Development of novel methodologies for the optimization of production processes of biopharmaceuticals.

Sviluppo di nuove metodologie per l'ottimizzazione dei processi di produzione di prodotti biofarmaceutici

Dottorando: Marco Barba (XXI ciclo)

Academic supervisor: dott. Fabio Polticelli

Industrial supervisor: Horst Bierau - MerckSerono S.p.A.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITIONS AND ABBREVIATIONS

ATR	Attenuated Total Reflectance
CD	Circular Dichroism
DS	Drug Substance
EDTA	Ethylene Diamine Tetra-Acetic Acid
EMEA	European Medicine Agency
EPR	Electron Spin Resonance
FDA	Food and Drug Administration
FT-IR	Fourier Transform Infrared
HPLC	High Pressure Liquid Chromatography
IE-HPLC	Ion Exchange High Pressure Liquid Chromatography
MVDA	Multivariate Data Analysis
NIR	Near Infrared
NMR	Nuclear Magnetic Resonance
O-PLS	Orthogonal Partial Least Squares
PAT	Process Analytical Technologies
PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Base
PLS	Partial Least Squares
PTM	Post Translational Modification
QC	Quality Control
RMSEE	Root Mean Square Error of Estimation
RMSEP	Root Mean Square Error of Prediction
ROESY	Rotational nuclear Overhauser Effect Spectroscopy
RP-HPLC	Reverse Phase High Pressure Liquid Chromatography
SE-HPLC	Size Exclusion High Pressure Liquid Chromatography
SOD	SuperOxide Dismutase
TFA	Trifluoroacetic Acid
TOCSY	Total Correlated Spectroscopy
UPLC	Ultra Performance Liquid chromatography
UV	Ultra Violet

SUMMARY

Process Analytical Technology (PAT) is defined by the FDA as a "System for designing, analyzing and controlling manufacturing through timely measurements of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality". The goal of implementing PAT is defined therein as enhancing the understanding and the control of a production process. This broad definition encompasses testing of raw material for batch consistency as well as online sensors that provide feedback for the process control. In this context, rapid methods placed at-line, i.e. close to the process, can be considered to be PAT applications since they will both contribute to the understanding how individual steps impact on product quality and will accelerate and facilitate process development and optimization decisions.

There are many tools available that enable process understanding for scientific, pharmaceutical development. These tools can provide effective and efficient means for acquiring information to facilitate process understanding and continuous improvement. From a physical, chemical and biological perspective, pharmaceutical products and processes are complex multi-factorial systems. Methodological experiments based on multivariate statistical principles provide useful means for the identification and study the effect and interaction of product and process variables. Traditional one-factor-at-time experiments cannot address these kinds of interactions. These tools enable the identification and evaluation of product and processes variables that may be critical to product quality and performance.

Thanks to the PAT initiative, spectroscopic sensors systems have gained interest for bioprocess monitoring because they allow rapid and non-destructive monitoring of product quality attributes. The improvements in spectrometers, detectors and optics have led to interesting applications related to PAT.

The main goal of the thesis is the development of novel methodologies based on spectroscopic techniques coupled with multivariate data analysis for the optimization of production process of biopharmaceuticals.

One of the approaches described here combines the strengths of various spectroscopic techniques, such as Circular Dichroism (CD), Infrared (IR), Raman, Fluorescence and UV-Visible-NIR measurements, to provide a more comprehensive description of a substance, the so-called "fingerprint". This, in combination with Principal Component Analysis (PCA) may be use to establish and define quality, equivalence, and comparability of substances while also providing a means to monitor processes and provide relevant information about molecular changes in product. Moreover, it can highlight the relationships between different properties, for example that between structure and aggregation, and a better understanding of the nature of a product. The potential of such a fingerprint has an impact in a wide variety of areas within biopharmaceutical research and development.

The protein used in this study is a homo-dimeric Fc-fusion protein. In particular, for this first part of the project, focused on PCA of multispectroscopic data, were used ten batches of drug substance produced with the current process called "process C". This bulk material has a concentration of "not less than" 160 mg/ml. Within this set, the batches differ with respect to hydrolysates that were used as feed during the fermentation process. In addition to these difference, three additional batches form the new "process D", after some minor optimisations with respect to "process C", were analysed. In order to generate a wider diversity of samples, with aspects of deterioration, solutions were diluted with water (instead of using a buffer) and then stored at room temperature for several weeks before spectroscopic analysis. The major consequence of this treatment is a change in buffer/additive concentrations, a change in pH, and a deterioration of solutions through aggregation, structural and chemical decomposition.

A comprehensive set of spectra of eleven batches thus were acquired, using a variety of techniques. Some twelve variants of five spectroscopies have been employed, covering the complete wavelength range from far-ultra violet to infrared and involving phenomena including absorption, fluorescence, Raman scattering, Rayleigh scattering and circular dichroism. Both concentrated stock and deliberately deteriorated dilute solutions were investigated. All of the techniques employed have yielded useful data of some forms in terms of identifying variance in the batches and are potentially complementary and cross-supporting.

Each set of spectra were subjected to multivariate data analysis, primarily PCA, to highlight patterns and differences between batches. Such analysis highlighted an apparent connection between the spectra and the history of batches regarding production date and hydrolysate type used. In particular, a series of anomalous absorptions in the visible wavelength region, together with potentially related fluorescent species, were identified. These may derive from contaminants, post-translational modifications (PTM) dependent on production conditions.

Another aim of the thesis was to assess the feasibility of obtaining quantitative data about degradation products of a therapeutic protein when employing Circular Dichroism and infrared spectroscopy in combination with multivariate data analysis, primarily Partial Least Squares (PLS) regression, and an extension of PLS, Orthogonal Partial Least Squares (O-PLS). This is a novel approach since the classical applications for CD and IR spectroscopy are the determination of secondary structure content of proteins. Also the use of multivariate statistical methods for the determination of secondary structure content is reported. Nevertheless, the present approach is to our knowledge the first one that seeks to exploit PLS in order to correlate CD and IR spectral data with quantitative data of common protein degradation forms.

In order to generate a suitable calibration matrix, a set of samples containing pre-defined levels of aggregates, oxidized forms, and free Fc, was generated. In order to ensure non-correlation of the degradation levels within the calibration matrix, the target concentrations therein were chosen according to an approach described by Brereton (2000). All the samples generated were then analyzed separately for each of the three degradation forms employing dedicated chromatographic QC assays in order to obtain accurate degradant levels. Furthermore, both CD (near and far UV) and IR spectra were measured. Both the QC and the spectroscopic data form the basis for the generation of various PLS/O-PLS models, i.e. based respectively CD or IR spectra alone, as well as CD and IR data combined.

The feasibility of employing PLS/O-PLS analysis to extract quantitative data for common protein degradation forms was successfully demonstrated for an Fc fusion protein. Both CD and IR spectra contained the relevant information, nevertheless, CD-based O-PLS models achieved a higher accuracy compared to that of IR-based models for predicting aggregate and oxidation levels, while the accuracy for free Fc levels could be equally well predicted. Combining CD and IR data improved the accuracy of the prediction for all degradation forms. In addition, we demonstrated that O-PLS models yielded to a better accuracy compared to that obtained with PLS models.

The last part of the thesis is based on the “protein design” methodologies. Aim of the present thesis is to study the scaffold stability of contryphan-Vn, a small peptide isolated from the venom of *Conus ventricosus* formed by only 9 residues and characterized by the presence of a single disulfide bridge, after substitution of 4 of 9 amino acids of its sequence.

Contryphans are bio-active peptides, isolated from the venom of marine snails of the genus Conus, which are characterized by the short length of the polypeptide chain and the high degree of unusual post-translational modifications. The cyclization of the polypeptide chain through a single disulphide bond, the presence of two conserved Pro residues and the epimerization of a Trp/Leu residue confer to Contryphans a stable and well defined structure in solution, conserved in all members of the family. The potential of Contryphans as scaffolds for the design of redox-active (macro)molecules was tested by engineering a copper binding site on two different variants of the natural peptide Contryphan-Vn, named Cupryphan and Arg-Cupryphan through the introduction of four His residues. The binding site was designed by computational modelling and the redesigned peptides were synthesized and characterized by optical, fluorescence, electron spin resonance and nuclear magnetic resonance spectroscopy.

The novel peptides, named Cupryphan and Arg-Cupryphan bind Cu2+ ions with a 1:1 stoichiometry and a Kd = 1.3(± 0.2) × 10-7 M and 1.0(± 0.4) × 10-7 M, respectively. Other divalent metals (e.g. Zn2+ and Mg2+) are bound with much lower affinity. In addition, Cupryphans catalyze the dismutation of superoxide anions with an activity comparable to other non-peptidic superoxide dismutase mimicks.

We tested the potential of conopeptides as scaffolds for the engineering of novel, metal based, biocatalysts starting from the simplest prototype of disulphide constrained conopeptides: the Contryphans. The results of the present work indicate that indeed this class of peptides could be successfully exploited to engineer novel, stable and redox active macromolecules.

.

RIASSUNTO

La produzione di prodotti biofarmaceutici è un processo complesso che si evolve costantemente durante l'intero sviluppo di un prodotto. Negli ultimi 30 anni sono stati fatti significativi progressi nello sviluppo di metodi di analisi delle caratteristiche chimiche (es. identità e purezza) di un prodotto. Nonostante ciò, alcuni aspetti chimico-fisici dei componenti di prodotti biofarmaceutici non sono ancora di facile comprensione. Di conseguenza, l'inerente e non rilevata variabilità dei materiali grezzi può manifestarsi anche nel prodotto finale.

In più l'esigenza di aumentare la produzione e/o la purezza di un determinato prodotto potrebbe richiedere lo sviluppo di un nuovo processo o l'ottimizzazione dello stesso. L'eterogeneità di un prodotto biofarmaceutico può derivare anche da variazioni nel processo di produzione, risultando nella presenza di varie forme di modificazioni post-traduzionali o prodotti di degradazione. Chiaramente, l'obiettivo primario dello sviluppo o ottimizzazione di un dato processo di produzione è quello di ottenere la massima purezza del prodotto, minimizzando quindi le sue eventuali forme di degradazione. Nondimeno, soprattutto per molecole sul mercato e usate nei "clinical trial", è molto importante assicurare la consistenza della molecola e dei suoi parametri critici di qualità, in modo da assicurare che qualsiasi cambiamento nel processo di produzione non influisca negativamente sulla sicurezza ed efficacia della stessa. Quindi, con l'obiettivo di accelerare i tempi di sviluppo, c'e' un chiaro bisogno di sviluppare nuovi metodi che riducano i tempi di analisi sui campioni e allo stesso tempo massimizzino l'informazione ottenuta.

Process Analytical Technology (PAT) è definita come un "sistema per lo sviluppo, analisi e controllo della produzione attraverso misure appropriate dei parametri critici di qualità dei materiali grezzi e dei processi stessi, con l'obiettivo di assicurare la massima qualità del prodotto finale". L'obiettivo della PAT è definito quindi come un aumentata comprensione e controllo dei processi di sviluppo. Questa definizione generale comprende test di consistenza sui materiali cosi come "sensori" *on-line* che forniscano informazioni per il controllo del processo. In questo ambito, metodi analitici rapidi posti in stretta vicinanza del processo possono essere considerati in conformità con quanto dettato dalla PAT, in quanto possono sia contribuire ad una maggiore comprensione di come ogni singolo step del processo possa influenzare la qualità del prodotto, sia accelerare e facilitare lo sviluppo e l'ottimizzazione dello stesso.

Esistono molti strumenti che pemettono una migliore comprensione dello sviluppo scientifico e farmaceutico. Questi possono fornire reali mezzi per acquisire informazioni volte a facilitare la comprensione di un dato processo e il suo continuo miglioramento. Da un punto di vista chimico e biologico, i prodotti biofarmaceutici e i processi di produzione sono complessi sistemi multifattoriali. Esperimenti metodologici basati su principi di statistica multivariata forniscono dei mezzi molto utili per l'identificare e studiare l'effetto e l'interazione tra prodotto e variabili del processo. I tradizionali esperimenti univariati (che analizzano un fattore alla volta) non sono in grado di identificare questo tipo di interazioni. Quindi, l'analisi multivariata permette di valutare quali variabili del prodotto e del processo possono essere cruciali per la qualità e la performance del prodotto finale.

Grazie alla PAT, la spettroscopia ha guadagnato molto interesse per i processi di monitoraggio in quanto permette analisi rapide e non distruttive per il prodotto. In più, i miglioramenti raggiunti negli strumenti, nei detector e nelle ottiche hanno condotto ad interessanti iniziative conformi alla PAT.

L'obiettivo principale di questa tesi è lo sviluppo di nuove metodologie basate su tecniche spettroscopiche in combinazione con analisi multivariata per l'ottimizzazione dei processi di produzione di biofarmaceutici.

Uno degli approcci descritti unisce la forza di varie tecniche spettroscopiche quali dicroismo circolare (CD), Infrarossi (IR), Raman, Fluorescenza e UV-Vis allo scopo di fornire una migliore ed esauriente descrizione di una molecola, il cosiddetto "fingerprint". Questo, in combinazione con la PCA (Principal Component Analysis) può essere utilizzato per stabilire la qualità, uguaglianza e eventuali differenze tra prodotti e allo stesso tempo fornire un mezzo per monitorare i processi di sviluppo e fornire informazioni rilevanti su possibili cambiamenti della molecola. Inoltre, può mettere in evidenza relazioni tra differenti proprietà (per esempio quella tra struttura e aggregazione) e una migliore comprensione sulla natura della molecola.

La molecola usata in questo studio è una proteina omodimerica di fusione con Fc. In particolare, per questa prima parte del progetto, incentrata sull'analisi di dati multi-spettroscopici attraverso PCA, sono stati utilizzati 10 batches di "drug substance" (DS) prodotti con il corrente processo di produzione chiamato "process C". Questo materiale ha una concentrazione non inferiore di 160 mg/ml. Tra questi campioni, ci sono delle differenze in termini di idrolizzato usato come nutrimento durante il processo di fermentazione. Oltre a questi campioni, sono stati analizzati altri 3 campioni provenienti dal nuovo "process D". Allo scopo di generare una più vasta varietà di campioni a livello di deterioramento, i campioni originali sono stati diluiti in acqua e conservati a temperatura ambiente per diverse settimane. Le conseguenze principali di questo trattamento sono il cambiamento delle concentrazioni di additivi/buffer e il conseguente cambio di pH, la formazione di aggregati e la decomposizione chimica e strutturale della molecola.

E' stato generato un esauriente set di spettri a partire da undici diversi lotti di proteina, utilizzando una grande varietà di tecniche, coprendo l'intero "range" dal lontano ultravioletto all'infrarosso e coinvolgendo fenomeni quali assorbimento, fluorescenza, Raman e Rayleigh scattering e dicroismo circolare. Sono stati analizzati sia i campioni originali sia quelli diluiti, ed ogni tecnica utilizzata ha dato origine a risultati utili per l'identificazione di eventuale variazione tra i campioni in analisi.

Ogni set di spettri è stato analizzati attraverso analisi multivariata, soprattutto PCA, allo scopo di evidenziare similarità e differenze tra i campioni. Questo tipo di analisi ha evidenziato una apparente connessione tra gli spettri e la storia del campione in termini di data di produzione ed idrolizzato usato. In particolare, sono state evidenziate una serie di bande di assorbimento inaspettate nella regione del visibile, insieme a probabili molecole fluorescenti potenzialmente correlate a questi fenomeni di assorbimento. Queste ultime potrebbero derivare da contaminanti, modificazioni post-traduzionali (PTM) o trasformazioni post-PTM dipendenti dalle condizioni di produzione.

Un ulteriore obiettivo della tesi è quello di dimostrare la possibilità di ottenere dati quantitativi su prodotti di degradazione di una proteina terapeutica utilizzando spettroscopia di dicroismo circolare e infrarossi in combinazione con analisi multivariata, in particolare la Partial Least Squares (PLS) regression e la O-PLS (orthogonal-PLS), un'estensione della classica PLS. Questo è approccio nuovo in quanto le applicazioni classiche del CD e IR sono la determinazione del contenuto di strutture secondarie di una proteina. Nondimeno, il presente approccio è, secondo la nostra conoscenza, il primo che cerca di sfruttare la PLS allo scopo di correlare la spettroscopia CD e IR con dati quantitativi di comuni forme di degradazione di proteine.

Allo scopo di generare una matrice di calibrazione idonea, sono stati preparati una serie di campioni contenenti quantità predefinite di aggregati, ossidati e Fc libero. Per assicurare la non correlazione di ciascuna forma di degradazione all'interno della matrice di calibrazione, queste concentrazioni sono state scelte seguendo quanto descritto da Brereton (2000). Tutti i campioni generati sono stati quindi analizzati separatamente per ognuna delle tre forme di degradazione utilizzando le rispettive e specifiche analisi cromatografiche, allo scopo di determinare in maniera accurata i livelli di tutte e tre le forme di degradazione. Inoltre, tutti i campioni sono stati analizzati attraverso spettroscopia CD e IR. Sia i dati spettroscopici sia i dati cromatografici rappresentano le basi per generare i vari modelli PLS/O-PLS, basati sia su dati CD e IR singolarmente, sia su dati CD e IR combinati insieme.

Questo approccio ha dimostrato la capacità di estrarre dati quantitativi su alcune forme di degradazione proteiche per una proteina di fusione con Fc. Sebbene entrambi gli spettri CD e IR contengano informazioni rilevanti, si è dimostrato che i modelli O-PLS basati su dati CD hanno dato una accuratezza maggiore rispetto a quella ottenuta attraverso i modelli O-PLS basati su dati IR nel predire i livelli di tutte e tre le forme di degradazione. Inoltre, si è dimostrato come l'accuratezza di predizione raggiunta coni modelli O-PLS è maggiore di quella ottenuta con i rispettivi modelli PLS.

Nell'ultima parte della tesi ci si è soffermati sull'utilizzo del "protein design" nelle biotecnologie. In questo ambito, l'obiettivo della tesi è quello di studiare la stabilità dello scheletro peptidico del contrifano-Vn, un piccolo peptide isolato dal veleno della lumaca di mare *Conus ventricosus* formato da solo nove aminoacidi e caratterizzato dalla presenze di un singolo ponte disolfuro, dopo la sostituzione di 4 su 9 residui aminoacidici della sua sequenza.

I contrifani sono piccoli peptidi bio-attivi isolati dal veleno di piccole lumache di mare del genere *Conus,* caratterizzati dalla piccola lunghezza della catena peptidica e dall'alto grado di modificazioni post-traduzionali insolite. La ciclizzazione della catena peptidica attraverso il singolo ponte disolfuro, la presenza di due residui conservati di Prolina (Pro) e l'epimerizzazione di residui di Triptofano (Trp)/Leucina (Leu) conferiscono una struttura ben definita e stabile in soluzione, conservata in tutti i membri della famiglia. È stato testato il potenziale dei contrifani come struttura per il *design* di macromolecole attive dal punto di vista redox attraverso l'ingegnerizzazione di un sito di legame per il rame su due varianti del peptide naturale contrifano-Vn, chiamate Cuprifano e Arg-Cuprifano, inserendo quattro residui di Istidina (His). Il sito di legame è stato modellato attraverso modelling computazionale e le due varianti peptidiche sintetizzate e caratterizzate attraverso spettroscopia UV-Vis, fluorescenza, EPR (Electron Paramagnetic Resonance) e NMR (Nuclear Magnetic Resonance).

I nuovi peptidi Cuprifano e Arg-Cuprifano legano ioni Cu^{2+} con stechiometria 1.1 e una $K_d = 1.3(\pm 0.2) \times 10\text{-}7$ M e $1.0(\pm 0.4) \times 10\text{-}7$ M, rispettivamente. Altri metalli divalenti (Zn^{2+} e Mg^{2+}) vengono legati con una affinità molto minore. Inoltre, i nuovi peptidi catalizzano la dismutazione di anioni superossido con un attività comparabile con quella altre molecole non peptidiche che mimano l'azione della superossido-dismutasi.

Quindi, sfruttando la stabilità dello scheletro peptidico del contrifano-Vn, la sua capacità di tollerare sostituzioni aminoacidiche che non influiscono sulla sua stabilità e struttura, questo lavoro ha dimostrato la possibilità di usare lo scheletro peptidico dei conopeptidi come base per l'ingegnerizzazione di nuovi biocatalizzatori e ingegnerizzare nuove e stabili macromolecole attive dal punto di vista redox

1 INTRODUCTION

The manufacture of biological products is a complex process that constantly evolves throughout the development of a product (Bierau et al., 2007). It consists of a series of unit operations, each intended to modulate certain properties of the materials being processed. To ensure acceptable and reproducible modulation, consideration should be given to the quality attributes of incoming materials and their process-ability for each unit operation. During the last 3 decades, significant progress has been made in developing analytical methods for chemical attributes (*e.g.,* identity and purity). However, certain physical and mechanical attributes of pharmaceutical ingredients are not necessary well understood. Consequently, the inherent, undetected variability of raw materials may be manifested in the final product (Food and Drug Administration, 2004).

Moreover, a number of constraints, such as increased yields, scale-up or increased purity may require the re-design or optimization of a given process. Heterogeneity in a biopharmaceutical product, at the beginning of its shelf life, arises from inherent variation within the production process, resulting in the presence of various forms of post-translational modifications or degradation products. Clearly, the foremost aim of designing or optimizing a production process is obtaining maximum purity of the product, i.e. minimization of its degradation forms. Nevertheless, especially for products on the market or used in clinical trials, product consistency, *i.e.* a consistent pattern of critical quality parameters, is equally important in order to ensure that any change in a manufacturing process does not adversely affect the safety or efficacy of a product (Bierau et al., 2007).

Process development accounts for a significant fraction of the cost (Harms et al., 2002) of bringing a drug to the market and there is a strong economic incentive for improved monitoring tools. In theory, early process development stages have the goal of understanding and optimizing the process by measuring as many parameters as possible. Later on in production, only critical parameters may be measured to enable proper control of the process and to ensure high quality and yield; however, in practice, the inadequacy of suitable sensors and tools for online monitoring has not allowed this ideal to be widely implemented (Harms et al., 2002).

Thus, with the aim to speed up development time, there is a clear need for establishing rapid methods that shorten the turnaround time for analyzing in-process samples and simultaneously maximize the information content obtained (Bierau et al., 2007) Today significant opportunities exist for improving pharmaceutical development, process analysis and process control. Unfortunately, the pharmaceutical industry generally has been hesitant to introduce innovative systems into the manufacturing sector for a number of reasons. One reason is regulatory uncertainty, which may arise from the perception that the regulatory system is rigid and unfavourable to the introduction of innovative systems. On the other hand, pharmaceuticals continue to have a prominent role in health care, therefore pharmaceutical manufacturing will need to employ innovation, and scientific knowledge (Food and Drug Administration, 2004).

1.1 PAT (PROCESS ANALYTICAL TECHNOLOGY)

In August 2002, the Food and Drug Administration (FDA) launched a new initiative entitled “Pharmaceutical CGMPs for the 21st Century: A risk based approach. With this initiative, manufacturers are encouraged to use the latest scientific advances in pharmaceutical manufacturing and technology.

Pharmaceutical manufacturing continues to evolve with increased emphasis on science and engineering principles. Effective use of the most current pharmaceutical science and engineering principles and knowledge can improve the efficiency of both manufacturing and regulatory processes. This FDA initiative is designed to do that by using an integrated system approach to regulate pharmaceutical quality. The approach is based on science principles for assessing and mitigating the risks related to poor product and process quality. The desired state of pharmaceutical manufacturing and regulation may be characterized as follows:

- Product quality and performance are ensured through the design of effective and efficient manufacturing processes

- Product and process specifications are based on the understanding of how process factors affect product performance

- Continuous real time quality assurance

Process Analytical Technology (PAT) is defined by the FDA as a "System for designing, analyzing and controlling manufacturing through timely measurements of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality" (Food and Drug Administration, 2004).

It is important to note that the term analytical in PAT includes chemical, physical, microbiological, mathematical and risk analysis conducted in an integrated manner.

The goal of implementing PAT is defined therein as enhancing the understanding and the control of a production process, with the following approach: quality cannot be tested into products; it should be built-in or should be by design. (Food and Drug Administration, 2004).

Quality is built into pharmaceutical products through a complete understanding of the intended therapeutic objectives and the pharmacological, toxicological, pharmacokinetic, chemical, physical and biopharmaceutic characteristic of a drug, and by design of manufacturing processes using principles of engineering, material science and quality assurance to ensure reproducible product quality and performance throughout a product's shelf life.

Using this approach of building quality into products, PAT highlights the necessity for process understanding and opportunities for improving manufacturing efficiency through innovation. A process is well understood when: i) all critical sources of variability are identified and explained; ii) variability is managed by the process; iii) product quality attributes can be accurately and reliably predicted over the design space established for materials used, process parameters, manufacturing and other conditions.
A desired goal of the PAT framework is to design and develop well-understood processes that will consistently ensure a predefined quality at the end of the manufacturing processes. (Food and Drug Administration, 2004).

1.2 PAT TOOLS

There are many tools available that enable process understanding for scientific, pharmaceutical development. These tools, can provide effective and efficient means for acquiring information to facilitate process understanding and continuous improvement. In the PAT framework, these tools can be categorized according to the following:

- Multivariate tools for design, data acquisition and analysis

- Process analyzers

- Process control tools

An appropriate combination of some, or all, of these tools may be applicable to an entire manufacturing process (Food and Drug Administration, 2004).

1.2.1 Multivariate tools for data analysis

From a physical, chemical and biological perspective, pharmaceutical products and processes are complex multi-factorial systems. There are many development strategies that can be used to identify optimal processes, and the knowledge acquired in these development programs is the basis for process design. A knowledge base can be very useful when it arises from scientific understanding of the multi-factorial relationships, *e.g.* between process and quality attributes. This benefit can be achieved by using multivariate approaches in conjunction with knowledge management systems.

Methodological experiments based on multivariate statistical principles provide useful means for the identification and study of the effect and interaction of product and process variables. Traditional one-factor-at-time experiments cannot address these kinds of interactions. Multivariate statistical tools enable the identification and evaluation of product and processes variables that may be critical to product quality and performance (Food and Drug Administration, 2004).

1.2.2 Process analyzers

Process analysis has advanced significantly during the last several decades, due to an increasing appreciation for the value of collecting processing data. Available tools have evolved from those that predominantly take univariate process measurements, such as pH, temperature and pressure, to those that measure biological, chemical and physical attributes in a non-destructive way (Food and Drug Administration, 2004).

Indeed some process analyzers provide non-destructive measurements that contain information related to biological, physical and chemical attributes of the materials. These measurements can be:

- at-line: Measurements where the sample is removed isolated and analyzed in proximity to the process
- on-line: Measurements where the sample is diverted from the manufacturing process
- in-line: Measurements where the sample is not removed from the process

Measurements collected from these process analyzers don't need to be absolute values of the attributes of interest. The capability to measure relative differences in material before and during processing can provide useful information in process development and control. However, multivariate analyses are often necessary to get critical process knowledge for real time control and quality assurance (Food and Drug Administration, 2004).

1.2.3 Process control tools

It is important to emphasize that a strong link between product design and process development is essential to allow the effective control of all critical quality attributes. Process monitoring and control activities are intended to monitor the state of a process and manipulate it to maintain a desired state (Food and Drug Administration, 2004).

Design and optimization of manufacturing process within the PAT framework can include:

- Identification and measure of critical material and process attributes related to product quality
- Design a process measurements system to allow real time or near real time (on-, in or at-line) monitoring of all critical parameters

- Design process controls that provide adjustments to ensure control of all critical attributes

- Develop mathematical relationships between product quality attributes and measurements of critical process attributes

Thus, within the PAT framework, a process end point is not a fixed time, but rather it's the achievement of the desired attribute. Obviously, this does not mean that process time is not considered (Food and Drug Administration, 2004).

1.3 PAT PRINCIPLES

1.3.1 Real time release

Real time release is the ability to evaluate and ensure the acceptable quality of in-process and/or final product based on process data. The PAT component of real time release includes a combination of assessed material attributes and process controls. Material attributes can be determined using process analytical methods. The combined process measurements and other test data collected during the manufacturing process can serve as the basis for real time release of the final product and to demonstrate that each batch conforms to established regulatory quality attributes.

1.3.2 Risk-based approach

Within an established quality system and for a particular process, one would expect an inverse relationship between the level of process understanding and the risk of producing a poor quality product. For a well-understood process, opportunities exist to develop less restrictive regulatory approaches to manage change (Food and Drug Administration, 2004).

1.4 SPECTROSCOPIC TECHNIQUES

Thanks to the PAT initiative, spectroscopic sensors systems have gained interest for bioprocess monitoring because they allow rapid and non-destructive monitoring of product quality attributes. Improvements in spectrometers, detectors and optics have led to interesting applications related to PAT (Harms et al., 2002; Adam et al., 1999; Vaidyanathan et al., 1999; Schugerl, 2001). More specifically several authors showed that spectroscopic techniques could be used to analyze glycosylation, aggregation or oxidation (Lu et al., 1995; Petty et al., 2005).

1.4.1 Circular Dichroism spectroscopy (CD)

Circular Dichroism (CD) is being increasingly recognised as a valuable technique for examining the structure of proteins in solution. It measures the optical activity of asymmetric molecules in solution by measuring their unequal absorption of left- and right- handed circularly polarised light.

Plane polarised light can be viewed as being made up of 2 circularly polarised components of equal magnitude, one rotating counter-clockwise (left handed, L) and the other clockwise (right handed, R) (fig. 1).

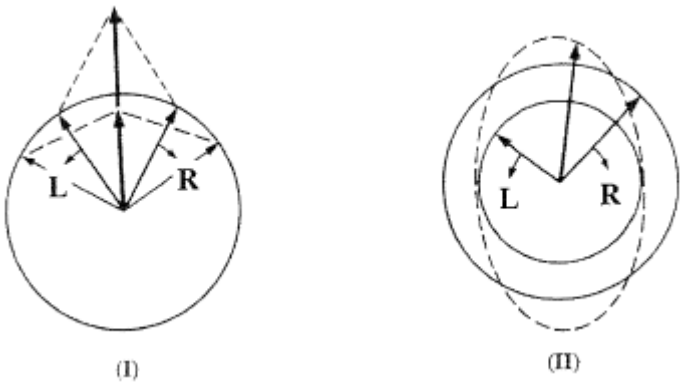

Figure 1. Origin of CD effect
(I) the L and R circularly polarized components have the same amplitude and when combined generate plane polarized radiation,; **(II)** the components are of different magnitude and the resultant (dashed line) is elliptically polarized

Circular Dichroism is observed when molecules absorb left and right circularly polarized light to different extent (Pelton and McLean, 2000). If, after passage through the sample, the L and R components are not absorbed or are absorbed to an equal extent, the recombination of L and R would regenerate radiation polarised in the original plane. However, if L and R are absorbed to different extents, the resulting radiation would be said to possess elliptical polarisation (Fig. 1) (Kelly et al., 2005). A CD signal will be observed only when a chromophore is chiral (optically active). A chromophore is chiral when:

- It is intrinsically chiral because of its structure, *e.g.* a C atom with 4 different substituents, or the disulfide bond which is chiral due to the dihedral angles of S-C-C-S chain of atoms.

- It is covalently linked to a chiral centre in the molecule

- It is placed in a asymmetric environment because of the three-dimensional structure adopted by the molecule

In proteins, the chromophores of interest include the peptide bonds, aromatic amino acids side chains, and disulphide bonds. If a number of chromophores of the same type are in close proximity, they can behave as a single absorbing unit that will give rise to characteristic spectral features. CD is measured in two spectra regions, the far-UV and near-UV region (Kelly et al., 2005).

Far-UV spectra. The far-UV (ca. 180-250 nm), also called the amide region, is dominated by contributions of the peptide bond. CD bands in the far-UV contain information about the configuration of the peptide bonds and therefore the secondary structure of the protein. It can be used to detect changes in secondary structure (New technologies catalogue, MerckSerono).

Alpha helix, beta-sheet and random coil structures each give rise to a characteristic shape and magnitude of the CD spectrum (fig. 2).

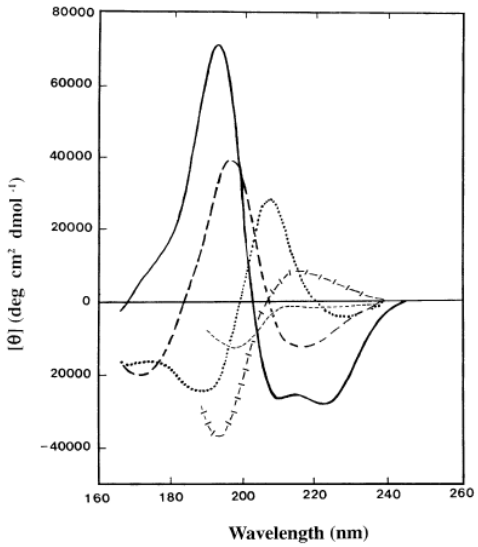

Figure 2. Far-UV spectra associated with various types of secondary structure
Solid line; α-helix; long dashed line, anti-parallel β-sheet; dotted line, type I β-turn; cross dashed line, extended 31-helix or poly (pro) II helix; short dashed line, irregular structure

The approximate fraction of each secondary structure can thus be determined by analyzing its far-UV CD spectrum as a sum of fractional multiples of such reference spectra for each structural type. (New technologies catalogue, MerckSerono).

Near-UV spectra. In the near-UV region (250-340 nm), also called aromatic region, CD bands arise from the absorption of aromatic amino acids. In addition, disulphide bonds give rise to minor CD bands around 250 nm. CD bands in the near-UV region are observed when aromatic amino acids are immobilized in a folded protein and thus transferred to an asymmetric environment. Each of the amino acids tends to have a characteristic wavelength profile (fig. 3).

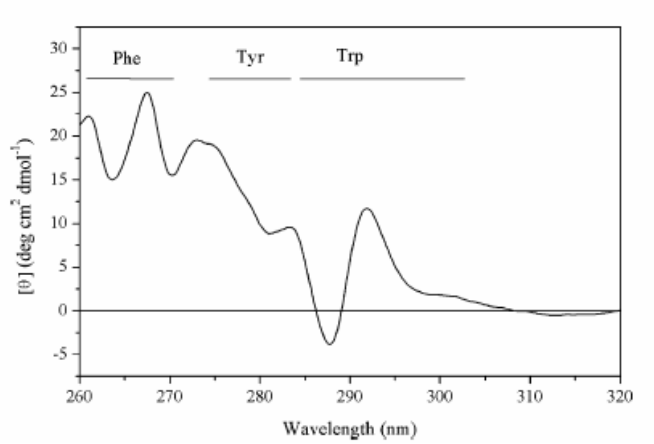

Figure 3. The near-UV CD spectrum
Wavelength ranges corresponding to signals from Phe, Tyr and Phe side-chains are indicated

Tryptophan (Trp) shows a peak close to 290 nm, with fine structure between 290 and 305 nm; Tyrosine (Tyr) a peak between 275 and 282 nm, with a shoulder at longer wavelengths often obscured by bands due to Trp; Phenylalanine (Phe) shows weaker but sharper bands with fine structure between 255 and 270 nm (Kelly et al., 2005). The actual shape and magnitude of a near-UV spectrum depends on the number of each type of aromatic amino acids present, their mobility and environment (H-bonding, polar groups and polarizability) (Kelly et al., 2005).

1.4.2 Fourier- transformed infrared spectroscopy (FT-IR)

Infrared (IR) spectroscopy is one of the most common spectroscopic techniques used by organic and inorganic chemists for the determination of secondary structure of proteins in solution. It is based on the wavelength and intensity of the absorption of infrared light by a sample. Infrared light is energetic enough to excite vibrational transitions of bonds in molecules. The energy of these molecular vibrations corresponds to the IR region of the electromagnetic spectrum (New technologies catalogue, MerckSerono).

Two experimental methods are available for measuring infrared spectra. In the dispersive method, infrared light is passed through the sample, and the absorption is measured. This method has been largely supplanted by Fourier transform method. Instead of viewing each component frequency sequentially, as in a dispersive IR spectrometer, all frequencies are examined simultaneously in Fourier transform infrared (FTIR) spectroscopy. Basically with Fourier transform infrared measurements, a beam of light is split in two, with half of the light going directly to the sample, and the other half of it is diverted to a moving mirror which then directs it to the sample. The difference in phase of the two waves creates a constructive and/or destructive interference and is a measure of the sample absorption. The waves are scanned over a specific wavelength region of the spectra, and multiple scans are averaged to create the final spectrum (Hammes, 2005). Among the fundamental vibrations (also known as normal modes of vibration), those that produce a net change in the dipole moment may result in an IR activity. The major types of molecular vibrations are stretching and bending (fig. 4).

