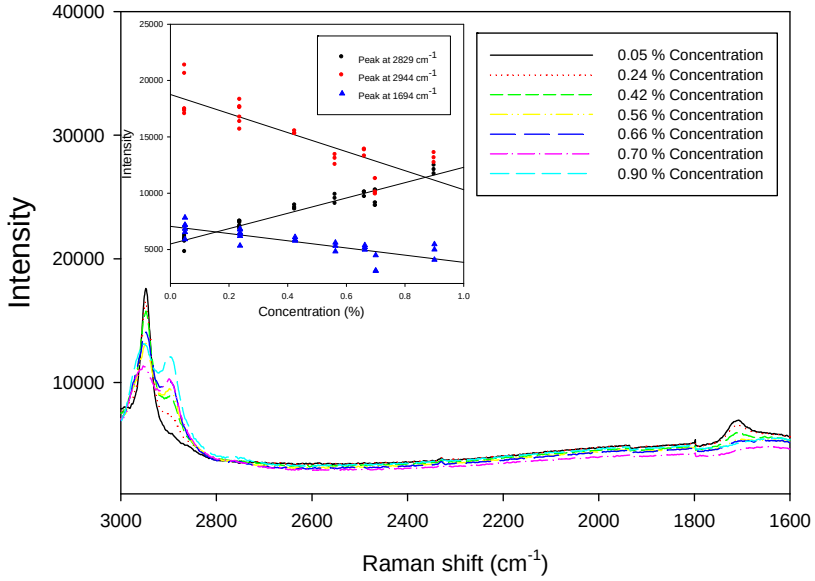
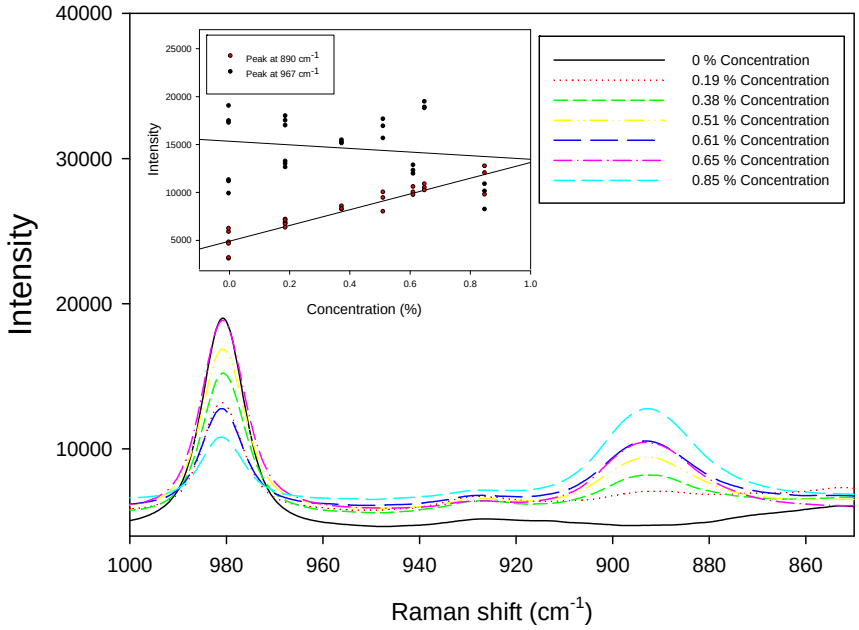


Figure 4.3 – Raman spectra of glycerol calibration profile at a) 890 and 967 cm^{-1} and b) 2829, 2944 and 1694 cm^{-1}.

a)



b)

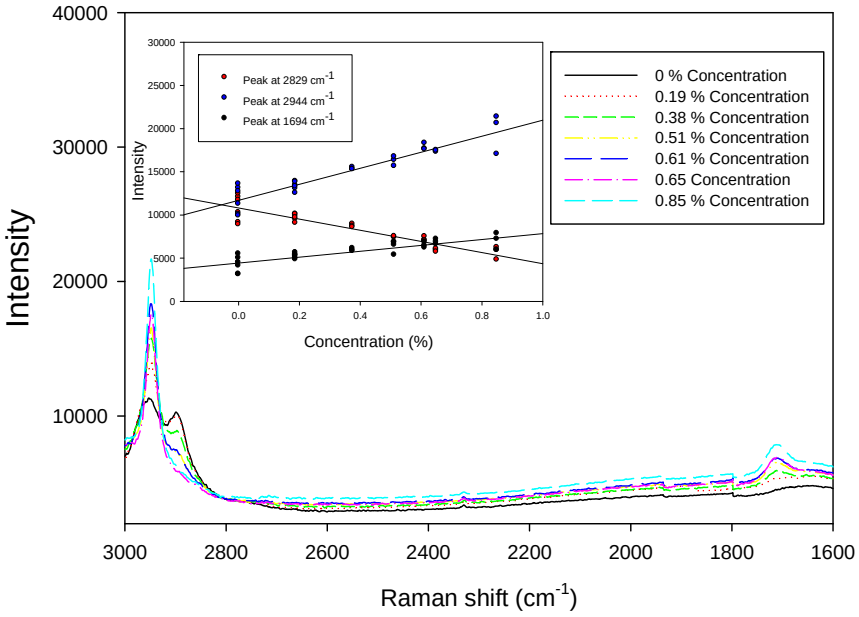


Figure 4.4 - Raman spectra of acetate calibration profile at a) 890 and 967 cm^{-1} and b) 2829, 2944 and 1694 cm^{-1}.

4.2.3 Evaluation of Raw Data

From an examination of the raw spectra (Figure 4.5), both the additive and multiplicative effects are apparent. The higher the concentration of the mixtures, the higher the absorbance in the spectra.

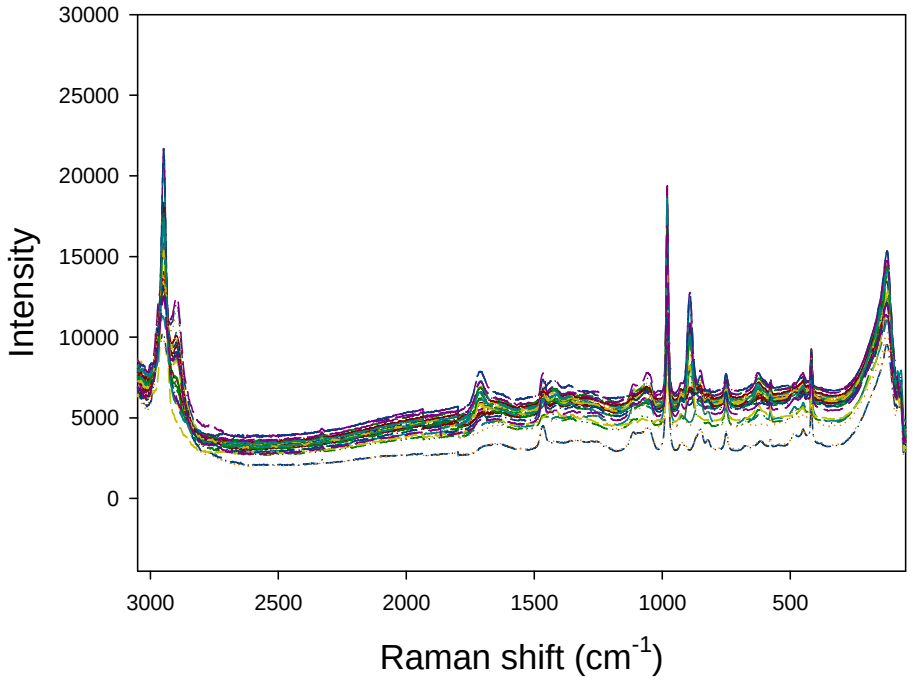


Figure 4.5 - Raman spectra of mixture test: glycerol, acetate and ammonium.

