Investigating Protein Modifications Using Vibrational Spectroscopy and Fluorescence Spectroscopy

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences.

2012

Victoria Louise Brewster

School of Chemistry

Contents

Index of Figures.

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Index of Tables

Index of Equations

List of Abbreviations

ALS- Asymmetric Least Squares

ANN- Artificial Neural Network

AUC- Analytical Ultra-Centrifugation

CCD- Charge Coupled Device

CD- Circular Dichroism

CG-MALS- Composition Gradient-Multi Angle Light Scattering

CHO- Chinese Hamster Ovary

CID- Collision Induced Dissociation

Con A- Concanavalin A

Cyt C- Cytochrome C

DFA- Discriminant Function Analysis

DLS- Dynamic Light Scattering

DNA- Deoxyribonucleic Acid

DSC- Differntial Scanning Calorimetry

DTGS- Deuterated Triglycine Sulfate

EMSC- Extended Multiplicative Scatter Correction

ESI-MS- Electrospray Ionisation-Mass Spectrometry

FCS- Fluorescence Correlation Spectroscopy

FDA- Food and Drug Adminsitration

FRET- Fluorescence Resonance Energy Transfer

FT-IR- Fourier Transform Infrared

GalNAc- N-Acetyl Galactosamine

GC-MS- Gas Chromatography Mass Spectrometry

GFP- Green Fluorescent Protein

GlcNAc- N-Acetyl Glucosamine

GuHCl- Guanidine Hydrochloride

HTS- High Throughput System

HbA- Haemoglobin A

HbS- Haemoglobin S (Sickle Cell Haemoglobin)

HCP- Host Cell Protein

ICH- International Conference on Harmonisation of Technical Requirements for Registartion of Pharmaceuticals for Human Use.

IgG- Immunoglobulin G

IR- Infrared

LC-MS- Liquid Chromatography-Mass Spectrometry

LV- Latent Variable

MALDI-MS- Matrix Assisted Laser Desorption Ionisation-Mass Spectrometry

MALS- Multi-Angle Light Scattering

Man- Mannose

MBPCA- Multiblock Principal Components Analysis

MCA- Micro-Cuvette Array

mmTf- Mono-mannosylated transferrin

MS- Mass Spectrometry

MVA- Multivariate anaysis

NIPALS- Non-Iterative Partial Least Squares

omTf- Olig-mannosylated Transferrin

PARAFAC- Parallel Factor Analysis

PAT- Process Analytical Technologies

PBS- Phosphate Buffered Saline

PC- Principal Component

PCA- Principal Components Analysis

PDB- Protein Data Bank

PLS-DA- Partial Least Squares- Discriminate Analysis

PLSR- Partial Least Squares Regression

PMF- Peptide Mass Fingerprint

PNGase- Peptide N-Glycosidase

PTM- Post Translational Modification

Q-TOF-MS- Quadrupole-Time of Flight-Mass Spectrometry

RER- Rough Endoplasmic Reticulum

RF- Random Forest

ROA- Raman Optical Activity

RMS- Root Mean Square

RNA- Ribonucleic Acid

RNase- Ribonuclease

Sav-Gol- Savitzky-Golay

SD- Standard Deviation

SEC- Size Exclusion Chromatography

SER(R)S-Surface Enhance (Resonance) Raman Spectroscopy

SLS- Static Light Scattering

SVM- Support Vector Machine

TEV- Total Explained Variance

Tf- Transferrin

TFMS- Trifluromethansulfonic acid

TOF- Time-of-Fight

UV- Ultraviolet

UVRR- Ultraviolet Resonance Raman

2D- Two Dimensional

3D- Three Dimensional

Abstract

Protein based biopharmaceuticals are becoming increasingly popular therapeutic agents. Recent changes to the legislation governing stem cell technologies will allow many further developments in this field. Characterisation of these therapeutic proteins poses numerous analytical challenges. In this work we address several of the key characterisation problems; detecting glycosylation, monitoring conformational changes, and identifying contamination, using vibrational spectroscopy. Raman and infrared spectroscopies are ideal techniques for the *in situ* monitoring of bioprocesses as they are non-destructive, inexpensive, rapid and quantitative.

We unequivocally demonstrate that Raman spectroscopy is capable of detecting glycosylation in three independent systems; ribonuclease (a model system), transferrin (a recombinant biopharmaceutical product), and GFP (a synthetically glycosylated system). Raman data, coupled with multivariate analysis, have allowed the discrimination of a glycoprotein and the equivalent protein, deglycosylated forms of the glycoprotein, and also different glycoforms of a glycoprotein. Further to this, through the use of PLSR, we have achieved quantification of glycosylation in a mixture of protein and glycoprotein. We have shown that the vibrational modes which are discriminatory in the monitoring of glycosylation are relatively consistent over the three systems investigated and that these bands always include vibrations assigned to structural changes in the protein, and sugar vibrations that are arising from the glycan component.

The sensitivity of Raman bands arising from vibrations of the protein backbone to changes in conformation is evident throughout the work presented in this thesis. We used these vibrations, specifically in the amide I region, to monitor chemically induced protein unfolding. By comparing these results to fluorescence spectroscopy and other regions of the Raman spectrum we have shown that this new method provides improved sensitivity to small structural changes.

Finally, FT-IR spectroscopy, in tandem with supervised machine learning methods, has been applied to the detection of protein based contaminants in biopharmaceutical products. We present a high throughput vibrational spectroscopic method which, when combined with appropriate chemometric modelling, is able to reliably classify pure proteins and proteins ‘spiked’ with a protein contaminant, in some cases at contaminant concentrations as low as 0.25%.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables(“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.

Acknowledgements

My First and biggest thanks go to my supervisor, Roy Goodacre, for his constant support, guidance and encouragement. His understanding and compassion during difficult times were invaluable. I would also like to express my sincere thanks to all members of the Goodacre group, in particular Lorna Ashton and Elon Correa, for their assistance, advice and friendship. A special thank you must go to David Cowcher, whose scientific knowledge, kindness, and ability to make caffeinated beverages and procure cake based snacks has been greatly appreciated.

I would like to thank the BBSRC and Avacta for funding this work. Thanks go to Avacta Analytical for use of instrumentation and to the staff of Avacta, namely, Simon Webster, Charlotte Dodd and Graham Spence, for help and advice. I'd especially like to thank Simon for his ideas and encouragement during this work. I'm grateful to those who kindly provided samples, without which this project would not have been possible. Thanks to Malcolm Saxton of Novozymes, Andrew Martin and Sabine Flitsch of the University of Manchester and Chris Jones of NIBSC. Other collaborators to be thanked include, Zahra Hamrang, Alan Dixon, Ewan Blanch, Christian Johannsen, Dorota Roberts and Sarah Newton.

I would like to thank all of my family and friends, particularly my parents for all their support over the last four years. Finally, thank you to Daniel Blenkinsop for proof reading, and more importantly, being a constant source of inspiration and support. Your care and patience has kept me going through difficult times, both professional and personal, and I can honestly say I could not have done this without you.

Chapter 1: Introduction.

1.1 Background.

Biotechnology is becoming an increasingly important area of the pharmaceutical market, generating an annual revenue in excess of 50 billion dollars (Greer, 2008). A wide variety of biopharmaceuticals can be produced from various cell lines including: yeast, bacteria, mammalian and plant cells. These biological pharmaceuticals can range from small molecules, such as antibiotics, to large protein therapeutics. One of the earliest examples of protein biopharmaceuticals was the use of recombinant DNA technology to modify *Escherichia coli* for the production of human insulin, which was closely followed by the development of human growth hormone and human blood clotting factor (Goddard, 1991). There are now over 50 therapeutic proteins approved by the FDA, the most routinely prescribed being erythropoietin, used to treat anaemia, and an additional 500 proteins are under development (Greer, 2008). A major area of the biopharmaceutical market is the production of antibodies, a specific group of proteins which aid the immune system, for example in fighting bacteria and viruses. More recently, legislation controlling stem cell technologies has been relaxed, allowing further developments of biopharmaceuticals.