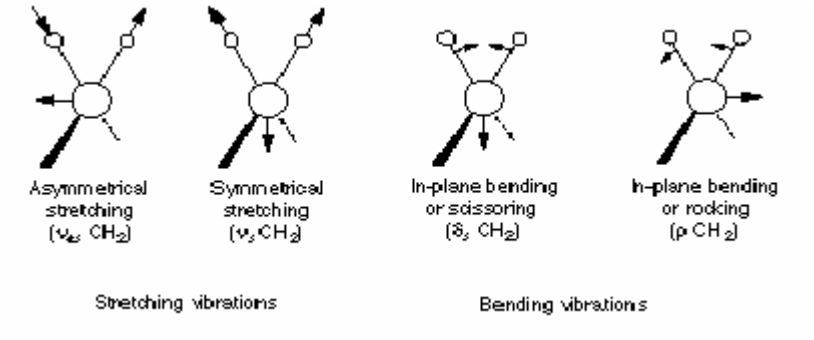

Figure 4. Major vibrational modes for a non-linear group, CH_2

The carbonyl group of amide bond in proteins is particularly useful for the determination of secondary structure (Hammes, 2005). The stretching normal mode, amide I mode, of the carbonyl has been shown to have a specific frequency associated with α-helices, β-sheets and other characteristic structures. According to the spectra-structure database of infrared frequencies (Hammes, 2005) the major absorbance peaks can be assigned to specific chemical structures (Hammes, 2005). The C=O stretch vibration of peptide bonds can be identified around 1650 cm^{-1}. It corresponds to the amide I vibrational state. The N-H bending is located at 1550 cm^{-1}. It corresponds to the amide II vibrational state. The N-H bending and the C-N stretching absorb IR light around 1250 cm^{-1} (fig. 5).

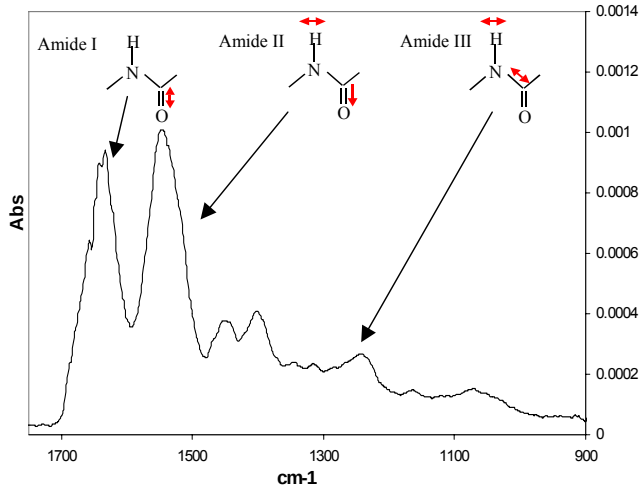

Figure 5. Infrared spectra of proteins

The vibrational spectrum of the amide bond for a protein is complex because of the many amide bonds present in multiple environments (Hammes, 2005).The amide I band is the most intense absorption band in proteins. It is primarily governed by the stretching vibration of C=O (70-85%) and C-N groups (10-20%). Its frequency is found in the range between 1600 and 1700 cm^{-1}. The exact band positions are determined by the backbone conformation and hydrogen-bonding pattern (New technologies catalogue, MerckSerono).

For structure determination in proteins, analysis of amide I band is the method of choice, because it is the strongest band, and it almost exclusively arise from C=O stretching and therefore directly correlated to the backbone conformation (New technologies catalogue, MerckSerono).

1.4.3 Raman spectroscopy

IR and Raman spectroscopy are the main methods to measure the vibrational spectrum of proteins. When a molecular vibration is accompanied by a change of dipole moment, the mode is IR active, and when the vibration is accompanied by a change of polarizability, the mode is Raman active (Kitagawa et al., 2002). In addition to absorbing light, samples also scatter light. The amount of scattered light is at maximum at 90° from the direction of the incident light. Most of the scattered light is at the same frequency of the incident light. This is called Rayleigh scattering (Hammes, 2005). At the molecular level, the electric field of the light perturbs the electron distribution, but no transitions between energy levels occur so that the molecule immediately returns to its unperturbed state. This scattering is inversely proportional to the fourth power of the wavelength so that the scattering is much greater at shorter wavelengths. Rayleigh scattering is observed at all wavelengths and the intensity of the scattered light is related to the polarizability of the molecule.

A small fraction of the molecules return to a different vibrational energy level after scattering. The vibrational energy level can be either higher or lower than the initial state. As a result of this change in energy level, some of the scattered light will be at a lower or a higher frequency than the incident light. This is called Raman scattering (Hammes, 2005) (fig. 6).

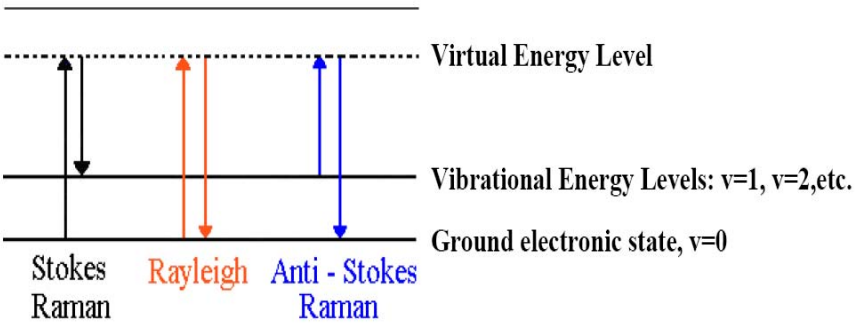

Figure 6. Origin of Rayleigh and Raman effect

The Rayleigh line is very intense, but much less intense scattering can be detected at a lower frequency than the incident light, because the incident light was used to promote the molecule to a higher vibrational energy level. These are called Stokes line. An even smaller fraction of the scattered light occurs at a higher frequency than the incident light, because energy is added to the incident light by the movement of the molecule to a lower vibrational state. These are called anti-Stokes line (Hammes, 2005) (fig. 6).

Raman spectroscopy has two principal advantages over infrared spectroscopy. First, a permanent dipole is not required. Second, a Raman spectrum can be easily obtained in aqueous solution, because the Raman scattering of water is very weak (Kitagawa et al., 2002). The primary disadvantages are that because the intensity of the Raman lines is very weak, an intense light source and high concentrations of the molecule are required (Hammes, 2005). Another important characteristic of Raman spectroscopy is the possibility of resonance enhancement. When the excitation wavelength of Raman scattering approaches an absorption maximum of an electronic transition of a molecule, some of the molecular vibrations gain strong enhancement of Raman intensity. This is called resonance Raman spectroscopy (Kitagawa et al., 2002).

Raman and infrared spectroscopy should be regarded as complementary. Since infrared spectroscopy is dependent on the permanent dipole moment and Raman spectroscopy on the polarizability, usually a vibrational transition is observed either in the infrared or in Raman scattering, but not in both (Hammes, 2005).

The frequencies of the vibrational motions of the amide I, II and III modes reflect the structure of the main polypeptide mainchain. Typical frequencies for α-helix, β-sheet and non-regular structures are shown in table 1 (Nielsen)

Secondary structure	Amide I (cm^{-1})	Amide III (cm^{-1})
α-helix	1645-1658	1260-1305
β-sheet	1665-1680	1230-1245
β-turn	1663-1678	1258-1300
Randm coil	1660-1665	1242-1255

Table 1. Characteristic amide bands in Raman spectra of proteins.

The side-chain vibrational modes of aromatic amino acids have been established by Harada and Takeuchi (Kitagawa et al., 2002). The Raman bands around 1360 and 1340 cm^{-1} are called the Trp doublet and their relative intensity reflect the hydrophobic/hydrophilic environment of Trp residues, being larger for more hydrophobic environments (Kitagawa et al., 2002). The Raman bands around 860 and 833 cm^{-1} are called the

Tyr doublet, the relative intensity of which is sensitive to the microenvironment of Tyr residues (Kitagawa et al., 2002).

1.4.4 UV-Vis absorption spectroscopy

The most common type of spectroscopy involves light in the visible and ultraviolet (UV) region of the spectrum interacting with molecules. This interaction causes electrons to shift between their allowed energy levels (Hammes, 2005). The most important aspect of the quantum mechanical calculation is the determination of how much light is absorbed by the sample. This is embodied in the Lambert-Beer law, which gives the relationship between the light intensity entering the solution, I_0, and the light intensity leaving the solution, I:

$$Log(I_0/I) = A = \varepsilon cl \qquad (1)$$

Here A is defined as the absorbance, ε is the molar absorptivity or extinction coefficient, c is the concentration of the absorbing sample and l is the thickness of the sample trhough which the light passes. The extinction coefficient is a different constant for each wavelength and it is characteristic of the molecule (Hammes, 2005).

A very important property of the light absorbing solutions is that if multiple absorbing species are present, the total absorbance is simply the sum of the absorbance of the individual species (Hammes, 2005). UV spectroscopy can be used to confirm the identity and to assess the purity of recombinant proteins and their peptide fragments. The near-UV (250 to 350 nm) absorbance spectrum of a protein is almost entirely a function of its aromatic amino acids (tryptophan, tyrosine and phenylalanine). Because each protein has a unique amino acid sequence, the particular aromatic amino acids content of each protein results in a unique spectrum in the near-UV region (Coligan et al., 1995).

The spectra of the aromatic amino acids phenylalanine, tyrosine and tryptophan are shown in figure 7. It can be seen that electronic transitions in tyrosine and tryptophan are responsible for the absorption peak at around 280 nm. It can also be seen that all three amino acids absorb strongly at shorter wavelengths. Measurements of spectra of proteins without aromatic amino acids show strong absorbance at around 192 nm, due to the electronic transitions associated with peptide bonds (Hammes, 2005).

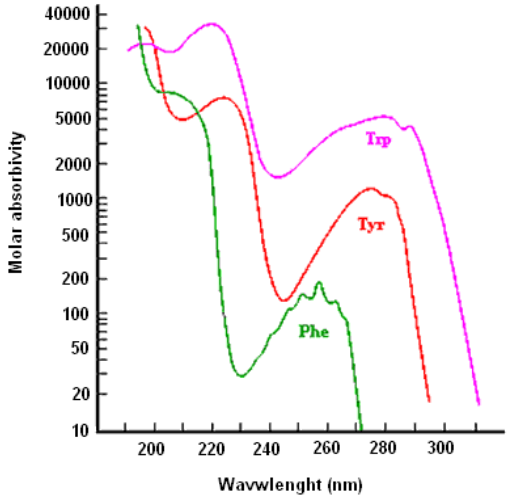

Figure 7. Spectral characteristics of chromophores in proteins

The spectra of aromatic amino acids and peptide bonds are strongly influenced by their local environments. Thus, the spectra of peptide bonds and aromatic residues buried inside are somewhat different from those on the exterior of the protein. This environmental sensitivity of protein spectra can be used to obtain information about protein structure (Hammes, 2005).

1.4.5 Fluorescence spectroscopy

Fluorescence is a very common phenomenon in biology. Fluorescence is the emission of light associated with electrons moving from an excited state to the ground state. It is a more useful tool for studying biological process than absorbance and can be considerably more sensitive than absorbance so that much lower concentrations can be detected (Hammes, 2005). This means, however, that small traces of fluorescent impurities in the solvent are readily detected and can lead to misinterpretation of the spectra (Coligan et al., 1995). In order to understand fluorescence, an energy diagram such as that shown in figure 8 is useful.

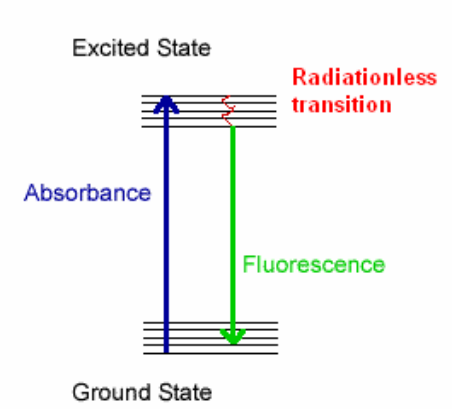

Figure 8. Schematic energy level diagram for absorbance and fluorescence

In this diagram are shown two electronic energy levels, each with their vibrational energy levels, due to the vibrations of atoms within the molecule. Electrons are in their ground electronic energy state at room temperature, and the molecule it is in its lowest vibrational level. They can be excited by light to the next electronic energy level, but can be in many different vibrational levels in the excited state. After this excitation, the molecule will return to the lowest vibrational level of the second electronic energy level very rapidly, through a radiationless transition (production of heat). Electrons in the second energy level can then decay to the electronic ground state. The light emission produced is is named fluorescence (Hammes, 2005).

The emitted light is always at a longer wavelength than the absorbed light, since the energy change associated with emission can never be greater than the energy change associated with absorption and the number of photons emitted can never exceed the number of photons absorbed (Hammes, 2005). The efficiency of fluorescence is characterized by the quantum yield, Q, that is the fraction of photons absorbed that are eventually emitted:

$$Q = n \text{ photons emitted} / n \text{ photons absorbed} \quad (2)$$

The measurement of fluorescence is more complex than the measurement of absorption since the sample must be excited at a specific wavelength and the fluorescence observed at a different wavelength. The emission spectrum is obtained by keeping the excitation wavelength constant and observing the emission over a range of

wavelengths. The emission spectrum of figure 9 was determined by exciting tryptophan with light at the absorption maximum of 275 nm (Hammes, 2005).

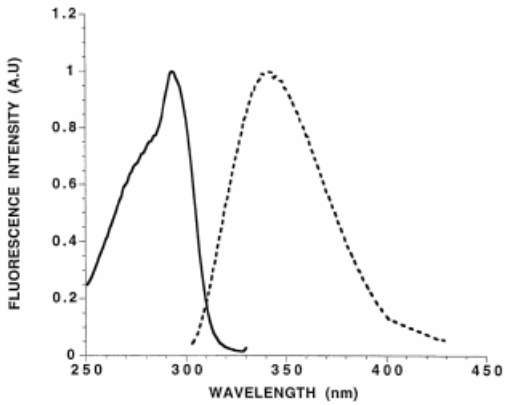

Figure 9. Fluorescence spectrum of Trp
The solid curve is the fluorescence excitation spectrum and the dashed curve is the fluorescence emission spectrum. The fluorescence intensity is normalized to one in order to display the spectral shape of the excitation and emission curves.

.

1.5 CHEMOMETRICS ANALYSIS

For a number of good reasons, spectral measurement is often the method of choice in qualitative and quantitative analysis of chemical mixtures. It is relatively easy to generate a good deal of data in a short time by proper use of spectroscopy. Getting useful results from a set of spectral data is not always straightforward, however. Determining the amounts of the components of a mixture can often be problematic without a prior separation step because of the overlap of spectral responses. Identifying the components of a mixture can also be challenging because of the similarity of many spectral responses. Often, the solution to these problems has been to increase spectral resolution or, as in the case of the quantitative analysis, to enhance the spectral resolution by means of a prior separation step. Many of these spectroscopic fixes to the problem of extracting results from data work less well than one might expect, given the apparent "information" in a spectral scan. For this reason, spectroscopists have increasingly turned to chemometrics for help in dealing with spectral data.

Svante Wold and Bruce R. Kowalski introduced the term chemometrics in the early 1970s. Chemometrics is rapidly advancing in industry. In the year 2003, the regulative authorities in USA—the Food and Drug Administration—indirectly recommends chemometrics contained as a vital component in process analytical technology (PAT) in guidance for the pharmaceutical industry (Munck, 2007). There is no doubt that an investment in process monitoring by new multivariate instruments combined with training of staff in chemometric data analysis has an extremely short payback time in almost all kinds of industries (Munck, 2007). A reasonable definition of chemometrics remains as: “how do we get chemical relevant information out of measured chemical data, how do we represent and display this information, and how do we get such information into data” (Chau et al., 2004). There are two main reasons why chemometrics developed so rapidly at that time: (1) large piles of data not available before could be acquired from advanced chemical and (2) advancements in microelectronics technology within that period. The abilities of chemists in signal processing and data interpretation were enhanced with the increasing computer power (Chau et al., Wiley, 2004).

Masses of data are produced by measuring many variables on ensembles of chemical samples, or by frequently recording many signals from an industrial process in order to track its behaviour (Eriksson et al., 2006). The data explosion necessitates the use of appropriate tools for extracting meaningful information from the large amount

of raw data. It is no longer efficient to analyze data by simply looking at them or by plotting them in simple graphs. More sophisticated computer-based methods are needed (Eriksson et al., 2006).

Two multivariate projection methods are very useful:

- Principal Component Analysis (PCA)
- Partial least squares projection to latent structures (PLS)

Multivariate data analysis (MVDA) provides a toolbox of versatile data analytical tools. There are three basic problem types to which these multivariate tools can be applied:

- Overview of data table
- Classification and/or discrimination among groups of observations
- Regression modelling between two blocks of data (X and Y)

1.5.1 Principal component analysis (PCA)

In the early stage of a project, when little is known about a problem, a simple overview of the information in a data table is often required. Such overview can be obtained with principal component analysis (PCA) (Eriksson et al., 2006). PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data (Smith, 2002). PCA produces a summary showing how the observations are related and if there are any deviating observations or groups of observation in the data. In addition, PCA also provides an understanding of the relationships among the variables: which variables contribute similar information to the PCA model and which provide unique information about the observations. PCA describes the correlation structure in the X block of data (Eriksson et al., 2006). Often an initial PCA of a data set reveals grouping among the observations. This may indicate the need for further PCA modelling of each subgroup ("class") in order to understand the features of the different groups. Subsequently, it is possible to use the separate class models for classification. By "classification" it is meant the use of the established class models for classifying new observations (Eriksson et al., 2006).

Principal component analysis forms the basis for multivariate data analysis. As shown in figure 10, the starting point for PCA is a matrix of data with N rows (observations) and X columns (variables), here denoted by data matrix X. The observations can be analytical samples, chemical compounds, process time points of a process, batches from a batch process and so on (Eriksson et al., 2006). In order to characterize the properties of the observations, one measures variables. These variables may be of spectral origin (*e.g.* NIR, IR, CD, UV, NMR), chromatographic origin (e.*g.* HPLC, UPLC, GC) or they may be measurements from sensors in a process (*e.g.* pH, temperature, flows).

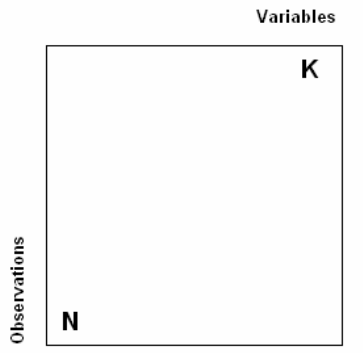

Figure 10. Notation used in PCA

The most important use of PCA is indeed to represent a multivariate data table as a low-dimensional plane, usually consisting of 2 to 5 dimensions. Statistically, PCA find lines and planes in the K-dimensional space that best approximate the data in the least square sense. A line or a plane that is the least squares approximation of a set of a data points makes the variance of the co-ordinates on the line or plane as large as possible (fig. 11)(Eriksson et al., 2006).

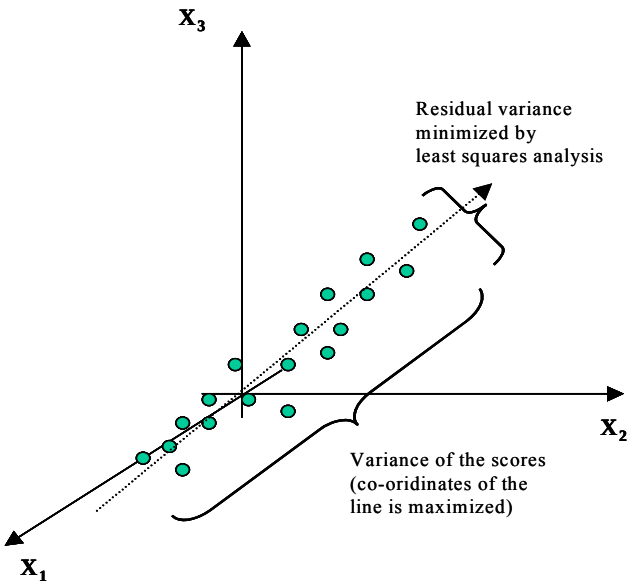

Figure 11. A geometric interpretation of PCA

Prior to PCA, data are often pre-treated, in order to transform data into a form suitable for the analysis. Variables often have substantially different numerical ranges. A variable with a large range has a large variance and a variable with small range has a small variance. Since PCA is a maximum variance projection method, a variable with a large variance is more likely to be expressed in the modelling than a low-variance variable (Eriksson et al., 2006). In order to give to all variables equal weight, the data have to be standardized. Such standardization is also known as “scaling” or “weighting”.

After scaling, the data are ready for the computation of the first principal component (PC1). This component is the line in the K-dimensional space that best approximates the data in the least squares sense, and represents the maximum variance direction in the data (Eriksson et al., 2006) (fig. 12). Usually one component is not sufficient to model the variation of a data set. Thus a second principal component (PC2) is calculated, which is orthogonal to the first (fig. 12). It reflects the second largest variation in the data and improves the approximation of X-data as much as possible (Eriksson et al., 2006). The two principal components define a plane, which can be visualized graphically (fig. 12). By projecting all the observations onto this plane, it is possible to investigate the structure of the data set. The co-ordinate values of each observation on this plane are called scores and hence the plotting is known as score plot (Eriksson et al., 2006). The score plot shows how the observations are projected into the plane. In a PCA model, we wonder which variables are responsible for the patterns seen among the observations. We would like to know which variables are influential and also how the variables are correlated. Such knowledge is given by the principal component *loadings.* These loading vectors are called p_1, p_2, p_3….(Eriksson et al., 2006). The loadings unravel the magnitude (large or small correlation) and the manner (positive or negative correlation) in which the measured variables contribute to the scores (Eriksson et al., 2006).

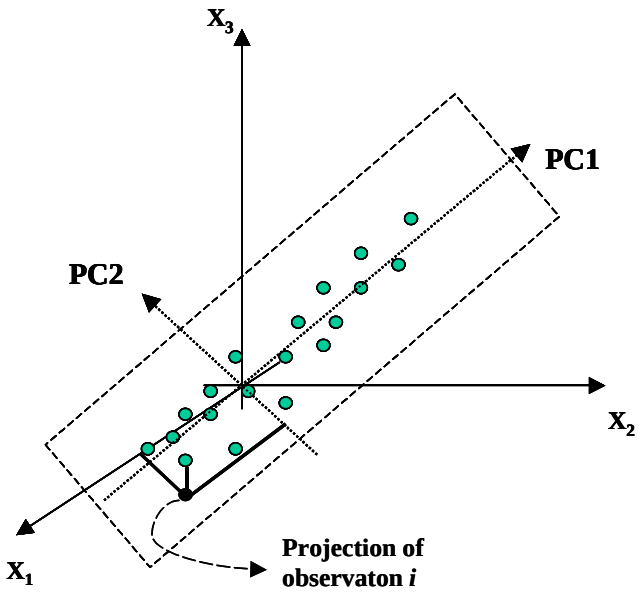

Figure 12. Principal components of PCA
Two PCs form a plane, which can be visualized graphically. Each observation can be projected onto this giving a score for each

1.5.2 Partial least squares projection to latent structures (PLS)

PLS regression is a recent technique used to connect two data matrices, X and Y, to each other by a linear multivariate model (Eriksson et al., 2006). It has been successfully applied to the quantitative analyses of UV, near-infrared, chromatographic and electrochemical data (Haaland et al., 1988). It is particularly useful when we need to predict a set of dependent variables from a very large set of independent variables (i.e., predictors). Thus, the goal of PLS regression is to predict Y from X and to describe their common structure. This prediction is achieved by extracting from the predictors a set of orthogonal factors called latent variables which have the best predictive power. (Abdi, 2007).

PLS derives its usefulness from its ability to analyze data with many, noisy and even incomplete variables both in X and Y. For parameters related to the observations, (samples, compounds, items), the precision of a PLS model improves with the increasing number of relevant X-variables (Eriksson et al., 2006). As in PCA, each observation can be represented graphically. The big difference in PLS is that each row of a data table corresponds to two points rather than one, one in the X-space and one in the Y-space. The task for the data analysis is to describe the relationship between the position of the observations in the predictor space (X) and their positions in the response space (Y) (Eriksson et al., 2006).

After scaling, it is time to calculate the first PLS principal component. This component is a line in the X-space that well approximates the point-swarm and provides a good correlation with the y-vector. The projections of the observations onto this line on the X-space give the score values for each observation. The scores of all the observations from the X-score vector, and this score can then be used to acquire an estimate of Y (Eriksson et al., 2006). The differences between measured and estimated response data are called residuals. The y-residuals represent the variation that is left unexplained by the first PLS component. A good model has small residuals. As in PCA, usually one PLS component is not sufficient to model the variation in the y-data. The predictive ability of the model is then improved calculating a second PC, a line in the X-space orthogonal to the first component.

In analytical chemistry, PLS is mainly used for multivariate calibration. Here X contains N spectra digitized at K wavelengths, and Y the analyte concentrations of the N training set samples. The resulting model can then be used to predict the analyte concentrations from the spectra of new samples (fig.13). Since the whole spectra is used, the multivariate PLS model results in better predictive precision, and also much improved selectivity in comparison with traditional univariate calibration (Eriksson et al., 2006).

1.5.2.1 Orthogonal Partial Least Squares (O-PLS)

Orthogonal Partial Least Squares is an extension of PLS. The objective of OPLS is to divide the systematic variation in the X-block into two model parts: one part which models the co-variation between X and Y, and a second part which captures systematic variation in X that is unrelated (orthogonal) to Y. O-PLS removes variation from X (descriptor variables) that is not correlated to Y (property variables). The non-correlated systematic variation in X is removed, making interpretation of the resulting PLS model easier and with the additional benefit that the non-correlated variation itself can be analyzed further. (Trygg and Wold, 2002).

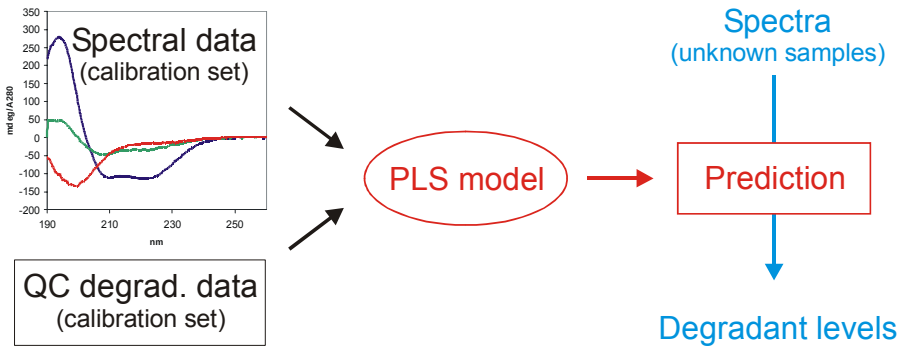

Figure 13. PLS regression model

1.5.3 Multivariate calibration

Spectroscopic sensors require calibration to determine the proportional relationship between the spectra measured and the component concentration or properties that are to be estimated. The signal of spectroscopic techniques in fact is not specific to the product to be measured. In complex mixtures the spectral features of the different components overlap. Due to this overlapping, univariate analysis such as Lambert-Beer's law cannot be applied to determine the relationship between spectra and the component concentration. Multivariate calibration has historically been a major corner stone of chemometrics as applied to analytical chemistry (Brereton, 2000).

The traditional approach to calibration is univariate. Analytical samples with known concentration (Ci) of an analyte are used as the basis for a calibration curve. In summary, traditional calibration means that known reference concentrations of an analyte are related to data arising from another measuring technique. The relevant calibration relationship is usually established by means of regression analysis (Eriksson et al., 2006). Traditional calibration normally utilizes only one predictor variable. Usually, more than one measured x-variable is available, especially when samples have been analyzed by spectroscopic methods. In such cases multivariate calibration can be used; this involves the use of all the information in a battery of X-variables (x1, x2… xk) to predict the analyte concentration (y) (Eriksson et al., 2006). In multivariate calibration we employ not just one selected signal from a spectrum, but rather the entire spectrum. Translating a spectrum in discrete variables is called digitization. Each spectrum is digitized at K wavelengths, and K variables (signal) are the results. These data are then analyzed with multivariate projection methods like PLS (Eriksson et al., 2006).

A typical multivariate calibration experiment involves spectra of several compounds, recorded at several concentration levels (Brereton, 1997). In many real-world situations there are a large number of components in a mixture. In order to build up models in laboratory, experiments involving mixing together compounds in different proportions and recording the resultant spectra are commonly employed (Brereton, 1997). One of the major problems arises in designing an adequate training set (Brereton, 2000). It is essential that the compounds forming the training set are uniformly distributed over mixture space (Brereton et al., 1998). An unbalanced design may results in poor predictions for unexpected and fairly subtle combinations of component concentration (Brereton, 1997). In practical cases, a series of mixtures may be made in the laboratory and a model developed using a multivariate approach such as PLS to predict concentrations of compounds from spectra using experimental mixtures. However, if the laboratory experiments don't adequately cover the mixture space, predictions could be dramatically in error under certain situations (Brereton, 1997). A key concept is that of orthogonality. Two factors are said to be orthogonal if they have a correlation factor of zero. This is equivalent to assess that the factors span each other's mixture space evenly. In order to have good predictions, it is necessary that any combination of factors is orthogonal (Brereton, 1997).

1.6 PROTEIN DESIGN

Engineering of metal centers has led to the development of a number of merging technologies with a wide variety of applications, including protein stability, control of protein structure and, therefore, control of protein activity (Hellinga, 1996). The ability to manipulate function rationally also offers the possibility of creating new proteins of biotechnological value (Marvin *et al*, 1997).

There are many reasons to pursue the goal of protein design. In medicine and industry, the ability to engineer protein hormones and enzymes to perform existing functions, or to perform entirely new functions, has tremendous potential (Street e Mayo, 1999). Great interest is directed towards mini metal-proteins and metal-enzymes, because of the wide range of applications in which can be used. Among these, the ones with oxidase and peroxidase activity are the most attractive. Metal binding sites are very interesting for *protein design* for many reasons (Regan, 1993). First of all, metal ions are simple ligands, the minimum requirements for a binding site is the right positioning of two to four amino acid side chains. Then, in natural proteins metal ions are responsible for a wide variety of functions: electron transfer, activation of molecular oxygen, protein structure stabilization, hydrolysis and structural rearrangements (Glusker, 1991).

Furthermore, a practical reason to choose metal binding sites as main candidates for design is the great number of spectroscopic techniques that can be used to detect the number and type of ligands involved and the binding site geometry (Regan, 1995). An important feature of metals chemistry is that one metal ion can carry out several functions depending on the nature of protein environment. For example, iron can act as electron transporter when is located in the eme group of a protein (in cytochromes) and as an oxygen transporter (in hemoglobin). In these two cases, the metal centre remains the same, but the reactivity is determined by the protein environment (Hellinga, 1996).

Protein engineering requires well-organized protein scaffolds. By engineering several disulfide bridges in short sequences (10-70 amino acids) nature has produced small proteins, toxins and protease inhibitors, which are able to adopt stable and biologically active structures. These natural mini-proteins constitute interesting candidates as core structures for protein design, since the disulfide bonds provide most of their stabilization energy, leaving a large part of the protein structure available for mutations (Vita *et al*., 1995). Among these, the most interesting, because of their scaffold stability, are a group of small toxins found in the venom of scorpions and of small sea predators, called *Coni.* Contryphans, a small peptides (8-9 residues) family found in the venom of *Coni,* are an example of well-organized protein scaffolds. Stability of Contryphans scaffold has been proved by engineering a member of this family, Contryphan-R, with the aim of design a mimic of the pharmacophore part of the unrelated globular polypeptide -conotoxin GVIA, which blocks N-type calcium channels. (Pallaghy e Nortorn, 2000). Contryphan-R is a 8 residues peptide. Residues D-Tyr4, Asn5 and Lys7 were replaced respectively by D-Tyr, Asn e Lys, generating a cyclic peptide called YNK-contryphan. The structure of this engineered contryphan, YNK-contryphan-R was shown to be similar to that of native contryphan-R confirming that the scaffold is robust with respect to the multiple substitutions (Pallaghy e Nortorn, 2000).

1.7 AIM OF THE THESIS

The main goal of this thesis is the development of novel methodologies based on spectroscopic techniques coupled with multivariate data analysis for the optimization of the production process of biopharmaceuticals. The approach combines the strengths of various spectroscopic techniques, such as Circular Dichroism (CD), Infrared (IR), Raman, Fluorescence and UV-Visible, to provide a more comprehensive description of a substance, the so-called “fingerprint”. This may be used to establish and define quality, equivalence, and comparability of substances while also providing means to monitor processes and provide relevant information on changes in a product. Moreover, it can highlight the relationships between different properties (for example those between structure, condition and aggregation) and provide a better understanding of the nature of a product.

Furthermore, the thesis presents the feasibility of obtaining quantitative data about degradation products of a therapeutic protein by employing Circular Dichroism and infrared spectroscopy in combination with multivariate data analysis (i.e., partial least squares, PLS) (Bierau et al., 2007), in order to be compliant with the PAT initiative launched by the FDA. This is a novel approach since the typical applications for CD and IR spectroscopy are the determination of secondary structure content of proteins (Kelly et al., 2005; Haris and Chapman, 1995). Also the use of multivariate statistical methods for the determination of secondary structure content has been reported (Rahmelow and Hubner, 1996; Oberg et al., 2004; Pribic, 1994). Nevertheless, the present approach is to our knowledge the first one that seeks to exploit PLS in order to correlate CD and IR spectral data with quantitative data of common protein degradation forms.

Regarding the “protein design”, aim of the present thesis is to study the scaffold stability of contryphan-Vn, a small peptide isolated from the venom of *Conus ventricosus* formed by only 9 residues and characterized by the presence of a single disulfide bridge, after substitution of 4 of the 9 amino acids of its sequence. The potential of Contryphans as scaffolds for the design of redox-active (macro)molecules was tested by engineering a copper binding site on two different variants of the natural peptide Contryphan-Vn, named Cupryphan and Arg-Cupryphan through the introduction of four His residues. The binding site was designed by computational modeling and the redesigned peptides were synthesized and characterized by optical, fluorescence, electron spin resonance and nuclear magnetic resonance spectroscopy.

2 RESULTS AND DICUSSION

2.1 MULTI-SPECTROSCOPIC CHARACTERIZATION OF DRUG SUBSTANCE

One of the aims of the thesis was to develop a new approach that combines the strengths of various spectroscopic techniques, such as Circular Dichroism (CD), Infrared (IR), Raman, Fluorescence and UV-Visible spectroscopy, to provide a more comprehensive description of a substance, the so-called "fingerprint". This may be used to establish and define quality, equivalence, and comparability of substances while at the same time providing means to monitor processes and provide relevant information on changes in product characteristics. The potential of such a fingerprint has impact in a wide variety of areas within biopharmaceutical research and development.

2.1.1 CD spectroscopy analysis

All the eleven batches of drug substance were analyzed by CD spectroscopy, both in the near and far-UV region.. Near-UV CD analysis were performed both on the concentrated (> 160 mg/ml), and on the diluted samples. Far-UV analysis were performed only on diluted samples. For instrumental parameters see Materials and Methods section.

Near-UV spectra of concentrated samples. In figure 14 are shown the near-UV spectra of concentrated samples. All the spectra were normalized using the correction factor described in Materials and Methods section.

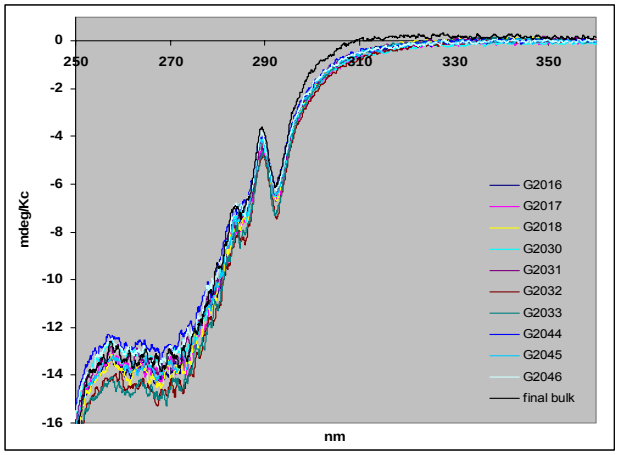

Figure 14. Near-UV CD spectra of concentrated samples

From the analysis of figure 14, it is evident that all the CD spectra of concentrated samples have similar form and exhibit CD bands between 250-310 nm, typical of proteins containing aromatic amino acids residues, but lacking of positive bands between 290-310 nm often associated to tryptophan residues. However, the normalised spectra are not absolutely identical, particularly in the 250-280 nm region (often associated with Phenylalanine) and in the tail between 300-350 nm (associated with disulfide bridges). The only spectrum that looks different from the others is the "final bulk" spectrum. In fact, this one shows a different shape in the region from 300 to 330 nm. This batch is the one arising from the new process called "process D".