4.2.4 Suitable Fit of X/Y Variables

Estimating the optimal number of factors is one of the key steps in PLS. For this case the PLS model was explained using SIMCA software with three factors in order to account for 99% of the variance in the system (Figure 4.6). The plot displays the cumulative R^2 and Q^2 for the Y matrix, after each component. R^2Y (cum) is the percent of the variation of the entire Y explained by the model and Q^2 (cum) is the percent of the variation of the entire Y that can be predicted by the model (Table 4.1).

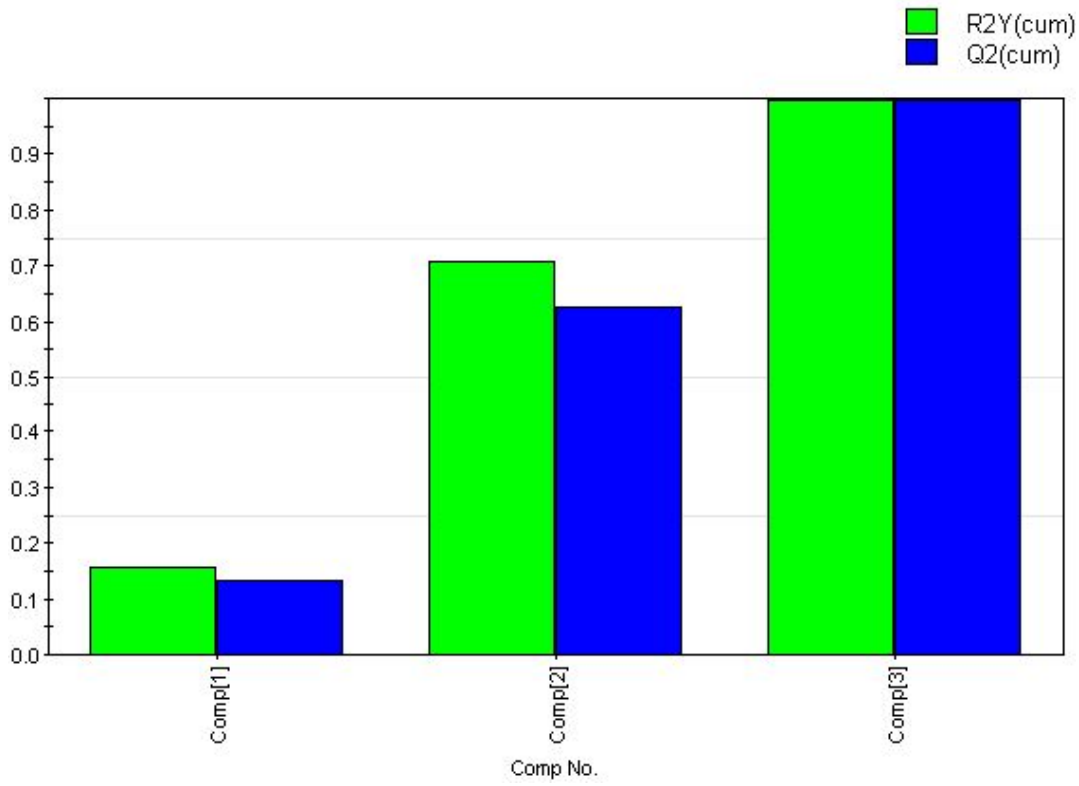


Figure 4.6 - Goodness of the fit: cross validation suggests three factors. This component finds the direction in the X-space that improves the description of the X-data as much as possible, while providing a good correlation with the Y-residuals.

Table 4.1 - PLS model information.

Var ID (Primary)	M1.R2VY Adj(cum)	M1.R2VY (cum)	M1.Q2VY (cum)	Stdev(Y)	RSD(Y)	Stdev(Y) WS	RSD(Y) WS	DFR	N
Acetate	0.996989	0.99735	0.994924	0.285541	0.0156695	1	0.0548767	22	26
Ammonium	0.997903	0.998155	0.997813	0.0747303	0.00342218	1	0.0457938	22	26
Glycerol	0.997158	0.997499	0.987099	0.29023	0.0154725	1	0.0533112	22	26
	Comp	Obs.	Y-miss(tot)						
	3	26	0						

Figure 4.7 shows the $\mathbf{t_1}$ vs. $\mathbf{t_2}$ plot of the X scores, which can be interpreted as a window into the X space. This shows how the X spaces (X conditions) are located with respect to each other. This plot shows the possible presence of outliers, groups, and other patterns in the data, as is the case with the samples 16 and 13. In this case, the mixture proportions are 0.0, 0.3 and 0.7 (glycerol, ammonium and acetate), respectively and these points are possible outliers. To understand and interpret this behavior, we can examine the corresponding T2 range plot (Figure 4.8). The T2 range plot displays the distance from the origin in the model plane (score space) for each selected observation. Values larger than the 95% critical limit (horizontal green line) are possible outliers (0.05 significance level), and values larger than the 99% critical limit significance (horizontal red line, 0.01 significance level) can be considered as serious. Samples 16 and 13 present suspicious values, however, these data points were not truncated due to the goodness of the model obtained so far. It presence must be taken in consideration for further discrepancies or poor behavior in the model.

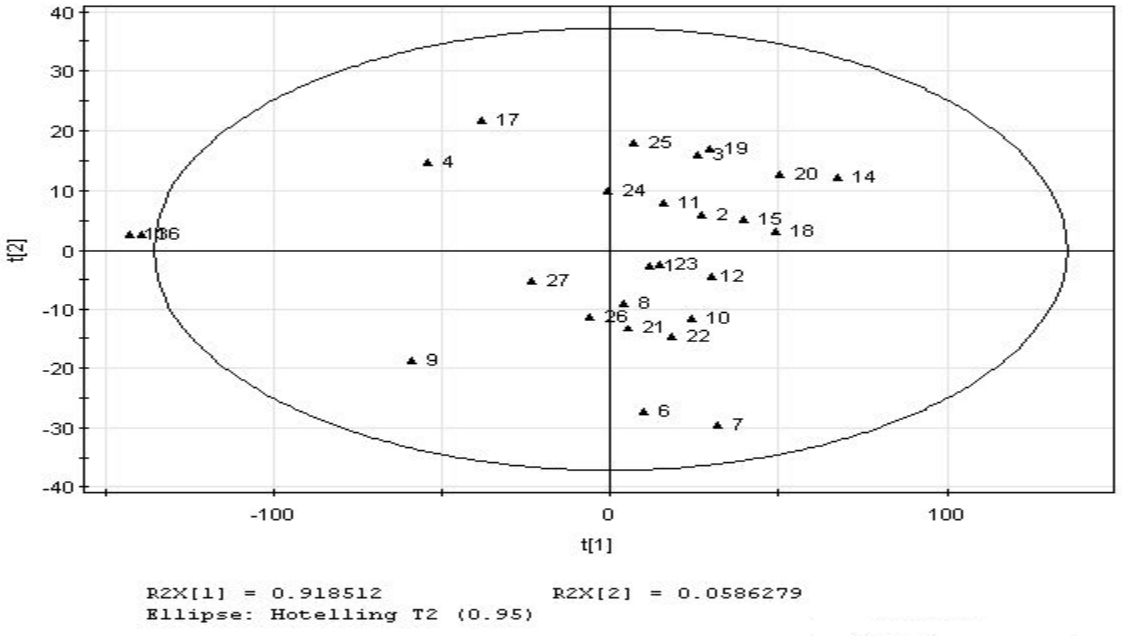


Figure 4.7 - Score t_1 vs t_2 using PCA-X analysis.