Effective monitoring of the complex biological systems used to produce biopharmaceuticals is essential both for product yield optimisation and quality assurance purposes. There is therefore a pressing need to develop robust process analytical technologies (PATs) for this purpose. This project will be concerned with developing optical spectroscopies, in particular Raman spectroscopy, as a non-invasive, rapid method of characterising proteins. This work will focus on detecting structural changes such as glycosylation and aggregation, which may occur in the time frame between protein translation, secretion, downstream recovery and administration.

Raman spectroscopy is a particularly suitable method for on-line analyses of bioprocesses, since it is non-destructive, inexpensive, rapid and quantitative. The con-

focal nature of the technique means that one can focus through a window in a fermenter, thus obviating the need for introducing additional probes into the bioreactor. The utility of Raman spectroscopic instrumentation is greatly increased by the ability to interface with microscopes for trace analysis, fibre optic probes for *in situ* identification and also other types of analytical instrumentation for sample clean up, such as liquid chromatography. The versatility of Raman spectroscopy is also increased by the lack of sample preparation required and the ability to profile samples through a variety of transparent materials. Unlike mid infrared spectroscopy, Raman spectroscopy provides the capability to analyse compounds in aqueous solution with minimal interference from water absorption, which is a critical factor when considering biological applications. The combination of all of the above properties makes Raman spectroscopy a very promising technique for analysis of biopharmaceuticals and the rapid monitoring of complex bioprocesses.

1.2 Aims and Objectives.

The major aim of this project is to develop the application of Raman spectroscopy for monitoring bioprocesses, concentrating in particular on structural changes of proteins such as post translational modifications. The hypothesis that Raman spectroscopy with appropriate chemometrics can be used to detect and quantify glycosylation in recombinant proteins will be tested using the approaches outlined below. Raman spectroscopy will be used initially to differentiate between protein and glycoprotein standards and glycoproteins which have been deglycosylated. In this cognate approach, measurements from various spectroscopies will be used to detect and quantify glycosylation, identify the structural changes brought about by glycosylation, and investigate how these changes affect stability and aggregation.

Spectroscopic data will be benchmarked against gold standard techniques such as matrix assisted laser desorption ionisation-mass spectrometry (MALDI-MS), to determine if the protein is glycosylated or not, fluorescence spectroscopy, to provide information on

structural changes of proteins, and static light scattering (SLS), to monitor protein aggregation.

In order to extract meaningful information from the data, spectra will be subjected to multivariate data analysis methods. In the first instance, principal components analysis (PCA) will be used to separate glycosylated and deglycosylated spectra. A partial least squares (PLS) model will also be used in the quantification of glycosylation status; assessing how much protein is glycosylated. In addition, the loadings from the PCA and PLS models will allow us to derive the most important features in the Raman spectra without bias, as it is currently unknown which vibrational modes will be the most selective for glycosylated proteins.

1.3 Vibrational spectroscopy.

Raman and infrared spectroscopy are both branches of vibrational spectroscopy. Both techniques involve a sample being illuminated with a radiation source. The photons from this radiation source can interact with the molecules in one of four ways, they can be reflected by the sample, transmitted through the sample, absorbed by the sample (as in infrared spectroscopy) or scattered by the sample (as in Raman spectroscopy).

1.3.1 Principles of vibrational spectroscopy.

1.3.1.1 Molecular vibrations.

If the electrical energy possessed by a molecule remains constant, its energy can be divided into a number of components, called degrees of freedom. The first three degrees of freedom describe the translation of the molecule in space. A further three are used to describe rotational movement, unless the molecule is linear and thus then only has two possible types of rotation, and hence only two vibrational degrees of freedom, since the rotation about the intermolecular axis is not a proper rotation (Banwell and McCash, 2006). So, for linear molecules the number of vibrational degrees of freedom and also the number of possible vibrations is given by 3N-5, where N is the number of atoms in the

molecular. For all molecules which are not linear the number of vibrational degrees of freedom is given by 3N-6 (Smith and Dent, 2005, Atkins, 1998).

When a molecule vibrates there is a combination of changes in the position of atoms within the molecule, called the vibrational coordinate. There are six main types of molecular vibrations; stretching (symmetric or asymmetric), scissoring, rocking, wagging and twisting (Figure 1.1) (Harris and Bertolucci, 1978).

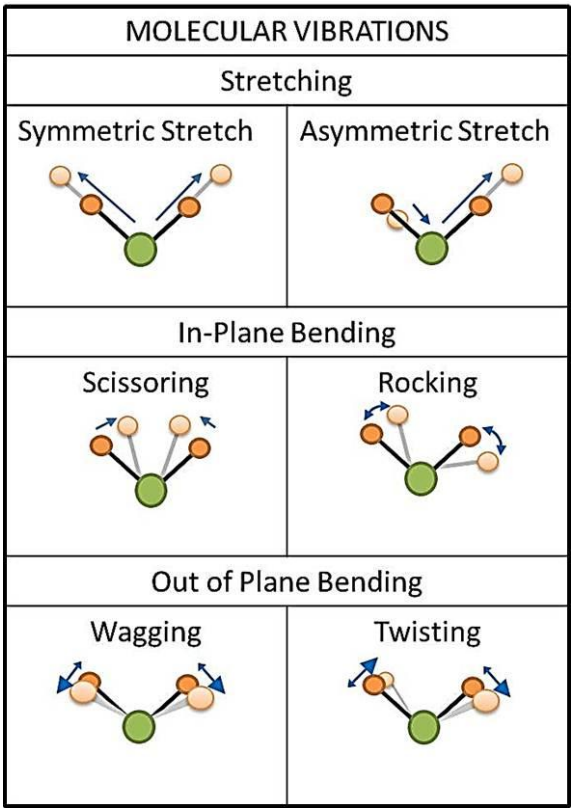

Figure 1.1: Cartoon Diagrams Depicting the Six variations of Molecular Vibrations.

When two or more bonds of similar energies are close together in a molecule, the vibrations can interact. Observed vibrations in these cases, relate to the vibrations of groups of atoms. By contrast, where there is a large energy difference between the vibrations of different bonds and the atoms are well separated, the molecular vibrations of individual atoms will have distinct effects on spectra and can be considered separately.

1.3.1.2 Polarisability.

All electromagnetic radiation including infrared and visible light has electric and magnetic fields associated with it, which will interact with molecules and may change the molecular properties. Some molecules already have an electric dipole moment due to charge separation, and this dipole interacts with the electric field of the incident light. Other types of molecules may acquire a temporary dipole, induced as they vibrate within the electric field. Methane molecules, CH_4, for example, which are normally non-polar may gain a temporary dipole as their C-H bonds stretch. Some molecules neither have nor acquire dipole moments but can still be polarised. This happens because the electron cloud of the molecule is distorted by the electric field. The polarisability (α) of the molecule can be calculated using the equation below (Smith and Dent, 2005):

$$\mu = \alpha E \qquad \textbf{Eq. 1.1}$$

where μ is the induced dipole, E is the electric field of the incident photon and α is the polarisability; a measure of how easily the electron cloud around a molecule can be distorted. As the effect on the electron cloud applies in all directions, the change in the dipole of a molecule can be described in each of the Cartesian co-ordinates, *x, y* and *z*. In order to illustrate the effect of linearly polarised radiation on molecular polarisability, all three co-ordinates must be considered**.** To allow for this, the polarisability components are usually labelled $\alpha_{xx,}$ $\alpha_{yy,}$ α_{yx} etc, where the first subscript character refers to the polarisability of the molecule and the second to the polarisation of the incident light, therefore a more suitable equation is (Long, 2002):

$$\mu_x = \alpha_{xx}E_x + \alpha_{xy}E_y + \alpha_{xz}E_z \qquad \textbf{Eq.1.2}$$

Similar expressions exist for μ_y and μ_z, thus the polarisability of a molecule is best described by the tensor (Long, 2002):

$$\begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix} = \begin{bmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \qquad \textbf{Eq.1.3}$$