Near-UV spectra of diluted samples. In general, all the spectra acquired on the diluted solutions were normalised both for pathlength and concentration, using the directly comparable 4°C store solutions concentrations. The samples pH values are listed in table 2, and can be compared with the stated pH 5.0 of the original concentrated stock solutions.

Batch	pH	
	4°C stored	R.T stored
A87G2016	7.4	8.5
A87G2017	7.4	8.4
A87G2018	7.6	7.7
A87G2030	7.8	8.2
A87G2031	7.6	8.2
A87G2032	6.7	7.3
A87G2033	7.6	7.9
A87G2044	7.7	8.0
A87G2045	7.8	6.9
A87G2046	7.6	8.4
Final bulk	8.0	7.6

Table 2. pH values of 4°C and room temperature stored solutions

It is clear that dilution led to an increase of the samples pH. Moreover, it is noticeable that the pH of the ambient stored solutions generally increased more than the corresponding ones stored at 4°C. Batch A87G3032 consistently displayed the lowest pH for both modes of storage. The spectra of diluted samples (fig. 15) were generally very reproducible within the noise level. We can assess that all the CD spectra look highly comparable, both for concentrated and for diluted ones. Only one sample looks different; the final bulk (process D) among the concentrated spectra.

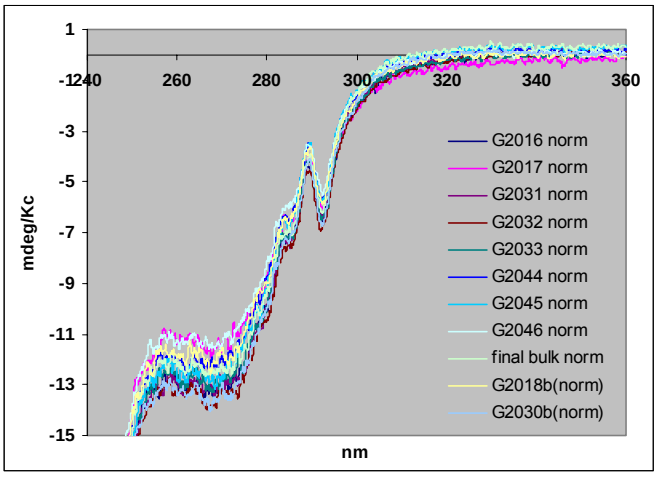

Figure 15. Near-UV CD spectra of diluted sample

Far-UV spectra. In figure 16 are shown the far-UV CD spectra normalized by using a specific correction factor. As expected, due to the excessive absorbance of buffer components, the spectra become dominated by noise for wavelength lower than 205 nm, and therefore spectra were truncated at this wavelength limit. As it can be seen, all the far-UV spectra look noisy in the whole region. By visual inspection, it can be noticed that all the far-UV

CD spectra are very reproducible, sharing the same spectral shape in the whole spectral range. The forms of CD spectra are consistent with a protein adopting a predominantly anti-parallel β-sheet secondary structure.

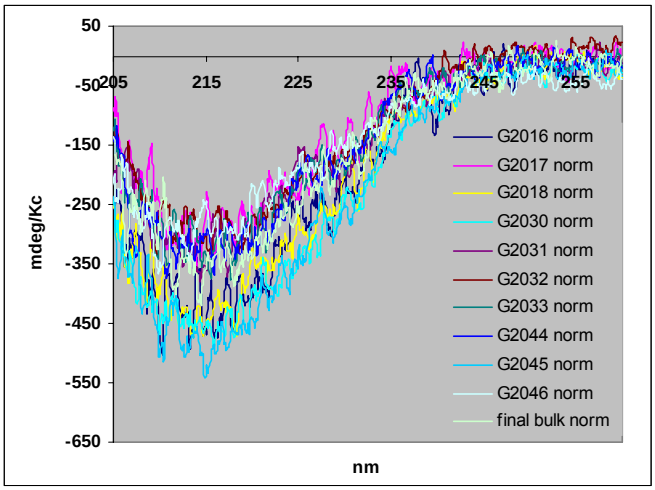

Figure 16. Far-UV CD spectra of diluted samples

2.1.2 PCA analysis of CD spectra

To have further information about the real comparability of the spectra, and to eventually detect difference features of the spectra not detectable by only visual inspection, the CD data, both near and far-UV, were analyzed by PCA analysis,

Near-UV PCA analysis of concentrated sample. In figure 17 is shown the score plot of near-UV data of concentrated samples. Normalized data were used, and these were "mean-centred" before PCA analysis.

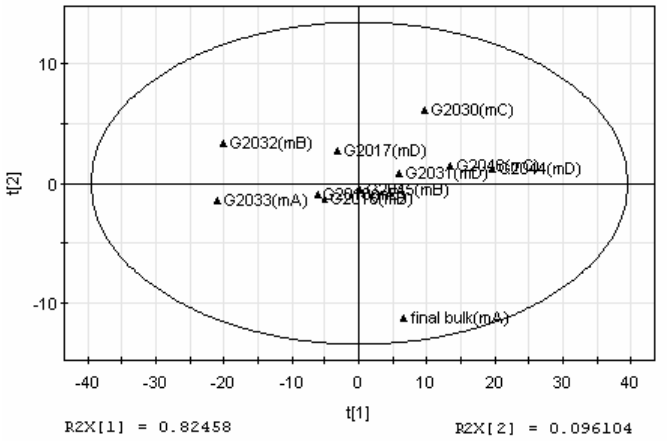

Figure 17. PCA score plot of near-UV CD spectra of concentrated samples

The score plot shows that data points form a main central cluster, in which the data scatter mainly along PC1, with A87G2044, 46, 30 and 31 at one extreme and A87G2032 and 33 at the other. However, the final bulk and A87G2030 batches are distinguished as opposite extremes in PC2. To understand which variables, or in other words which part of the spectrum contributes more to the patterns observed among the spectral data in the score plot, the loading plots of figure 18 must be analysed.

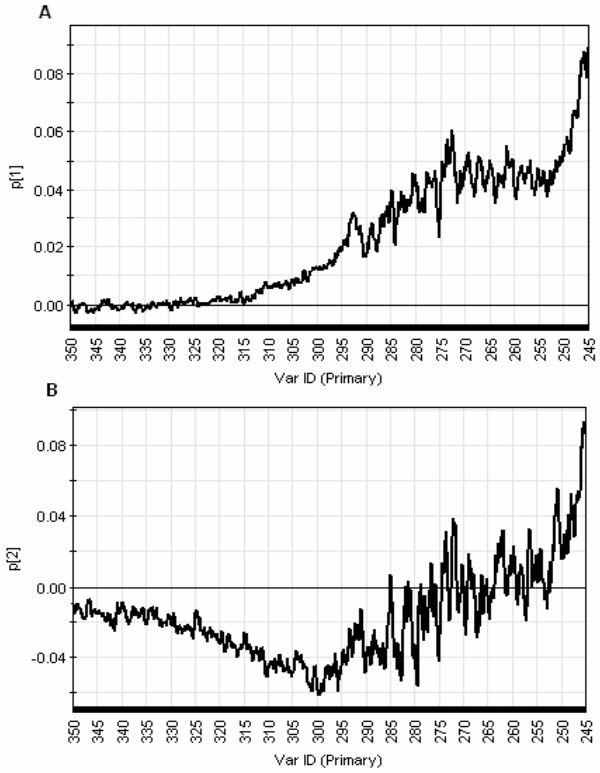

Figure 18. PCA loading plots of CD near-UV spectra of concentrated samples.
A) PCA loading plot of first component p_1; **B)** PCA loading plots of second component p_2

In these loading plots, the farther the variables values are from zero (black line), the higher is the importance of these variables in the model.

The loading plot p_1 (fig. 18 A) shows which variables are responsible for the pattern seen among the observations in the score plot along PC1. From loading plot p_1, it is evident that PC1 mainly represents a broad CD band at 275 nm with a CD signal rising to below 240.

The loading plot p_2 shows instead which variables are important for the pattern of data points along the t_2 component of the score plot. From the figure 18 B, it can be noticed that PC loading plot p_2 represents a very broad CD band centred around 285 nm and tailing towards visible wavelength values. These features are consistent with phenylalanine and disulfide bridges respectively.

It can be hypothesised that batches A87G2044, 46, 30 and 31 may subtly differ from A87G2032 and 33 in the environment of phenylalanine components, with a subtle difference in disulfide environment being apparent between final bulk and the other batches.

Near-UV PCA analysis of diluted samples. . In figure 19 is shown the score plot of near-UV data of diluted samples. As for concentrated samples, normalized and "mean-centred" data were used for PCA analysis.

The score plot shows that data points form a main central cluster, in which the data scatter equally along PC1 and PC2 Batch G2017 is separated from this cluster with respect PC2. Furthermore, it can be noticed that PC1 explains 80% of variation among data points.

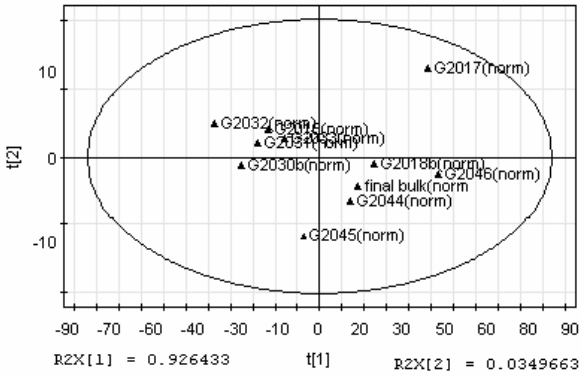

Figure 19. PCA score plot of near-UV CD spectra of diluted samples.

Analysing the loading plots (fig 20 and 21), it is possible to have further information about which wavelengths are responsible for the observed data distribution along PC1 and PC2.

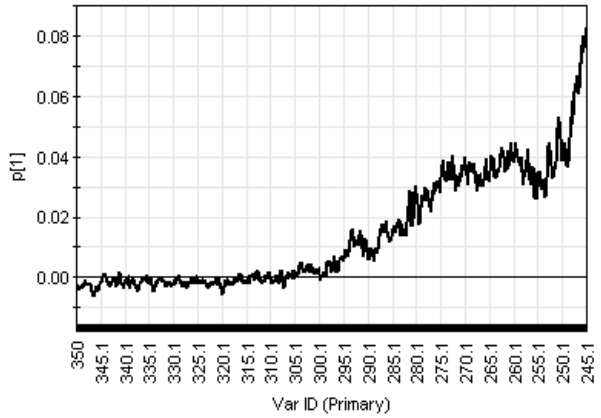

Figure 20. PCA loading p_1 plot of near-UV CD spectra of diluted samples

The loading plot p_1 (fig. 20) shows that PC1 mainly represents a broad CD band around 295-250 nm, which looks like an inverted mean spectrum of the drug substance, and may represents changes in concentration.

In contrast, loading plot p_2 (fig. 21) indicates that PC2 represents a tailing CD band from 290 to 240 nm and may represents either a baseline drift or extended disulphide CD bands.

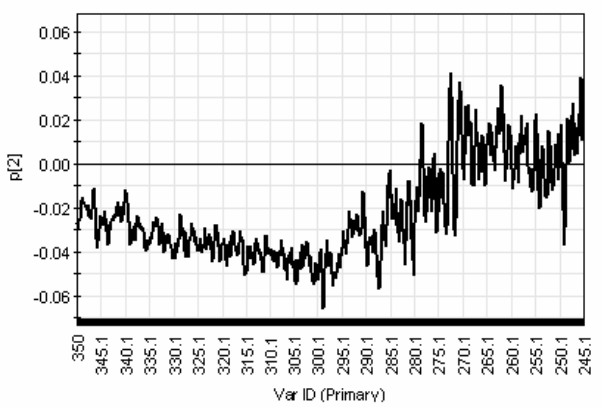

Figure 21. PCA loading plot p_2 of near-UV CD spectra of diluted samples

PCA analysis confirms that all the diluted spectra are comparable to each other throughout the near-UV range, with only the G2017 spectrum showing little differences in the region from 295 to 250 nm.

Far-UV PCA analysis. As for the near-UV data, far-UV spectra were analysed with PCA analysis, using the normalized and mean-centred data. In figure 22 is shown the score plot of far-UV data.

As expected from the visual inspection of the spectra (fig. 16), data form a big single cluster, in which is not observable any specific pattern. Only data G2046 seem slightly separated from the cluster along PC2. This is an indication of the complete comparability of the far-UV spectra. The score plot also indicates that 79% of the variation of data points is explained by the PC1.

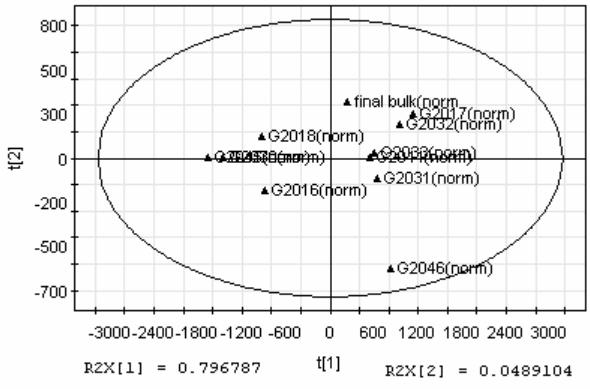

Figure 22. PCA score plot of far-UV spectra

Inspecting loading plot p1 (fig. 23, A) it is evident that PC1 looks like an inverted far-UV CD spectrum for this drug substance, and may therefore be correlated either to concentration variations or simple loss of secondary structure.

Loading plot p2 (fig. 23, B) indicates that the pattern of data points seen along t2 component of the score plot is mainly due to instrumental noise, equally distributed in the whole far-UV range. Consequently, the delineation of

Batch G2046 as an extreme of the distribution with respect to PC2 may be connected to a relative higher level of noise in this spectrum.

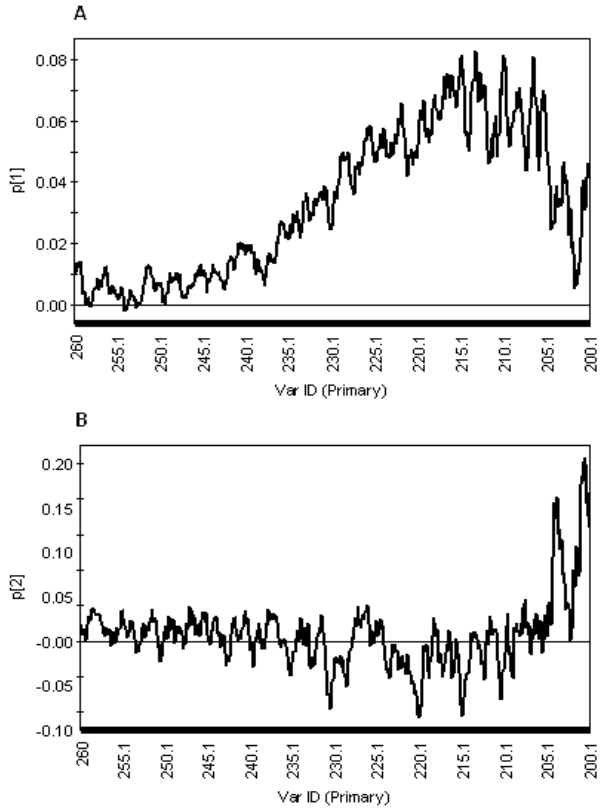

Figure 23. PCA loading plots of far-UV spectra
A) PCA loading plot of first component p_1; **B)** PCA loading plots of second component p_2

2.1.3 FT-IR spectroscopy analysis

FT-IR measurements were performed only on the concentrated drug substance. For instrumental parameters and normalization process see Materials and Methods section. In figure 24 are shown the spectra obtained after correction for the buffer spectrum.

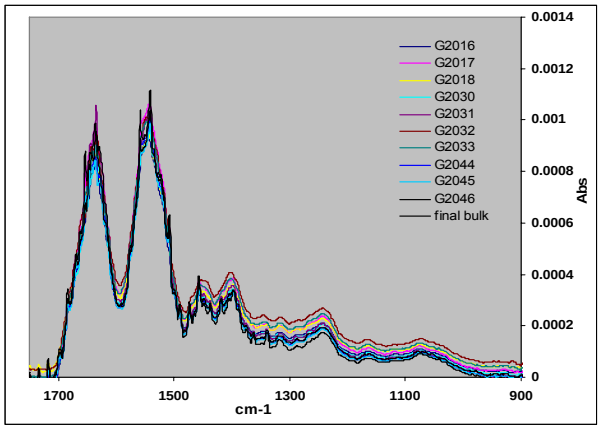

Figure 24. FT-IR spectra of concentrated samples

We can see that all the FT-IR spectra look highly comparable in terms of vibrational bands, but exhibit some variations in baseline level, as it is frequently observable when performing ATR measurements. In particular, all FT-IR spectra exhibit strong Amide I and II bands at ca. 1640 and 1540 cm^{-1} respectively together with a visible shoulder at ca. 1675 cm^{-1} also associated with the Amide I band, all consistent with a protein having a predominantly anti-parallel β-sheet structure.

2.1.4 PCA analysis of FT-IR spectra.

To further investigate the structural/spectral comparability of these batches, FT-IR spectra were analyzed by using PCA, as already done for CD spectral data.

From the score plot of figure 25 it can be seen how data points are arranged in a single cluster, indication of spectral comparability among all the FT-IR spectra. The first two principal components are sufficient to describe 92% of the overall variance in the spectra.

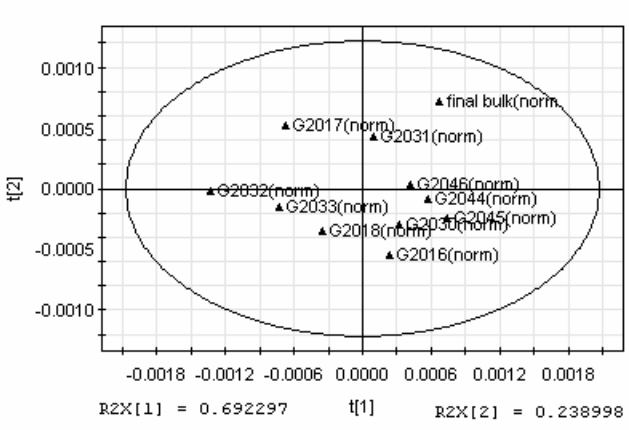

Figure 25. PCA score plot of FT-IR spectra

The loading plot p_1 (fig. 26) indicates that PC1 mainly represents a baseline shift across the whole spectral region, this is often seen in FTIR-ATR spectra and is connected with different aggregation of the proteins on the ATR crystal.

In contrast loading plot p_2 (fig. 27), encompasses a counteracting shift together with distinct absorbance at Amide I and II band positions. The peaks in PC2 may represent concentrations effects and/or subtle changes in secondary structure. Overall, batch G2031, 17 and final bulk are separated from the others by PC2, which may be associated with subtle changes in structure. In addition, batch G2032 is distinguished by the other batches by PC1, which is consistent with different aggregation on the ATR crystal.

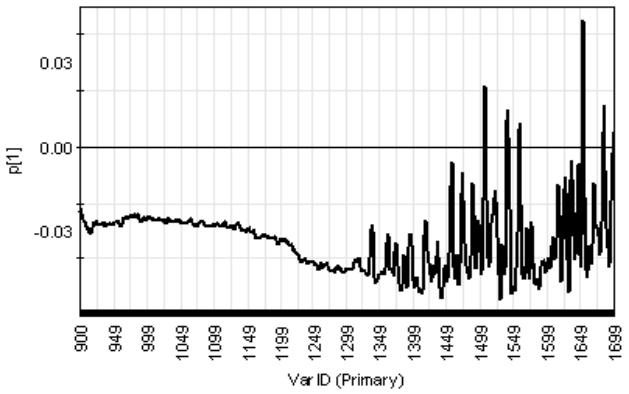

Figure 26 PCA loading plot p_1 of FT-IR spectra

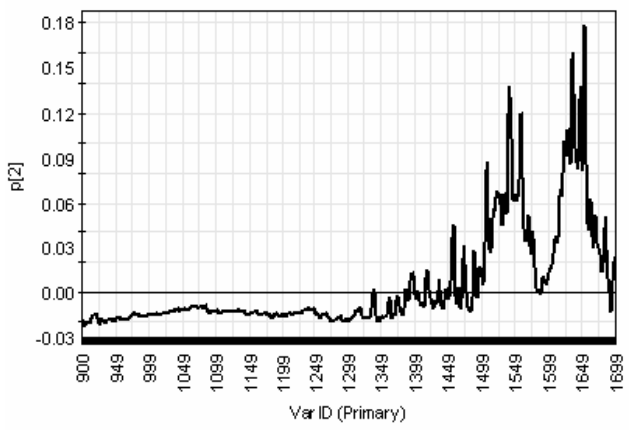

Figure 27 PCA loading plot p_2 of FT-IR spectra

2.1.5 UV-Vis spectroscopy analysis

UV-Vis measurements were performed under different conditions. At first, the analysis was performed using the concentrated samples (> 160 mg/ml). Then, these samples were diluted to about 0.5 mg/ml in water, generating two replicates of each sample. Then, one replicate was stored at room temperature, and the other at 4°C and analyzed separately.

UV-Vis measurements of concentrated samples. The UV-Vis spectra of concentrated samples are reported in figure 28. Spectra were normalized using a specific correction factor (see Materials and Methods section).

All the spectra are extremely reproducible in form, indicating high comparability of the different batches obtained with different processes. Analyzing more in detail the spectra, the only evident difference is a slight raising of the baseline for batch G2030, probably due to scattering effects,

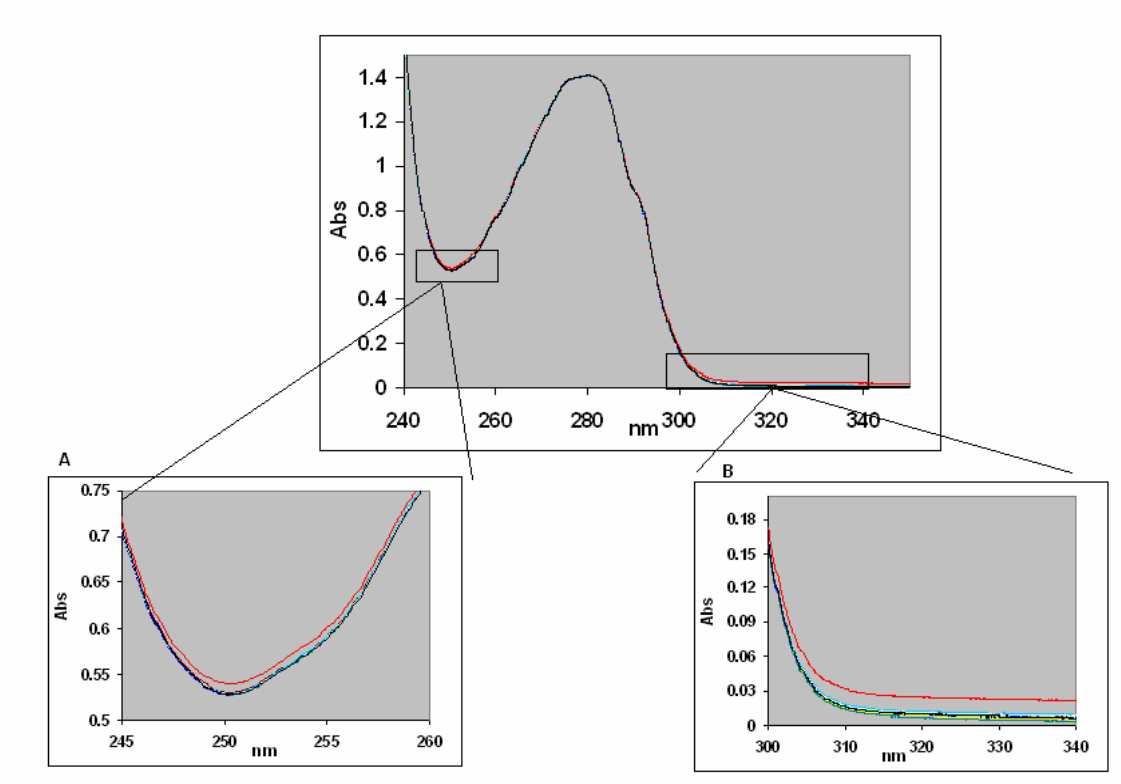

Figure 28. UV-Vis spectra of concentrated samples
A) Enlarged view of 240 –260 nm region; **B)** Enlarged view of 300-340 nm region

UV-Vis measurements of diluted samples (room temperature). The UV-Vis spectra of diluted samples (fig. 29) stored at room temperature shows some differences with respect to the spectra obtained by using the concentrated analytes. It is evident in fact how a number of spectra show the appearance of a little shoulder at ca. 275 nm, while some others show a completely different shape in the region from 270 to 240 nm, consistent with sample deterioration.

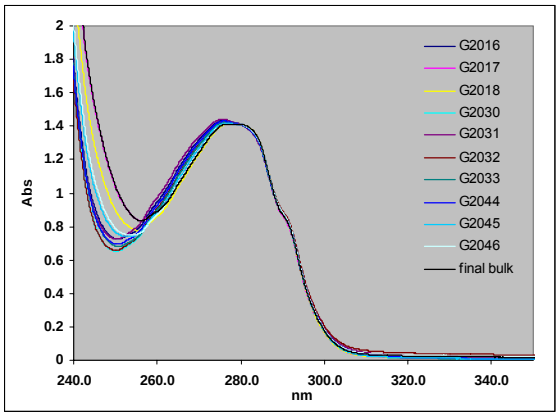

Figure 29. UV-Vis spectra of dilute samples (room temperature)

The reasons of these changes could be, on one hand, the storage conditions, that may facilitate degradation processes such as aggregation, and, on the other hand, the dilution in water, since replacing the original buffer solution could affect the structure of the drug substance in solution.

UV-Vis measurements of diluted samples (4°C). UV-Vis spectra of diluted samples stored at 4°C are shown in figure 30. It's immediately clear that spectra are much more comparable with respect to the spectra of the same samples stored at room temperature.

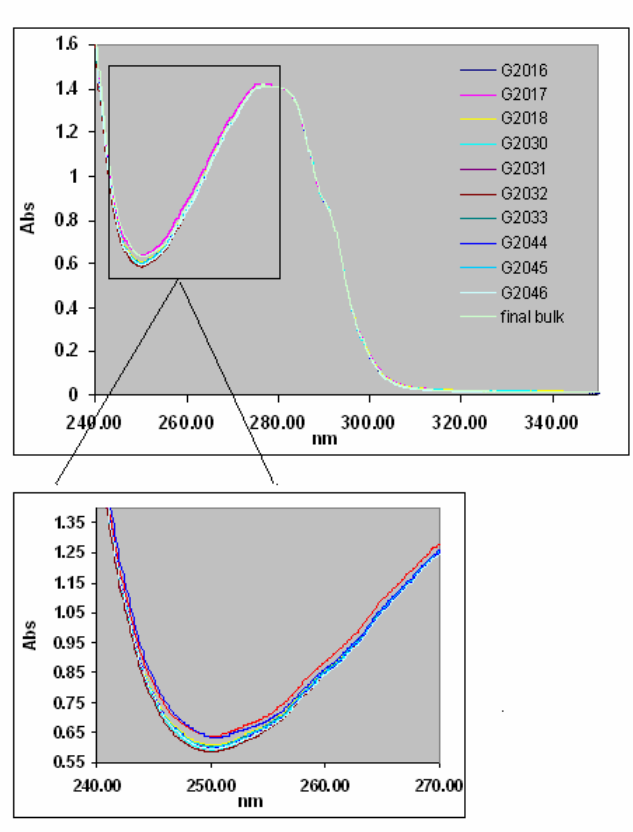

Figure 30. UV-Vis spectra of diluted samples (4°C)

However, the spectra are not identical and, in particular, the final bulk and G2017 batches show a raised absorbance from 240 to 270 nm, where the two spectra cross each other, in a manner often associated with scattering of particulates.

2.1.6 PCA analysis of UV-Vis spectra

PCA of concentrated samples. PCA analysis of spectra obtained by analyzing the concentrated samples (fig. 31) confirms the difference already seen by visual inspection of the spectra (fig. 28).

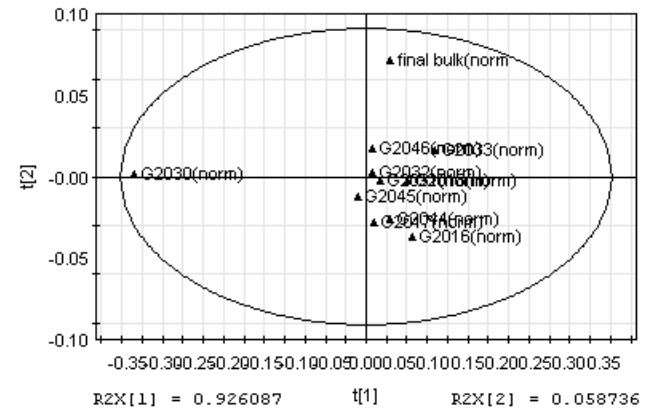

Figure 31. PCA score plot of UV-Vis spectra of concentrated samples

From the score plot it is evident that data points arrange themselves in a large central cluster, excepting final bulk marginally distinguished by PC2 component, and G2030 which is an outlier from the central cluster with respect to the PC1. This pattern confirms the difference seen between G2030 and the other spectra. To further investigate the reasons of the different pattern of G2030 and final bulk data points detected in the score plot, once again the loading plots must be analyzed(fig. 32).

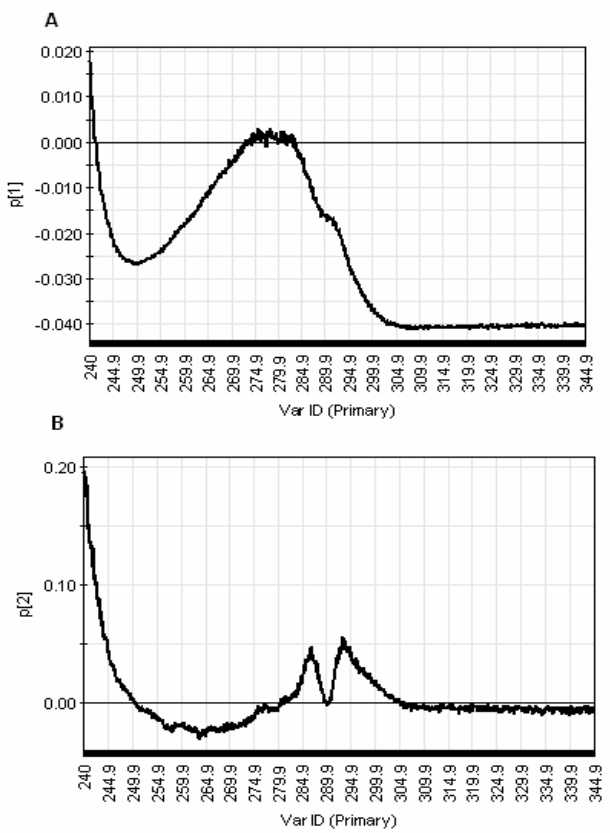

Figure 32. PCA loading plots of UV-Vis spectra of concentrated samples
A) Loading plot of p1 component; **B)** Loading plot of p2 component

From the inspection of loading plot p_1 (fig. 32, A) it is evident that PC1 represents a raised baseline at long wavelength, consistent with the reason why batch G2030 is an outlier. The loading plot p_2 (fig. 32, B) shows that PC2, which distinguishes final bulk, has a form consistent with a change in the relative absorbance of tryptophan with respect to phenylalanine and tyrosine residues.

PCA of diluted samples (room temperature). As expected, data points of UV-Vis diluted spectra do not form a single cluster in the score plot (fig. 33), but distribute themselves with a characteristic pattern reflecting the spectral difference seen in figure 30.

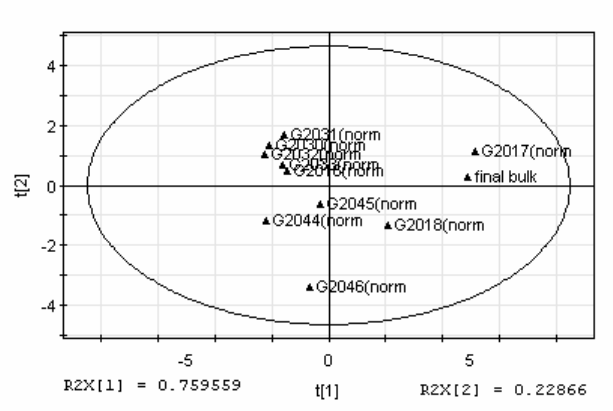

Figure 33. PCA score plot of UV-Vis spectra of diluted samples (room temperature)

Batches G2017 and final bulk are clear outliers with respect to PC1. Batches G2046, G2018, G2044 and G2045 are delineated by PC2. PC1 and PC2 account for 97% of the total variance. Such complicated pattern can be understood only analyzing the respective loading plots of figure 36.

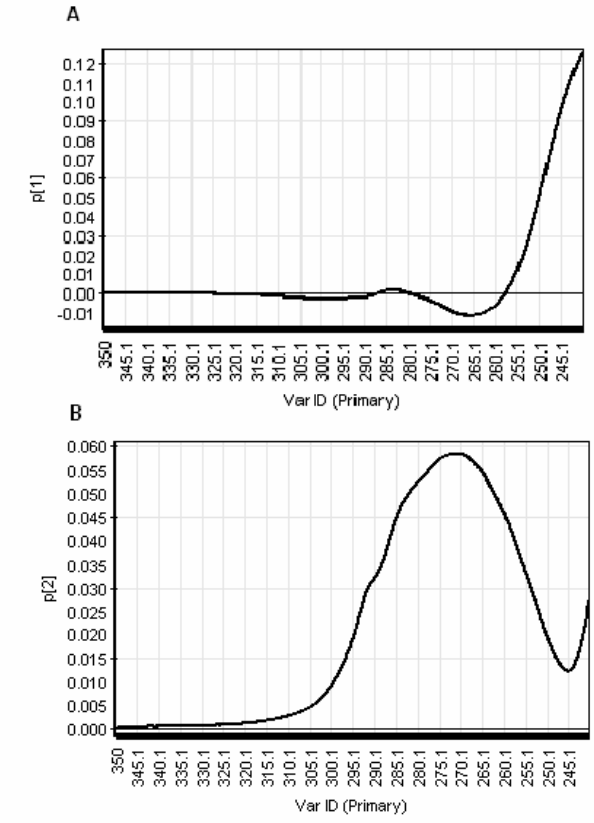

Figure 34. PCA loading plots of UV-Vis spectra of diluted samples (room temperature)

Inspecting loading plot p_1 (fig. 34, A) it is evident that PC1 represents a raised absorbance in the region from 240 to 270 nm, perhaps corresponding to particulate scattering, together with a loss of absorbance at 265 nm.

Loading plot p_2 (fig. 34, B) highlights instead that PC2 looks like a mean UV-Vis spectrum of the drug substance and may reflect a loss of concentration although the correspondence is not perfect and PC2 could also include a perturbation of the aromatic residues such as tyrosine and/or phenylalanine.

Concluding, it is clear from these analysis that the spectra do not completely correspond to each other and that deterioration of the solutions has occurred.

PCA of diluted samples (4°C). The differences observed in the dilute samples stored at room temperature are present at very lower extent in the UV-Vis spectra of the 4°C stored diluted samples.

The data points in the score plot (fig. 35) are in close proximity to each other, although not as close as in the concentrated case, excepting final bulk and G2017 which are completely outliers from this cluster with respect to PC1. These two outliers are also separated from each other by PC2. These two principal components explain 98% of the total variation.

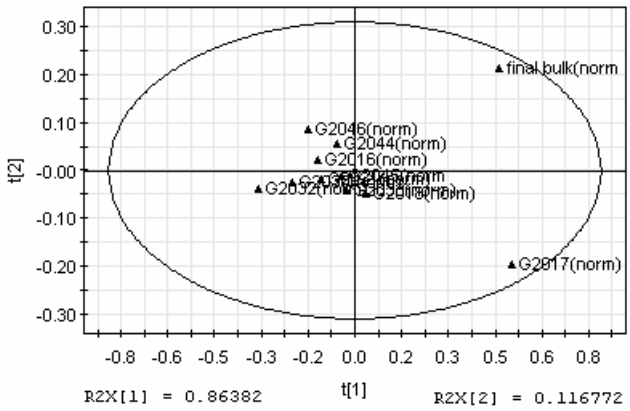

Figure 35. PCA score plot of UV-Vis spectra of diluted samples (4°C)

Loading plot p_1 (fig. 36, A) indicates that PC1 mainly represents a raised absorbance at short wavelengths, with two observable bands at 242 and 264 nm.

Loading plot p_2 (fig. 36, B) in contrast represents a raised absorbance at 264 nm, but a loss at short wavelength below 250 nm. Notably, phenylalanine has a major absorption band around the 264 nm region, and thus this amino acid residue could be involved in the outlying positions of G2017 and final bulk, although the differences are small.