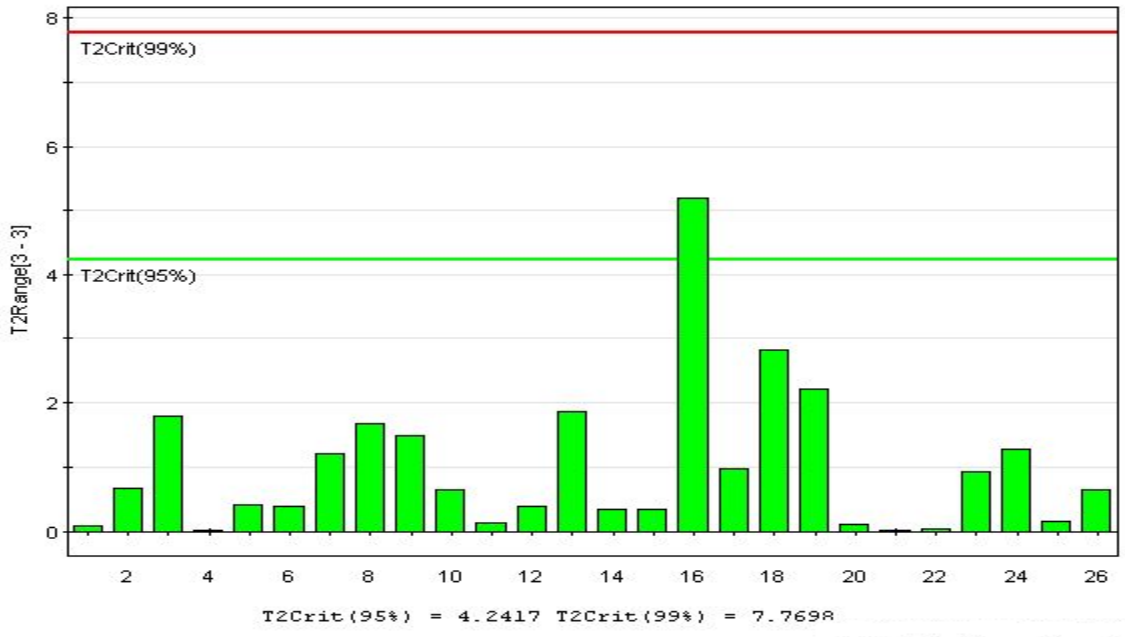


Figure 4.8 - Hotelling T2 range from component three to the last.

To obtain a good number of factors in a model, plot such as the cross-validation predictive residual sum of squares plot (PRESS), or the loadings plot (not shown) can be examined in order to choose an optimal model that can explain the information without modeling data noise.

Cross-validation (CV) is a practical and reliable way to test the significance of a PCA or a PLS model. With CV, the basic idea is to keep a portion of the data out of the model development, develop a number of parallel models from the reduced data, predict the omitted data from the different models, and finally compare these predicted values with the actual ones. The squared differences between predicted and observed values are added to form the PRESS, which is a measure of the predictive power of the tested model (Equation 4.1). Figure 4.9 shows how the coefficients of each Y varies within three factors near to an average equal to zero and shows that three factors are adequate to obtain a suitable model [4].

$$\textbf{PRESS} = \sum (x_{ik} - \hat{x}_{ik})^2 \qquad \textbf{4.1}$$

One of the most important goals is, of course, to predict the behavior along time of new fermentations runs. Below is the test for predictive ability of the model. Figure 4.10 displays the relationship between measured ("observed") amounts and the corresponding value predicted by the PLS model. Note that the results show an excellent fit with the three different components of each mixture.

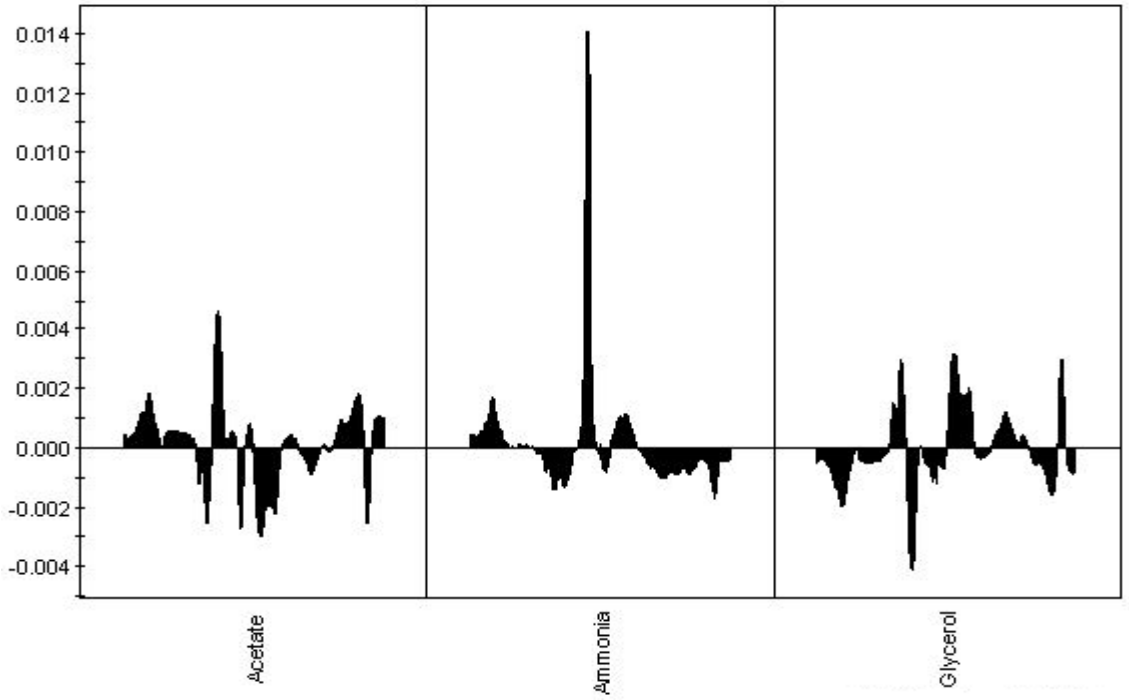


Figure 4.9 - Coefficients residuals by three factors.

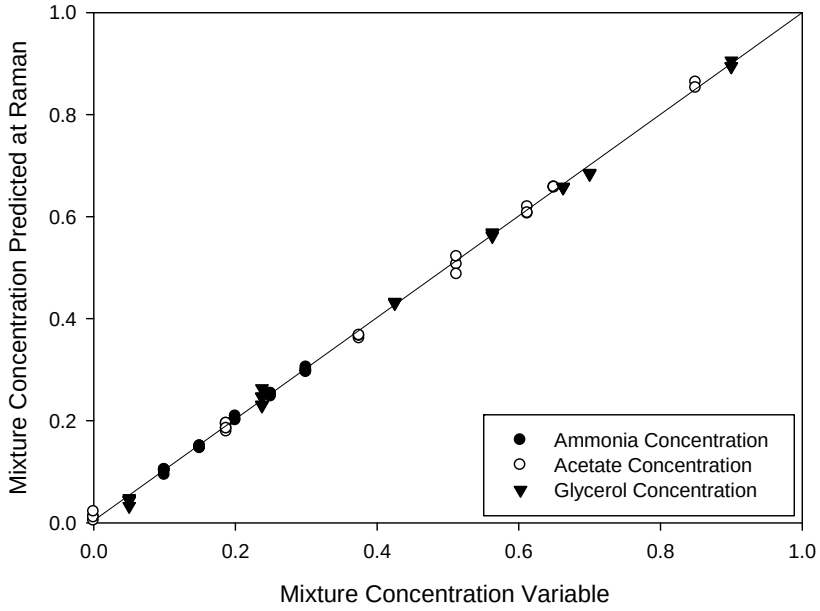


Figure 4.10 - Calibration model fit plot of the proportion mixture using Raman spectroscopy

4.2.5 Predictions and Validation

To determine the reliability of the analytes under investigation and not unique to the calibration data set only, it is necessary to include several validation steps. Cross-validation uses the data model to predict individual each sample used in the training data base (biased systematic errors estimate the difference between the average of reference values for a set of samples and the average of the instrumental value for the same samples, Figure 4.11). The purpose is to determine which sample provides more variability to the model or if there are possible outliers. Test-set validation is used to challenge the model goodness using the model to predict samples reserved for validation (an unbiased systematic error estimates the difference between the average of the mean and the sample set which was not included in the model).