Equation 1.3 is valid for linear Raman spectroscopy in which the Raman scattering efficiency is linearly related to the laser power. However, Raman spectroscopy can also be nonlinear, when more than one photon interacts with a molecule at the same time, resulting in a nonlinear relationship between the magnitude of the scattering and the laser power. In this case, the dependence of the polarisability against the electric field strength is better defined as the power series, where *β* and *γ* are the hyper-polarizability:

$$\mu = \alpha E + 1/2\ \beta E^2 + 1/6\ \gamma E^3 \ldots\ldots \qquad \textbf{Eq.1.4}$$

1.3.2 Infrared Spectroscopy.

The basic selection rule of infrared spectroscopy states that for a molecule to absorb infrared radiation, *i.e.* be IR active, there must be a change in the dipole moment of the molecule during a vibration (Banwell and McCash, 2006). A simple example of a molecule which is IR active is carbon monoxide, which has a just one fundamental vibrational frequency of 2168 cm^{-1}. The molecule has a permanent dipole which will change as the molecule stretches, meaning that it will interact with IR radiation to produce an absorption peak close to 2168 cm^{-1} in the IR spectra. Absorption will only occur when the frequency of the incident radiation matches the frequency of the vibration or alternatively when the energy of the incident photon is equal to the energy gap between the ground and excited states of the molecule (Kealey, 2002). The molecule may be promoted to a vibrationally excited state by the incident radiation, causing the molecule to gain energy and the radiation to lose energy; it is this loss in energy that is detected as heat.

1.3.2.1 Fourier Transform – Infrared Spectroscopy.

Fourier transform-infrared spectroscopy (FT-IR) involves a different measurement technique for collecting spectra compared to a dispersive instrument. In traditional infrared spectroscopy, the amount of energy absorbed by a sample at different wavelengths is measured using a polychromatic beam, which changes wavelength over time using a monochromator (Chalmers, 2001). However, a FT instrument allows the measurement of all wavelengths simultaneously using a Michelson interferometer. A Michelson interferometer, depicted in Figure 1.2, splits the incident radiation into two separate beams using a half silvered mirror. One beam is then reflected off a fixed mirror and the other from a movable mirror. The position of the movable mirror can be altered, changing the relative distances travelled by the two beams and, therefore introducing a difference into their journey times. The two beams are then recombined and interference between the two beams allows measurement of their temporal coherence, a measurement of how well the radiation interferes with itself at different points. The light is then passed through a sample where it is absorbed, transmitted, scattered or reflected, the transmitted light is then detected. The signal is measured at various different time delays, produced by altering the position of the moving mirror and then a mathematical Fourier-transform is performed, resulting in spectrum that is similar to the conventional dispersive spectrum of the sample (Chalmers, 2001, Banwell and McCash, 2006).

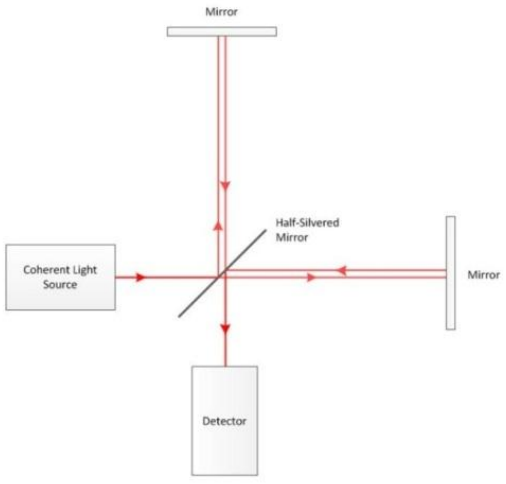

Figure 1.2: Schematic Diagram of Michelson Interferometer.

FT-IR spectroscopy has many advantages over dispersive IR spectroscopy. FT-IR is a much faster technique, as the information at all wavelengths/frequencies is gathered simultaneously. As the time taken to record a spectrum is less, there will be less shot noise and hence a higher signal to noise ratio in an FT instrument, this is known as the Fellgett advantage (Banwell and McCash, 2006). The Jaquinot advantage states that an FT instrument will have a more efficient throughput of radiation than a dispersive instrument; this is because a dispersive instrument focusses light through a slit, thus limiting the total amount of energy passing through the system. By contrast, in an FT instrument the two parallel beams need only be focused at the sample and detector, so all the light will pass through the instrument (Banwell and McCash, 2006, McCreery, 2000).

1.3.3 Raman Spectroscopy.

Raman spectroscopy arises from the inelastic scattering of photons, which was initially proposed theoretically by Smekal in 1923 (Smekal, 1923). The Raman effect was first observed experimentally by Raman and Krishnan in 1928 and was reported in the journal *Nature* the same year (Raman and Krishnan, 1928).

As Raman spectroscopy is concerned with the scattering of light, unlike infrared spectroscopy, there is no need for the photons to have a specific energy which matches the energy gap between the ground and vibrational excited states of the analyte. (Kealey, 2002). In Raman spectroscopy, when a sample is irradiated with monochromatic radiation (in the UV/visible or near infrared regions) the photons will be scattered by the molecules in the sample, causing the energy of the inelastically scattered photons to change by one vibrational unit of energy ($\Delta v = \pm 1$). For a molecule to be Raman active there must be a change in the polarisability (α) of the molecule during the induced molecular vibrations.

The scattering of photons occurs because the electric field associated with the radiation causes the electron cloud around the nuclei of a molecule to become polarised for a short

time, i.e. there is a change in the polarisability of the molecule (Smith and Dent, 2005). This electron cloud distortion causes a type of elastic scattering called Rayleigh scattering which is shown in terms of the energy levels in Figure 1.3. If nuclear motion is also induced during the scattering process, then energy is either transferred to the scattered photon from the molecule or from the incident photon to the molecule, so the energy of the scattered photon is different to that of the incident photon, hence the scattering is inelastic, Raman scattering. Raman scattering is normally very weak and typically only one photon in every 10^6- 10^8 photons are Raman scattered, but this problem is overcome in modern spectroscopic instruments by the use of high powered lasers, sensitive radiation detectors and more efficient filters optics. (Banwell and McCash, 2006).

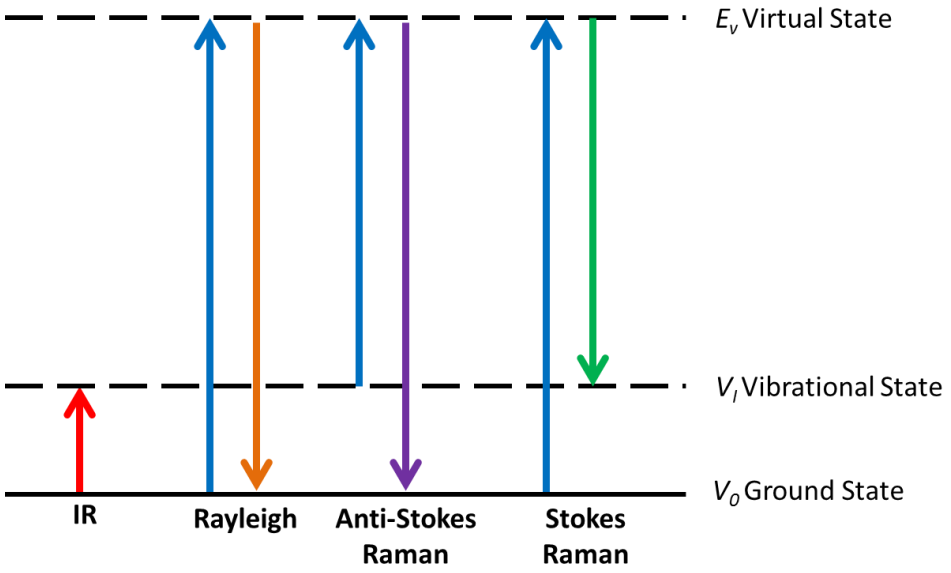

Figure 1.3: Jablonski energy diagram showing the Raman scattering effect, showing Rayleigh (elastic) scattering and both Stokes and anti-Stokes Raman (inelastic) scattering.