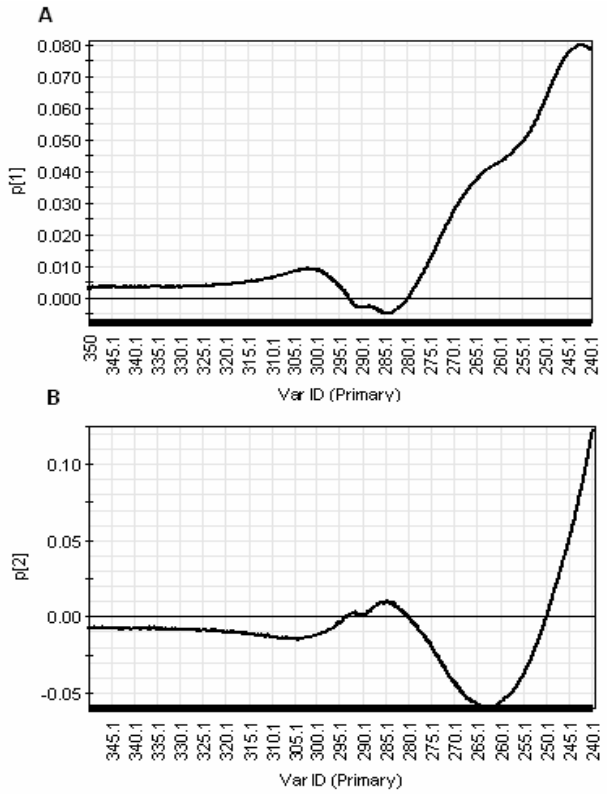

Figure 36. PCA loading plots of UV-Vis spectra of diluted samples (4°C)

2.1.7 UV-Vis-NIR spectroscopy

UV-Vis-NIR measurements were performed under different conditions. At first, the analysis was performed using the concentrated samples (> 160 mg/ml). Then, these samples were diluted to about 0.5 mg/ml in water.

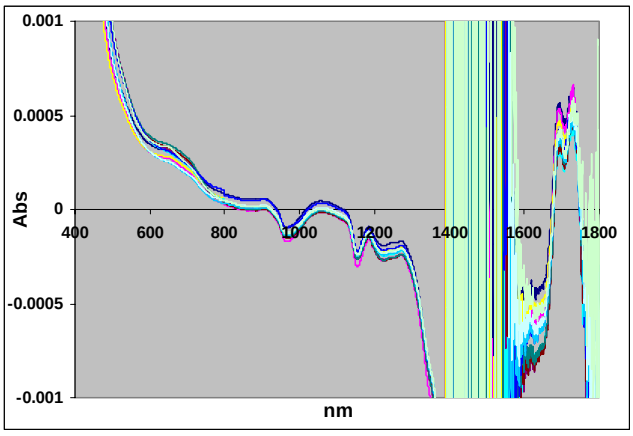

Figure 37. Vis region of normalized UV-Vis-NIR spectra of concentrated samples

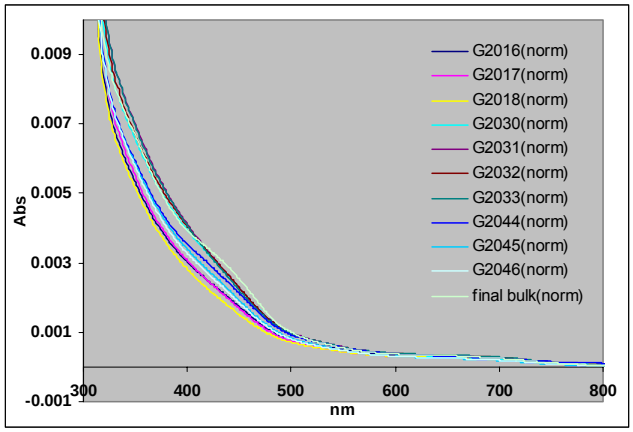

Figure 38. NIR region of normalized UV-Vis-NIR spectra of concentrated samples

UV-Vis-NIR of concentrated samples. UV-Vis-NIR spectra of concentrated samples are shown in figures 37 and 38. These spectra were normalized using a specific correction factor (see Materials and Methods section).

The expected effect of excessive absorbance of water in the 1400-1800 nm region is evident, with the spectra being dominated by noise in this region; for this reason the 1400-1800 nm region has been disregarded in the analysis. First, it is apparent that all the spectra show absorption bands in the 400-800 nm region (fig. 37), although of varying magnitude. These bands are not expected for a pure native (glyco)protein solution and may be related to the "yellow-protein" phenomenon often observed for biotechnology products. Spectroscopic features are also observed in the 800-400 nm NIR region (fig. 38). The origin of these features is debated and under further investigation, but could be related to solvent-protein effects (which would be consistent with the apparent "negative absorption" characteristics relative to the water baseline).

UV-Vis-NIR of diluted samples. The UV-Vis-NIR spectra of diluted samples are shown in figure 39. These spectra were normalized using a specific correction factor (see Materials and Methods section).

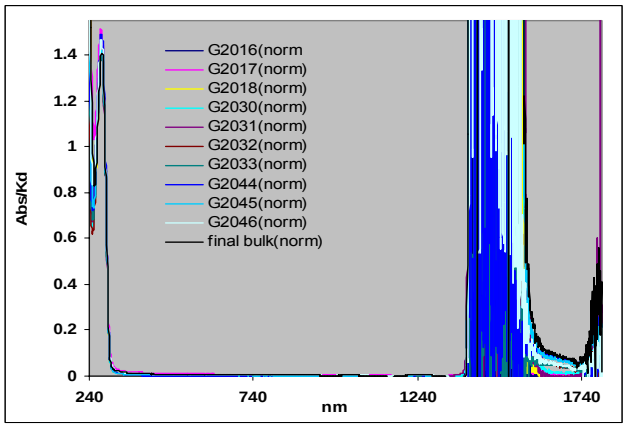

Figure 39. UV-Vis-NIR normalised spectra of diluted samples

As for concentrated samples, the expected effect of excessive absorbance by water in the 1400-1800 nm region is evident, with the spectra being dominated by noise in this region; for this reason the 1400-1800 nm region has been disregarded in the analysis. Figures 40, 41 and 42 show a close-up of the UV, visible and NIR region respectively.

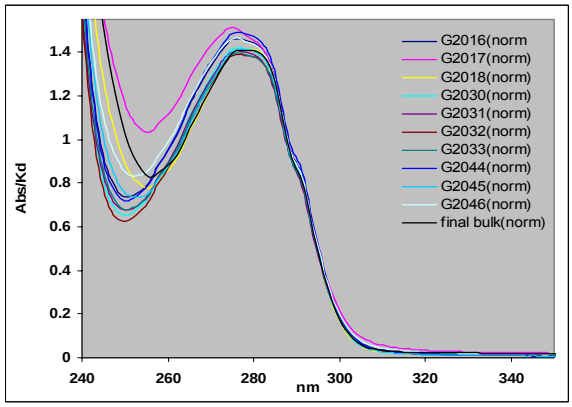

Figure 40. NUV region of UV-Vis-NIR spectra of diluted samples

It is evident that all the batches show a diversity of peak heights and shape in the UV region (fig .40), as might be expected from the UV-Vis spectra discussed above. In addition, all batches show raised tail of absorption extending across the 300-800 nm region (fig. 41). Some of the batches show a series of small peaks in the NIR region (800-1400 nm)(fig. 42). However the molecular origin of these features cannot be established and they will not be considered further .

The clear absorption bands seen in the visible region for the concentrated solution are not evident for the diluted solutions, due to ca. 300-fold reduced concentration, which brings them below the detection level. However, the absorption tails observed in the visible region may be related to the bands observed in the concentrated solutions..

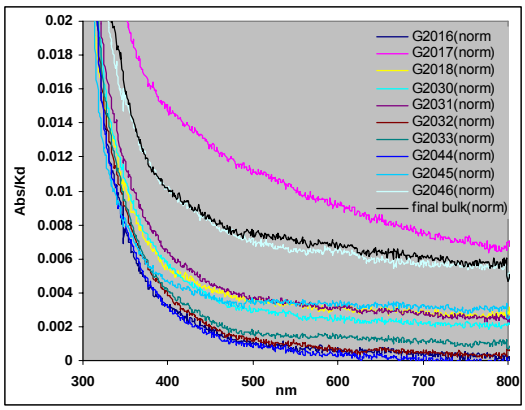

Figure 41. Visible region of UV-Vis-NIR spectra of diluted samples

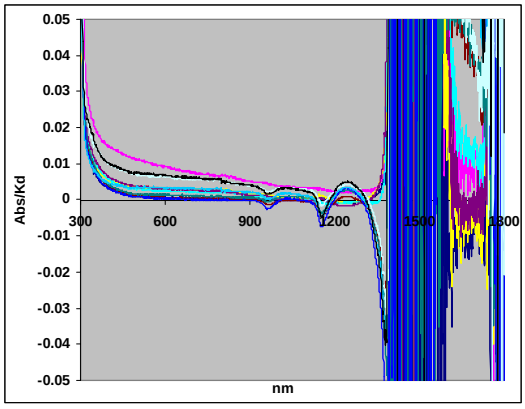

Figure 42. NIR region of UV-Vis-NIR spectra of diluted samples.

PCA analysis of UV-Vis-NIR spectra of concentrated samples. In the score plot of figure 43, spectra are predominantly separated by PC1, with only minor separation by PC2. Only batch of A87G2017 is separated from the group by PC2. It is apparent that the first two PC's are sufficient to explain 98% of the total variation in the spectra.

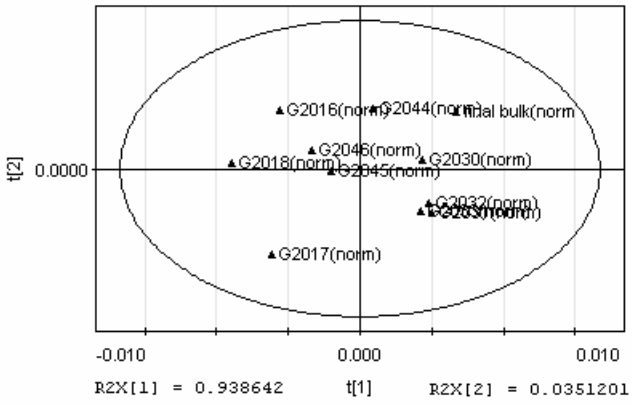

Figure 43. PCA score plot of UV-Vis-NIR spectra of concentrated samples

Inspecting the loading plot p_1 (fig. 44, A) it is evident that PC1 mainly represents absorption bands around 400-450 nm, and a tail extending across the visible region, consistent with the spectral features commented above regarding the "yellow-protein" phenomenon.

Loading plot p2 (fig. 44, B) indicates instead that PC2 represents a further increase in absorbance at 450 nm together with a decrease at 400 nm and shorter wavelengths. Given that the batches A87G2016-18 and A87G2030-33 are delineated by PC1 (opposite extremes) and this is, in turn, associated with the appearance of additional bands at 400-450 nm and a tail extending across the visible region, it can be hypothesised that these spectral features could be related to the period in which the batch was produced and/or the change in hydrolysates used.

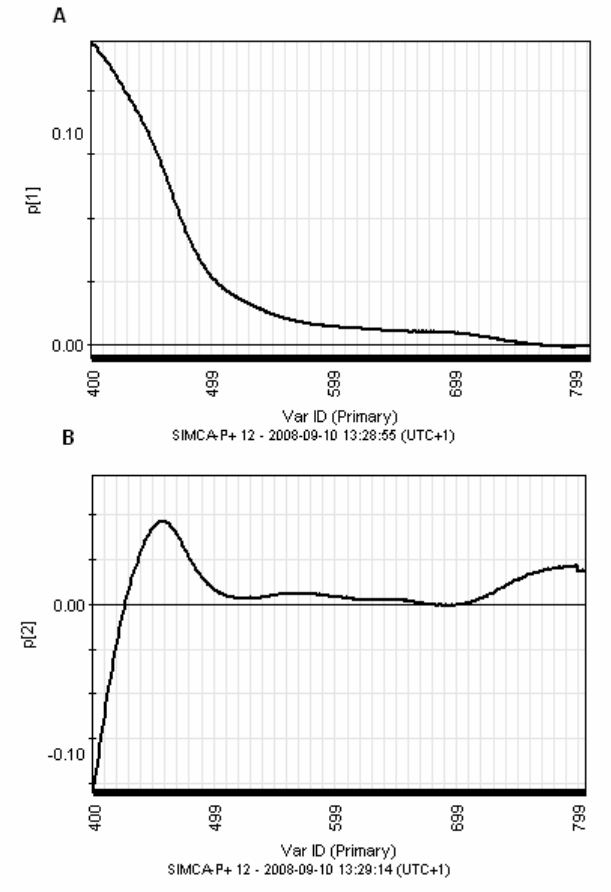

Figure 44. PCA loading plots of UV-Vis-NIR spectra of concentrated samples

PCA analysis of UV-Vis-NIR spectra diluted samples. The spectra of diluted samples clearly contain information of two scales, a large absorption around 280 nm, and a small absorption elsewhere. Thus, the two spectral regions were considered separately. First, the 340-800 nm region was subject to PCA analysis (fig. 44). The spectra are delineated mainly by PC1, with batches A87G2017 and A87G2044 at either extremes, and with minor separation by PC2, with A87G2045 more detached from the other batches.

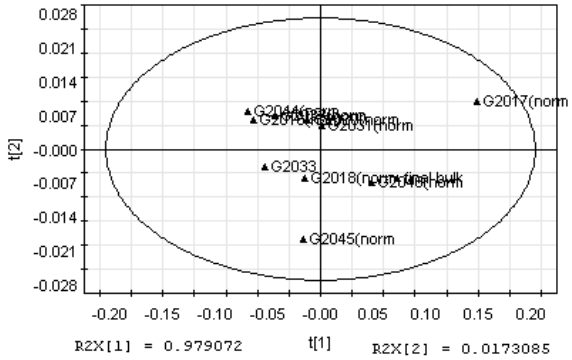

Figure 45. PCA score plot of Vis region of UV-Vis-NIR spectra of diluted samples

The first PC is sufficient alone to explains 97,9% of the total variation. Inspection of the loading plot p_1 (fig. 45) indicates that PC1 represents a gradual rising absorbance with decreasing wavelengths, similar to what seen for Rayleigh scattering. Batch A87G2017 displays the greatest contribution of this spectral feature while batch A87G2044 displays the lowest contribution, as can be also observed in the spectra (fig. 41).

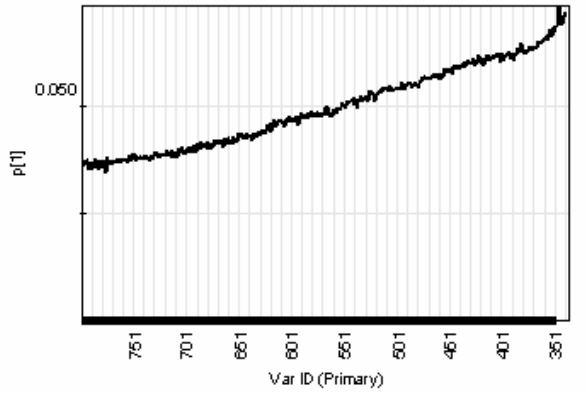

Figure 46 Vis region PCA loading plot p1 of UV-Vis-NIR spectra of diluted samples

The near-UV region of 240-340 nm (fig. 40) was also subjected to PCA. This region has already been covered by the specific NUV absorption spectra reported above. However, these spectra were acquired at different time and NUV region from UV-Vis-NIR spectra was treated separately.

The score plot of this region is illustrated in figure 47. Batches A87G2017 and G2033 are clear outliers to the others batches with respect toPC1, although batches A87G2018 and final bulk are also delineated by PC2.

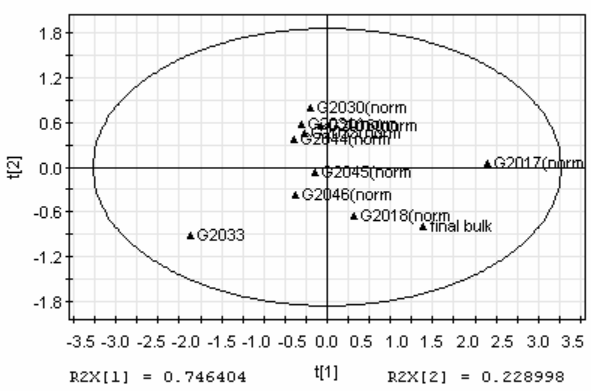

Figure 47. NUV PCA score plot of UV-Vis-NIR spectra of diluted samples

Almost 98% of the total variance is explained by the first two PCs. The loading plot p_1 (fig. 48) shows that PC1 represents an increased absorbance at short wavelength, perhaps corresponding to Rayleigh scattering. The loading plot p_2 (fig. 49) indicates instead that PC2 is a broad band centred around 270 nm, akin to an aromatic residue absorption, especially tyrosine and phenylalanine. This is very similar, but not identical, to that deduced for NUV absorption spectra described above. In this case PC2 is less associated with the mean spectrum of drug substance (i.e. concentration) and has more a character of specific classes of aromatic residues. This may be a consequence of all the UV-Vis-NIR spectra being recorded close in time to each other, whereas the NUV absorption spectra were run over a slightly more extended period of few days, allowing concentration variations through aggregation to become apparent in the samples.

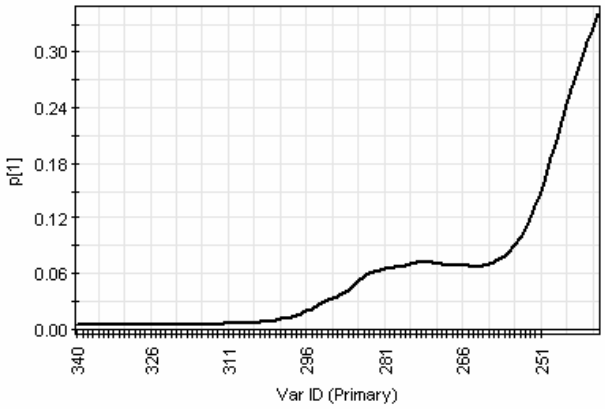

Figure 48. NUV PCA loading plot p_1 of UV-Vis-NIR spectra of diluted samples

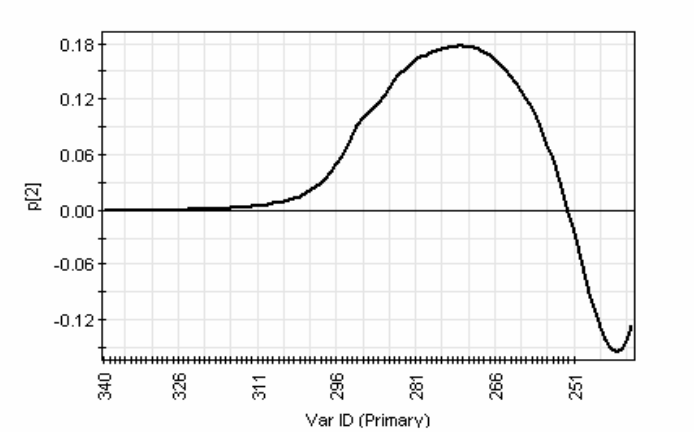

Figure 49. NUV PCA loading plot p_2 of UV-Vis-NIR spectra of diluted samples

2.1.8 Raman scattering

Solution Raman scattering of concentrated solutions (fig. 50) proved to be of a very poor quality, predominantly due to massive fluorescence artefacts. The spectra described below were obtained from constant volume aliquots of the solutions that were freshly allowed to dry onto glass microscope slides, and using the concentration of each sample as normalization factor.

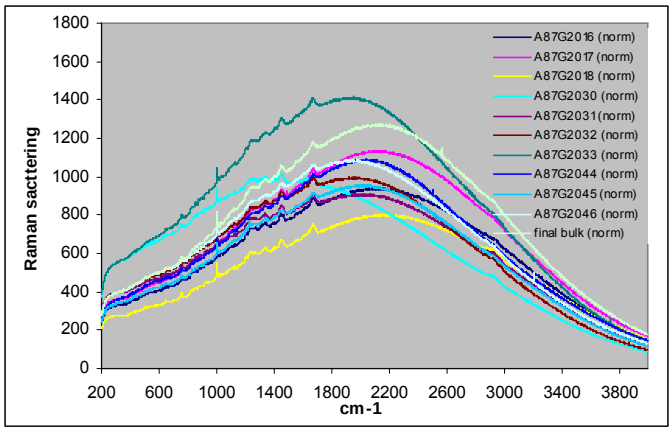

Figure 50. Raman scattering of concentrated samples

It can be immediately seen that the Raman scattering is dominate by fluorescence and that the fluorescence varies from batch to batch, both in intensity and wave number. The Raman spectrometer employed a 632.8 nm incident laser, an intense light source capable of exciting any absorbing species with absorption bands in the 600 nm area. Notably, the fluorescence doesn't occur at a constant wavenumber, but varies across the batches. Based on the position of fluorescence maxima, batches can be divided into 5 groups as summarized in table 3.

Fluorescence Maximum (cm^{-1})	Batches
1400	A87G2030
1950	A87G2032 A87G2033 A87G2046
2000	A87G2031 A87G2044 A87G2045
2250	A87G2016 A87G2017 final bulk
2300	A87G2018

Table 3 Fluorescence maxima in Raman spectra

2.1.9 PCA of Raman scattering spectra

The score plot of Raman scattering normalised spectra is presented in figure 51. The spectra are predominantly delineated by PC1 and PC2,explaining together 98% of total variance in the spectra.

In particular, batch A87G2033 is an extreme with respect to PC1 and A87G2030 with respect to PC2. Notably, batches A87GG2016, 17 , 18 and final bulk are discriminated from the other batches by PC2. So, the batches broadly cluster as shown in table 6.

The loading plots shows that principal components are, as might be expected, dominated by fluorescence, with PC1 representing the broad fluorescence centred at 1950 cm^{-1} (fig. 52, A), and PC2 introducing a shift of this peak to higher (with negative scores) or lower (positive scores) wavenumber (fig. 52, B).

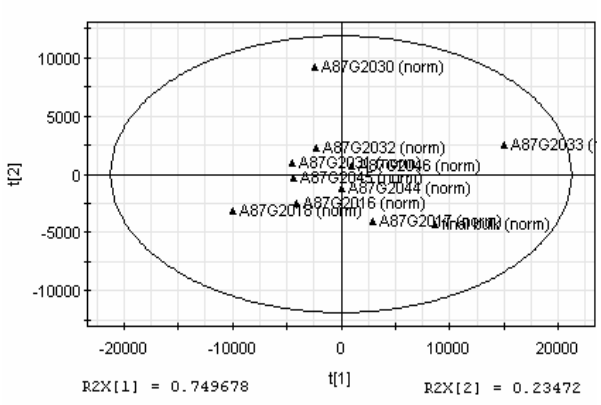

Figure 51. PCA score plot of Raman spectra

Batch A87G2033 is therefore seen as an outlier due to its higher fluorescence around 1950 cm^{-1}, with batches A87G2016, 17, 18 and final bulk being separated by having a fluorescence peak at higher wavenumber. Batch A87G2030 is unique to have a fluorescence band at a particularly low wavenumber (~ 2250 cm^{-1}). However, in each case it can be seen that the fluorescence seems to be the sum of two or more bands

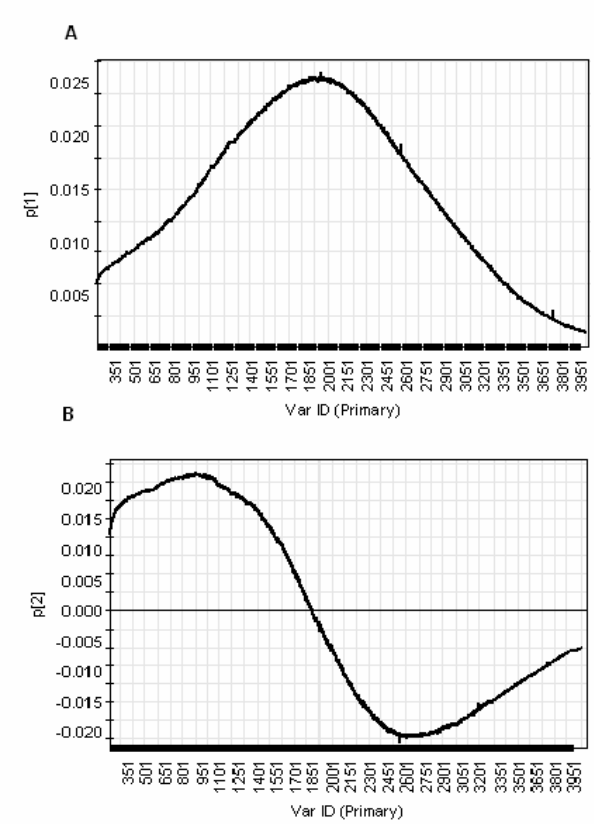

Figure 52. PCA loading plots of Raman spectra
A) loading plot p_1; **B)** loading plot p_2

2.1.10 Fluorescence mapping and emission

Fluorescence mapping of concentrated samples. An example of the series of emission scans making up an excitation-emission map are presented in figure 53. Each emission spectrum corresponds to a different excitation wavelength, progressing from 260 to 595 nm as the emission peaks progress to longer wavelength.

Due to the excessive absorbance of the solutions, and hence the fluorescence self-masking effect, there is no signal detected when excited at wavelength shorter than 300nm. Consequently, the normal tryptophan fluorescence at 300-350 nm that would typically be observed with excitation in the 270-300 nm region is unobservable.

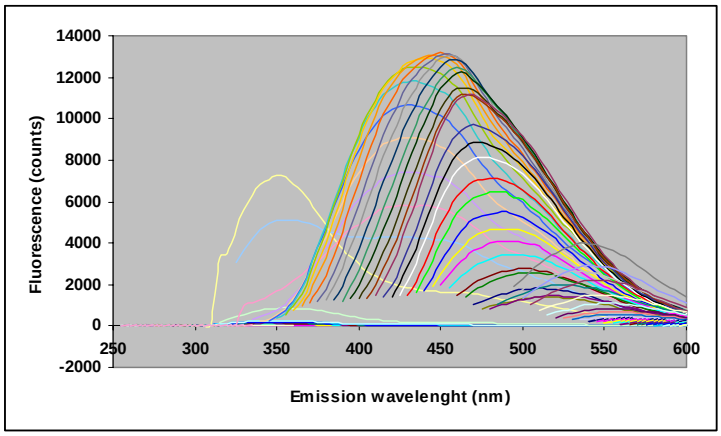

Figure 53. Excitation-emission map of drug substance

However, very distinctive fluorescence bands in the visible region 355-600 nm are evident, arising from excitation with light of wavelengths longer than 300 nm. These visible region fluorescence bands occur for all the samples and can be separated into at least four separate excitation-emission pairs, which partially overlaps to each other. The corresponding normalised emission spectra for all the batches are illustrated in figure 54 (Ex 330, 360, 400 and 480 nm).

Such visible fluorescence is not an expected feature of pure, native glycoproteins. Fluorescence bands in these region are often related to the "yellow protein" phenomenon often observed for biotechnology products. Moreover, they may be related to the visible absorption bands described in the UV-Vis-NIR absorption section above.

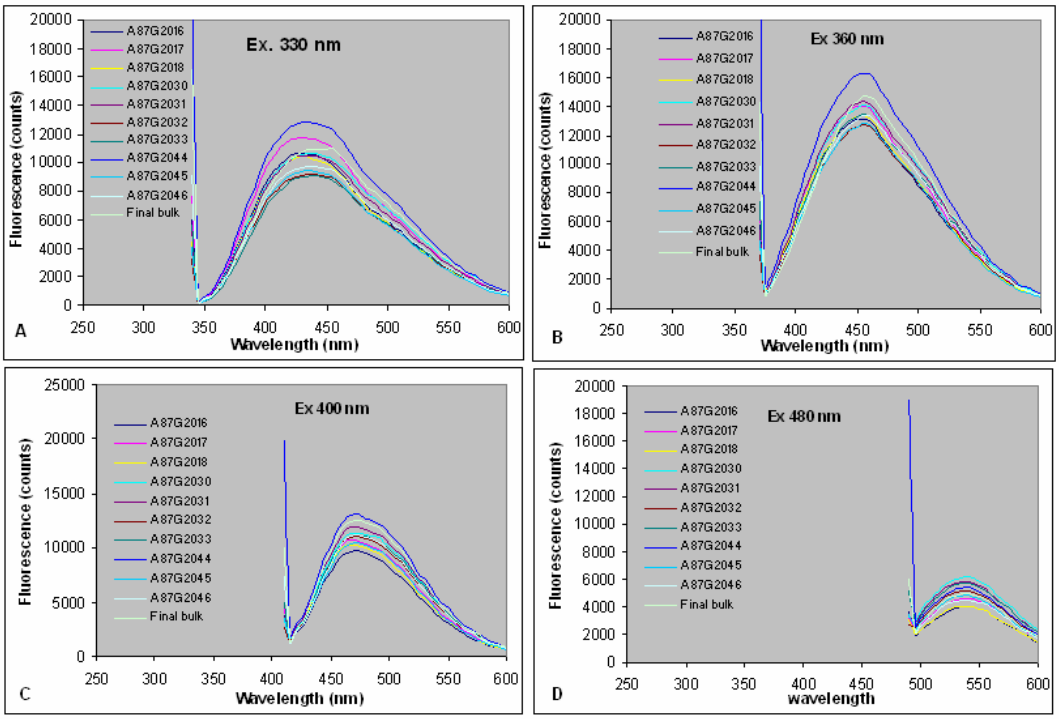

Figure 54. Fluorescence emission spectra at different excitation

A methods of viewing fluorescence excitation-emission maps is a colour coded contour map over the excitation-emission domain. Such maps for each batch, based on the normalised spectra are shown from figures 55 to 65. The localisation of the fluorescence in the excitation-emission wavelength domain is evident in the maps of all samples and correspond to a broad fluorescence across the visible region with excitation in the 330-480 nm region.

It should be borne in mind that the fluorimeter light source (450 W Xenon arc lamp) possesses some sharp modulations (spikes) in the intensity of its radiation at some wavelengths in the 450-500 nm region. The increased excitation at these wavelengths will give rise to apparent sharp increases in fluorescence superimposed on the broad fluorescence. So, the sharp, "fine structure" to the fluorescence with varying excitation wavelength in the 450-500 nm range should be ascribed to lamp intensity effects rather than real molecular excitations.

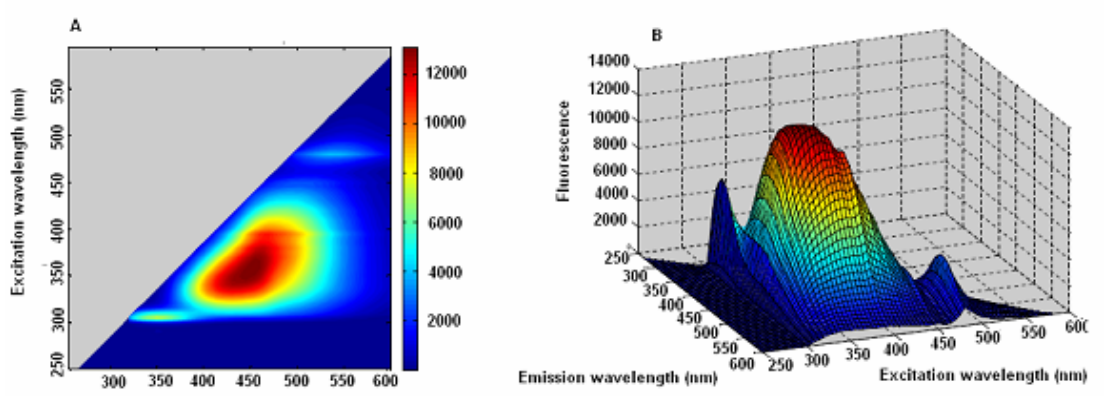

Figure 55. Fluorescence excitation-emission maps of A87G2016
A) Excitation emission contour map; **B)** excitation-emission surface map

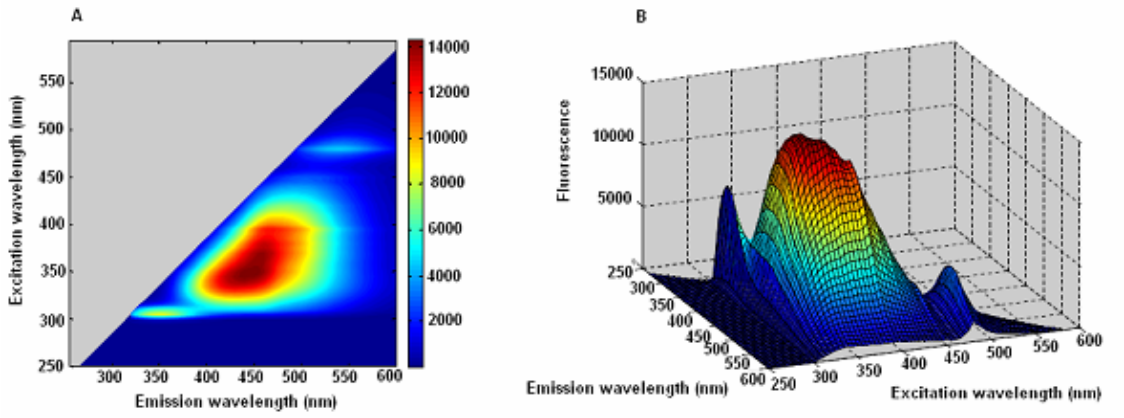

Figure 56. Fluorescence excitation-emission maps of A87G2017
A) Excitation emission contour map; **B)** excitation-emission surface map

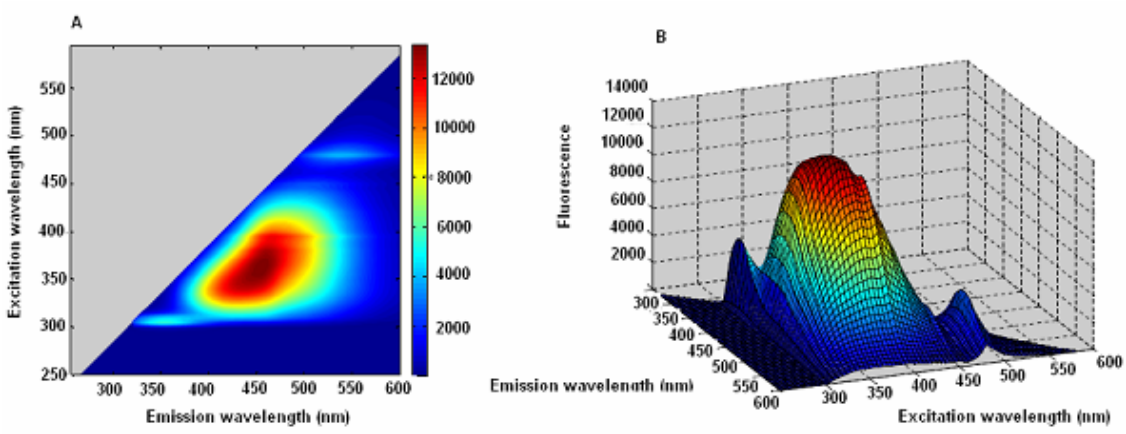

Figure 57. Fluorescence excitation-emission maps of A87G2018
A) Excitation emission contour map; **B)** excitation-emission surface map

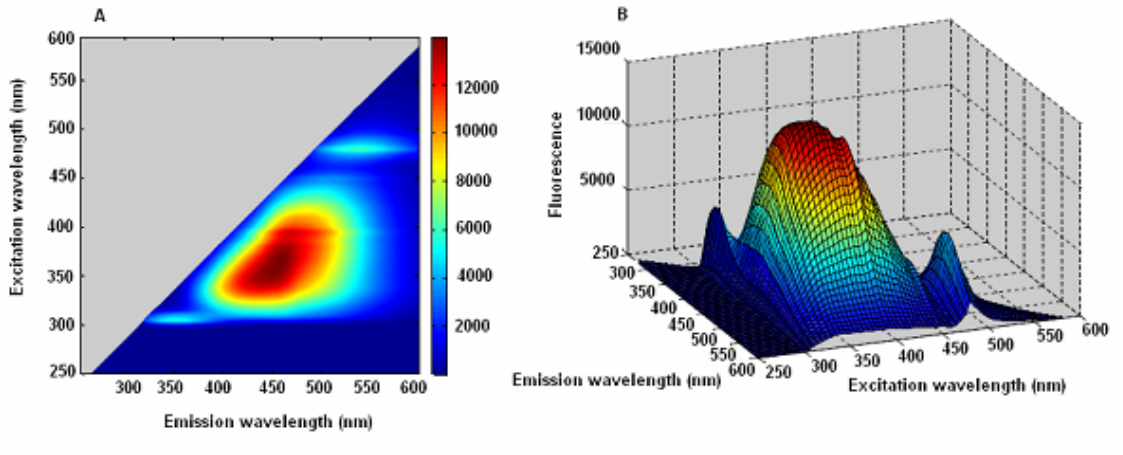

Figure 58. Fluorescence excitation-emission maps of A87G2030
A) Excitation emission contour map; **B)** excitation-emission surface map