Figure 4.11 represents three different fermentations (n=28), for the characteristic fluorescence of the GFP protein. A low frequency noise is observed, however, the scaled and centered pretreatment in the PLS model analysis shows fewer physical behavior at the validation, making the model prediction suitable. HPLC reference method that is designed or widely acknowledged as having the highest quality is compared with the model and plotted ($y = x$, Figure 4.12).

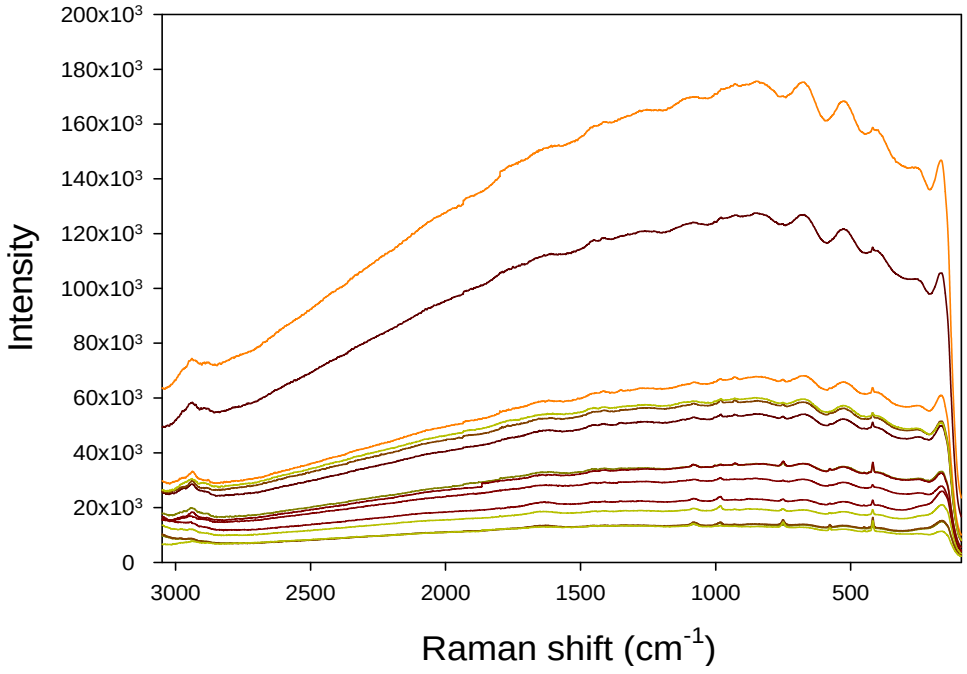
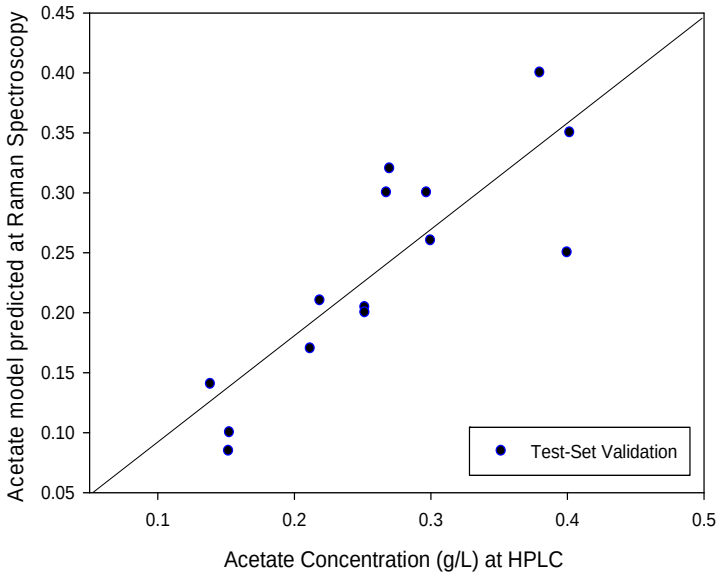


Figure 4.11 - Whole broth fermentation spectra using Raman spectroscopy.

4.3 Determination of Biomass Concentration using NIR Spectroscopy

At-line biomass concentration determination was investigated in order to gain information to assist in the application of a process "in-situ" NIRS [1]. Although a strong knowledge of the statistical technique is needed to extract information from such complex data set as those seen in bioprocesses, it is really necessary to understand a priori why specific regions of the spectra are most appropriate for model building compared to others [1, 5, 6]. Hall et al. among others established that the best two biomass ranges to derive the biomass model are 4700-4200 cm^{-1} and 5800- 5300 cm^{-1} [1].

a)



b)

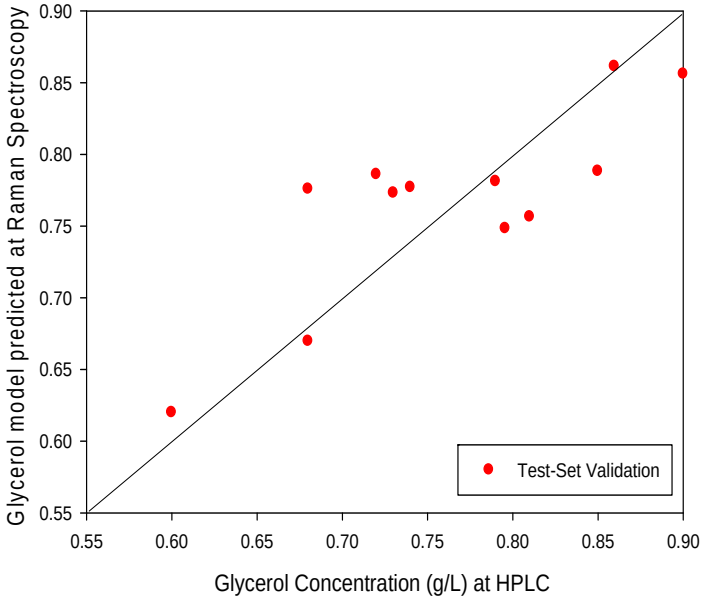


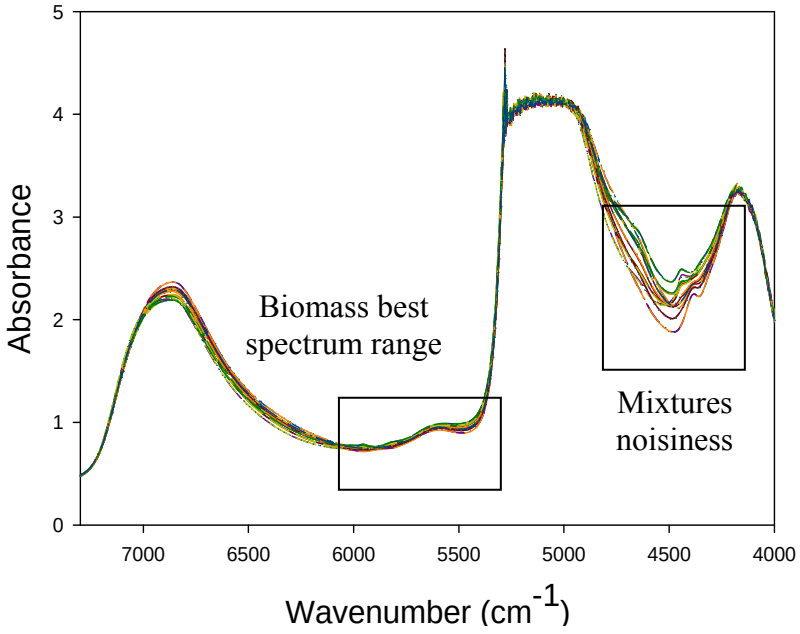
Figure 4.12 - Relationship between observed and predicted values for (a) acetate and (b) glycerol concentrations (proportional units) for the test set samples.