There are two different types of Raman scattering, Stokes and anti-Stokes. Stokes scattering occurs when a molecule in its ground vibrational state is promoted to a higher energy excited state by the absorption of energy. Some molecules are already present in an excited state and energy is transferred from the molecule to the photon, thus the molecule is demoted from the higher excited energy state to the ground state, this is known as anti-Stokes scattering. At room temperature, Stokes scattering is more intense

than anti-Stokes scattering, due to the fact that only a small number of molecules are present in an excited state. Typically Raman scattering is recorded between 100 and 4000 cm^{-1} Raman shift from the V_0 line, so only Stokes scattering is observed. Observation of anti-Stokes scattering can be preferred in certain circumstances, such as when there is a large amount of fluorescence present (McCreery, 2000).

If a molecule is promoted to an upper vibrationally excited state by electronic excitation from the laser, the excess energy may be lost through a mechanism known as fluorescence emission (see section 1.4.1). For fluorescence to occur, the incident radiation must populate an excited electronic state. The lower the wavelength of the incident radiation the more energy is put into the system, therefore it is more likely that fluorescence will occur. If fluorescence is observed in a sample, then little or no Raman signal will be observed, this can be overcome by using a longer wavelength laser (or shorter in the case of UV-resonance Raman), photo-bleaching the sample (exposing the sample to the laser for a time period), or collecting the anti-Stokes spectrum.

1.3.3.1 Interpretation of Raman Spectra.

Raman scattering is measured in terms of the shift in energy from the energy of the exciting radiation, which is expressed as either Raman shift/cm^{-1} or wavenumber/cm^{-1}. It is possible to identify frequency ranges for the most common functional groups in which infrared absorption or Raman scattering will occur. According to Hooke's law, vibrational frequency is related to the force constant k, which is dependent on the strength of the bond, and the mass of the atoms attached to the bond. It can therefore be deduced that stronger bonds with lighter atoms will have a higher vibrational frequency, whereas heavy atoms with weak bonds will have relatively low vibrational frequencies.

The intensity of the different bands in a Raman spectrum varies with the nature of the vibration being studied, for example, symmetric vibrations cause the greatest changes in the electron cloud around the molecule and therefore give the greatest Raman scattering (Smith and Dent, 2005). The most environmentally sensitive Raman bands, such as

those from OH and NH vibrations tend to be broad and weak due to hydrogen bonding, whereas bands arising from the structural backbone of a molecule (C-H, C-C) are often strong and sharp (Banwell and McCash, 2006).

Raman scattering occurs because there is a change in the polarisability (α) of a molecule caused by molecular vibrations. This change can be described using the polarisability derivative, $d\alpha/dQ$, where Q is the vibrational normal co-ordinate. The scattering intensity (I) of a Raman active molecule is proportional to the square of $d\alpha/dQ$, or in other words proportional to the square of the induced dipole moment (Banwell and McCash, 2006). Hence if the polarisability of a molecule does not change with the molecular vibration then $d\alpha/dQ = 0$, so the resulting Raman band will have a low intensity.

The intensity of a band in a Raman spectrum can be calculated using the following equation known as the Kramer-Heisenberg-Dirac equation (Smith and Dent, 2005):

$$I = K\, l\, \alpha^2 \omega^4 \qquad \textbf{Eq. 1.5}$$

Where: K is constant which varies between instruments but always contains the speed of light, I is the power of the laser power, α is the polarisability and ω is the frequency of the incident radiation.

1.3.3.1.1 Multivariate Analysis of Vibrational Spectroscopic Data.

Raman spectral data, particularly that collected from biological systems, are often complex and contain many overlapping bands, making the spectra difficult to interpret by visual examination of the data set alone, hence multivariate data analysis strategies are often employed. Multivariate data are data which consists of the results of observations of many different variables (in Raman data, wavenumber shift) for a number of different samples (Brereton, 2005). Each variable may be described as constituting a different dimension. If the number of variables is n, each object can be described as existing at a

unique position referred to as *n*-dimensional hyperspace (Otto, 1999). This dimensional hyperspace can be difficult to visualise, so multivariate analysis can simplify the dimensionality (Ellis and Goodacre, 2006). Further details on the data analysis methods used in this work can be found in Chapter 2, Section 2.3.

1.3.3.2 Variations of Raman spectroscopy.

1.3.3.2.1 Resonance Raman spectroscopy.

Resonance Raman spectroscopy is achieved when a sample is excited with a frequency of light which is within the molecular absorption bands of a chromophore within the sample. This excitation is in resonance with the electronic transitions. Scattering enhancements as high as 10^5 have been observed using this technique (Chi and Asher, 1998, Smith and Dent, 2005). As well as vibrational information, electronic information about a molecule can also be deduced from the intensities of bands in the resonance Raman spectrum and from the energy separations in overtone progression.

1.3.3.2.2 SERS and SERRS.

Surfaced enhanced Raman scattering relies on the analyte being in close proximity to, or adsorbed onto, a roughened metal surface or a colloidal solution, usually of silver or gold. The interaction of the electric field of the incident light with the metal gives rise to a surface plasmon (an oscillating electric field) which is believed to enable an electromagnetic enhancement effect (via surface-plasmon polariton resonances), for analytes adsorbed or in close proximity to the metal surface (Smith and Dent, 2005). Additional signal enhancement is also thought to arise from a chemical or charge transfer effect often associated with adsorption of the analyte onto the metal surface. The overall Raman signal enhancement can be as much as 10^8 over conventional Raman scattering, and even higher (10^{14}) since single molecule spectra have been reported (Nie and Emory, 1997). In addition, further enhancement can be achieved through a combination of the RR effect with SERS, known as surface-enhanced resonance Raman scattering (SERRS), which makes use of reporter dye molecules to allow SERRS to be observed at specific incident wavelengths (Faulds et al., 2008, Graham and Faulds, 2008).

1.3.3.2.3 Raman Optical Activity (ROA).

Another variation on traditional Raman spectroscopy is Raman optical activity (ROA) which uses circularly polarised light to observe the changes in the spectra of optically active, chiral molecules (Zhu et al., 2006). ROA relies on interference between the scattered photons and the optical activity tensors of a chiral molecule. This leads to small differences in the relative intensities of right and left handed circularly polarised scattered light, which will give information on the chirality of the molecule (Barron, 2004).

1.4 Other Analytical Techniques Used in This Thesis.

1.4.1 Fluorescence Spectroscopy.

Fluorescence Spectroscopy, also known as spectrofluorometry, is a branch of spectroscopy which analyses the fluorescence emissions from a sample. A sample is illuminated with a light source, typically UV light, which may promote the molecules in the sample to an electronically excited vibrational state. The excess vibrational energy may be lost by series of intermolecular collisions, known as 'radiationless' transitions, in which the vibrational energy is converted to kinetic energy which appears as heat within the sample. The resultant loss of energy causes the molecule to move to a lower vibrational energy state within that electronic state. When the excited molecule reaches the excited state (E_1) it can emit radiation, known as fluorescence, and return to the ground state (E_0). The emitted

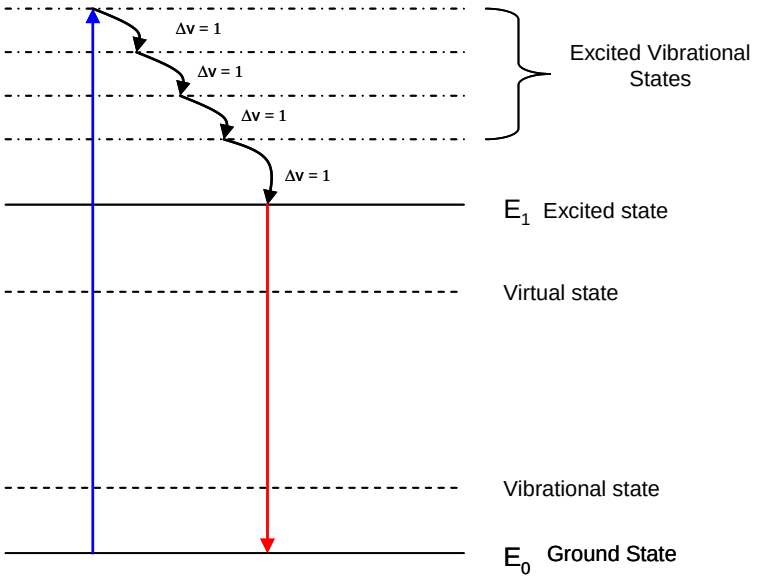

Figure 1.4: Schematic of fluorescence emission.

fluorescence is usually of a lower frequency than the laser used to excite the molecule. By analysing the emission spectrum, structural details of the sample can be determined (Banwell and McCash, 2006, Sharma, 1999).