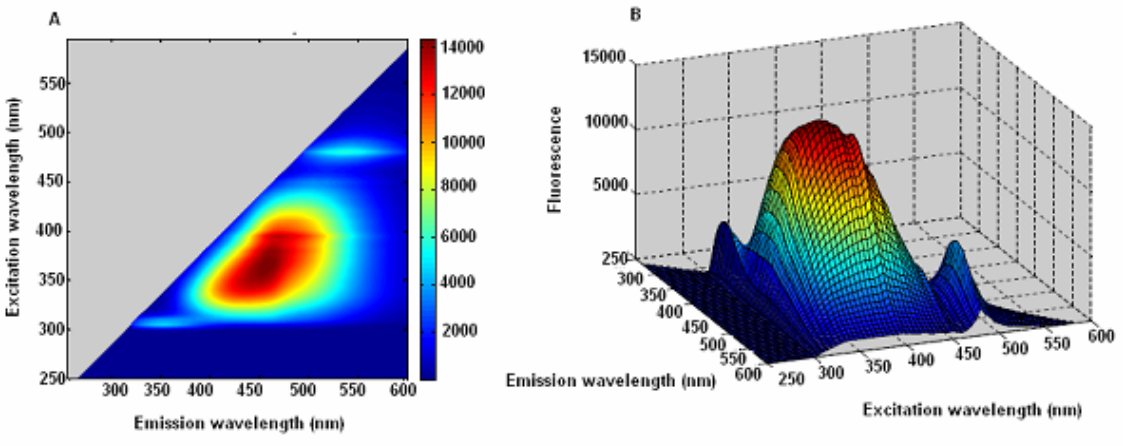

Figure 59. Fluorescence excitation-emission maps of A87G2031
A) Excitation emission contour map; **B)** excitation-emission surface map

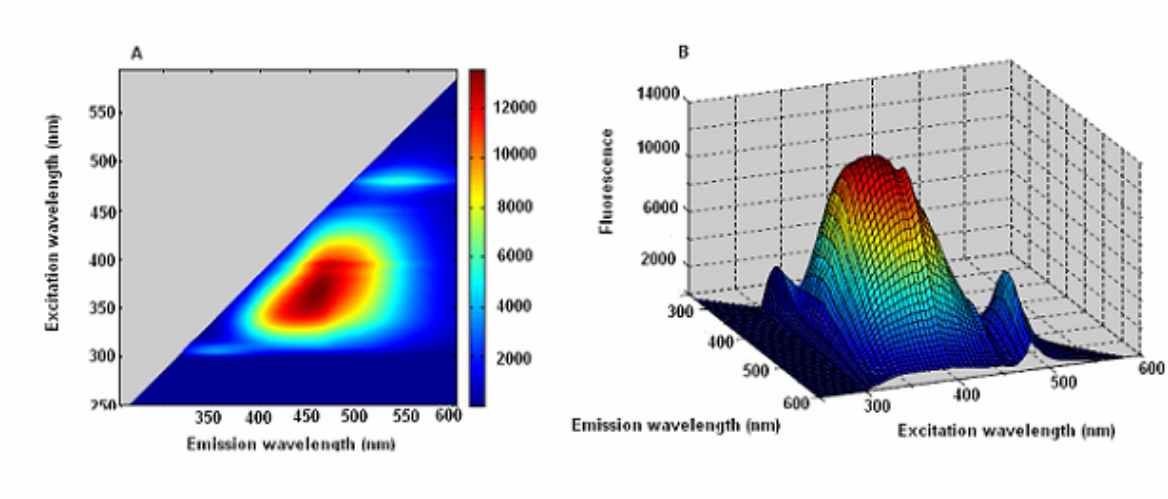

Figure 60. Fluorescence excitation-emission maps of A87G2032
A) Excitation emission contour map; **B)** excitation-emission surface map

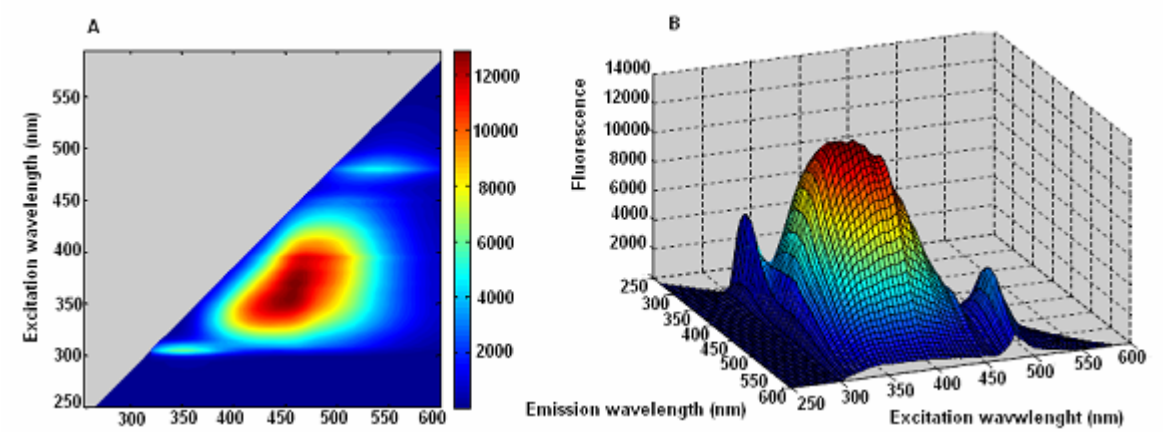

Figure 61. Fluorescence excitation-emission maps of A87G2033
A) Excitation emission contour map; **B)** excitation-emission surface map

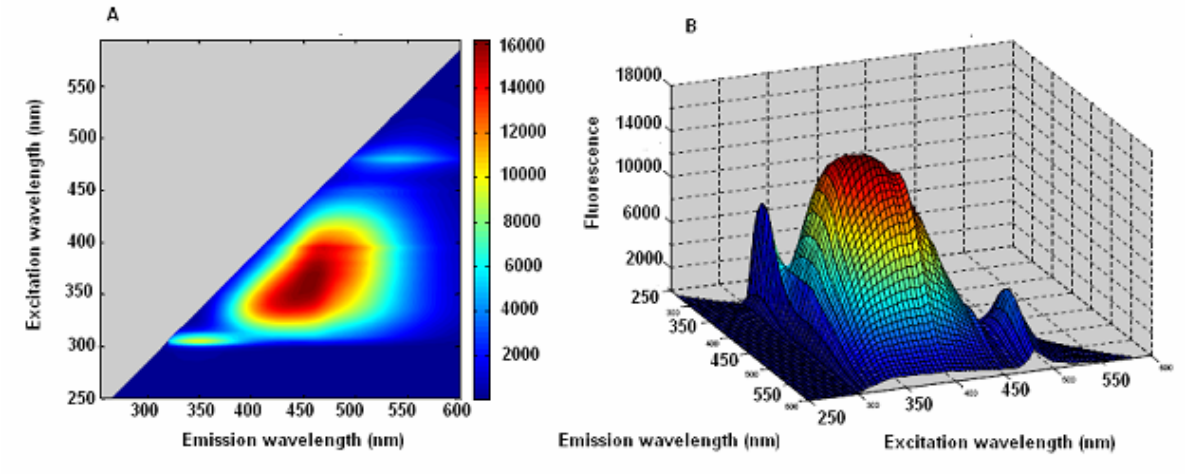

Figure 62. Fluorescence excitation-emission maps of A87G2044
A) Excitation emission contour map; **B)** excitation-emission surface map

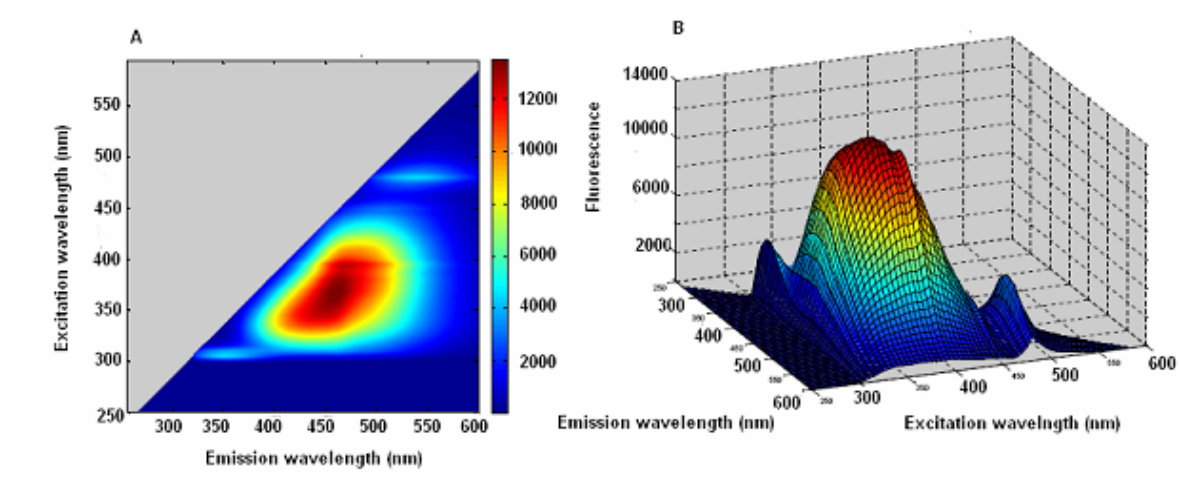

Figure 63. Fluorescence excitation-emission maps of A87G2045
A) Excitation emission contour map; **B)** excitation-emission surface map

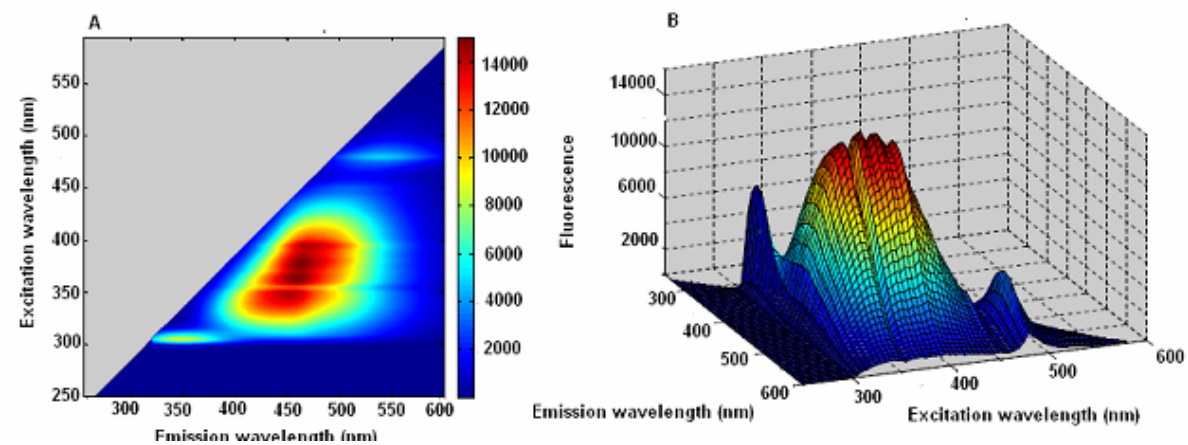

Figure 64. Fluorescence excitation-emission maps of A87G2046
A) Excitation emission contour map; **B)** excitation-emission surface map

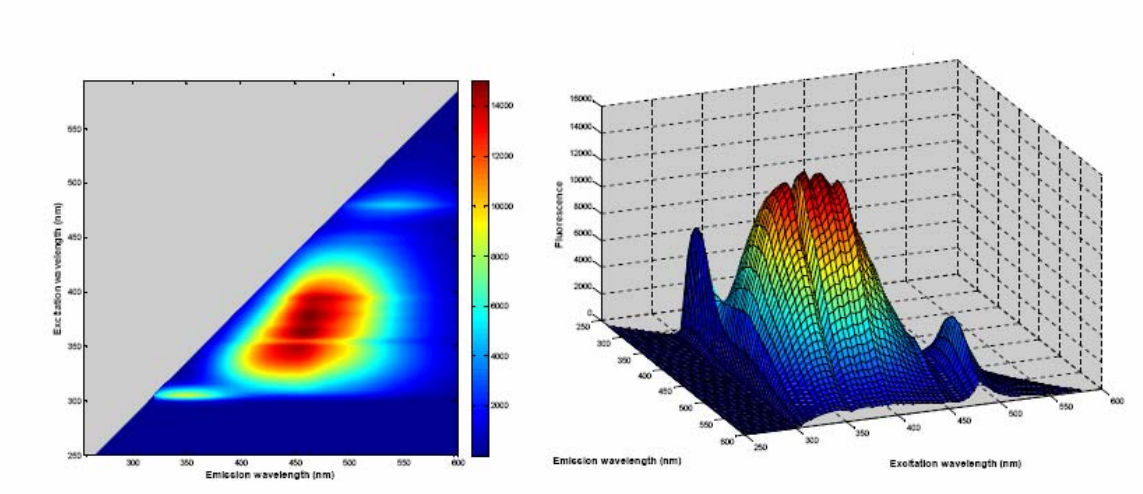

Figure 65. Fluorescence excitation-emission maps of final bulk
A) Excitation emission contour map; **B)** excitation-emission surface map

Fluorescence mapping of diluted samples. An example of the series of emission scans making up an excitation-emission map are presented in figure 66. Each emission spectrum corresponds to a different excitation wavelength, progressing from 260 to 595 nm as the emission peaks progress to longer wavelength.

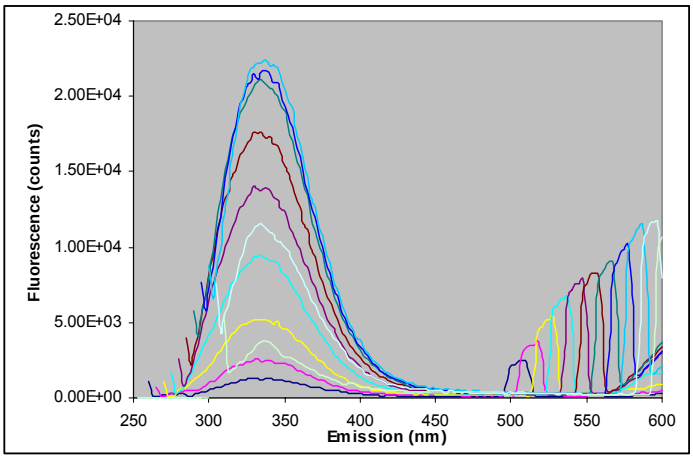

Figure 66. Fluorescence map of diluted samples

Unlike in the concentrated solution case, diluted samples do not have excessive absorbance in the NUV wavelength region, and thus the usual tryptophan absorption at 300- 350 nm is observable.

In addition, the effect of Rayleigh and Raman scattering of excitation light detected by the emission monochromator second-order dispersion are apparent at 500-600 nm. See the section "Fluorimetry Second-order scatter mapping" below for more details.

A methods of viewing fluorescence excitation-emission maps is a colour coded contour map over the excitation-emission domain. Such maps for each batch, based on the normalised spectra are shown in figure 67 to 72. The fluorescence maps of all the batches excepting final bulk have similar form, although there is some variation in fluorescence intensity. In contrast, final bulk exhibits a greater Rayleigh and Raman scattering band in the 500-600 nm region. The tryptophan fluorescence region (Ex: 250-300 nm, Em: 265-480 nm) and the Rayleigh-Raman scattering region (Ex: 250-300 nm, Em: 480-600 nm) are treated as separate maps for the purposes of pattern recognition.

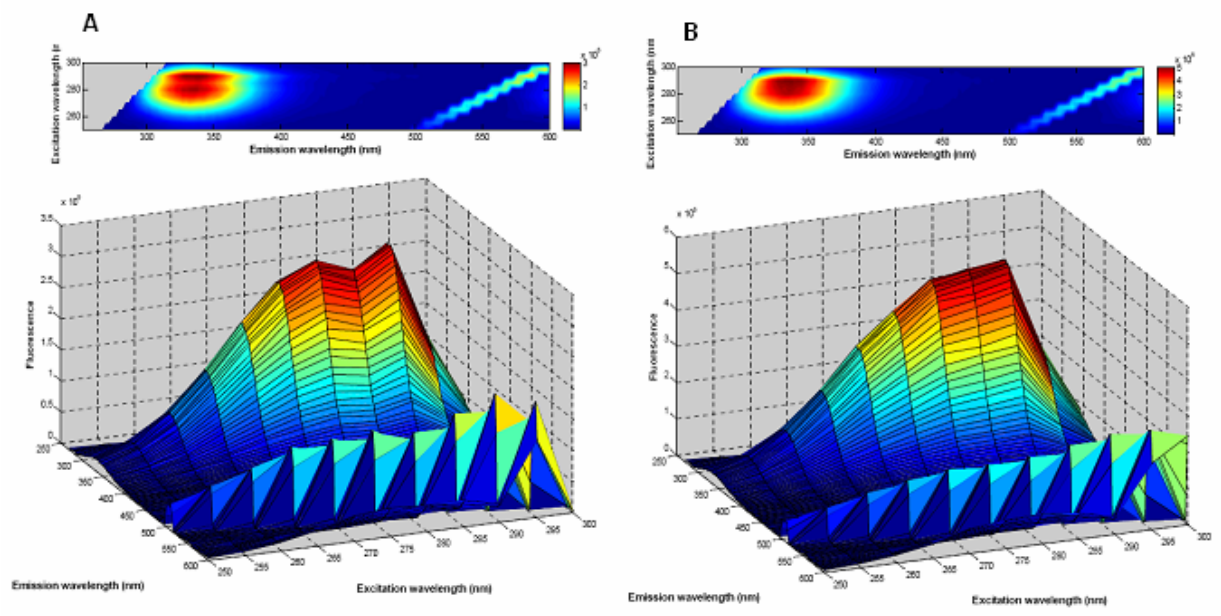

Figure 67. Fluorescence excitation-emission maps of A87G2016 and A87G2017
A) Excitation emission contour and surface map of A87G2016; **B** Excitation emission contour and surface map of A87G2017

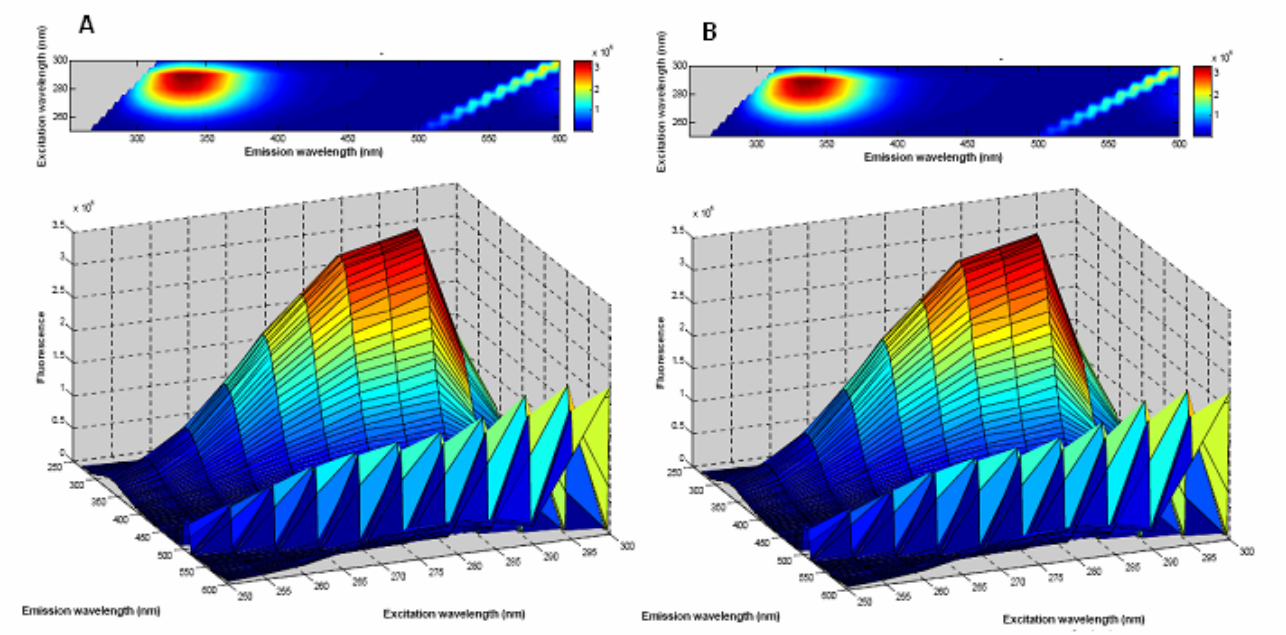

Figure 68. Fluorescence excitation-emission maps of A87G2018 and A87G2030
A) Excitation emission contour and surface map of A87G2018 **B)** Excitation emission contour and surface map of A87G2030

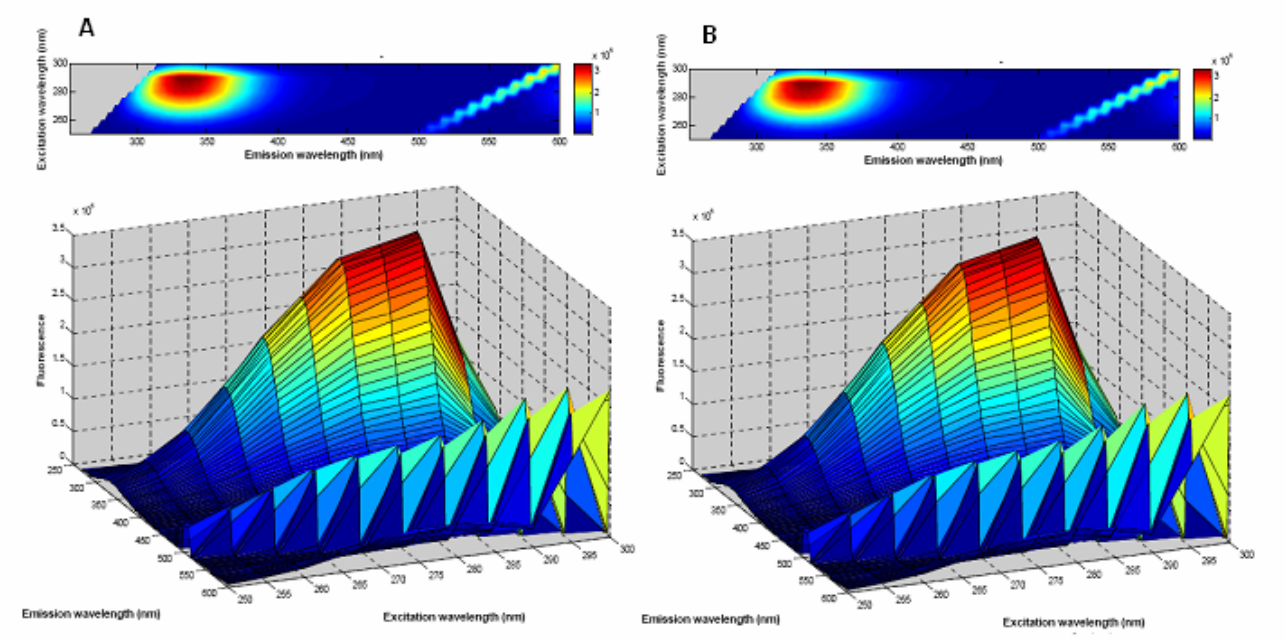

Figure 69. Fluorescence excitation-emission maps of A87G2031 and A87G2032
A) Excitation emission contour and surface map of A87G2031 **B)** Excitation emission contour and surface map of A87G2032

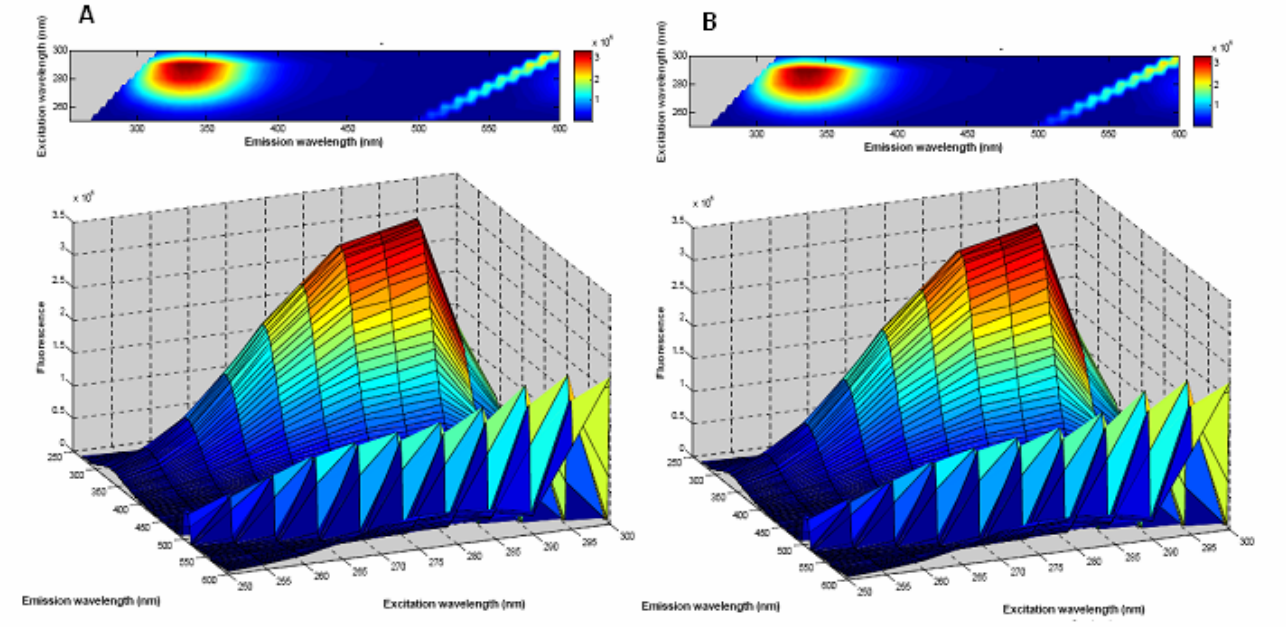

Figure 70. Fluorescence excitation-emission maps of A87G2033 and A87G2044
A) Excitation emission contour and surface map of A87G2033 **B)** Excitation emission contour and surface map of A87G2044

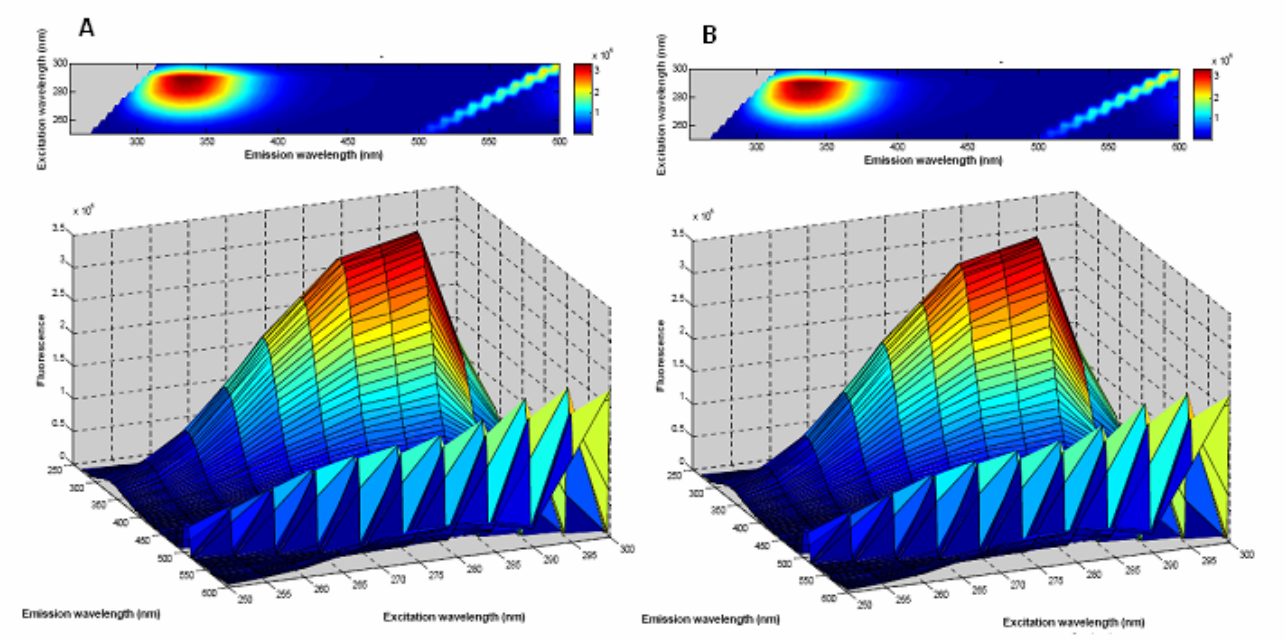

Figure 71. Fluorescence excitation-emission maps of A87G2045 and A87G2046
A) Excitation emission contour and surface map of A87G2045 **B)** Excitation emission contour and surface map of A87G2046

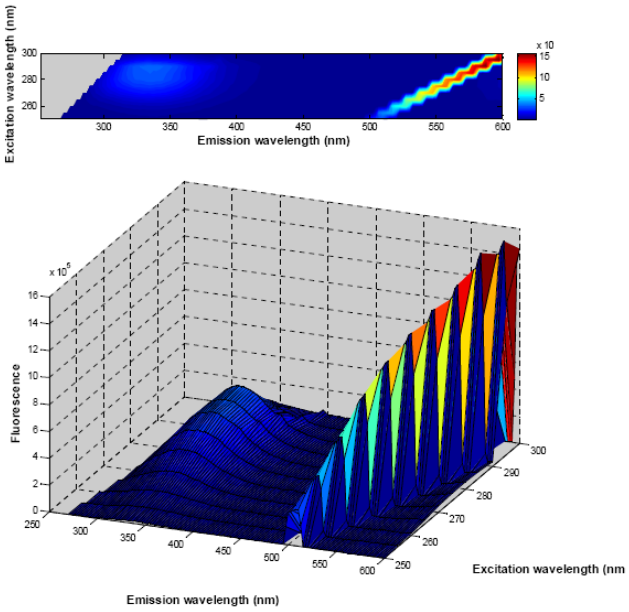

Figure 72. Fluorescence excitation-emission maps of final bulk

PCA of concentrated samples fluorescence maps. By concatenating successive emission scans a fluorescence excitation-emission map can be transformed to a format amenable to PCA and other multivariate analysis. The PCA score plots of the normalised fluorescence maps in the excitation: emission 305:460 315:600 range are illustrated in figure 73.

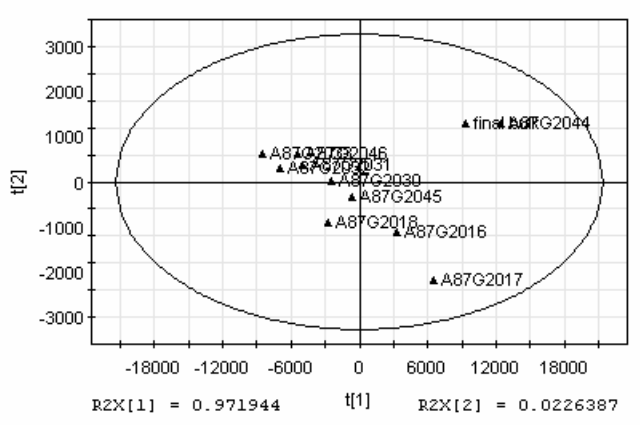

Figure 73. PCA score plot of fluorescence map of concentrated samples

From the score plot it can be seen that batches A87G2044 and final bulk are clearly outliers with respect to PC1, whereas batches A87G2016, 17 and 18 are discriminated from the others by PC2. the first two components are sufficient to explain 97% of total variance in the map.

Form the loading plot p_1 (fig. 74, A), it can be seen that PC1 represents a broad fluorescence band in the 320-400 nm region, while PC2 (fig. 74,B) represents a decrease in the fluorescence from excitation with short wavelength, and an increase in the fluorescence from long wavelength excitation.

Consequently, batches A87G2044 and final bulk can be seen to be outliers by having a generally higher fluorescence in the emission region from 320 to 400 nm, while batches A87G2016, 17 and 18 are discriminated from the other batches in having a relatively higher fluorescence from short wavelength excitation and a lower fluorescence frorn long wavelength excitation: this is consistent with the presence of an additional (or an increased quantity) of an additional fluorophore in these batches that absorb at short wavelength.

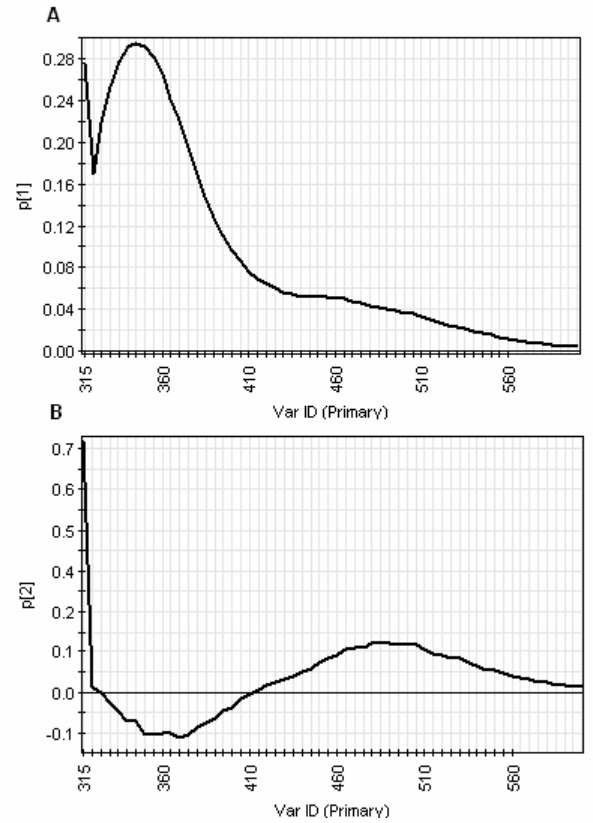

Figure 74. PCA loading plots of fluorescence map of concentrated samples
A) Loading plot p_1**; B)** loading plot p_2

PCA of diluted samples fluorescence maps. From the score plot of figure 75 it can be seen that the batches are primarily delineated by PC1, with batch A87G2046 at one extreme and A87G2017 to the other, while final bulk is separated from the other batches by PC2. these first two components are sufficient to explain 97% of overall variance in the maps.

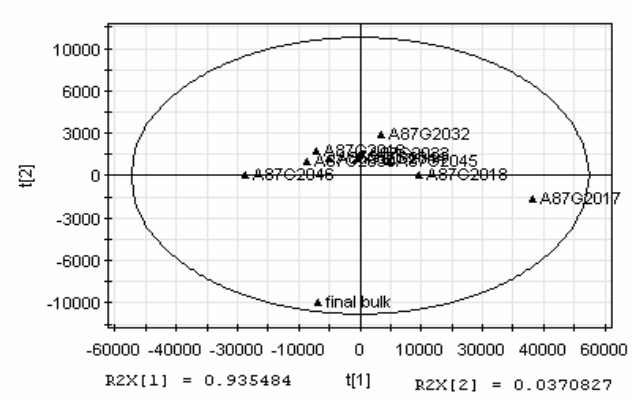

Figure 75. PCA score plot of fluorescence maps of diluted samples

From the loading plots of figure 76, it can be seen that PC1 (fig. 76, A) represents a broad fluorescence band in the 300-380 nm fluorescence emission region, corresponding to tryptophan fluorescence, while PC2 (fig. 76, B) essentially increase the fluorescence for 285 nm excitation at the expense of that from shorter wavelength excitation. Thus, the separation of the batches in terms of PC1 relates to the magnitude of the tryptophan fluorescence.