Extreme vertices mixture design programmed in Minitab software [3] was proposed to determine the characteristic frequency range of biomass in the NIR spectrum. Nine different samples (in triplicates) were prepared without cells to identify noisiness in the infrared spectrum to base on the medium (experimental design was shown in Figure 4.13). The experimental results indicated us that in the range of 4700-4200 cm^{-1} there was overlapping between the analyte mixture (glycerol, acetate, ammonia) and the biomass concentration, while within the 5800- 5300 cm^{-1} range, no significant presence of analytes was detected (Figure 4.13).

4.3.1 Description of the Primary Data Set

The primary data consisted of 32 samples (observation, row vectors) prepared and analyzed at-line, both with the gravimetric (dried cell) and UV reference methods. NIR spectra were acquired from each sample of mixtures in triplicate. The spectra were collected in the 8000-4000 cm^{-1} range but was analyzed only between 5800- 5300 cm^{-1}, using a nominal resolution of 8 cm^{-1} and giving a total of 260 X- variables (absorbance), NIR spectral wavenumbers, and two Y variables (UV absorbances and dry cell concentrations).

(a)



(b)

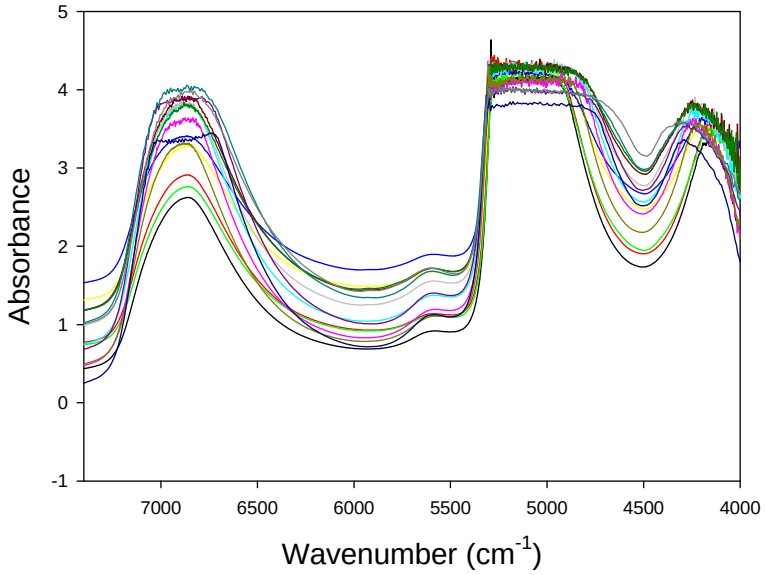


Figure 4.13 - NIR spectra of (a) mixture test samples and (b) whole broth (unmanipulated) used for the model calibration.

4.3.2 Suitable Fit of X/Y Variables

Tables 4.2 and 4.3 display the cumulative R^2VX, SEV (standard error of CV), SEC (standard error of the model) and SEP (standard error of prediction) for each X/Y variables, where the model can explain 99% with two factors.

Table 4.2 - Factor contribution to the OD concentration predicting model from two factors to the last.

	Variance	Percent	Cumulative	SEV	Press Val	rVal	SEC	Press Cal	rCal
Factor 1	644.90	98.84	98.84	14.49	2308.77	0.73	12.27	1354.55	0.85
Factor 2	6.67	1.02	99.86	7.92	690.66	0.93	5.65	255.31	0.97

	OD(Absorbance)	Obs.	Y-miss(tot)	SEP	rSEP
	0.39 - 70	32	0	5.56	0.95

Table 4.3 - Factor contribution to the dry cell concentration predicting model from two factors to the last.

	Variance	Percent	Cumulative	SEV	Press Val	rVal	SEC	Press Cal	rCal
Factor 1	1161.74	98.82	98.82	7.43	1048.70	0.67	6.72	768.68	0.76
Factor 2	12.77	1.09	99.90	4.22	338.16	0.90	3.31	174.83	0.95

	OD(Absorbance)	Obs.	Y-miss(tot)	SEP	rSEP
	0.13 - 40	32	0	2.58	0.94

4.3.3 Predictions and Validation

This section presents the relation between optical densities and dry cell concentrations with the FT-NIR spectra. To determine the righteousness of the model, a leave out one cross

validation set and test-set validation was required. A full random fermentation was taken for characterization of the biomass band in the infrared spectrum. The experimental results are presented in Figure 4.14, NIR spectra from this fermentation was not included in the model calibration. Figures 4.15 and 4.16 present the goodness of the model using the optical density and gravimetric tests, respectively.

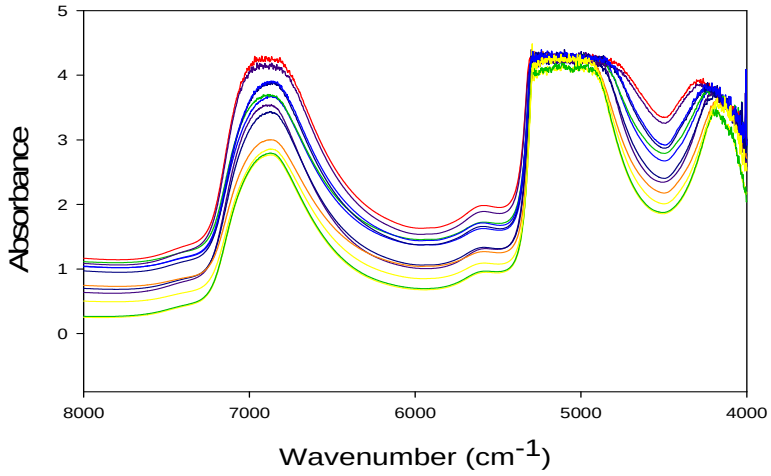


Figure 4.14 - Raw culture spectra used to validate the model.

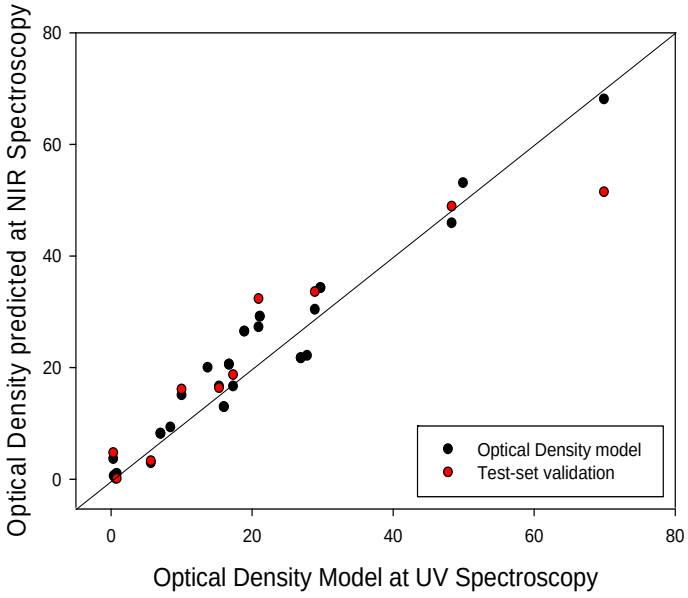


Figure 4.15 - OD fit plot of both: the model and the test-set validation at microbial fermentation by NIR Spectroscopy.