The fluorescence spectrum of a protein sample is composed of emissions from individual aromatic amino acid residues of tryptophan and tyrosine. There may also be some small contributions from phenylalanine and the disulfide bonds. The fluorescence emission of tryptophan is known as solvatochromatic, meaning the emitted radiation varies in wavelength depending on the solvent environment. Monitoring the fluorescence of a protein is a useful probe of its conformational state. As tryptophan is hydrophobic, it is usually found in the core of a folded protein. When the protein's structure is disordered, the tryptophan becomes exposed to the aqueous atmosphere, and changes in the polarity of the tryptophan's local environment will cause a change in the fluorescence emission spectra (Sharma, 1999, Royer, 2006).

1.4.2 Static Light Scattering.

In static light scattering (SLS) the amount of light scattered by a macromolecule, such as a protein, is measured in order to obtain information on the size of the particles in the sample. In a light scattering experiment, solutions of analyte are illuminated with laser light and a detector is used to measure the scattering intensity, usually at an angle of 90 degrees. In multi-angle light scattering (MALS), numerous detectors are used to measure the intensity of scattered light at many different angles (Wyatt, 1993).

The intensity of scattered light is dependent on the polarisability of the molecules, which is dependent on the molecular weight of the sample. Therefore, larger particles with higher molecular weights will give higher intensities of scattered light. Hence, by simply plotting the intensity of scattered light, it is possible to crudely track whether a sample's average particle size is increasing or decreasing. A more precise indication of particle size is achieved by calculating the weight average molecular weight from the scattering intensity, using Zimm's formula (Wyatt, 1993, Wen et al., 1996):

$$\frac{K\,c}{R\theta} = \frac{1}{M_W P(\theta)} + 2A_2 c \qquad \textbf{Eq.1.6}$$

where: M_w is the weight average molecular weight, c is the sample concentration (in g/mL), A_2 is the second viral coefficient, $R\theta$ is the Rayleigh ratio (excess intensity of scattered light at angle θ), $P(\theta)$ is the angular dependence of scattered light and K is an optical parameter, calculated using equation 1.7, in which n_o is the refractive index of the solvent, dn/dc is the refractive index increment of the solution, N_A is Avagadroes's number (6.023×10^{23}) and λ is the wavelength of the light source.

$$K = 4\Pi^2 n_0^2 (dn/dc)^2 N_A\ \lambda^4 \qquad \textbf{Eq. 1.7}$$

1.5 Protein structure.

The primary level of structure in a protein is the linear sequence of amino acids and any other covalent bonds, such as disulfide bonds. (Stryer et al., 2002). Protein secondary structure involves the folding of regions of the polypeptide chains, most commonly into α-helices and β-sheets. α-helices occur when the amino acids arrange themselves in a regular helical conformation in which the carbonyl oxygen of each peptide bond is hydrogen bonded to the hydrogen on the amino group of the fourth amino acid away (Figure 1.5). By contrast, in a β-sheet, hydrogen bonds form between the peptides bonds either in different polypeptide chains or in different sections of the same polypeptide chains (Figure 1.6).

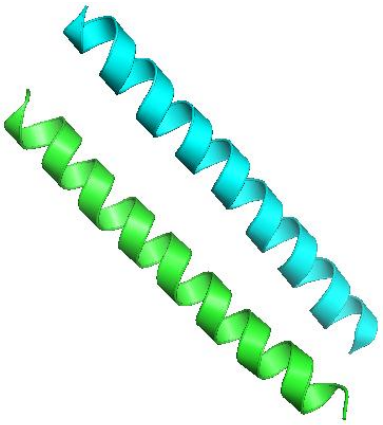

Figure 1.5: Cartoon diagram of α-Helical Structures.

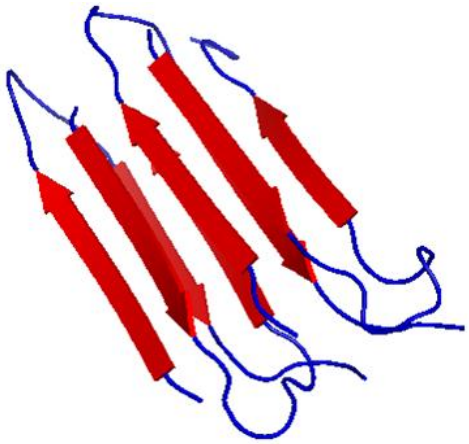

Figure 1.6: Cartoon Diagram of β-Sheet Structures.

The third level of structure found in proteins refers to the spatial arrangement of amino acids that are far apart in the linear sequence as well as those residues that are adjacent. The protein is folded by hydrophobic interactions, and tertiary structure is maintained by electrostatic forces, hydrogen bonding and disulfide bonds (Stryer et al., 2002).

1.5.1 Recombinant Proteins.

A recombinant protein is a protein which is coded for by recombinant DNA; DNA which is engineered by combining or inserting one or more DNA strands into a host. Recombinant DNA technology, through the ability to isolate, analyse and change genes, has now become common place. In recombinant protein expression mRNA is coded for by recombinant DNA, which has been inserted into a host cell. The cell will produce proteins based on this DNA.

Since the development of recombinant DNA technology many therapeutic proteins, which were previously harvested from animals, have become mass produced recombinant biopharmaceuticals. The most notable being human growth hormone, which stimulates growth and cell reproduction; human insulin, a hormone that metabolises glucose and follicle stimulation hormone which regulates reproductive processes (Goddard, 1991). Many other proteins which were not harvestable from animals or humans have since

become available as recombinant protein drugs. Such drugs include, tissue plasminogen activator, a protein which breaks down blood clots, and erythropoietin, a glycoprotein which controls red blood cell production (Goddard, 1991).

1.5.2 Post translational modifications.

After a protein pharmaceutical has been transcribed and translated by the recombinant DNA and mRNA, other chemical modifications may occur, these are known as post translational modifications. These modifications include glycosylation, phosphorylation and methylation, amongst many others (Table 1.1). The majority of biopharmaceuticals will contain a modification of some sort, whether it is intentional, or unintentional. These modifications are important in generating heterogeneity in proteins and in utilizing identical proteins for different cellular functions in different cell types (Greer et al., 2008).

Post translational modifications to a protein drug can affect the stability, immunogenicity (ability of a substance to produce an immune response), pharmacokinetics (the fate of a substance after administration) and efficacy (capacity of a substance to produce an effect). Understanding post translational modifications will not only help resolve these clinical problems but may also help refine the fermentation and purification processes, increasing the yield of the biopharmaceutical product.

Table 1.1: Common post Translational modifications

Addition of a functional Group
Acetylation
Acylation
Alkylation
Amidation (deamidation)
Carbamylation
Carboxylation
Formylation
Glycosylation (N/O linked)
Glycation
Hydrosylation
Lipoylation
Methylation
Phosphorylation (dephosphorylaton)
Sulfation
Methionine oxidation
Addition of other peptides
SUMOylation
Ubiqutinination
Changes to amino acids
Deamidation (glutamine, asparagine)
Citrullination (arginine)
Eliminylation (serine, threonine,cysteine)
Structural changes
Disulfide bridges
Proteolytic cleavage

1.5.2.1 Glycosylation.

Glycosylation is the covalent linking of short chains of carbohydrates to the peptide chain of a protein and is probably the most common of the post translational modifications (Apweiler et al., 1999), in fact over a third of the biotechnology derived proteins are glycosylated in their native form (Greer et al., 2008). It is widely agreed that glycosylation patterns and their heterogeneity are very important to both the biological activity and clinical efficiency of protein therapeutics, most notably erythropoietin and tissue plasminogen activator, in which correct glycosylation is needed for biological activity (Berman, 1985). In addition, glycosylation has been proved to play a major role in the biological activity of recombinant antibodies (Greer, 2007). The carbohydrate moiety of an immunoglobulin molecule can play many roles in the activity of the protein; it may target the molecule to a specific location or perform a structural function such as supporting the three-dimensional shape of the active site. Furthermore, it has been shown recently that protein glycosylation can greatly increase the stability of pharmaceutical proteins both *in vitro* and *in vivo* by increasing the internal non-covalent forces which hold the protein in its folded form, and destabilising the unfolded state of the protein (Sola and Griebenow, 2009). The production of glycosylated biopharmaceuticals by host-expression systems suffers technical difficulties in the form of very low expression yields and glycosylation heterogeneity (Chang et al., 2007).