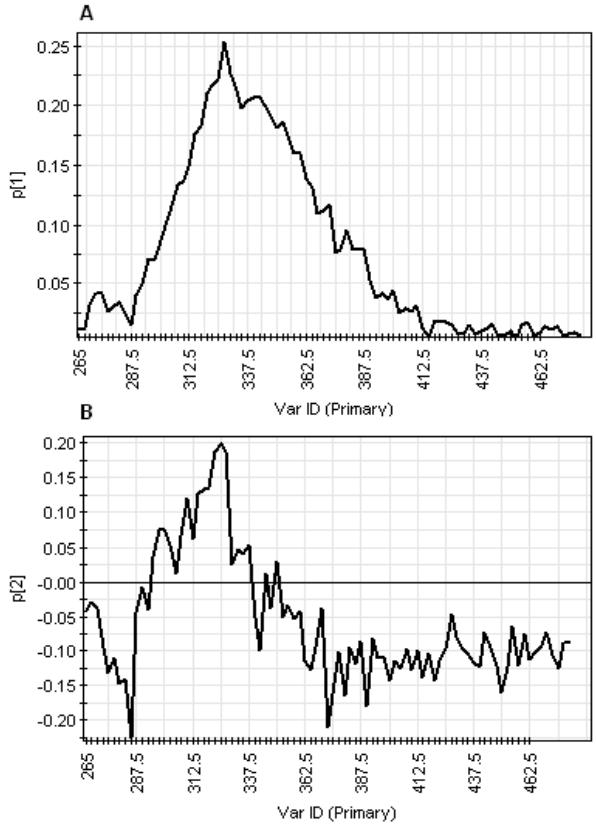

Figure 76. PCA loading plots of fluorescence maps of diluted samples

2.1.11 Fluorimetry Second-order scatter mapping

Following from the previous section, the Rayleigh-Raman scattering region of the fluorescence maps can be explored.

The Rayleigh-Raman scattering detection derives from the fact that diffraction gratings used in monochromators actually disperse not just light of wavelength λ, but also all the related orders λ/n, where n= 1, 2, 3…, at a coincident angle, although less efficiently. Thus, when the emission monochromator is set to a wavelength λ nm, it will also pass an amount of the second order light of wavelength λ/2 nm.

Consequently, if the excitation monochromator is set to a wavelength λ_{ex} nm and this is then Rayleigh-or Raman scattered by the sample, it will pass through the emission monochromator when it is set to $\lambda_{em} = 2\ \lambda_{ex}$, appearing to be an "emission" at a wavelength close to 2 λ_{ex} nm. For example, when exciting at 250 nm, this can be scattered and passed at apparently 500 nm by the emission monochromator to the detector.

Figure 79 presents the normalised emission scans form the fluorescence maps of batch A87G2016 and final bulk. It can be seen that the scattering band between 500-600 nm (for the scattering of 250-300 nm light) are much larger for final bulk than A87G2016.

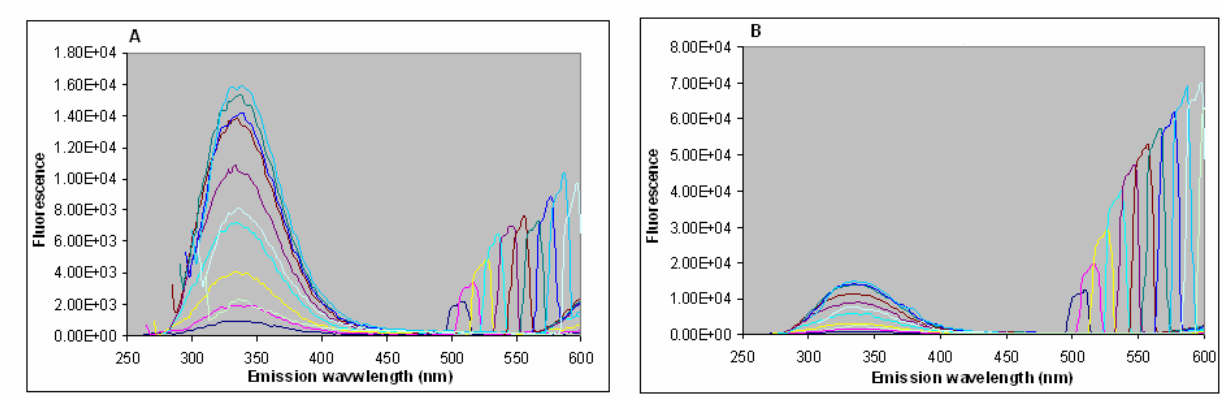

Figure 77. Fluorescence maps of diluted samples
A) fluorescence map of A87G2016; **B)** fluorescence map of final bulk

The score plot of the normalised fluorescence maps in the scattering region is presented in figure 78. As expected, the final bulk is an extreme outlier with respect PC1. PC1 is the only PC necessary to account 99.9% of total variance and has the form of Rayleigh and Raman scattering features observed for final bulk (fig. 77).

The Rayleigh scattering derives predominantly form particulates in the solution, including aggregated protein, whereas the Raman scattering, at a slightly lower wavenumber, will have a large contribution from water.

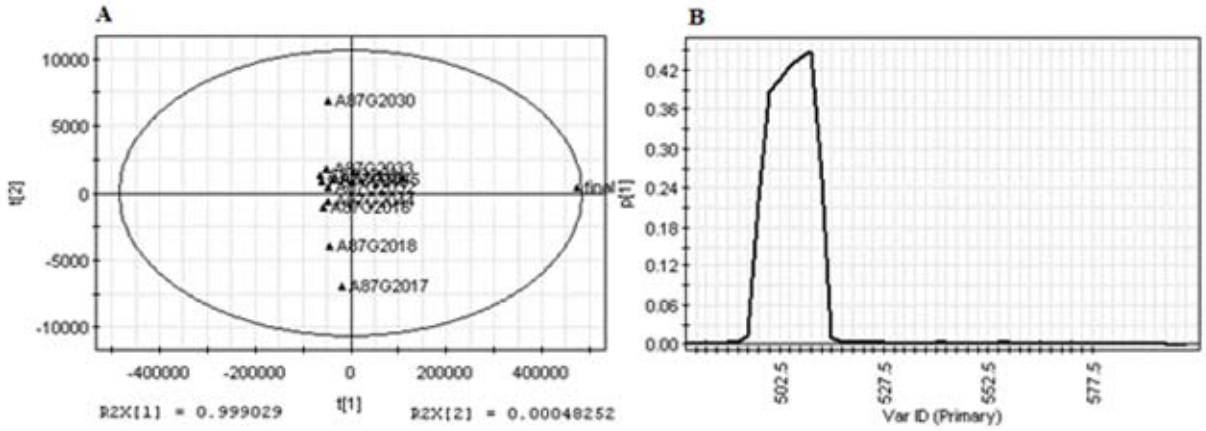

Figure 78. Fluorescence maps of scattering region
A) PCA score plot of fluorescence scattering region**; B)** PCA loading plot p_1 of fluorescence scattering region

2.2 QUANTIFICATION OF DEGRADATION PRODUCTS BY USING CD AND IR IN COMBINATION WITH MVDA

The aim of the present thesis was to assess the feasibility of obtaining quantitative information about degradation of biopharmaceuticals when employing spectral data, here exemplified by CD and IR in combination with PLS regression. In order to generate a suitable calibration matrix, a set of samples containing pre-defined levels of aggregates, oxidized forms and free-Fc, was generated as described in the Materials and Methods section. Table 4 shows the ranges covered by the calibration matrix for each degradation form.

Degradation forms	*%*
Free-Fc	0-22
Aggregates	0-15
Ox forms	0-48

Table 4. Ranges covered by the calibration matrix

All the samples generated were then analyzed separately for each of the three degradation forms employing dedicated chromatographic QC assays in order to obtain accurate degradant levels. Furthermore, both CD (near and far-UV) and IR spectra were measured. Both QC and spectroscopic data form the basis for the generation of various PLS models, based respectively on CD and IR spectra alone, as well as CD and IR data combined.

2.2.1 CD spectra of matrix samples

All matrix samples were analyzed by CD spectroscopy, both in the near and far-UV region. For instrumental parameters see Materials and Methods section. The obtained spectra are shown in figures 79 and 80.

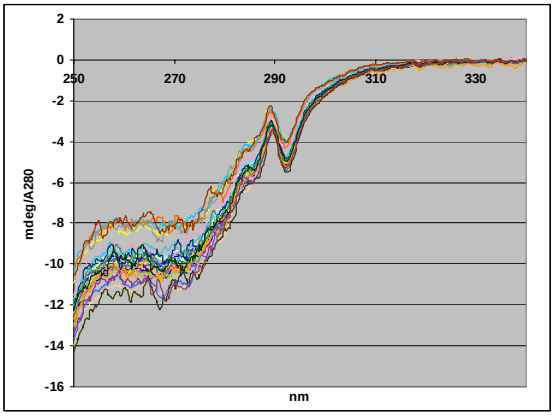

Figure 79. Near-UV CD normalised spectra

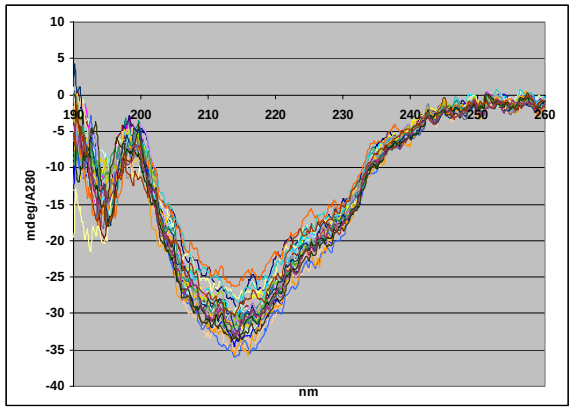

Figure 80. Far-UV CD normalized spectra

Data are usually pre-processed before PLS analysis. CD spectra, both near and far-UV, were normalized dividing the obtained signal by the UV absorption value measured at 280 nm. This normalization allows to keep only information about structural changes. So, as it can be seen from the graphs, by normalizing the spectra spectral differences arising from the different concentration of each sample were eliminated. Looking at the near-UV spectra, it can be noted that all of them are very consistent to each other, but, despite the normalization process, there is a set of four spectra lightly separated from all the other. On the other hand, far-UV spectra are all very similar, with all the spectra forming a single group.

2.2.2 IR spectra of matrix samples

IR spectra o f matrix samples are shown in figure 81. Here are showed the raw spectra, spectra obtained without any subtraction of the buffer spectrum. For all the instrumental settings, see Materials and Methods section

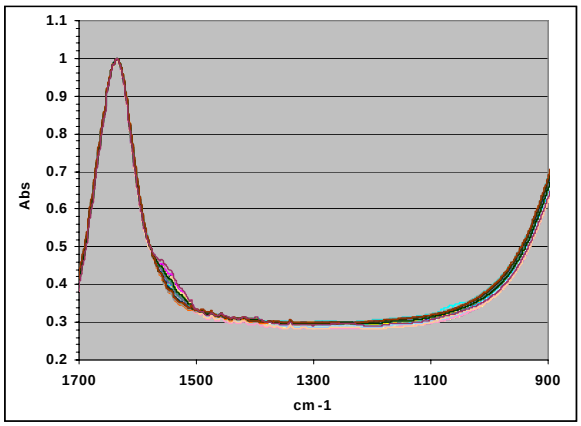

Figure 81. IR spectra of matrix samples

In order to eliminate differences among the spectra due only to the different concentration of each sample, IR spectra were normalized dividing the obtained signal by the maximum absorption value measured in the 1700-1600 cm^{-1} region of the IR spectrum (measured at ~ 1635 cm^{-1})

Looking at the spectra, it can be noted that all the spectra are very consistent to each other along the whole spectral region. Also in the Amide region (1700-1500 cm^{-1}), the most important region for structure determination in proteins directly correlated to the backbone conformation, the spectra are consistent; furthermore, are evident signals in the amide II region (1600-1500 cm^{-1}), the second most important band in protein structure determination.

2.2.3 QC analysis of matrix samples

All matrix samples were analyzed by the dedicated QC chromatographic assays described in Material and Methods section. In this way, for each matrix sample the observed value of each degradation forms were obtained and used for the generation of PLS models.

In table 5 are reported the QC values obtained for aggregates (SE-HPLC), oxidized forms (peptide mapping + UPLC) and free- Fc (IE-HPLC) for all the matrix samples. These values are compared with the corresponding theoretical values obtained generating the calibration matrix.

Samples	*% Aggregates*		*% Oxidised*		*% Free Fc*	
	(SE-HPLC)	*Matrix*	*UPLC*	*Matrix*	*IE-HPLC*	*Matrix*
Samples 1	1.79	0.00	11.03	3.21	3.28	2.65
Samples 2	3.75	0.88	10.77	6.32	0.22	0.00
Samples 3	33.20	13.34	as matrix	48.18	as matrix	6.63
Samples 4	10.79	3.55	24.84	19.26	3.46	3.53
Samples 5	8.97	3.10	20.50	14.90	2.01	2.05
Samples 6	6.36	5.16	5.61	0.00	4.30	3.85
Samples 7	4.85	0.00	17.05	13.23	3.17	2.91
Samples 8	7.16	2.95	15.03	10.66	0.20	0.00
Samples 9	28.90	10.60	as matrix	45.93	as matrix	10.53
Samples 10	6.09	2.97	11.04	5.37	2.74	2.46
Samples 11	3.74	1.54	7.10	3.71	3.56	3.06
Samples 12	3.67	0.86	11.98	6.18	1.31	1.28
Samples 13	8.57	2.79	21.44	15.13	1.81	1.39
Samples 14	12.50	5.23	25.41	18.90	3.92	3.47
Samples 15	5.62	3.82	6.36	1.87	4.65	3.79
Samples 16	3.51	0.34	13.16	7.44	2.78	2.46
Samples 17	4.13	2.25	7.66	3.25	0.46	0.36
Samples 18	8.70	4.56	as matrix	8.25	as matrix	11.34
Samples 19	1.97	0.56	7.29	2.01	as matrix	1.10
Samples 20	2.21	1.41	4.80	0.00	1.13	1.40
Samples 21	16.30	0.00	6.02	31.38	1.75	1.44
Samples 23	13.68	0.00	as matrix	40.73	as matrix	22.41
Samples 24	4.75	0.00	13.35	9.00	0.28	0.00
Samples 25	33.94	1.25	as matrix	45.81	as matrix	7.56
Samples 26	5.40	15.22	4.55	0.00	6.53	5.07
Samples 27	0.76	4.25	4.22	0.00	3.06	2.34
Samples 28	2.84	0.00	12.22	7.54	0.49	0.00
Samples 29	15.99	0.00	25.36	9.72	0.25	0.00
Samples 30	20.35	10.61	as matrix	48.19	as matrix	13.26
Samples 31	0.81	0.00	4.42	0.00	0.61	0.00

Table 5. Observed vs. theoretical degradation values

In those samples for which it was not feasible to perform QC analysis, due to concentration problems, the QC values were substituted by the corresponding theoretical values calculated in the calibration matrix .

Looking at the table, it is evident a difference between theoretical matrix and observed QC values for all the three degradation forms. In particular, for aggregates, QC values of samples 3, 9, 21. 23. 25 and 29 are much higher with respect to the theoretical ones. For oxidized forms, QC value of sample 21 is much lower than the theoretical one, while QC value of sample 29 is higher than the theoretical one. Differences observed for free-Fc between QC and theoretical values are less pronounced than those observed for the other two degradation forms.

2.2.4 PLS/O-PLS model of CD spectra

PLS model. The results of calibration model based on CD measurements are illustrated in the plots of figure 82, where the levels of each degradation forms, measured by QC methods, are plotted against the predicted levels obtained by PLS regression of spectroscopic data. In other words, these plots show how well the PLS-predicted values for the matrix samples correspond with the reference values from the QC assays. The closer the data points scatter around the diagonal line, the better the correlation, hence the higher the quality of the model.

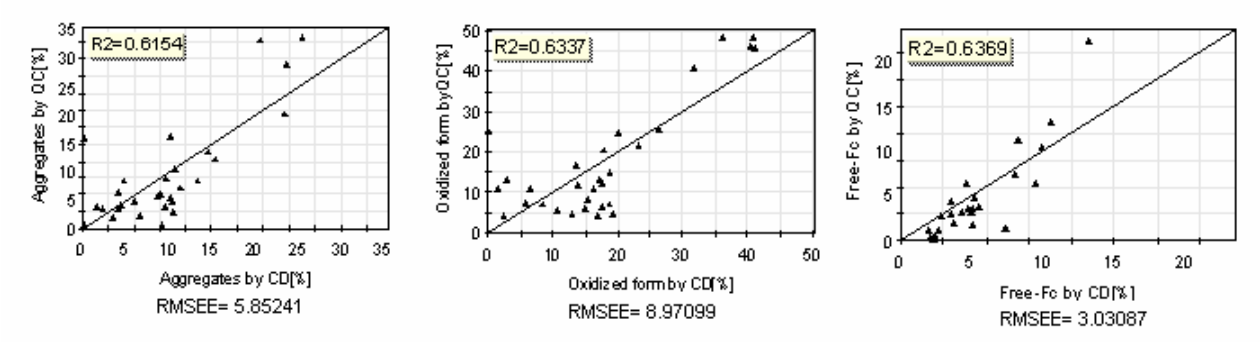

Figure 82. QC vs. PLS-predicted degradation values for near and far UV combined

Another way to assess the quality of the model is looking at the RMSEE (Root Mean Square Error of Estimation) values, representing the degree of confidence of how well QC data can be estimated for the calibration matrix samples.

The fact that for all the three degradation forms a trend is observable, or in other words that the data points scatter around the diagonal, indicates that CD contain quantitative information about these degradation forms. Comparing the RMSEE values obtained with PLS model based on both the near and far-UV regions, with the values obtained by using near and far-UV regions separately (fig. 83), it is evident that with the model based only on the near-UV region (340-250 nm) we obtained an increase of the accuracy of the model.

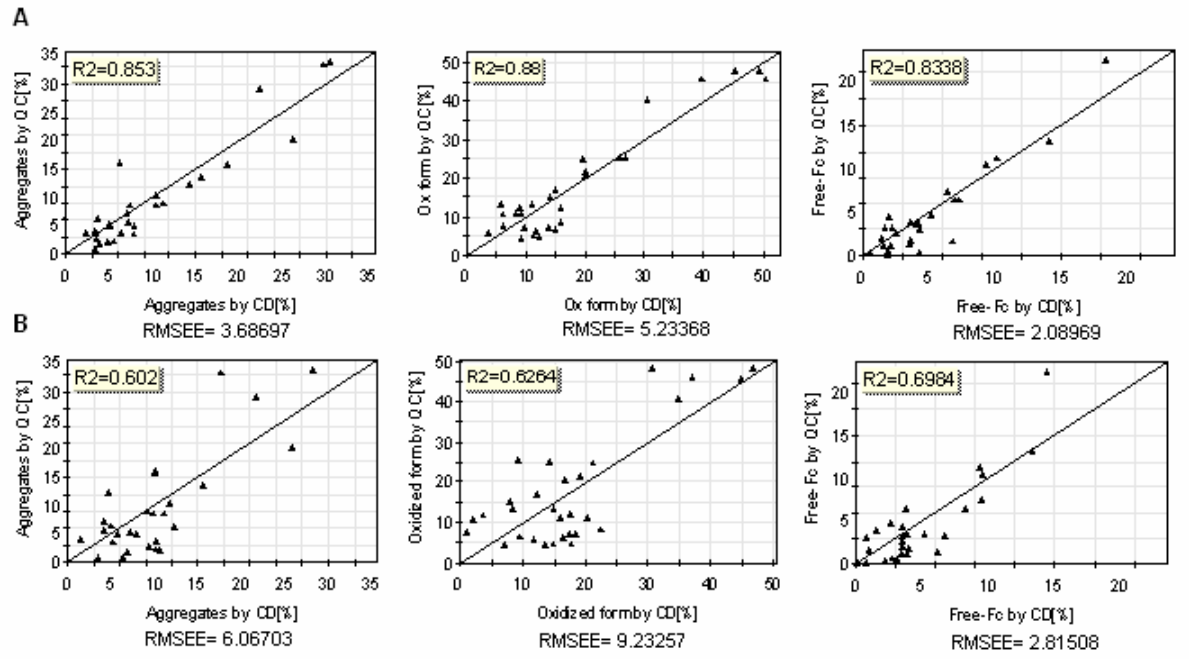

Figure 83. QC vs. PLS-predicted degradation values for near and far-UV
A) RMSEE values of near-UV based PLS-model; **B)** RMSEE values of far-UV based PLS-model

In the case of free-Fc, pronounced spectral differences between intact and degraded forms can be expected due to the differences in primary and tertiary structure content. However, the presence of aggregates or oxidized forms did not result in pronounced spectral differences that were easily detectable visually. Thus, the use of multivariate data analysis becomes unavoidable to extract quantitative information. In order to numerically assess the quality of prediction of various PLS models generated, we used the so-called "cross-validation" method, where all 31 matrix samples and, in addition, five "unknown" samples (not contained in the calibration matrix, containing random amounts of each degradation forms) were leaved out one at a time from the employed PLS model. The results are expressed as RMSEP (Root Mean Square Error of Prediction). RMSEP is a measure to assess the quality of the prediction of the PLS model applied to the spectra. It expresses the average difference between spectroscopic and QC data. A lower RMSEP value indicates a better prediction. In figure 84 are showed the RMSEP values obtained with CD-based PLS models.

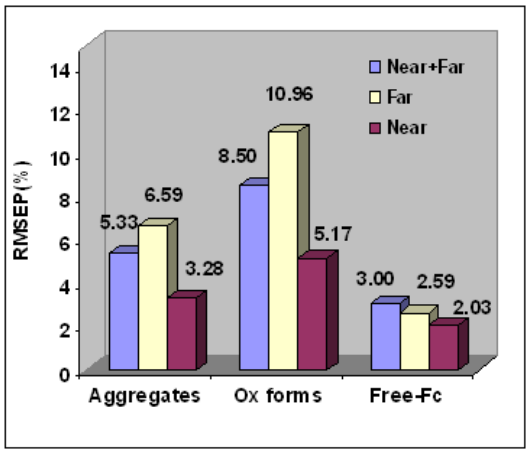

Figure 84. RMSEP values for CD-based PLS models

As expected, the most reliable prediction was obtained for levels of free-Fc, a lower prediction quality was found for aggregate levels, while for oxidation levels the RMSEP values ranged from 5.1 to 10.9 %. It can be noticed that, using only the near-UV region, were obtained for all the three degradation forms lower RMSEP values, and hence better prediction, than those obtained using both near and far-UV regions or only the far-UV region. Furthermore, using only the far-UV region, the prediction obtained for aggregates and oxidized forms is the lowest (compared with near-UV and near + far-UV models), while for free-Fc levels the prediction is quite comparable with that obtained with the "all range" model.

O-PLS model. Results of O-PLS calibration model are presented in the scatter plots of figure 85. These are the same plots showed for the PLS model, but in this case the O-PLS predicted values are plotted against the QC data obtained for each degradation forms.

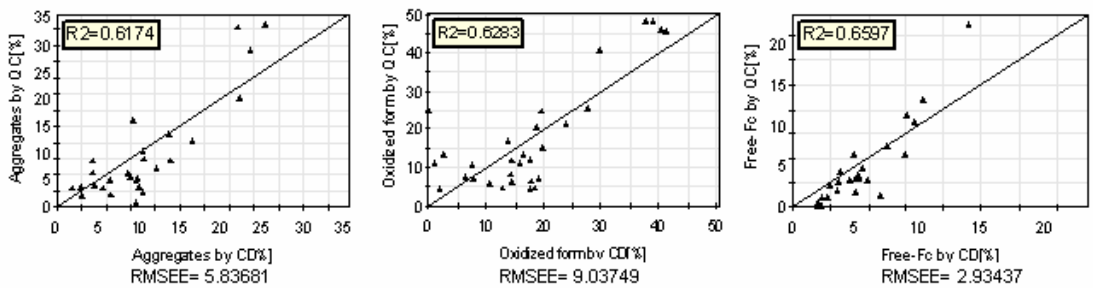

Figure 85. QC vs. O-PLS predicted degradation values for near and far-UV combined

Comparing these plots with those of PLS model (fig. 82 and fig. 83), we can observe that applying O-PLS regression, we obtained an improvement of the quality of the model, since the RMSEE values are lower than the respective values obtained for each degradation form by using PLS regression, and the data points scatter around the diagonal closer than in the previous model.

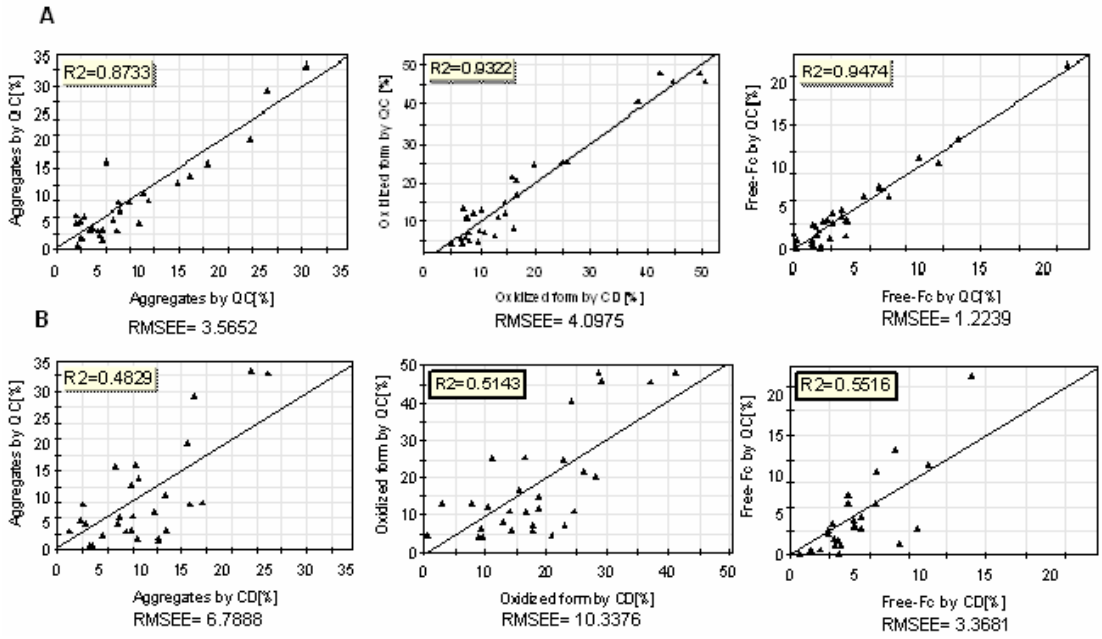

Figure 86. QC vs. O-PLS predicted degradation values for near and far-UV
A) RMSEE values of near-UV based O-PLS model; **B)** RMSEE values of far-UV based O-PLS model

Furthermore, with O-PLS a further improvement of the model quality was also obtained using the near-UV based model (fig 86). As for CD-based PLS models, cross validation process was applied to determine the prediction accuracy of the O-PLS models generated.

In figure 87,A are shown the RMSEP values obtained with O-PLS models. Furthermore, the prediction values obtained from the near-UV based O-PLS model are compared with those obtained with the corresponding near-UV based PLS model (fig. 87, B)

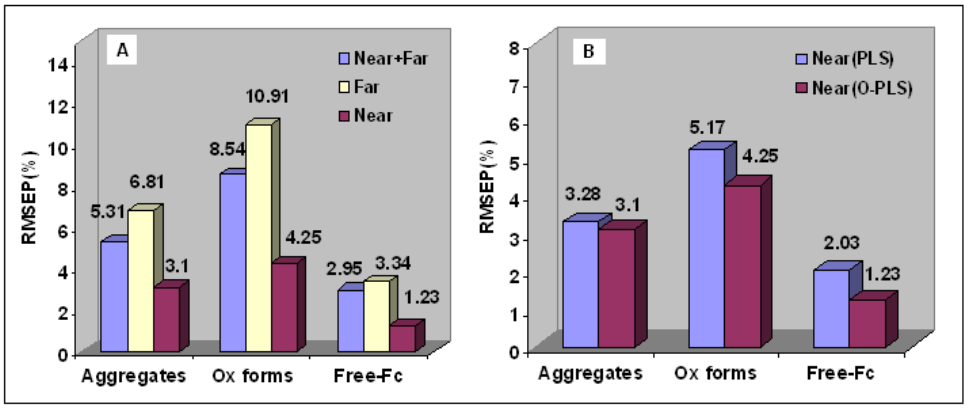

Figure 87. RMSEP values of CD based O-PLS models
A) RMSEP values of CD based O-PLS model: **B)** PLS vs. O-PLS degradation values.

Once again, the most reliable prediction was obtained for free-Fc levels, a lower prediction quality was found for aggregate levels, while for oxidation levels the RMSEP values ranged from 4.2 to 10.9 %. As in the PLS model, using only the near-UV region, we obtained, for all the three degradation forms lower RMSEP values than those obtained using near and far-UV region combined.

Furthermore, comparing the lower RMSEP values obtained with the PLS and O-PLS models (fig 87, B), it is clear how using O-PLS we obtained an increase of the predictive power of the model. These data suggest that, in the determination of all the three degradation forms, near-UV based models perform better than near/far-UV combined based model.

2.2.5 PLS/O-PLS model of FT-IR spectra

PLS model. The results of calibration model based on FT-IR measurements are illustrated in the plots of figure 88. Two different PLS models were generated, one using the 1700-900 cm^{-1} IR region (fig. 88, A); the other one was generated using only the so-called "Amide region" from 1700 to 1500 cm^{-1} (fig. 88, B).

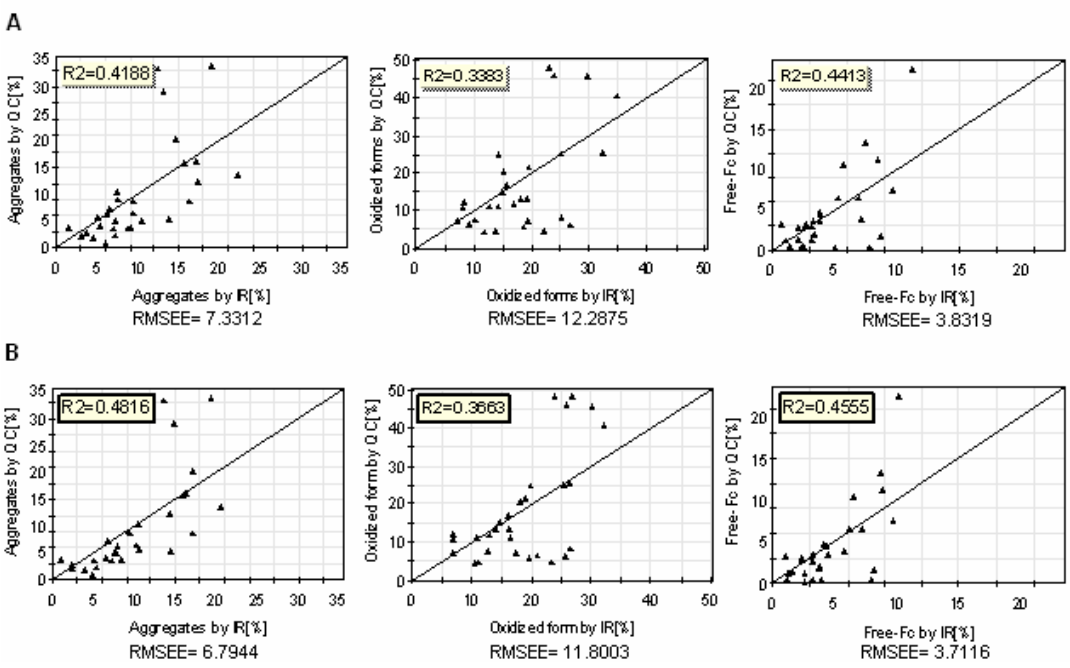

Figure 88. QC vs. PLS predicted degradation values for IR spectra
A) PLS model results based on 900-1700 cm-1 region; **B)** PLS model results based on 1500-1700 cm-1

Comparing the RMSEE values obtained from the two different PLS models, we can assess that the PLS model based only on the amide region (1500-1700 cm^{-1}) has lower RMSEE values, hence better quality, than the model obtained using the whole IR region. However, with respect to the near-UV based PLS model, the RMSEE values are still higher for all the three degradation forms, indicating that, when determining these contaminants, CD-based PLS model performs better than IR-based PLS model.

In order to determine the predictive power of the various PLS models generated, we followed the same approach used for CD models. So the 31 matrix samples and 5 "random" samples were used for the cross-validation process.

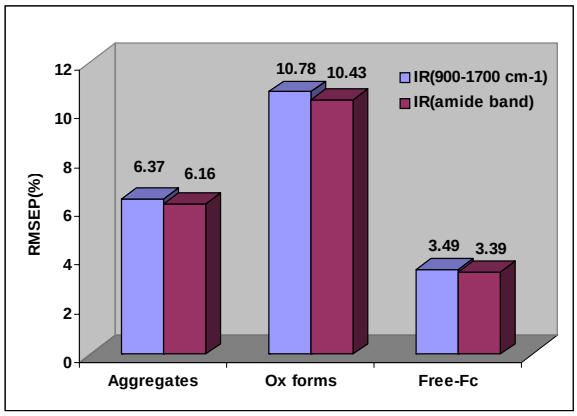

Figure 89. RMSEP values of IR-based PLS models

The RMSEP values for the two models generated are shown in figure 89. As we can see, the amide region-based models yielded a better prediction for all the three degradation forms compared to the whole region-based models. As in CD based models, the better prediction was obtained, for levels of free-Fc, a lower prediction accuracy for aggregates, while for oxidation levels the RMSEP values are higher. In general, it can be affirmed that the prediction values of the two IR-based PLS models are all quite comparable.

O-PLS model. The results of the calibration model based on FT-IR measurements are illustrated in plots of figure 90. As for the PLS models, two models were generated, one using the whole region (fig 93, A) and the other one using only the amide region (fig. 90, B).

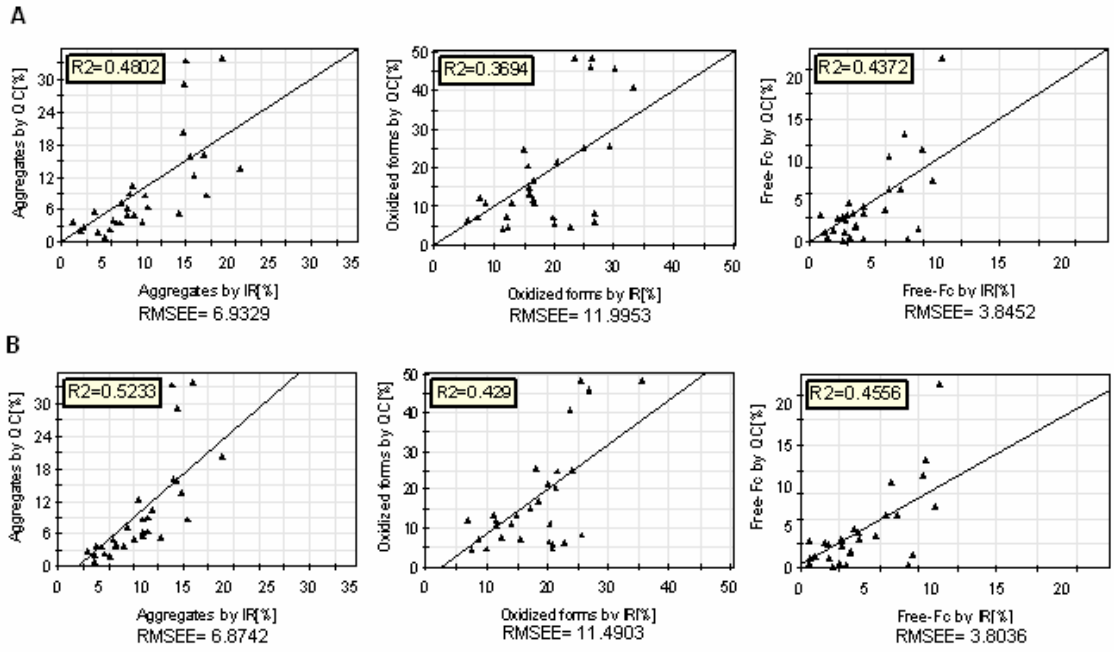

Figure 90. QC vs. O-PLS predicted degradation values for IR spectra
A) O-PLS model results based on 900-1700 cm-1 region; **B)** O-PLS model results based on 1500-1700 cm-1

Analysing at the scatter plots shown above, we can see that the two different models perform in a comparable way, with slightly lower values for aggregates and oxidized forms in the amide region-based model, while for free-Fc the results are slightly lower in the whole region-based model.

Comparing the RMSEE values obtained with O-PLS with those obtained with PLS, regarding aggregates and oxidized forms, both O-PLS models perform better than the corresponding PLS models. For free-Fc, the PLS and O-PLS models based in the whole region yield very similar results, while for the amide region, PLS model performs better than the respective O-PLS model.

In conclusion, we can affirm that models based on the amide region, both PLS and O-PLS, give rise to a better quality than the corresponding models based on the whole region.

Furthermore, the two amide region-based models (PLS and O-PLS) perform in a very similar way, with slightly lower RMSEE values for aggregates and oxidized forms in the O-PLS model. In the free-Fc case, we can notice that the results obtained with the two models are comparable. On the other hand, the quality of the IR-based models is lower with respect to that of CD-based models, since the RMSEE values for all the three degradation forms obtained in the IR-based model are much higher than those obtained with CD spectral data. As for PLS models, to determine the predictive power of the various IR-based O-PLS models generated, we followed the cross validation approach.