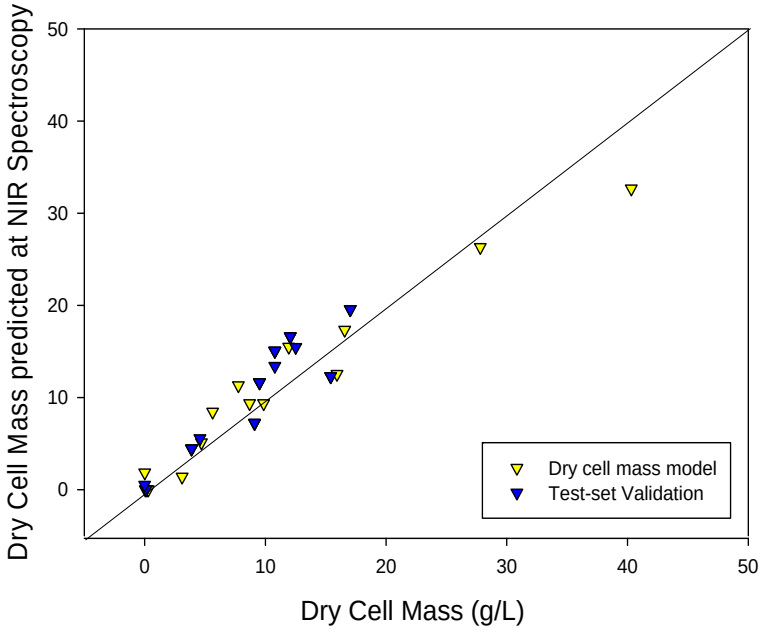


Figure 4.16 - Dry cell concentration fit plot of the model and the test-set validation at microbial fermentation by NIR Spectroscopy.

4.4 Determination of Protein Concentration using FT-IR Spectroscopy

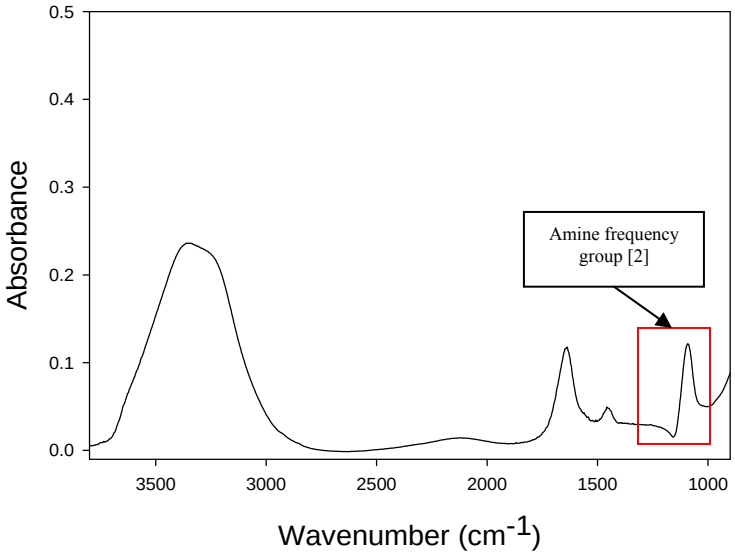
The transformed *E. coli* used in this research possessed the *bla* plasmid which provides physiological advantages such as antibiotic resistance and recombinant GFP expression by induction of arabinose. Fluorometry spectroscopy was used as a characteristic indicator of GFP presence. Fluorescence was examined for increased intensity over the period of expression of GFP along fed-batch fermentations of the *E. coli*. Although fluorometry is not commonly used for protein characterization because of the non-fluorescence characteristics of most biopharmaceutical proteins, for GFP a pronounced increase was detected along time after the induction step during the fermentations (Figure 3.5).

Raman spectroscopy is promising for bioprocess monitoring applications due to the low interference from water. For that reason Lee et al. discuss in their literature review (section 2.5), the use of Raman spectroscopy by modelling the effects of scattering of the protein synthesis accumulation (phenylalanine) in a microbial fermentation. However, they assumed no effects from fluorescence. In our case the use of Raman spectroscopy is limited by the ability of GFP to fluorescence. Fluorescence probability (of 1 in 103-105) vs. probability of Raman scattering probability (1 in 107-1010) represents significant background interferences and for that reason the use of other techniques is required in our case. Mc Govern et al. demonstrated the use of FT-IR techniques in combination with chemometrics for determining the synthesis protein (α2 IFN) accumulation in recombinant *E. coli* bioprocess samples (section 2.5). They showed a linear correlation in the cell paste and protein synthesis by PLS analysis. However, the study only focused in the dry cell paste and the supernatant samples (each sample took around 20 minutes for drying steps), they did not analyzed the culture broth. The need of protein synthesis modelling in the whole culture broth without sample pre-treatment presents a challenge to overcome.

Experimental results of the FT-IR spectra from those bioprocess samples were also complex (Figures 4.17 and 4.18) and because of the multitude of cellular components and proteins, all with their own molecular vibrations capable of absorbing by the mid-infrared range (4000-600 cm^{-1}). The infrared spectra of the protein exhibit strong amide I absorption bands at 1200-1000 cm^{-1} associated with the characteristic stretching of C-O and C-N and the bending

of the N-H bond [9]. This information shall be used to create a model for the determination of the protein concentration in a fed-batch reactor. This experiment presents a practical view to use multivariable tools for the determination of protein synthesis as a fermentation product.

(a)



(b)

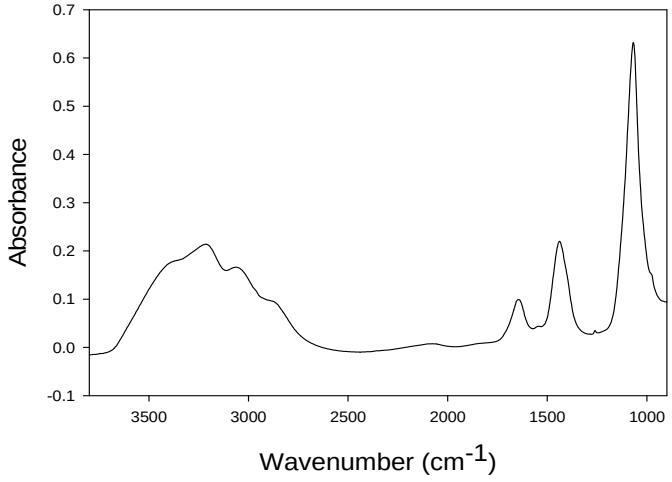


Figure 4.17 - FT-IR spectra (raw data) of (a) green fluorescent protein (GFP) purified using hydrophobic interaction chromatography (HIC) (b) GFP purified and lyophilized.

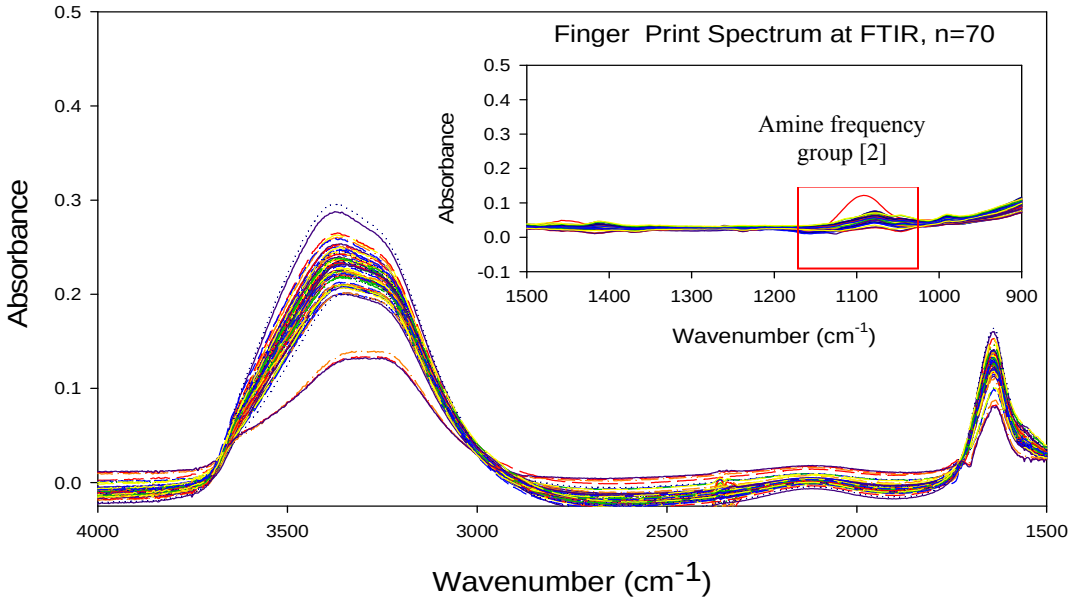


Figure 4.18 - FT-IR spectra (raw data) of whole culture broth with varying concentrations of GFP, n=70.