The covalent bond which forms between a sugar residue and a protein is called a glycosidic bond and can be one of two types O-linked or N-linked. In an O-linked glycoprotein the glycan is attached to the peptide chain through the OH group of serine or threonine side chains (Figure 1.7). N-linked glycans are linked to proteins via a glycosidic bond to the NH_2 group of an asparagine residue but only when it appears in the sequence; asparagine- X- serine or threonine, where X can be any amino acid except proline (Hames and Hooper, 2000) (Figure 1.8).

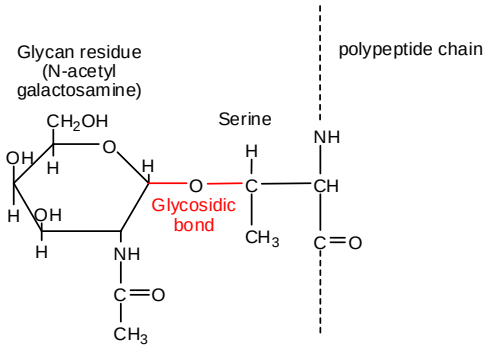

Figure 1.7: O-linked glycosylation.

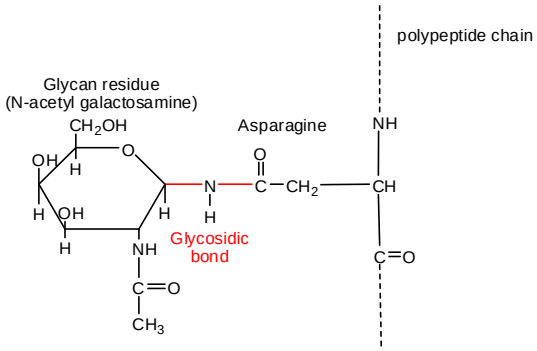

Figure 1.8: N-linked glycosylation.

O-linked glycans are synthesised by the sequential addition of monosaccharide units as it passes through the Golgi complex of the cell. N-acetlygalactosamine (GalNAc) is transferred to serine or theronine amino acids by the protein GalNAc transferase, other monosacharides such as galactose, fructose and N-acetylglucosamine are then added. N-linked glycosylation occurs as the translated protein passes through the rough endoplasmic reticulum (RER). All N-linked glycans are based on a common pentasaccharide core structure which consists of three mannose residues and two GlcNAc residues and an R group, which can be divided into two types; high mannose (where the R group is a variable number of mannose residues) and complex (where the R group can be a variety of number of monosacchrides). In N-linked glycosylation the oligosaccharide is made prior to attachment to the peptide backbone, by adding monosaccherides to a lipid carrier, dolichol phosphate which is attached to the cytosolic face of the RER membrane. Cleavage of the high energy pyrophosphate bond, which links the sugars to the dolichol phosphate provides the energy for transfer of the oligosaccharide to the protein (Hames and Hooper, 2000).

1.5.3 Other structural changes.

As well as post translation modifications, there are other protein structural changes which can affect the efficiency of a protein therapeutic. It is widely known that the hydrophobic forces, hydrogen bonds and disulfide bonds which keep proteins in their tertiary structure play a large role in the activity of the protein, therefore one of the most problematic structural changes in biopharmaceutical proteins is the unfolding of the native protein into

a disordered structure (Konermann, 2004). Protein unfolding can be induced by fluctuations in pH and temperature, or the introduction of a chaotropic agent such as guanidine or urea, all of which cause disruption of the intra-molecular non-covalent forces that stabilise the molecule. The unfolding of a pharmaceutical protein is problematic as it may lead to loss of activity, aggregation and decreased solubility.

Protein aggregation is another common issue encountered in the production of biopharmaceuticals. Aggregation occurs when hydrophobic protein molecules are attracted to each other and bonds form between them, often forming large insoluble particles. In some proteins an associated state is the native form and aggregation is necessary for protein activity. However in the majority of therapeutic proteins aggregation is undesirable as it may affect the immunogenicity of the drug and in larger aggregate particles can lead to an adverse effect upon administration (Cromwell et al., 2006). It is possible to take steps during cell culture, product purification and formulation processes that will minimise the occurrence of protein aggregation or remove aggregates from the final product.

A further problem which may be encountered with protein production is mutagenesis, when a change occurs in the DNA sequence, and therefore a change in the amino acid sequence of a peptide chain. This change in primary structure can lead to the incorrect folding of proteins into their secondary structure, and hence will have an impact on protein activity (Goddard, 1991). In some cases mutagenesis can be intentional in order to improve the stability of the protein, for example, the removal of chemically susceptible amino acids such as asparagine (Grimsley et al., 2009) or the removal of amino acid sequences which are commonly subjected to proteolytic cleavage (Markert et al., 2003).

1.6 Biopharmaceutical characterisation.

Biopharmaceutical characterisation involves gaining a comprehensive understanding of the chemical structure, biological properties, product stability and degradation pathways of the drug (Greer, 2008). Analysis of protein pharmaceuticals is more complex than that of small molecule drugs due to the heterogeneity of the protein, with variations in secondary and tertiary structure and post translational modifications, all of which play a role in activity (Goddard, 1991). In 1999 the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) published guidelines for the characterisation of biopharmaceuticals, which are summarised in Table 1.2 (ICH, 1999).

Table 1.2: Summary of the ICH guidelines for 'specification and acceptance criteria for biotechnological/biological products' (ICH, 1999)

Structural Characterisation	Physiochemical Properties
Amino acid sequence	Molecular weight/size
Amino acid composition	Isoform pattern
Terminal amino acid sequence	Extinction co-efficient
Peptide map	Electrophoresis patterns
Sulfhydryl groups and disulfide bridges	Liquid chromatography patterns
Carbohydrate structure of glycoproteins	Spectroscopic profile

Mass spectrometry (MS) has been routinely applied to the characterisation of biopharmaceuticals of varying size, from small synthetic peptides to conjugated antibody molecules (Greer, 2008). The most commonly utilised forms of mass spectrometry for biopharmaceutical analysis are ESI-MS (electrospray ionisation) and MALDI-MS (matrix assisted laser desorption ionisation), which are capable of ionising and detecting intact proteins of up to 150KDa and 500KDa, respectively (Baldwin, 2005). These methods are useful for initial confirmation of molecular weight to determine whether a drug has the correct anticipated *m/z*, and any mass differences may give an indication as to the type of modifications present. Mass spectrometry has been used to produce peptide maps by initially digesting the whole protein by enzymatic methods and producing a peptide mass fingerprint (PMF) using either MALDI-MS (Padliya and Wood, 2008, Henzel et al., 2003) or ESI-MS (El-Aneed et al., 2009). In cases where the PMF is too complex, LC-MS has

been used to separate peptide fragments (Rapp et al., 2009). MS peptide mass fingerprints do not confirm the actual sequence of amino acid residues, however, sequential ion data can be gained by using ESI-MS/MS (Liu, 2008, Cannon and Jarman, 2003), triple quadrupole instruments (Josephs and Sanders, 2004) or Q-TOF-MS (quadrupole-time of flight-MS) (Mouls et al., 2007).