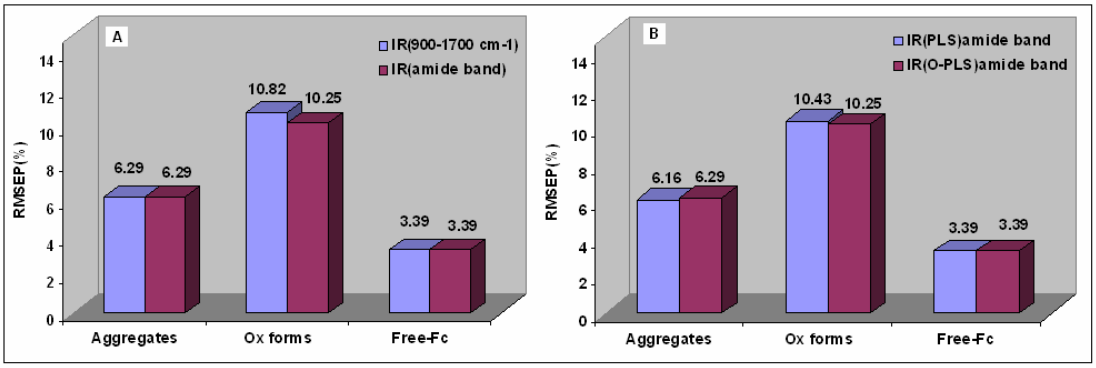

Figure 91. RMSEP values of IR-based O-PLS models
A) RMSEP values of IR based O-PLS model; **B)** PLS vs. O-PLS degradation values.

The RMSEP values for the O-PLS models are shown in figure 91. As in the PLS models, the amide region-based model yielded a slightly better prediction for all the three degradation forms as compared to the whole region-based models. As for the CD based models, the better prediction was obtained, for levels of free-Fc, a lower prediction accuracy for aggregates, while for oxidation levels RMSEP values are much higher with respect to the others. Comparing the PLS and O-PLS models based on the amide region (fig 91, B), we can see that the results are quite comparable.

In conclusion, in the prediction of all the three degradation forms, the amide region based PLS/O-PLS models perform better than the model based on the whole spectral region. However, applying O-PLS instead of PLS to the IR spectra doesn't yield a great advantage in terms of quality of the prediction as when it has been applied to the CD-based models. Furthermore, in the determination of the degradation forms studied, the obtained accuracy of CD-based O-PLS models was higher than the accuracy obtained with IR-based O-PLS models.

2.2.6 PLS/O-PLS model of CD and IR combined

In addition to the PLS/O-PLS models generated using CD and IR spectra separately, models based on CD and IR data combined were generated, in order to enrich the information content of the (combined) spectrum by combining signals generated from different physical parameters, and hence increase the accuracy of prediction.

With that goal, combined spectra were generated appending CD and IR data in a single array and scaled to each other (see Materials and Methods section) to ensure that they contribute in a similar way in the model building process. Both for PLS and O-PLS models, different combinations of the IR and CD spectral range were used and PLS/O-PLS model were generated for each of these combinations.

PLS models. In figure 92 are shown the results of the calibration models based on CD and IR combined spectral data. Four different PLS models were created by combining in one model the complete region of each spectroscopic technique, and in the others different parts of each spectral range. In particular, are illustrated plots for the PLS model based on the entire CD (near + far-UV) and IR range (900-1700 cm^{-1}) combined (fig. 92, A) and for the model based on near-UV and IR amide region combined (fig. 92, B).

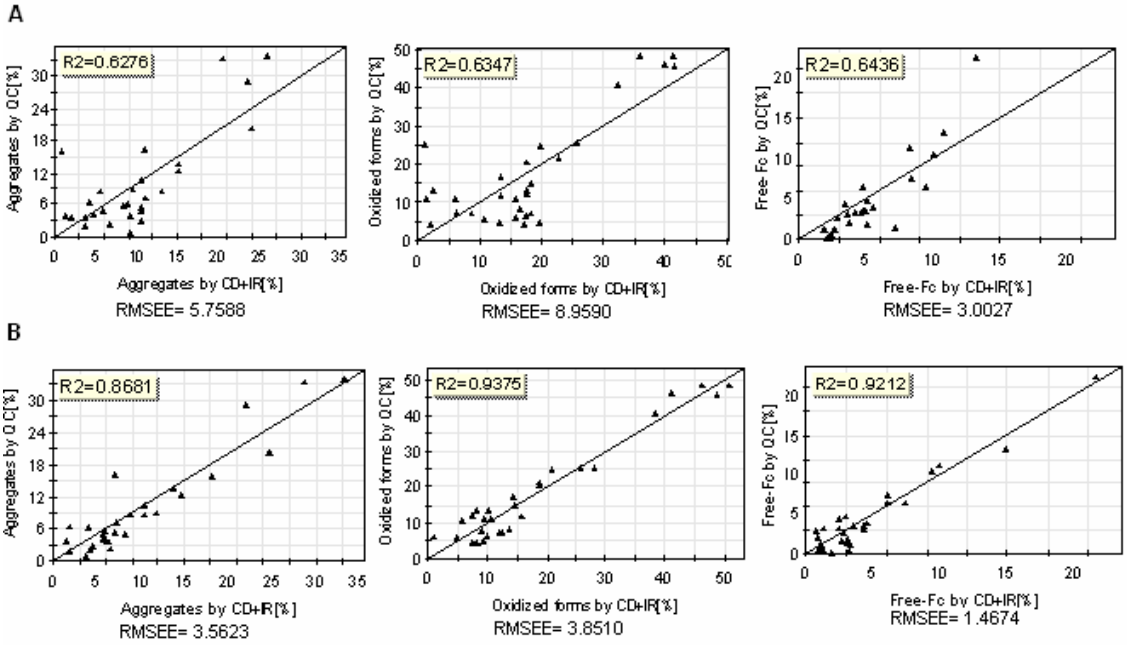

Figure 92. QC vs. PLS predicted degradation values for CD and IR spectra combined
A) PLS model results based on CD (near + far-UV) and IR (900-1700 cm^{-1}) combined; **B)** PLS model results based on CD (near-UV) and IR (1500-1700 cm^{-1}) combined

Combining the whole CD and IR spectral range in a single PLS model, the RMSEE values obtained for all the degradation forms are lower compared to the values obtained with the PLS model based only on the whole IR region, while respect to the whole CD region-based model the combination of the two spectral regions did not yield a real improvement of the RMSEE values. Further increase of the quality of the model was obtained combining the two spectral regions which gave the best RMSEE values both for CD and IR. Thus, near-UV CD region and IR amide region (1500-1700 cm^{-1}) were combined and a single PLS model was generated. The RMSEE values obtained (fig. 92, B) clearly show that, combining these spectral regions, yields to a real improvement of the PLS model quality, since the RMSEE values for aggregates, oxidized forms and free-Fc were lower than the respective values obtained with CD and IR single PLS models. The other PLS models generated, combining different parts of the spectra, together with the relative RMSEE values showed in table 6, further confirmed that the best quality was achieved by combining only the CD near-UV region with the IR amide region

	RMSEE values (%)		
PLS models	***Aggregates***	***Ox forms***	***Free-Fc***
CD (near+far) + IR (all range)	5.75	8.95	3.00
CD (near-UV) + IR (amide bands)	3.70	4.50	1.72
CD (near-UV) + IR (all range)	6.04	8.83	3.15
CD (near+far) + IR (amide bands)	3.22	5.68	1.84

Table 6. RMSEE values of CD and IR combined-based PLS models

To further confirm the improvement of the quality achieved combining the two spectral region in a single PLS model, the classical cross validation method already used to validate CD and IR single PLS models, was carried out. The obtained RMSEP values for all the PLS models generated are listed in table 7.

	RMSEP values (%)		
PLS models	***Aggregates***	***Ox forms***	***Free-Fc***
CD (near+far) + IR (all range)	5.25	8.50	2.97
CD (near-UV) + IR (amide bands)	3.15	4.21	1.35
CD (near-UV) + IR (all range)	5.76	8.01	5.09
CD (near+far) + IR (amide bands)	5.29	8.49	2.99

Table 7. RMSEP values of CD and IR combined-based PLS models

In figure 93 the RMSEP values obtained with the PLS model based on the whole CD and IR spectral region (fig .93, A) and those obtained with the model based only on CD near-UV and IR amide region (fig. 93, B), are compared with the values obtained employing CD and IR separately.

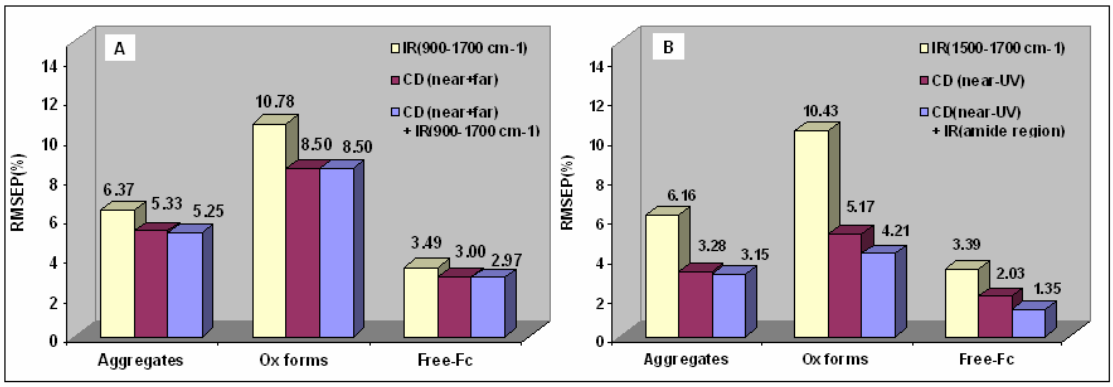

Figure 93. RMSEP values of CD and IR combined PLS models
A) RMSEP values of CD(near + far) and IR(900-1700 cm^{-1}) combined PLS model: **B)** RMSEP values of CD (near-UV) and IR (1500-1700 cm^{-1}) combined PLS model.

In the PLS model based on CD (near +far) and IR (900-1700 cm-1) combined data, there is an increase of the prediction accuracy for all the three degradation forms as compared to the accuracy obtained with IR-based PLS model. In comparison with the CD-based model, the new results can be considered very comparable, with only a little improvement of the prediction accuracy.

Further improvement of prediction was obtained combining near-UV CD region and IR amide region in a single model. As shown in fig. 93, panel B, RMSEP values of the combined-PLS model are better not only compared with the IR-model results, but also in comparison the CD-based model results, confirming that these two are the most useful spectral regions to predict the content of protein degradation forms. As in the previous models, the most reliable prediction was obtained for levels of free-Fc, a lower prediction quality was found for aggregate levels, while for oxidation levels the RMSEP values ranged from 4,21 to 5.17 %.

O-PLS model. As for PLS, various O-PLS models were generated based on combinations of different parts of the IR and CD spectral range, whose RMSEE values are shown in table 8.

	RMSEE values (%)		
O-PLS models	***Aggregates***	***Ox forms***	***Free-Fc***
CD (near+far) + IR (all range)	5.52	9.08	2.69
CD (near-UV) + IR (amide bands)	2.78	2.12	0.89
CD (near-UV) + IR (all range)	6.04	8.83	3.15
CD (near+far) + IR (amide bands)	3.22	5.68	1.84

Table 8. RMSEE values of CD and IR combined-based O-PLS models

For comparison with PLS, in figure 94 are illustrated the results of O-PLS calibration model based on all CD and IR spectral range combined data (fig. 94, A), and on near-UV and IR amide region combined (fig 94, B).

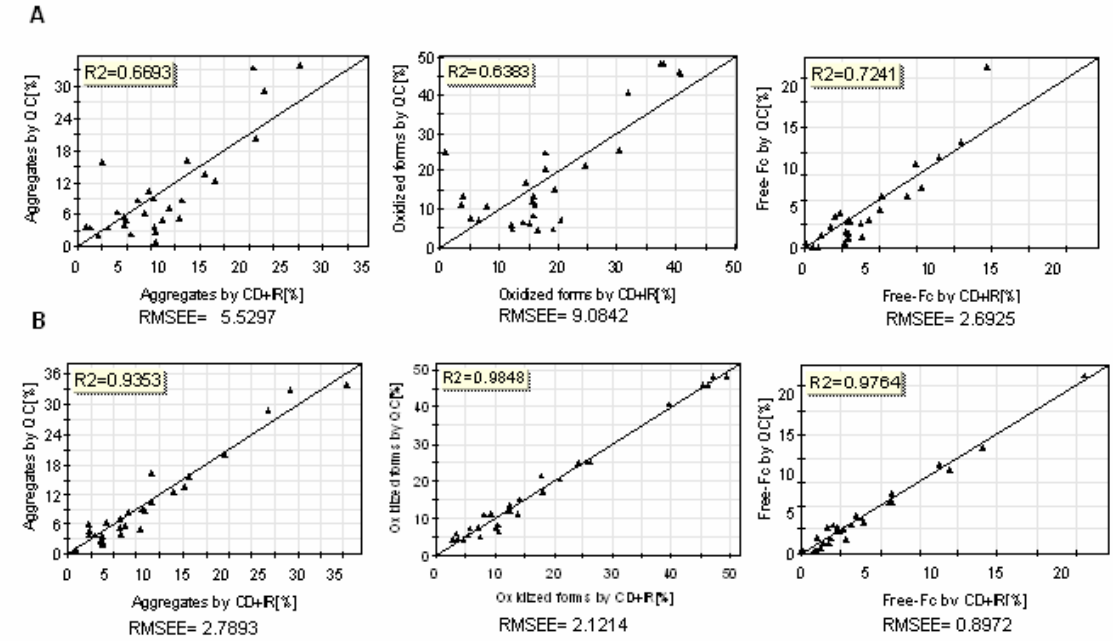

Figure 94. QC vs. O-PLS predicted degradation values for CD and IR spectra combined
A) O-PLS model results based on CD (near + far-UV) and IR (900-1700 cm^{-1}) combined; **B)** O-PLS model results based on CD (near-UV) and IR (1500-1700 cm^{-1}) combined.

From the scatter plots of figure 97, A, it can be noted that for the "all range"-based model the results are highly comparable with those obtained with the corresponding PLS model. More precisely, values obtained for aggregate and free-Fc levels are slightly lower compared with the same PLS values, while for oxidized form levels, O-PLS value is lightly higher than the PLS value. We can assess that, for "all range" model, using O-PLS instead of PLS doesn't yield a great advantage, in terms of increasing of prediction accuracy.

As in the previous models, the best quality was obtained combining CD near-UV region with IR amide region. In fact, not only the RMSEE values are much lower than those arising from the "all range" model, but the O-PLS model quality obtained is even better than that obtained in the corresponding PLS model, for all the degradation forms.

Further confirmation of the usefulness of applying O-PLS regression to CD and IR combined data, was given by the usual cross validation process. As for the previous models, cross validation was applied to all the O-PLS models generated (see table 9).

	RMSEP values (%)		
O-PLS models	***Aggregates***	***Ox forms***	***Free-Fc***
CD (near+far) + IR (all range)	4.95	8.29	2.76
CD (near-UV) + IR (amide bands)	2.40	3.32	0.92
CD (near-UV) + IR (all range)	5.56	8.05	2.89
CD (near+far) + IR (amide bands)	3.73	6.98	1.71

Table 9. RMSEP values of CD and IR combined-based O-PLS models

In figure 95 the RMSEP values obtained with the PLS model based on the whole CD and IR spectral region (fig. 95, A) and those obtained with the model based only on CD near-UV and IR amide region (fig. 95, B), are compared with the values obtained employing CD and IR separately.

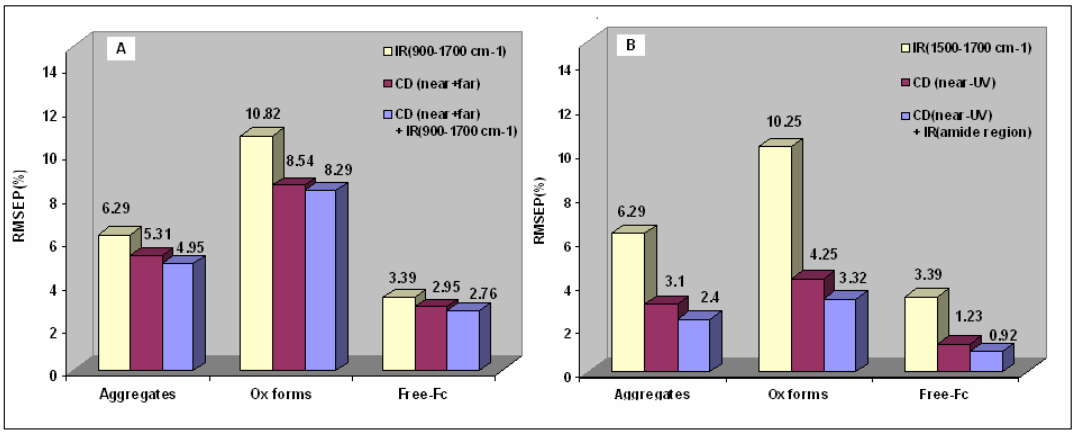

Figure 95. RMSEP values of CD and IR combined O-PLS models
A) RMSEP values of CD(near + far) and IR(900-1700 cm^{-1}) combined PLS model: **B)** RMSEP values of CD (near-UV) and IR (1500-1700 cm^{-1}) combined PLS model.

As for PLS regression, also O-PLS models of CD and IR data combined yielded to an increase of prediction accuracy for all the degradation forms. As we can see from figure 98, this increase was obtained both for the "all range" model and for the model based on near-UV and IR amide region. Once again, the lower RMSEP values were obtained with near-UV/IR amide region O-PLS model, as for PLS combined model. As expected, the better prediction accuracy was obtained for free-Fc levels, with a RMSEP values of 0.92 %. A slightly higher value was obtained for aggregates, but still lower than 3 % (2.4%), and for oxidized forms the obtained RMSEP value was 3.32 %, showing the major improvement compared to the values obtained using CD and IR separately.

This data further confirm that combining CD and IR data in a single model produce a real increase of the prediction accuracy of all the three degradation forms being analyzed and that this improvement is even greater when only the CD near-UV region and the IR amide region are combined. Furthermore, employing O-PLS instead of PLS further reduce the RMSEP values obtained, improving even more the predictive power of the multivariate analysis.

2.3 ENGINEERING OF METAL-BINDING PROTEINS BASED ON CONOPEPTIDES SCAFFOLD

2.3.1 Design of the metal binding site

Contryphan-Vn was used as a template for the design of a metal binding, disulphide constrained peptide. Analysis of the amino acid sequence of Contryphans reveals that amino acids at positions 2, 6 and 8 in the Contryphan-Vn sequence are not conserved in other Contryphans, despite a fairly good conservation of the overall molecular structure (Eliseo et al., 2004). Thus, residues Asp2, Lys6 and Trp8 were selected to be substituted by three histidine residues. Further, an additional His residue was added at the C-terminal end of the peptide, outside the disulphide ring, generating the new amino acid sequence shown in Figure 96, B.

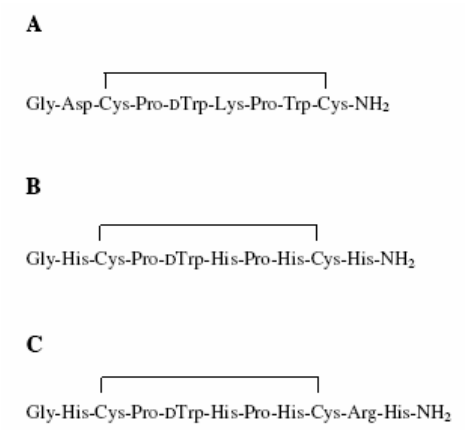

Figure 96. Amino acid sequence of Cupryphans
Amino acid sequence of Contryphan-Vn (A) and of the redesigned peptides Cupryphan(B) and Arg-Cupryphan (C).

The feasibility of formation of a copper binding site in the novel peptide, was confirmed by modelling its three-dimensional structure based on Contryphan-Vn structure (PDB code 1NXN;) (Eliseo et al., 2004). The side chain conformation of the four histidine residues was adjusted to form a metal binding site with tetragonal geometry, taking as a reference the copper binding site of bovine superoxide dismutase (PDB code 2SOD), (Tainer et al., 1982). Stereochemical compatibility of metal binding site formation was confirmed by energy minimization in explicit solvent using CHARMM (Brooks et al., 1983). The final model is shown in Figure 97 A.

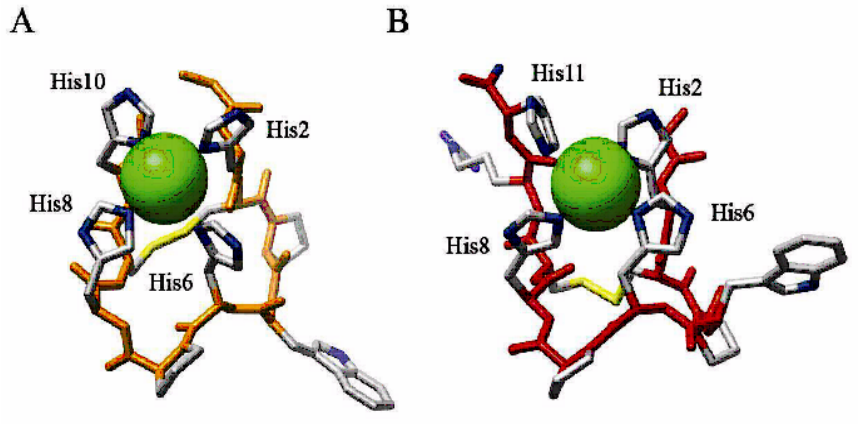

Figure 97. Energy minimized three-dimensional model of cupryphans
A) Cupryphan; **B)** Arg-Cupryphan. The copper ion is represented as a green sphere. The figure was prepared using UCSF Chimera (Nicholls et al., 1991)

2.3.2 Cupryphan metal binding ability probed by optical and fluorescence spectroscopy

Evidence of Cu2+ binding to the peptide was first obtained by fluorescence quenching experiments. Addition of copper quenches the fluorescence emitted by the unique Trp residue of the peptide (DTrp5), as shown in figure 98 (Vita et al., 1995). Fluorescence quenching is dependent on metal concentration, and reaches a maximum at a $[Cu^{2+}]$/[peptide] ratio higher than 1:1. All of these characteristics are consistent with an energy-transfer mechanism from the tryptophan residue to the copper bound at the binding site. The engineered peptide was named Cupryphan from the contraction of the words cupreous and Contryphan. Fluorescence quenching due to copper binding to Cupryphan was used to determine the binding affinity of Cupryphan for Cu^{2+}. The maxima of fluorescence intensity curves at 352 nm as a function of copper concentration were fitted with Scatchard transformation (fig. 99), obtaining a copper dissociation constant of 1.3 (± 0.2) × 10^{-7} M. The good fit of the data with a linear function indicated that only one class of copper binding site(s) is present in Cupryphan. In addition, fluorescence data indicated the presence of only one binding site per molecule. This is evident from the X-axis intercept value (2.06 × 10-5 M) in the Scatchard transformation plot, which represents the total Cu^{2+} bound at infinite Cu^{2+} concentration (i.e., when $[Cu^{2+}]$bound/$[Cu^{2+}]$free=0). In fact, this value correlates well with the total peptide concentration value (i.e. 2.1 × 10^{-5} M).

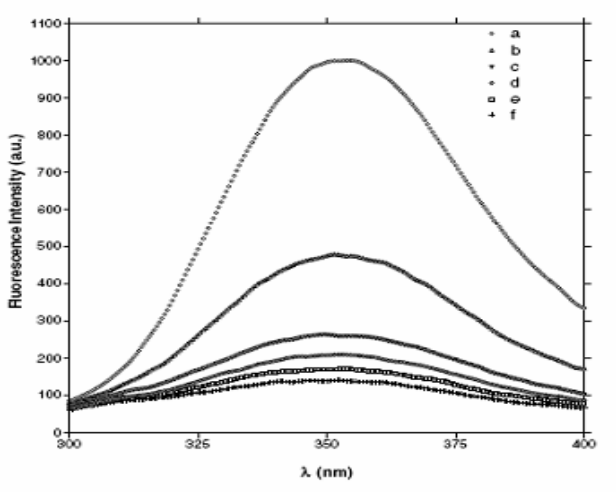

Figure 98. Emission fluorescence spectra of Cupryphan in presence of Cu^{2+} ions
a) apo Cupryphan (2.1×10^{-5} M); **b)** a + $CuCl_2$ 5.5×10^{-6} M; **c)** a + $CuCl_2$ 1.0×10^{-5} M; **d)** a + $CuCl_2$ 1.5×10^{-5} M; **e)** a + $CuCl_2$ 2.0×10^{-5} M**; f)** a + $CuCl_2$ 2.5×10^{-5} M**.**

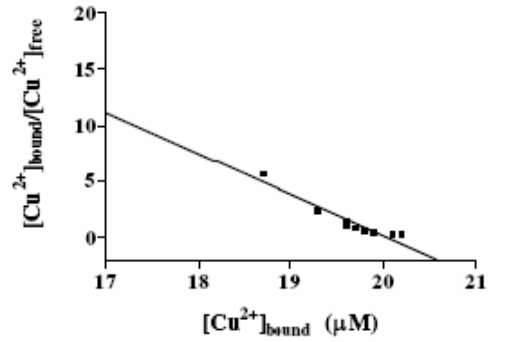

Figure 99. Kd determination of Cu^{2+} to Cupryphan by fluorescence quenching experiments
Is shown the Scatchard transformation of fluorescence intensities at 352 nm in the presence of increasing amounts of Cu^{2+} ions.

Further evidence of copper-peptide binding was obtained by optical spectroscopy analysis. Addition of a stoichiometric amount of $CuCl_2$ to the peptide (1.0 mM final concentration) induces the appearance of absorption bands with maxima at 312 and 580 nm (fig. 100), the first indicative of a Cu^{2+}-histidine charge-transfer (Cupane et al., 1994), the second due to electronic transitions of copper d-d orbitals and typical of copper complexes with nitrogen ligands (Cupane et al., 1994; Bryce et al., 1965). The maximum and the shape of the absorption band centred at 580 nm did not change with pH in the 6.0-9.5 range (data not shown), indicating a high pH stability of holo Cupryphan. The calculated molar extinction coefficient of Cupryphan at 580 nm was 50 M^{-1} cm^{-1}. No absorption bands appear in the optical spectrum of Contryphan-Vn after addition of copper ions (data not shown), indicating that the natural peptide does not bind copper ions.

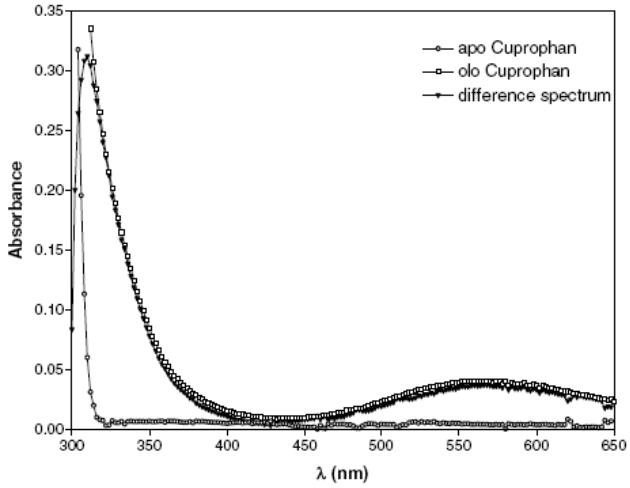

Figure 100. Optical spectra of Cupryphan
Apo (o) and holo (□) Cupryphan (1 mM in 50 mM sodium acetate buffer, pH 6.5). The difference spectrum is also shown (▼) to evidentiate the appearance of a band at 312 nm, indicative of a Cu^{2+}-histidine charge-transfer

To investigate the possible competition of other metal ions in the binding to the peptide, tryptophan fluorescence quenching experiments by copper ions were carried out in the presence of approx. a 5 fold excess of Zn^{2+} or Mg^{2+} (ion and peptide concentrations were 100 and 21 μM, respectively). The calculated copper dissociation constant was in both cases 1.4 (± 0.1) × 10^{-7} M, a value very similar to the one determined in the absence of Zn^{2+} or Mg^{2+}. This result indicates that Zn^{2+} and Mg^{2+} do not appreciably outcompete copper for binding to Cupryphan. In this regard, it must be noted that addition of Zn^{2+} and Mg^{2+}does not lead to quenching of tryptophan fluorescence, thus a direct determination of the dissociation constant of these two ions for Cupryphan was not possible using fluorescence spectroscopy.

2.3.3 EPR spectroscopy characterization of Cupryphan

The EPR spectrum of Cupryphan, recorded at liquid nitrogen temperature (fig. 101, A), revealed a homogeneous signal, due to the presence of only one copper species bound to the peptide. Coordination geometry appears to be axial, with values of the spectroscopic parameters ($g_{\|}$ and $A_{\|}$) typical of hexacoordinate complexes with an axial symmetry. Moreover, in the $g_{\perp}$ region (fig. 101, B) 5 to 7 superhyperfine lines are observed with position and coupling constant (≈ 12 G), typical of copper(II) complexes with at least 2 magnetically equivalent histidine residues on the coordination plane.

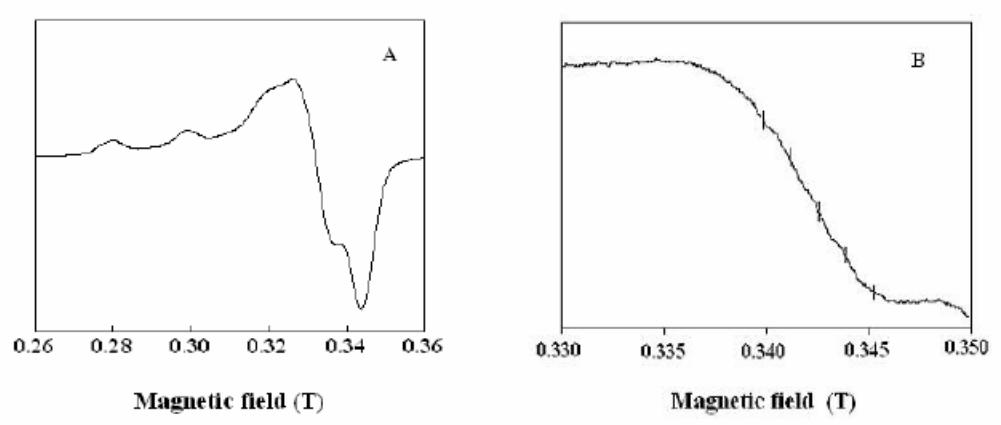

Figure 101. EPR spectra of Cupryphan
A) 9.556 GHz; Power =20 mW; modulation=10 G; T=115 K; **B)** 9.556 GHz; Power = 20 mW; modulation = 4 G; T=115 K. The vertical lines indicate the minimum number of superhyperfine lines observable.

The presence of a single class of binding sites per molecule was further corroborated by saturation experiments in the 20 to 200 mW power range (not shown). In fact the signal line shape did not change upon increasing the power, suggesting that it likely arises from a single copper coordination environment.

2.3.4 NMR spectroscopy characterization of Cupryphan

First of all, the copper interaction with the native Contryphan-Vn peptide was investigated. The assignment of the 1H spectrum of Contryphan-Vn is reported elsewhere (Eliseo et al., 2004). The addition of $CuCl_2$ to a 3.1×10^{-4} M solution of Contryphan-Vn (peptide: Cu ratio 1:0.27) caused a decrease of the intensity of all ^1H amino acid signals whereas the corresponding linewidths were not significantly changed (fig. 102).

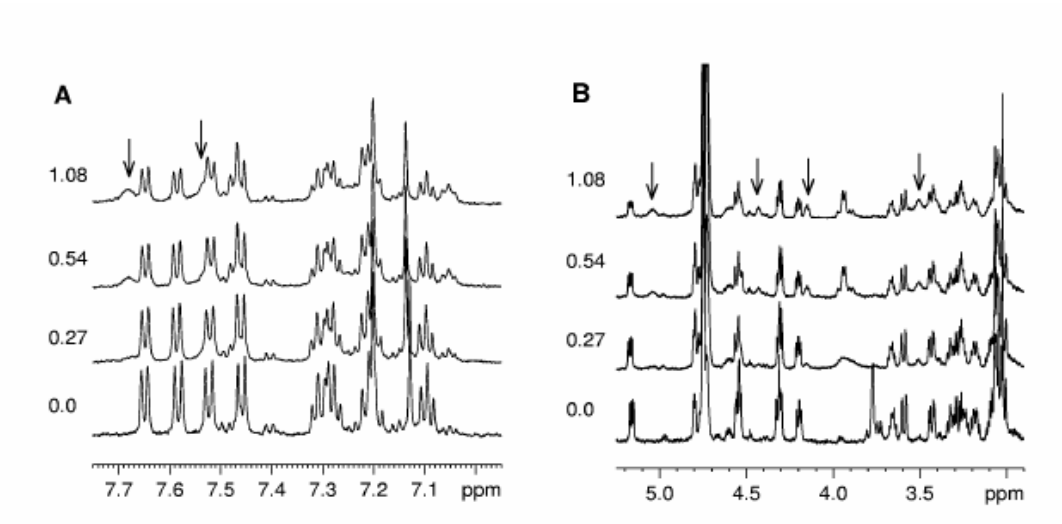

Figure 102. Titration of Contryphan-Vn with $CuCl_2$ monitored by ^1H-NMR
The molar ratio CuCl2/Contryphan-Vn is reported on the left side of each spectrum. The arrows indicates the new signals probably belonging to a Cu^{2+}/peptide complex. **A)** 7.7-7.0 ppm region of 1H spectrum; **B)** 5.2-2.9 ppm region of 1H spectrum.

Only the α-CH2 signal of Gly at 3.78 ppm disappeared due to the large line broadening. The increase of the Cu^{2+} concentration resulted in a further decreasing of amino acid signal intensities along with a partial signal broadening (fig 102). Moreover, new signals, well evident at the 1:1.08 petide:Cu^{2+} ratio appeared together with α-CH2 signal of Gly at 3.94 ppm. The new signals probably belong to a Cu^{2+}/peptide complex. Substantial signal broadening due to fast paramagnetic relaxation impeded the application of common 2D NMR experiments for spectral assignment and structure determination of this Cu^{2+}/peptide complex. The NMR data suggest that the interaction between native Contryphan-Vn and Cu^{2+} is characterized by low affinity since only a partial decrease of the native peptide signals was observed even at 1:1 peptide:Cu^{2+} ratio.

Amino acid	Atom	Chemical shift (ppm)				
		Cα	Cβ,β'	Cγ,γ'	Cδ	Others
Gly1	^{1}H	3.090				
	^{13}C	44.9				
His2	^{1}H	4.646	3.190; 3.095		7.004	7.713 (Hε)
	^{13}C	54.7	29.7		118.8	137.5 (Cε)
Cys3	^{1}H	4.806	3.089; 2.623			
	^{13}C	55.0	38.9			
Pro4	^{1}H	4.429	2.091; 1.324	1.600 0.782	3.240; 3.383	
	^{13}C	61.3	33.3	22.3	48.3	
D-Trp5	^{1}H	5.117	3.459; 3.097		7.253	7.642 (Hε); 7.253 (Hζ3); 7.147 (Hη); 7.506 (Hζ2)
	^{13}C	54.2	28.9		126.2	119.9 (Cε); 123.0 (Cζ3); 120.2 (Cη); 113.1 (Cζ2)
His6	^{1}H	4.947	3.147; 2.793		7.018	7.434 (Hε)
	^{13}C	49.2	34.3		118.6	137.3 (Cε)
Pro7	^{1}H	4.308	2.445; 1.965	2.043	3.826; 3.275	
	^{13}C	63.8	30.5	25.9	49.5	
His8	^{1}H	4.615	3.094; 3.030		6.944	7.683 (Hε)
	^{13}C	54.4	29.7		118.6	137.3 (Cε)
Cys9	^{1}H	4.321	3.104			
	^{13}C	55.2	39.0			
His10	^{1}H	4.730	3.325; 3.155		6.891	7.571 (Hε)
	^{13}C		28.3		117.0	137.3 (Cε)

Table 10. Assignments of ^{1}H and ^{13}C resonances of Cupryphan in D2O at 300K, pH 8.0

The copper interaction with Cupryphan was studied by the addition of different aliquots of $CuCl_2$ to a 2.7×10^{-4} M solution of Cupryphan. In order to study this interaction a complete NMR assignment of Cupryphan was performed. The ^{1}H and ^{13}C NMR spectral assignment of Cupryphan, obtained by means of 2D experiment, is reported in table 10. As in the case of Contryphan-Vn (Eliseo et al.,2004) Pro4 residue in Cupryphan has the *cis* conformation, while Pro7 was found in the trans conformation.