4.4.1 Description of the Primary Data Set

The primary data set consists of 70 samples (observation, row vectors) prepared and analyzed at-line, with fluorimetry as the reference method. FT-IR spectra were acquired from each sample of mixtures in triplicate. The spectra were obtained in the 4000 - 900 cm^{-1} range but were analyzed only at 1200 - 1000 cm^{-1}, giving a total of 519 X- variables (absorbance), FT-IR spectral wavenumbers, and one Y variable (fluorescence absorbances).

4.4.2 Suitable Fit of X/Y Variables

In the multivariate calibration, all data were scaled to unit variance. The X-block compromised all 519 fingerprint variables and the Y-block the response (fluorescence absorbance). Table 4.4 displays the cumulative R^2VX, SEP, SEV and SEC for each X/Y variable. These data indicate tell us that with only two factors, 87% variability would be accounted for.

Table 4.4 - Factors contribution to the recombinant protein (GFP) concentration predicting model from two factors to the last.

	Variance	Percent	Cumulative	SEV	Press Val	rVal	SEC	Press Cal	rCal
Factor 1	0.020	58.25	58.25	122.58	1053468.1	0.990	119.11	964790.75	0.991
Factor 2	0.010	28.30	86.55	114.46	917038.5	0.992	110.76	821936.69	0.993

	API (RFU)	Obs.	Y-miss(tot)	SEP	rSEP
	58-3959	70	0	117.40	0.991

4.4.3 Predictions and Validation

A full random fermentation was used for characterization of the amine band in the infrared spectrum. Figure 4.19 plots the PLS estimates versus the true measured value of GFP in the whole broth. It resulted in linear fits, which were very close to the expected proportional fits ($y = x$).

The applicability of the model was confirmed using a test data set. In this case, the properties of ten samples (which had not been included in the calibration models) were predicted using

the calibration models. The test-set raw data and test-set validation is shown in Figures 4.19 and 4.20, respectively. The calibration models were in accordance with the protein concentration predicted. The leave-one-out strategy was also used to determine if there where some outliers, but it provided good results, as well as those obtained by the prediction of the test-set.

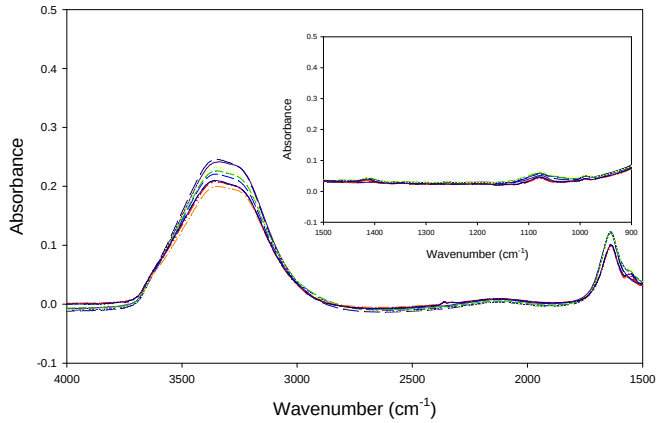


Figure 4.19 - Raw culture broth spectra in order to validate the model.

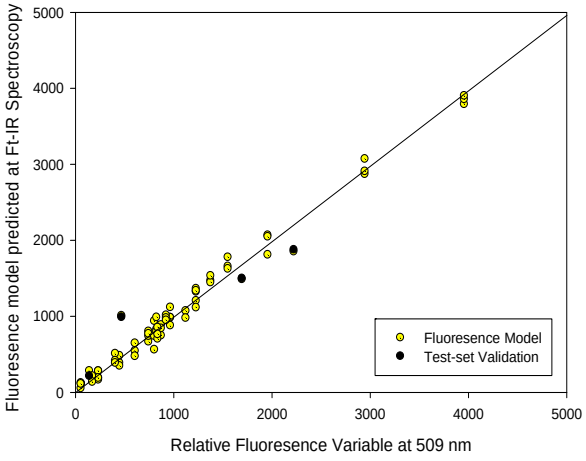


Figure 4.20 - Estimates and validation of the levels of GFP in whole broth by FT-IR spectroscopy.

4.5 References

[1] Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B.G.; Macaloney, G. Near-Infrared Spectroscopic Determination of Acetate, Ammonium, Biomass and Glycerol in an Industrial *Escherichia coli* Fermentation. *Applied Spectroscopy,* **1996**, 50, 102-8.

[2] Moisheky, Z.; Melling, P. J.; Thomson, M.A. In situ real-time monitoring of a fermentation reaction using a fiber optic FT-IR probe. *ADVANSTAR,* **2001**, 1-5.

[3] Montgomery, D.C. *Design and Analysis of Experiments;* sixth Ed., John Wiley & Sons, Hoboken, NJ, 2005, pp 444- 52.

[4] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikstrom, C.; Wold, S. *Multi- and Megavariate Data Analysis Part 1*. Umetrics AB, Ch. 16, 2006, 337-60.

[5] Arnold, S.A.; Gaensakoo, R.; Harvey, L.M.; McNeil, B. Use of At-line and In-situ Near-Infrared Spectroscopy to Monitor Biomass in an Industrial Fed-Batch *Escherichia coli* Process. Wiley Periodicals, Inc. 2002, 405-13.

[6] Cimander, C.; Mandenius, C.F. Online monitoring of a bioprocess based on a multi-analyser system and multivariate statistical process modelling. *J. Chem. Technol. Biotechnol.,* **2002**, 77, 1157-68.

[7] Lee, H.L.T.; Boccazzi, P.; Gorret, N.; Ram, R.J.; Sinskey, A.J. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, **2004,** 35, 131-7.

[8] McGovern, A.C.; Ernill, R.; Kara, B.V.; Kell, D.B.; Goodacre, R. Rapid analysis of the expression of heterologous protein in *Escherichia coli* using pyrolysis mass spectrometry and Fourier transform infrared spectroscopy with chemometrics: application to α2-interferon production. *J. Biotechnology,* **1999**, 72, 157-67.

[9] Lun-Vien, D.; Colthup N.B.; Fateley, W.G.; Grasselli, J.G. *The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecule*. New York, Academic Press, Ch. 10-15, 1991, 155-261.

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Determination of Acetate and Glycerol Concentration using Raman Spectroscopy

Through the work reported in section 4.2, a substantially better understanding of the substrate and by-products modelling using chemometrics and its comparative performance in a fed-batch fermentation using a HPLC as a reference was accomplished. The ability to determine the concentrations of acetate and glycerol during the fermentation without a pre-treatment of samples and in less time provides an important step in the monitoring of the bioprocess with a potential to improve process control and automation and a processing cost savings (Table A.1).