1.6.1 Post translational modifications.

The ICH guidelines require that all modifications to protein pharmaceuticals be well characterised in terms of structure of the attached functional group, modification site and quantification of modified proteins (ICH, 1999). Mass spectrometry is again the gold standard technique for this purpose (Greer et al., 2008). ESI-MS and MALDI-MS can be used for initial molecular weight measurements to determine if any modifications are present (Liu, 2008, Fenn et al., 1990, Baldwin, 2005). This may also give an indication as to the types of modification present, i.e. a mass difference of 40 Da would suggest acetylation or trimethylation, whereas a mass increase of 1000 Da would suggest the addition of a glycan or another protein. More sophisticated MS technologies such as ESI-CID-MS/MS and Q-TOF-MS have been used in order to determine specific sites of modifications (Liu, 2008, Mouls et al., 2007).

Other, non-MS based methods for characterising post translational modifications include 2D gel electrophoresis based techniques, in which antibodies specific to the modification being studied have been used to identify and quantify modifications (Kaufmann and Fussenegger, 2001). Rajkumar *et al*. demonstrated a 2D gel method for detecting glycosylation and phosphorylation in proteins from epithelial cell membranes using three different fluorescent dyes (Ge et al., 2004). Affinity chromatography based approaches have also been used to identify modifications prior to MS analysis. These approaches include immunoprecipitation with modification-specific antibodies (Sun et al., 2005) and affinity capture using anchor molecules (Elortza et al., 2003). To increase the specificity of these methods, the chemical derivatisation and selective tagging of glycosylation and phosphorylation has also been demonstrated (Lambert et al., 2005).

1.6.1.1 Glycosylation Status.

Glycosylation is the most common (Apweiler et al., 1999) and also the most complex of the post translational modifications (Greer, 2007). This complexity is largely due to the fact that glycoproteins are a mixture of 'glycoforms', meaning the same polypeptide backbone is attached to various different glycans (Taylor, 2006). From a regulatory standpoint, it is accepted that some degree of variation occurs in the carbohydrate moiety of engineered glycoproteins, however this heterogeneity must be consistent between production batches and the content and structure of the glycans must be well characterised (ICH, 1999).

The first glycoprotein primary structure determination by mass spectrometry was carried out over 30 years ago (Morris et al., 1978). Since then, advancing ionisation and mass separation techniques have made MS the primary method for glycoprotein characterisation. Modern soft ionisation methods (ESI and MALDI) have made it possible to analyse intact glycoproteins (Morris, 1980, Fenn et al., 1990). However, full structural characterisation of glycoproteins requires determination of branches, linkages and the determination of same-mass sugar isomers (ICH, 1999), for which GC-MS using electron impact ionisation is still a commonly utilised technique (Dell and Morris, 2001).

Q-TOF instruments have been used extensively in the carbohydrate sequencing of glycans. The glycosidic bonds between sugars vary in strength depending on the monosaccharide units, and therefore some saccharides will fragment more easily than others. This has been exploited in collision activation MS/MS where fragmentation patterns have been used to determine carbohydrate sequences (Teng-umnuay et al., 1998) and also glycosylation sites (Zheng et al., 2009).

Structure determination of glycans released from glycoprotiens (or glycolipids) has been achieved in a qualitative and quantitative fashion using reverse-phase LC-MS (Anumula, 2006). Ion mobility MS has been used to successfully distinguish between structural isomers of the same sugars based on the differences in the 3D shape of the ions

(Clowers et al., 2005, Shvartsburg et al., 2009). Finally. tandem MS with collision induced dissociation (CID) has also been shown to be capable of differentiating between isomeric forms of saccharides (Chai et al., 2001).

1.6.2 Aggregation.

The most routinely used analytical method for the detection and quantification of protein aggregation is size exclusion chromatography (SEC) (Cromwell et al., 2006, Garcia-Fruitos et al., 2011). However, due to problems encountered through non-specific binding to the column, matrix free techniques such as analytical ultra-centrifugation (AUC) and light scattering based approaches are now being favoured (Arakawa et al., 2007). AUC is a method which monitors the concentration distribution of solutes within a centrifuge cell as the rotator is spinning. AUC is able to detect and quantify aggregates, as fractionation is dependent on particle size (Berkowitz, 2006, Gabrielson et al., 2010). Protein aggregates can also be detected using light scattering, as the intensity of the scattered light is proportional to the molecular weight of the protein molecules (Wyatt, 1993). There are a number of light scattering based techniques used in the detection of biopharmaceutical protein aggregation, including static light scattering (SLS) (Roberts et al., 2011, Demeule et al., 2007), dynamic light scattering (DLS) (Wang et al., 2012) and multi-angle light scattering (MALS) (Sahin and Roberts, 2012).

1.6.3 Conformation and Stability.

Many of the methods used to monitor conformation and stability in biopharmaceuticals are spectroscopic techniques. Fluorescence spectroscopy, as described in section 1.4.1, can be used to probe conformational changes by monitoring changes in tryptophan and tyrosine fluorescence (Sharma, 1999, Royer, 2006). In addition to monitoring the native intrinsic fluorescence, fluorescence spectroscopy can be used to track changes in the fluorescence emission of probe dyes which bind to the hydrophobic regions of a protein (Samanta et al., 2011, Bhattacharya et al., 2011). A variation on fluorescence spectroscopy which has also been used for stability studies is fluorescence resonance energy transfer (FRET). FRET utilises variations in fluorescence brought about by energy

transfer between two chromphores to monitor structural changes as a function of physical conditions (Serrano et al., 2012). This method has been particularly useful in conformational studies measuring the distances between the different domains of a multi-domain protein (Truong and Ikura, 2001).

There are many examples of the use of vibrational spectroscopy to detect changes in protein structure, in particular determining changes in the α-helix or β-sheet content of a protein (Wen, 2007, Kong and Yu, 2007). More information on the uses of infrared and Raman spectroscopy in biopharmaceutical analysis is given in section 1.6.4. Circular dichroism (CD) spectroscopy is much more common in therapeutic protein analysis than vibrational spectroscopic methods, and is also used primarily for secondary structure analysis (Li et al., 2011) .

There are a number of non-spectroscopic methods for profiling the stability of proteins. The most notable is differential scanning calorimetery (DSC), a thermo-analytical technique which measures the amount of heat required to increase the temperature of a sample. DSC allows calculations of the transition midpoint temperatures (T_m) for protein unfolding and, along with fluorescence spectroscopy, is one of the gold standard techniques for profiling protein stability (Walters et al., 2009, Konermann, 2004, Bruylants et al., 2005).

1.6.4 Applications Vibrational Spectroscopy in Biotechnology.

Vibrational spectroscopy has been applied to the study of protein conformation both in solution (Tuma, 2005, Barth, 2007) and in solid states (Sane et al., 2004, Kong and Yu, 2007). It is particularly useful for qualitatively and quantitatively assessing secondary structure and providing structural information about proteins with very little secondary order, where methods such as X-ray crystallography and circular dichorism are not very informative.

The use of FT-IR spectroscopy for the classification of protein secondary structure was first demonstrated in the 1950's (Elliot and Ambrose, 1950, Ambrose and Elliot, 1950) and emerged in the late 1980's as a popular tool for the detection of α-helix and β-sheet structures within proteins (Jackson and Mantsch, 1995, Byler and Susi, 1986, Jackson et al., 1989). Since then methods of quantifying the levels of α-helix, β-sheet and disordered structure have been developed by combining chemometrics with IR data (Hering et al., 2002, Yu, 2005). Similar approaches have been established using Raman, SERS and ROA (Das et al., 2011, Barron et al., 2002, Yamamoto, 2012, Shashilov and Lednev, 2010, Oladepo et al., 2012).

Traditionally, interpretation of protein spectra has centred on the analysis of the intensity and position of the amide bands, which arise due to the coupled vibrational modes of the peptide back bone. As these bands tend to be fairly broad, interpretation is usually based on spectral decomposition and deconvolution of the amide bands into its component bands (Sane et al., 1999, Ganim et al., 2008). The main contributor to the amide I band, which occurs at around 1650 cm^{-1}, is the C=O stretching modes from the carbonyl groups of the peptide. As these carbonyl groups act as acceptors of hydrogen bonds in the secondary structure of a protein, the position of the amide I band is strongly dependant on secondary structure. In addition, vibrational coupling between motions of different peptides within an ordered structure will cause excitionic coupling of amide I modes, producing a further spectral shift (Wen, 2007). A further application of deconvolution of the amide I band into its component peaks has been to help predict the conformational changes of tumor necorsis factor in solution (Tuma et al., 1995) and to study the fibrillation in globular proteins, specifically the insulin fibrillation mechanism (Huang et al., 2006). Raman spectroscopic analysis of insulin has also provided information on protein aggregation, showing the transition of the native α-helical structure into β-sheets by reduction of the disulfide bonds (Zheng et al., 2004).