The first $CuCl_2$ addition (peptide:Cu^{2+} ratio 1:0.25) caused a drastic broadening of Hδ and Hε proton signals of the imidazole rings of all four histidine residues: the signal half width increased by factor of four or six depending on the signal (fig. 103). On the other hand, signals of Trp and other amino acid residues remained unchanged or only slightly broadened. Stepwise addition of $CuCl_2$ led to a further broadening of His signals and

to a decrease of the intensity of other ^{1}H amino acid signals without any appreciable broadening, similarly to what observed for Contryphan-Vn titration (fig. 102)

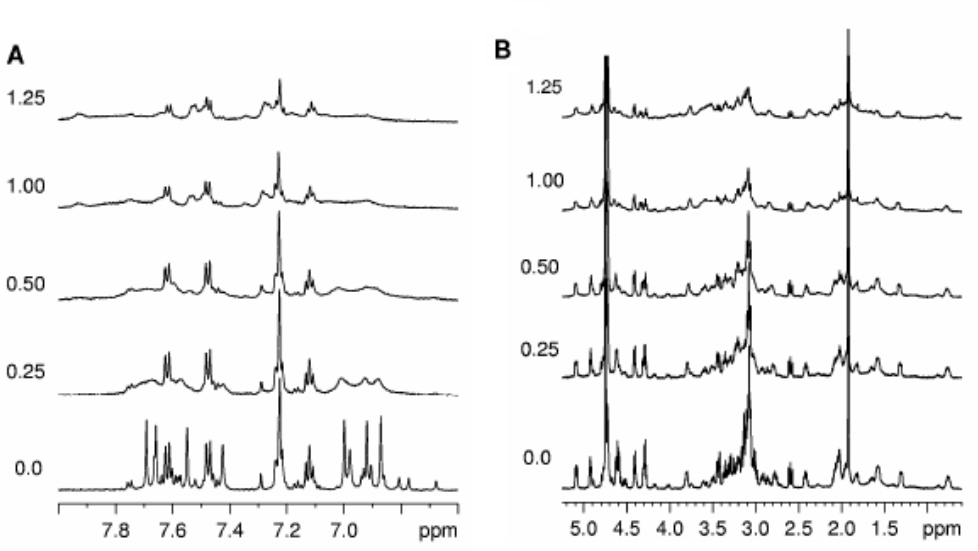

Figure 103. Titration of Cupryphan with $CuCl_2$ monitored by ^{1}H NMR
The molar ratio$CuCl_2$/Cupryphan is reported on the left side of each spectrum. A) 8.0-6.6 ppm region of 1H spectrum; B) 5.2-0.5 ppm region of 1H spectrum.

These results suggest that at low Cu^{2+}:Cupryphan ratios (<1) a peptide-Cu^{2+} complex is formed with a specific interaction between the metal and all the four His residues of the peptide as indicated by the broadening of histidine signals. The relatively fast ligand exchange in the copper coordination sphere (Gaggelli et al., 2005) allows the interaction between Cu^{2+} and other amino acid residues especially at high Cu^{2+}:Cupryphan ratios (>1) when the specific binding is already saturated. As in the case of Contryphan-Vn, this non-specific interaction is manifested by the stepwise intensity decrease of all the amino acid signals upon the addition of $CuCl_2$ to the peptide solution .

The selectivity of Cupryphan for copper was further assessed by NMR titration of the apo form of Cupryphan with Zn^{2+}. Stepwise additions of $ZnCl_2$ to a solution of Cupryphan (0.94 mM) give rise to a decrease of ^{1}H NMR signals of Cupryphan and a simultaneous increase of a new set of signals that can be ascribed to a Cupryphan-Zn^{2+} complex (fig. 104). The ^{1}H and ^{13}C NMR assignment of Cupryphan-Zn complex was obtained by means of 2D NMR experiments.

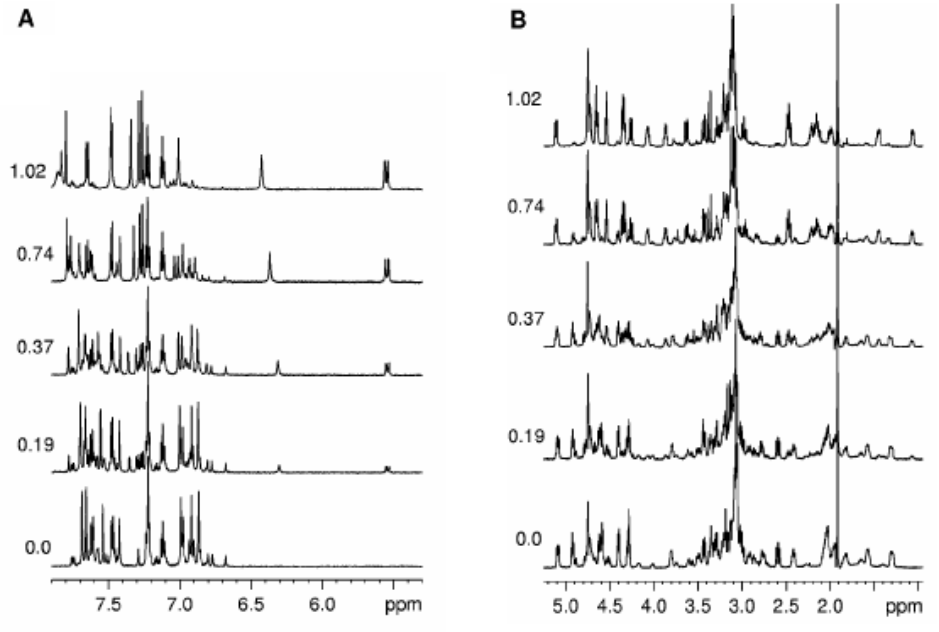

Figure 104. Titration of Cupryphan with $ZnCl_2$ monitored by ^1H NMR
The molar ratio $ZnCl_2$/Cupryphan is reported on the left side of each spectrum. A) 8.0-5.3 ppm region of ^1H spectrum; B) 5.2-1.0 ppm region of ^1H spectrum.

The ^1H NMR titration allows an estimation of the Kd, obtained by a non linear fitting of the experimental data relative to Cupryphan/Zn-Cupryphan ratios (see Methods) at different total Zn^{2+} concentrations. A Kd value of 9 (± 4) × 10^{-6} M was obtained: almost two orders of magnitude higher than that determined for copper binding to Cupryphan (see fluorescence quenching experiments), confirming the selectivity of Cupryphan for copper with respect to other divalent cations.

In order to obtain some structural information on the Cupryphan-Zn^{2+} complex, a ROESY experiment was performed. However, the NH protons signals of all amino acid residues are rather broad due to the relatively high exchange rate with H2O/HDO at the experimental pH value and peptide concentration. Consequently, the ROESY correlations between NH and side chain protons of amino acids were not observed, making impossible a valid structural analysis of peptide conformation based solely on NMR data.

2.3.5 Design and characterization of the Arg-Cupryphan variant

In the attempt to improve the metal binding affinity of Cupryphan, a second variant was designed in which an Arg residue was inserted between Cys9 and His10 residues of Cupryphan (fig. 96, C). The rationale behind this choice was to confer higher flexibility to the C-terminal portion of the peptide allowing an easier access of His10 to the copper site while lowering at the same time the pKa values of the His residues by a general electrostatic effect of the Arg guanidinium group. The three-dimensional structure of this second peptide was also modelled and the stability of the copper binding site confirmed by energy minimization (fig. 97, B).

The novel peptide, named Arg-Cupryphan, was characterized by fluorescence and NMR spectroscopy. In detail, fluorescence quenching experiments confirmed binding of copper with a 1:1 stoichiometry and a Kd = 1.0 (± 0.4) × 10^{-7} M, a value slightly lower than that determined for Cupryphan. A slightly different copper coordination environment of Arg-Cupryphan with respect to Cupryphan was also confirmed by a 20 nm blue shift of the optical absorption peak in the visible region (centred at 560 nm as compared to 580 nm for Cupryphan) and by a two-fold increase of the molar extinction coefficient (ε_{560} = 100 $M^{-1}cm^{-1}$ as compared to ε_{580} = 50 $M^{-1}cm^{-1}$ for

Cupryphan). The ^{1}H NMR experiments performed on Arg-Cupryphan (6.8×10^{-4} M) upon addition of different aliquots of $CuCl_2$ showed the same changes observed for Cupryphan, i.e. a significant broadening of Hδ and Hε signals of all four histidine residues and a progressive decrease of the intensity of the other ^{1}H amino acid signals (data not shown).

2.3.6 Determination of superoxide dismutase activity of Cupryphans

The possibility that the copper ion bound to Cupryphans could be reversibly reduced, a prerequisite for any copper mediated catalytic activity, was tested by assaying the ability of Cupryphan to dismutate superoxide anions. Superoxide dismutase activity of Cupryphans was determined using the pyrogallol enzymatic assay. A solution of superoxide dismutase 0.1 μM in 20 mM Tris HCl buffer pH 8.2 was used as a standard, and a solution of Cupryphan or Arg-Cupryphan 6.8×10^{-4} M in 50 mM sodium acetate buffer pH 6.5 was used to determine Cupryphans superoxide dismutase activity.

The rate of autoxidation of pyrogallol was recorded in the presence of superoxide dismutase, reaching a 50% inhibition of the reaction rate after addition of superoxide dismutase 2.15×10^{-9} M. The same experiment was performed in the presence of Cupryphan, reaching 50% inhibition

at a peptide concentration of 8.5×10^{-5} M. Taking as a reference the enzymatic activity of superoxide dismutase (3.92×109 M^{-1} s^{-1}) (Rotilio et al., 1972), and from the concentration of Cupryphan required to attain 50% inhibition of pyrogallol autoxidation, a superoxide dismutase activity rate for Cupryphan of approx. 1×10^{5} M^{-1} s^{-1} was calculated (table 11). Surprisingly, Arg-Cupryphan superoxide dismutase activity was much lower than that of Cupryphan (approx. one order of magnitude lower (table 11). A possible explanation is that Arg10 exerts a "disturbing" effect on the electrostatic field generated by the copper ion, which leads to a lower productivity of the superoxide-copper encounter rate, a position specific effect well known to occur in superoxide dismutase charge mutants (Polticelli et al., 1995, 1998). Nonetheless both peptides were found to be redox active, highlighting the ability of the engineered copper sites to reversibly cycle between oxidized and reduced redox states..

	[SOD]	[Cupryphan]	[Arg-Cupryphan]
50% inhibition	2.15×10^{-9} M	8.5×10^{-5} M	8.0×10^{-4} M
K_1, K_2	3.92×10^{9} M^{-1} s^{-1}	1.0×10^{5} M^{-1} s^{-1}	9.4×10^{3} M^{-1} s^{-1}

Table 11 Superoxide dismutase activity of Cupryphans

3 CONCLUSIONS

3.1 MULTISPECTROSCOPIC CHARACTERIZATION OF DRUG SUBSTANCE

The primary aim of this study was not to investigate the drug substance *per se*, but rather to develop a database of spectra upon which correlations and SuperSpectra (combining all spectroscopic techniques) could be built. Nonetheless, given the huge amount of data collected, it is worthwhile reflect on the properties of the drug substance observed.

3.1.1 Concentrated stock solutions

Firstly, the batch to batch variance of the original concentrated stock solutions was very small, other than concentrations difference. Beyond these concentration difference, each of the spectroscopic employed showed the batches to be very similar to each other. Potentially, the most significant results is the observation of anomalous absorption in the visible wavelength region, which correlate with the batch production date and changes in hydrolysates used. Anomalous fluorescence in the visible wavelength region was also observed, which too clustered according the production date and hydrolysates. Substantial fluorescence was also observed during Raman scattering spectroscopy, which varied across samples, that also cluster according to production date and hydrolysates.

Consequently, at least three of the spectroscopic techniques used gave spectra that correlate with the history of the batches. The anomalous absorption and fluorescence signals would appear to derive from more than one moiety and may be related to chromophoric impurities, post-translational modifications (PTM) or post-PTM transformations yielding chromophoric components. In addition, NUV Circular Dichroism and FTIR ATR spectroscopies gave indications that there may be a subtle variations in the aromatic residue environment, disulphide geometry and/or structure. It is to be stressed that all the variations observed were small in magnitude, although detectable with high quality instrumentation.

3.1.2 Ambient stored diluted solutions.

The diluted solutions were deliberately prepared to encourage deterioration in the samples, although each solution was handled and prepared identically in so doing, a goal that was achieved.

The main outcomes were the increase in the pH on dilution and subsequent storage which, in part, led to a change in the spectra of the solutions compared to the concentrated stock they derived from. Evidence of scattering from aggregates was obtained from many of the spectroscopies. Likewise, many of the techniques indicated a variance in the quantity of protein present. Notably, for all excepting batch A87G2016, UV absorption spectra clustered according to the history of the batch in terms of production date and hydrolysates used. It is therefore feasible that the deteriorations of diluted samples at ambient temperature may be linked in part to the subtle differences of the original batches. Again, NUV Circular Dichroism indicated that there may be a subtle variations in the aromatic and disulphide environments between samples.

3.1.3 Individual spectroscopies

All the spectroscopies performed as might be expected, with their own well known advantages and disadvantages. All gave some useful information regarding the nature and variability of the batches.

Near-UV Absorption. Near-UV absorption proved a useful means of assessing concentration, providing the samples had maintained integrity, otherwise the absorbance is misleading. The ability to obtain reliable UV-absorption spectra on protein concentrations of ca. 200 mg/ml concentration, while challenging was achieved with consistency. The subtle differences in spectral shape as well as magnitude identified by UV-Absorbance spectroscopy, while apparently real and reproducible, are small and may not well-determined by spectrometers of lesser performance than that employed here. Nonetheless, it is potentially able to assess the role of hydrolysates in the follow-through PTM's, impurities and extraneous chromophores.

Visible Absorption. Visible absorption would normally be expected to reveal little for pure (glyco)proteins as, in principle, they should have no visible chromophores. However, the appearance of a "yellow protein" phenomenon is common in biotechnology products, especially at high concentrations. The source of this colouration is many–fold and no one contributing factor can be assigned to it. However, it is clear that visible absorption spectroscopy has identified components that are visible chromophores and are able to contribute to this phenomenon. Moreover, for the batches studied, the visible chromophoric components appear to correlate with the production date and change in hydrolysates used. Again, the subtle differences in spectral shape as well as magnitude identified by UV-Absorbance spectroscopy, while apparently real and reproducible, are small and may not well-determined by spectrometers of lesser performance than that employed here.

Near-IR Absorption. NIR absorption proved to be intriguing. The presence of water in the solutions masks much of the NIR wavelength region due to its massive absorption, thus limiting the use of this technique. However, a series of small features were apparent when the drug substance was present, but not otherwise, although they were of similar absolute magnitude irrespective of whether drug substance was concentrated or diluted. Moreover, these varied with whether or not the samples had already been exposed to UV-Vis-NIR light for dilute drug substance. The nature of these features is under further investigation to establish if they are artefacts or of true molecular origin.

Far-UV circular dichroism. Far-UV CD demonstrated the consistency of the secondary structure of drug substance in dilute solution, even if hindered by the excessive absorbance of the buffer at wavelengths below 200 nm. For the purpose of this study, spectra were acquired to illustrate instrumental factors as noise, which was indeed evident in the spectra. With more accumulations, or appropriate smoothing, the signal-to-noise ratio could be improved, but the wavelength limitation would remain. Due to the massive absorbance of the peptidic backbone in solution of greater than 200 mg/ml, far-UV CD was not an option for the original concentrated stock solutions.

Near-UV Circular dichroism. Near-UV CD was, however, achievable on both concentrated and diluted solutions. Reproducible spectra were obtained in both cases. The technique is capable to identifying subtle changes in the environments of aromatic residues and disulphide linkages within a protein.

FT-IR Attenuated Total Reflection. FTIR-ATR is potentially an information rich technique which is sensitive to virtually all molecules entities of interest in biopharmaceuticals, including both peptidic and carbohydrate components of glycoproteins. As expected, due to the excessive absorption of water, portions of the wavenumber range are nonetheless inaccessible. However, considerable detail in the Amide I and II regions were accessible. These regions potentially indicated subtle differences between batches. In addition, FTIR-ATR is sensitive to surface phenomena, such as binding of protein on the ATR crystal; variation in this was observed for the batches, suggesting that not all batches interact with their surrounding equivalently.

Fluorescence mapping. Fluorescence mapping proved to be an extremely powerful technique, both for concentrated and dilute solutions. For concentrated solutions, visible fluorescence mapping was able to establish the presence of visible fluorescence components that varied across the batches and was correlated with the history of the batch. Due to self-masking effects of concentrated solutions, the study of near-UV tryptophan fluorescence for concentrated solutions is not feasible. However, for dilute solutions, near-UV fluorescence characterised variations in tryptophan emissions which, in turn may be related to structural change and

accessibility of the protein. By taking advantage of the second-order scattering sensitivity of fluorescence mapping, the Rayleigh and Raman scattering of dilute solutions was amenable to characterisation.

Raman scattering. Raman scattering was hindered by massive fluorescence background from the original concentrated stock solutions. However, spectra were acquired on evaporated samples. These still exhibited massive fluorescence, which was found to vary with batches. Correcting for this fluorescence, Raman may be a potentially powerful technique for samples where fluorescence can be ameliorated.

Concluding, the database of spectra collected, using a variety of techniques, provides an extensive catalogue of information on which to base further SuperSpectra development. Some twelve variants of five spectroscopies have been employed, covering the complete range from far-ultra violet to infrared and involving phenomena including absorption, fluorescence, Raman scattering, Rayleigh scattering and circular dichroism. In total, 1533 spectra have been acquired on this study, amounting to some 750000 data points for multivariate data analysis. Both concentrated (ca. 200 mg/ml) stock and deliberately deteriorated dilute (ca. 0.6 mg/ml) solutions were investigated. All of the techniques employed have yielded useful data of some form in terms of identifying variance in the batches and are potentially complementary and cross-supporting.

Each set of spectra was subjected to multivariate data analysis, primarily Principal Components Analysis, to highlight patterns and differences between batches. such analysis highlighted an apparent connection between the spectra and the history of the batch regarding production date and /or hydrolysate used. In particular, a series of anomalous absorptions in the visible region, together with potentially related fluorescent species, were identified. These may derive from contaminants, post translational modifications (PTM) or post-PTM transformations dependent on production conditions.

3.2 QUANTIFICATION OF DEGRADATION PRODUCTS BY USING CD AND IR IN COMBINATION WITH MVDA

The feasibility of employing PLS/O-PLS analysis to extract quantitative data for common protein degradation forms was successfully demonstrated for an Fc fusion protein. Although both CD and IR spectra contained the relevant information, CD-based models achieved a higher accuracy compared to that of IR-based models for predicting aggregate and oxidation levels, while the accuracy for free Fc levels could be equally well predicted. Combining CD and IR data improved the accuracy of the prediction for all degradation forms.

CD analysis. We have demonstrated that CD can be used to predict some protein degradation forms in combination with multivariate analysis. In fact, PLS model based on near-UV region, gave, for all the three degradation forms lower RMSEP values, so a better prediction, than those obtained using PLS model based on near and far-UV region combined or on far-UV region. Furthermore, these prediction values have been further improved by applying O-PLS regression to the spectra. As for PLS analysis, using only the near-UV region, we obtained, for all the three degradation forms, lower RMSEP values than those obtained using near and far-UV region combined. Furthermore, comparing the lower RMSEP values obtained with the PLS and O-PLS models (fig 90, B), it's clear how using O-PLS we obtained an increasing of the predictive power of the model. These data suggest that, in the determination of all the three degradation forms, near-UV based models perform better than near/far-UV combined based model.

FT-IR analysis. Also FT-IR spectra were subjected to PLS/O-PLS analysis. Were generated two different PLS/O-PLS models, one using the 1700-900 cm^{-1} IR region, the other generated using only the so-called "Amide region" from 1700 to 1500 cm^{-1} . Regarding PLS, the RMSEP values for the two models generated indicated that the amide region-based model yielded to a slightly better prediction for all the three degradation forms compared to those obtained with the whole region-based model. As in CD based models, the better prediction was obtained, for levels of free-Fc, a lower prediction accuracy for aggregates, while for oxidation levels the RMSEP values are higher. In general, it can be affirmed that the prediction values of the two IR-based PLS model are all quite comparable.

The RMSEP values for the O-PLS models showed that, as in the PLS models, the amide region-based model yielded to a lightly better prediction for all the three degradation forms compared to the whole region-based models. As in CD based models, the better prediction was obtained, for levels of free-Fc, a lower prediction accuracy for aggregates, while for oxidation levels RMSEP values are much higher respect to the others. Comparing the PLS and O-PLS models based on the amide region we can see that the results are quite comparable. In conclusion, in the prediction of all the three degradation forms, the amide region based PLS/O-PLS models perform better than the model based on the whole spectral region. However, applying O-PLS instead of PLS to the IR spectra doesn't yield to a great advantage in terms of quality of the prediction as when it has been applied to the CD-based models. Furthermore, in the determination of the degradation forms studied, the obtained accuracy of CD-based O-PLS models was higher than the accuracy obtained with IR-based O-PLS models.

CD and IR combined. As already mentioned, in addition to the PLS/O-PLS models generated using CD and IR spectra separately, models based on CD and IR data combined were generated, in order to enrich the information content of the (combined) spectrum by combining signals that were generated from measuring different physical parameters, and hence increase the accuracy of prediction. Both for PLS and O-PLS models, were used different combinations of the IR and CD spectral range and for each of these combination were generated s PLS/O-PLS model. In the PLS model based on CD (near +far) and IR (900-1700 cm^{-1}) combined data, we had an increase of the prediction accuracy for all the three degradation forms compared to the accuracy obtained with IR-based PLS model, while in comparison with the CD-based model, the new results can be considered very comparable, with a little improvement of prediction accuracy.

Further improvement of prediction was obtained combining near-UV CD region and IR amide region in a single model. RMSEP values of this new-PLS model are better not only compared with the IR-model results, but also in comparison with RMSEP values of CD-based model, confirming that these two are the most useful spectral region to predict the content of protein degradation forms. As in the previous models, the most reliable prediction was obtained for levels of free-Fc, a lower prediction quality was found for aggregate levels, while for oxidation levels the RMSEP values ranged from 4,21 to 5.17 %.

As for PLS regression, also the O-PLS models of CD and IR data combined yielded to an increase of prediction accuracy for all the degradation forms comparing to the results obtained with the single models. This increase was obtained both for the "all range" model and for the model based on near-UV and IR amide region combined. Once again, the lower RMSEP values were obtained with near-UV/IR amide region O-PLS model, as for PLS combined model. As expected, the better prediction accuracy was obtained for free-Fc levels, with a RMSEP values of 0.92 %. A slightly higher value was obtained for aggregates, but still lower than 3 % (2.4%), and for oxidized forms the obtained RMSEP value was 3.32 %, showing the major improvement compared to the values obtained using CD and IR separately. This data further confirm that combining CD and IR data in a single model produce a real increase of the prediction accuracy of all the three degradation forms being analyzed and that this improvement is even greater when are combined not the all spectral range of each technique, but only the CD near-UV region and the IR amide region. Furthermore, employing O-PLS instead of PLS further reduce the RMSEP values obtained, improving even more the predictive power of the multivariate analysis.

Compared to conventional, discretely performed chromatographic assays, the proposed approach has the potential to considerably shorten the duration required to obtain information about product degradation even though the accuracies obtained during this preliminary study are inferior compared to those commonly achieved by chromatographic QC assays. Nevertheless, the required accuracy of an assay depends on the intended application. For example, in a process setting (development or control), speed and indications about trends may be more important and therefore the requirements in terms of accuracy may not be as demanding as for a final product release assay.

With regard to PAT, the proposed concept represents still a lower level of sophistication characterized by at-line measurements with sampling and manual intervention. Further work will aim at the refinement of the PLS model, so as to increase the accuracy of the prediction. This may be achieved by focusing on the most significant, i.e. the most predictive, regions of the spectra, which may in turn also shorten the time required for spectra acquisition or, in addition, may lead to the design of online sensors. Furthermore, appending additional spectral data, e.g. Raman or Raman optical activity, may further enhance the obtained specificity and accuracy.

3.3 ENGINEERING OF METAL-BINDING PROTEINS BASED ON CONOPEPTIDES SCAFFOLD

The increasing knowledge of the principles underlying protein structure and function and the need of novel environment friendly biocatalysts for biotechnological purposes has fuelled in the last few years the search for stable scaffolds that can be engineered to obtain (macro)molecules with given structural and/or functional properties. In this regard, conopeptides represent robust scaffolds highly tolerant to sequence mutation, as they have been developed by Conus species through a sort of combinatorial chemistry in order to target in highly specific and efficient manner the ion channels and receptors of their natural prey. In fact conopeptide scaffolds are rigid and stable, in order to bind with high affinity complementary molecular surfaces on their targets, and are tolerant to multiple sequence mutations, as they have been adapted during cone snails evolution to tolerate hypervariability of the intercysteine loops in order to bind with high specificity a huge variety ofdifferent functional targets.

We tested the potential of conopeptides as scaffolds for the engineering of novel, metal based, biocatalysts starting from the simplest prototype of disulphide constrained conopeptides: the Contryphans. The results of the present work indicate that indeed this class of peptides can be successfully exploited to engineer novel, stable and redox active macromolecules. In fact, Cupryphans bind copper in a stable manner and with a fairly low dissociation constant. In addition, the engineered metal center is able to catalyze redox reactions, as demonstrated by Cupryphans superoxide dismutase activity. A point that deserves further investigation is represented by the determination of the precise copper coordination environment at an atomic level. NMR data indicate that all the four His residues of the peptides coordinate the copper ion. However, the paramagnetic nature of copper doesn't allow the determination of the structure of Cupryphans in solution, which would greatly help in designing derivatives with increased metal affinity. Further, use of a zinc derivative of Cupryphan in NMR ROESY experiments did not yield the cross-peaks required for solution structure determination, making X-ray analysis the only choice for the atomic structure determination of Cupryphan.

As a final remark, the peptidic nature of Cupryphan, together with the ease of chemical synthesis makes it a promising starting point for the development of more efficient superoxidedismutase mimics and antioxidant molecules as a natural follow up of the present work.

REFERENCES

Abdi H. (2007). Partial least suares (PLS) regression *(www.utd.edu/~herve/Abdi-PLSR2007-pretty.pdf).*

Adam G.I.R., Sanders R., Jonsson J. (1999) The development of pharmacogenomic models to predict drug responses. *Trends in Biotechnology,* **19**, 277-316.

Bax, A., Davis, D. G. (1985) Practical aspects of two-dimensional transverse NOE spectroscopy. *J. Magn. Reson*. **63**, 207-213.

Bierau H., Barba M., Kornmann H., Giartosio C.E. and Jone C. (2007) Rapid methods to assess protein quality as an approach to PAT. *American Pharmaceutical Review,* **10,** (7): 17-22.

Braun, S., Kalinowski, H. O., and Berger, S. (1998) 2D NMR Spectroscopy with Field Gradients, in *150 and More Basic NMR Experiments*. 2nd ed., 476-528, Wiley-VCH, Weinheim, Germany.

Brereton R.G. (1997) Multilevel multifactor designs for multivariate calibration. *Analyst,* **122,** 1521-1529.

Brereton R.G., Munoz J.A. (1998) Partial factorial designs for multivariate calibration: extension to seven levels and comparison of strategy. *Chemom. Intell. Lab. Syst.,* **43,** 89-105.

Brereton R.G. (2000) Introduction to multivariate calibration in analytical chemistry. *Analyst,* **125,** 2125-2154.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus M. (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem*. **4**, 187-217.

Brown P.J. (1982) Multivariate calibration. *J. R. Statistic. Soc.* B, **44,** 287-321.

Bryce, G.F., and Gurd, F.R.N. (1965) Visible spectra and optical rotatory properties of cupric ion complexes of L-histidine-containing peptides. *J. Biol. Chem*. **241**, 122-129.

Chau F.T., Liang Z.Y., Gao J., Shao X.G. (2004) Chemometrics: from basics to wavelet transform. *Wiley & Sons Inc.*

Coligan J.E., Dunn B.E., Ploegh H.L., Speicher D.W. and Wingfield P.T. (1995) Current protocols in protein science. *Wiley & Sons Inc.*

Cupane A., Leone M., Militello V., Stroppolo M. E., Polticelli F. and Desideri A. (1994) Lowtemperature optical spectroscopy of native and azide-reacted bovine Cu,Zn superoxide dismutase. Astructural dynamics study. *Biochemistry* **33**, 15103-15109.

Eriksson L., Kettaneh-Wold N., Trygg J., Wilkstrom C., and Wold S. (2006) Multi- and megavariate data analysis, basic principles and application, 2nd edition, *Umetrics AB, Umea, Sweden.*

Eliseo, T., Cicero, D.O., Romeo, C., Schininà, M.E., Raibaudy Massilia, G.R., Polticelli, F.,Ascenzi, P., and Paci, M. (2004) Solution structure of the cyclic peptide contryphan-Vn, a Ca2+-dependent K+ channel modulator. *Biopolymers* **74**, 189-198.

Food and Drug Administration. (2004) Guidance for Industry-PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance", *(http://www.fda.gov/cvm/guidance/6419fnl.pdf).*

Gaggelli, E., Kozlowski, H., Valensin, D., and Valensin. G. (2005) NMR studies on Cu(II)- peptide complexes: exchange kinetics and determination of structures in solution. *Mol. BioSyst.* **1**, 79-84.

Gill S.C. and von Hippel P. H. (1989) Calculation of protein extinction coefficients from aminoacid sequence data. *Anal. Biochem*. **182**, 319-326.

Glusker J.P. (1991) Structural aspects of metal liganding to functional groups in proteins. *Adv. Protein Chem*., **42**, 1-44.

Gueron, M., Plateau, P., Decorps, M. (1991) Solvent signal suppression in NMR *Prog. NMR Spectrosc.* **23**, 135-209.

Guex, N., and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Haaland D.M., Thomas E.V. (1988) Partial least squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.,* **60,** 1193-1202.

Hammes G.G. (2005) Spectroscopy for the biological sciences, *Wiley & Sons Inc*.

Haris P.I. and Chapman D. (1995) The conformational analyses of peptides using Fourier transform IR spectroscopy. *Biopolymers,* **37**, (4), 251-263.

Harms P., Kostov Y. and Rao G. (2002) Bioprocess monitoring. *Curr. Opin. Biotecnol.,* **13,** 124-127.

Hellinga H. W (1996) Metalloprotein design. *Curr. Opin. Biotecnol*., **7**, 437-441.

Jorgensen W.L., Chandrasekhar, J., Madura, J., Impley, R.W., and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. **79**, 926–935.

Kelly S.M., Jess T.J. and Price N.C. (2005) How to study proteins by circular dichroism. *Biochim. Biophys. Acta,* **1751**,: 119-139.

Kitagawa T. and Hirota S. (2002) Raman spectroscopy of proteins. *Wiley & Sons Inc*.

Krell T., Horsburg M.J., Cooper A., Kelly S.M., Coggins J.R. (1996) Localization of the active site of type II dehydroquinases. Identification of a common arginine-containing motif in the two classes of dehydroquines. *J. Biol. Chem*., **271**,: 24492-24497

Kshirsagar, A. M. (1972) Multivariate Analysis. *Journal of Classification,* **10**,: 25-49. *Marcel Dekker, Inc., New York*.

Lu H.S., Chang D., Philo J.S., Zhang K., Narhi L.O., Liu N., Zhang M., Sun J., Wen J., Yanagihara D., Karunagaran D., Yarden Y., Ratzkin B. (1995) Studies on the structure and function of glycosylated and non-glycosylated neu differentiation factors. *J Biological Chemistry,* **270**, (9), 4784-4791.

Marvin J.S.,Corcoran E.E., Hattangadi N.A., Zhang J.V., Gere S.A., Hellinga H.W. (1997) The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl. Acad. Sci. U.S.A*., **94**, 4366-4371.

MacKerell, A.D. Jr, Bashford, D., Bellott, M., Dunbrack, R.L. Jr, Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E. III, Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modelling and dynamics studies of proteins. *J. Phys. Chem*. B **102**, 3586–3616.

Marklund, S., and Marklund, G. (1974) Involvement of the superoxide anion radical in the autoxidation of pyrogallol and a convenient assay for superoxide dismutase. *Eur. J. Biochem.* **47**, 469-474.

Munck L., (2007). A new holistic exploratory approach to Systems Biology by Near Infrared Spectroscopy evaluated by chemometrics and data inspection. *J. Chemom.,* **21,** 406–426.

New Technologies catalogue, the how and why of each technique. *Industria Farmaceutica SERONO S.p.A.*

Nicholls A., Sharp K.A., and Honig B (1991) Protein folding and association insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, **11**, 281-296.

Nielsen O.F. Raman spectroscopy. Methods for structural analysis of protein pharmaceuticals, 167-195.

Oberg K.A., Ruysschaert J.M., Goormaghtigh E. (2004) The optimisation of protein secondary structure determination with infrared and circular dichroism spectra. *Eur. J. Biochem*., **271,** 2937-2948.

Pallaghy P.K. e Norton R.S. (2000) The cyclic contryphan motif CPxXPXC, a robust scaffold potentially useful as an ω-conotoxin mimic. *Biopolymers*, **54**, 173-179.

Pelton J.T. and McLean L.R. (2000). Spectroscopic methods for analysis of protein secondary structure. *Anal. Biochem.,* **277,** 167-176.

Petty S.A., Adalsteinsson T., Decatur S.M. (2005) Correlations among morphology, B-sheet stability and molecular structure in prion peptide aggregates. *Biochemistry*, **44,** 4720-4726.

Polticelli, F., Bottaro, G., Battistoni, A., Carrì, M.T., Djinovic-Carugo, K., Bolognesi, M.,

O'Neill, P., Rotilio, G., Desideri, A. (1995) Modulation of the catalytic rate of Cu,Zn superoxide dismutase in single and double mutants of conserved positively and negatively charged residues. *Biochemistry* **34**, 6043-6049.

Polticelli, F., Battistoni., A., O'Neill, P., Rotilio, G., Desideri, A. (1998) Role of the electrostatic loop charged residues in Cu,Zn superoxide dismutase. *Protein Sci*. **7**, 2354-2358.

Pribic R. (1994) Principal component analysis of Fourier transform infrared and/or circular dichroism spectra of proteins applied in a calibration of protein secondary structure. *Anal. Biochem.,* **23,** 26-34.

Rahmelow K. and Hubner W. (1996) Secondary structure determination of proteins in aqueous solution by infrared spectroscopy: A comparison of multivariate data analysis methods. *Anal. Biochem*., **241**, (1), 5-13.

Regan L. (1993) The design of metal binding sites in proteins. *Annu. Rev. Biophys*. *Biomol. Struct*., **22**, 257-281.

Regan L. (1995) Protein design: novel metal-binding sites. *Trends Biochem. Sci*., **20**, 280-285.

Rotilio G., Bray R., and Fielden E.M. (1972) A pulse radiolysis of superoxide dismutase. *Biochim. Biophys. Acta* **286**, 605-609

Sarver R.W. and Krueger W.C. (1991) An infrared and circular dichroism combined approach to the analysis of protein secondary structure. *Anal. Biochem.,* **199,** 61-67.

Schugerl K. (2001) Progress in monitoring, modeling and control of bioprocesses during the last 20 years. *J. Biotechnology,* **85,** 149-173.

Street A.G. and Mayo S.L. (1999) Computational protein design. *Structure*, **7,** R105-R109.

Tainer J.A., Getzoff E.D., Beem K.M., Richardson J.J., and Richardson D.C. (1982) Determination and analysis of the 2-Å-structure of copper,zinc superoxide dismutase. *J. Mol. Biol*.**160**, 181-217.

Trygg J., Wold S. (2002) Orthogonal projections to latent structures (OPLS). *Journal of Chemometrics* **16**, 119–128.

Vaidyanathan S., Macaloney S., Vaughan J., Brian McNeil B., Harvey L.M. (1999) Monitoring of submerged bioprocessing. *Critical Review in Biotechnology*, **17,** 30-34

Vita C., Roumestand C., Toma F., Mènez A. (1995) Scorpion toxins as natural scaffolds for protein engineering . *Proc. Natl. Acad. Sci.,* **92,** 6404-6408.

Whelehan O.P., Earll M.E., Johansson E., Toft M., Eriksson L. (2006). Detection of varian cancer using chemometric analysis of proteomic profiles. *Chemometrics and Intelligent Laboratory Systems* **84,** 82–87.

I would like to thank Horst, for his invaluable support, and for have being not only a supervisor during this three years;
George and Delphine for all that I learnt from them, and for the great time spent in Oxford;
All my MerckSerono's friends (not colleagues!!): David, Rossella, Francesca, Erika, Alessandra, Sabrina, Anna, Irene, Cristoforo, Giordana, Aniello, Diego, Walter, Luigi, Katia, Carmelina, Laura… for the great fun!
It has been a pleasure to work with Mara, a great person!
The last but not the least, Fabio, my real guide in these journey, and my lovely parents…mum and dad!!