5.1.1 Conclusions

As a part of the determination of the substrate and by-products concentrations in a fed-batch reactor, a mixture experimental design was developed to determine the concentrations of glycerol ($10 < x$, g/L < 180) and acetate ($0 < x$, g/L < 180) through Raman spectroscopy to determine the significance of the mixture effect in the determination of other parameters (protein expression and biomass concentration). No evidence of previous literature on this subject was found, so our results and procedure apparently are the first to be reported.

The effects of the concentrations of glycerol and acetate were accomplished by Raman spectroscopy instead of NIR and FT-IR. Theoretically, glycerol and acetate may be determined in the FT-IR and Raman spectroscopies, but in the NIR it would be a difficult task; first because of the short radiation ranges of this spectroscopy and second because the FT-IR presents a strong -OH- group interference. In our case, where water is abundant, the modelling of the glycerol and acetate concentrations present some small discrepancies from batch to batch.

Raman spectroscopy was chosen for the determination of the glycerol and acetate concentrations because theoretically, glycerol and acetate result in good analyses in the Raman spectra without water interference. Experiences has demonstrated that a good experimental design is necessary to find a good calibration for the glycerol and acetate concentrations. In our case, our results demonstrated that the best approach to determine the analyte concentrations was by performing an experimental design, which incorporated a higher range of concentrations (Section 4.2). The experiment consisted in utilizing a mixture design with the best effective mixture combination to create a model followed by the validation with collected samples.

The modelling of the concentrations of glycerol and acetate showed excellent results, but it would only work for the same organism and fermentation conditions.

5.1.2 Recommendations

As discussed previously, an accurate modelling of the concentration of the glycerol and acetate during the fermentation was achieved; this included analysis of glycerol and acetate content in the samples using Raman spectroscopy and a comparison with HPLC as a reference for several runs. A calibration without the organism was needed and later used to predict the acetate and glycerol concentrations for the fermentation samples. Also, different spectroscopy techniques were used to test which one generated the best model for the different analytes. Findings suggested that the model has no application in other culture systems. It remains necessary to determine when a model fails for applicability purposes.

It would be interesting to seek a more robust model by considering the following fermentation conditions:

1) To use of a continuous stirred-tank bioreactor (CSTB) instead of a fed-batch mode for growth of *E. coli* at different specific growth rates (μ). That way, the experimental design would be more comprehensive and would cover a wider range of substrate and by-product concentrations.

2) Use different nutrient media and microorganisms, and determine through PCA which components provide the most variability.

3) Attach the spectrometers probes in-situ and at an industrial scale to determine incongruences between at-line and in-situ modes of data collection.

These experimental variations would give a better understanding of the bioprocess and would also help to determine the worst case scenarios and allow to determine which factors provide large variabilities to the system.

5.2 Determination of Biomass Concentration using NIR Spectroscopy

Through the work discussed in section 4.3 a substantially better understanding of the biomass modelling using chemometrics and its comparative performance in a fed-batch fermentation using an absorbance or a gravimetric technique as reference methods (Table A.1) shows the working time for the techniques tested in this research). The ability to determine the biomass concentration during the fermentation without a pre-treatment sample and at-line measurements provides an important improvement bioprocess monitoring, automation and control in automation.

5.2.1 Conclusions

As a part of the determination of the biomass concentrations in a fed-batch reactor, several runs were made with the purpose of creating a robust model for the quantification of the biomass during the fermentation using optical density and dry cell concentration as a reference. Both of the references were analyzed and plotted to ensure they showed the same behavior along time. A linear relation between the dry cell mass concentration (g/L dcm) and optical density (absorbance units) is held ($g / Ldcm = -0.217 + 0.579 \cdot OD_{600}$; R^2=83%). Once

the reference (Y, observation) data were correlated with the spectra (X, variable) we could observe a linear model where both reference methods were validated with excellent results. We found some differences between the dry cell concentration and optical density in the biomass model. The optical density instrument (UV) and the infrared spectroscopy (both are spectroscopic techniques and they are affected by water interferences) were comparable as opposed to the dry cell concentration method, which is a gravimetric technique which has a larger variability because of its more tedious procedure.

Because the modeling of the biomass concentration using NIR infrared is a well-utilized method (section 4.3), our research was focused on the determination on the effect of the component and the biomass. For these reasons, a mixture design was performed to find any presence in the component in the spectra. Experimental results showed the best-unaffected region area that was 5800- 5300 cm^{-1} of the NIR spectral region.

5.2.2 Recommendations

As discussed previously, an accurate modeling of the biomass concentration during the fermentations was achieved and included a quantification of biomass using NIR and dry cell concentration and optical density as a references. It remains necessary to scale up the analysis and observe if differences exist among the spectra and if the quantification of biomass would be affected by this scale up. It is also recommended to attach the NIR probe in-situ.

5.3 Determination of Protein Concentration using FT-IR Spectroscopy

Throughout the work described in section 4.4, a substantially better understanding of the modeling to predict recombinant protein concentration using chemometrics and its comparative performance in a fed-batch fermentation using fluorescence as a reference was accomplished. Table A.1 shows the working times for the methods that were used in this research (fluorometer and FT-IR) and compared to HPLC (a typical technique that would be used to determine GFP concentration that was not used in this work). These techniques consume more much time to quantify the protein expression during fermentation as compared to the infrared spectroscopy technique. Pretreatment of the samples is not required and the possibility of locating the probe in-situ offers infrared as one of the most promising techniques in bioprocess monitoring, automation and control.

5.3.1 Conclusions

As a part of the determination of the protein concentration in a fed-batch reactor, several runs were made with the purpose of creating a rigorous model for the quantification of protein expression during the fermentation using fluorescence spectroscopy as a reference method. Experimental results demonstrated that a strong model was achieved using FT-IR in the spectral range of 1200 - 1000 cm^{-1}.

Because the quantification of the protein by fluorescence is a special case, another way to determine the protein concentration between the spectra is the use of dynamic light scattering (DLS) equipment to scan the hydrodynamic bacterial surface area and to relate the increment of the average cell size (due to the accumulation of intracellular recombinant protein) with the DSL intensity. Because the DSL equipment was not able to discriminate between cellular binary fission and intracellular inclusion bodies, a modified approach for the determination of the particle sizes would have to be designed.

5.3.2 Recommendations

As discussed previously, an accurate modeling of the protein concentration during the fermentation was achieved that included a quantification of protein using fluorescence as a reference for several runs. It remains necessary to scale up the analysis and study if there might be differences in the spectra by the scale up. It is also recommended to attach the probe in-situ. It may be interesting to use another host with another quantification references (HPLC assays or an enzymatic assay) to understand if the use of FT-IR spectroscopy would be suitable to determine the protein concentration with the same accuracy that we were able to obtain using fluorimetry.

APPENDIX A

To provide an idea of the advantages that would be obtained by monitoring a bioprocess using NIR, FT-IR and Raman spectroscopies, the table below compares the estimated measurement times that would be expected in determining the variables that were tested in this work with the standard or traditional reference technologies. It is evident that using robust models derived from chemiometric analyses, there would be beneficial improvements in monitoring a bioprocess using spectroscopic techniques.

Table A.1- Comparison of measurement times for each analytical technique.

Analytical Techniques	Variables of Interest	Aliquot Samples for Calibration	Fermentation Samples	Working Time (min/samples)	Total time (hr)
Raman	Acetate and Ammonia	27	28	4	3.7
HPLC	Acetate and Ammonia	42	28	30	35.0
NIR	Biomass	N/A	44	2	1.5
UV	Biomass	N/A	44	2	1.5
Gravimetry	Biomass	N/A	44	30	22.0
FT-IR	Recombinant Protein	N/A	70	2	2.3
Fluorometry	Recombinant Protein	N/A	70	0.5	0.6
HPLC (proposed)	Recombinant Protein	27	70	20	32.3