The sensitivity of the amide III band (C-N stretching and N-H bending) to protein structure has been exploited in many studies of protein conformation, using infrared spectroscopy,

conventional Raman and UV resonance Raman (UVRR). Many infrared spectroscopy studies have involved tracking changes in the intensity and frequency of the amide III bands as a way to monitor protein denaturation by chemical and thermal means (Kong and Yu, 2007, Fabian and Mantsch, 1995, From and Bowler, 1998). This approach has also been applied to conformational changes in UVRR spectra of horse myoglobin and apomyoglobin after acid denaturation (Chi and Asher, 1999, Chi and Asher, 1998).

Raman spectroscopy and UVRR have been used to characterise the side chains of pharmaceutical protein molecules. An excellent example of this is the work carried out by Wen *et al*., in this study, Raman spectroscopic techniques were used to study cysteine side chains of the recombinant protein human interleukin-1 receptor antagonist (rhIL-1ra). Raman spectra of rhIL-1ra were able to show that the protein has four cysteine side chains, all in a free sulfhydryl state (Wen et al., 2008).

As previously mentioned, one of the most important areas of biopharmaceutical characterisation is the identification of post translational modifications. Raman spectroscopy has been used in this area to detect and quantify phosphorylation in the protein casein sucessfully, and also to determine the structural changes which occur upon dephosphorylation (Jarvis et al., 2007, Ashton et al., 2011). Raman microscopy has also been used in conjunction with computer modelling methods to determine the glycosylation status of the enzyme fungal β-N-acetylhexosamidas (Ettrich et al., 2007). SERS has been developed for the detection of acetylation, trimethylation, ubiqutination and phosphorylation in a variety of synthetic peptides and biological samples. Attempts were made to combine wavelet decomposition analysis with SERS spectra in order to yield information on the position of these modifications (Sundararajan et al., 2006). Torreggiani *et al.* combined Raman and FT-IR spectroscopies in order to characterise radical base modifications to the methionine residue of proteins (Torreggiani et al., 2011).

Both Raman spectroscopy and ROA have been used to characterise various glycoproteins and carbohydrates (Zhu et al., 2006, Zhu et al., 2005b). Raman

spectroscopy has been used to provide structural information about glycoprotein-C of the herpes simplex virus (Kikuchi et al., 1987) and α_1-acid glycoprotein from blood plasma (Kopecky et al., 2003), and also to monitor the binding of the glycoproteins found in antifreeze (Cui et al., 2005). SERS and ROA have been used to study interactions between pharmaceutical molecules, in particular anti-cancer drugs and P-glycoprotein, which plays a crucial role in mediating the drug resistance of cells (Fleury et al., 1999). Raman spectra of glycoproteins have been used to determine the amount of the monosaccharide *N*-acetylneuraminic (sialic) acid in a glycan; Oleinikov and colleagues showed that the Raman spectra of cell surface glycoproteins displayed variations in the position of the glycerol band between 873-830 cm^{-1}, depending on the amount of sialic acid present (Oleinikov et al., 1998).

Finally, advances in spectroscopic instrumentation and data analysis techniques have facilitated the application of vibrational spectroscopy to monitor complex bioprocesses. One of the first examples of this was in the on-line monitoring of the biotransformation by yeast, of glucose to ethanol, allowing accurate predictions to be made of glucose and ethanol concentrations simultaneously during the fermentation processes, without recourse to time consuming chromatography (Shaw et al., 1999). Following this, a range of fermentation broths containing the fungus *Gibberella fujikuroi* producing gibberellic acid were studied and Raman spectroscopy was able to accurately monitor the progress of the fermentation (McGovern et al., 2002). SERS has also proved to be extremely effective for the off-line monitoring of bioprocesses, enabling the quantification of secondary metabolites present in microbial fermentations (Clarke et al., 2005). FT-IR spectroscopy has been used predict levels of glucose, ethanol and ammonia in *Saccharomyces cerevisiae* culture (Schenk et al., 2007), and more recently has been successfully combined with chemometrics to quantify glucose and lactate in CHO cell cultures and also predict antibody titres in supernatants taken from mammalian cell cultures (Sellick et al., 2010, Rhiel et al., 2002).

Chapter 2: Materials and Methods.

2.1 Instrumentation.

2.1.1 Raman Spectroscopy.

2.1.1.1 Renishaw Raman Microscope.

Unless stated otherwise all Raman data it this thesis were collected using a Renishaw 2000 Raman microscope (Renishaw Plc., Old Town, Wotton-under-Edge, Gloucestershire, U.K.), equipped with a near-infrared 785 nm diode laser and a thermoelectrically cooled charge coupled device (CCD) detector. The spectrometer is coupled to an Olympus microscope with 50x and 20x objectives, providing a spectral footprint of around 2-4 microns. The diffraction grating gives a spectral range of 0-4000 cm^{-1} with a spectral resolution of 6 cm^{-1}. The laser power was approximately 27 mW at source and 2-4 mW at sample. The instrument was wavelength calibrated with a silicon wafer focused under the 50x objective and collected as a static spectrum centred at 520 cm^{-1} with 1s exposure, an offset correction was performed to ensure the position of the silicon band was 520.5 ± 0.1 cm^{-1}. The GRAMS WiRE software package (Galactic Industries Corp., 395 Main St., Salem, NH) running under Windows 95 was used for instrument control and data capture.

2.1.1.1.1 Tienta Spectra RIM™ Slides.

With the exception of Chapter 7, all samples analysed on the Renishaw Raman microscope were prepared on Tienta Spectra RIM™ slides, purchased from from Tienta Sciences Inc. (Tienta Sciences Inc, Indianapolis, IN, USA). These slides have a hydrophobic coating which causes the liquid samples to 'bead up' on the surface, increasing the concentration of protein in the spot. This can increase the Raman intensity by as much as 4 times (Kreimer et al., 2004, Tienta Sciences, 2004a) and can also help to reduce fluorescence emission (Kreimer et al., 2004, Tienta Sciences, 2004b). Samples were prepared by pipetting 2 µL aliquots of aqueous protein solutions onto Tienta Spectra

RIM™ slides and allowing the protein spots to dry at room temperature for approximately 1 h.

As a data quality control measure, the reproducibility of data acquired from these slides was tested. First, spectra were taken from four different protein spots (A-D), at six different positions (1-6) within each spot (shown in Figure 2.1), in order to assess the variation that occurs between and within protein spots. The PCA plot in Figure 2.2 describes the variation in the Raman spectra of the glycoprotein, RNase B collected from different positions within various spots on a Tienta slide. There appears to be no clustering of the data recorded from independent spots or observable trends in relation to sampling position within a spot. This strongly suggesting that the spectral output is independent of the sampling position between and within protein spots.

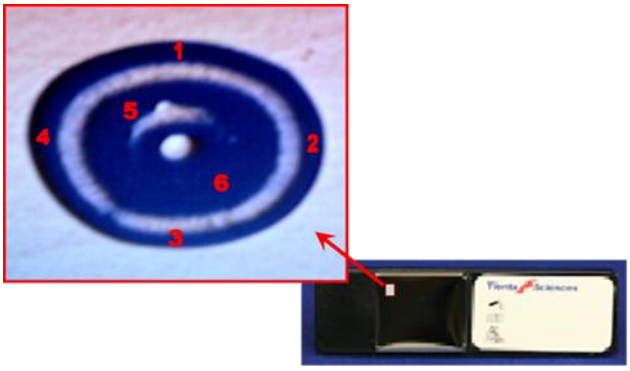

Figure 2.1: Photograph of Tienta Spectra RIM™ slide and 5 x objective microscope image of a protein spot on a Tienta slide (Red numbers indicate points from which measurements were recorded).

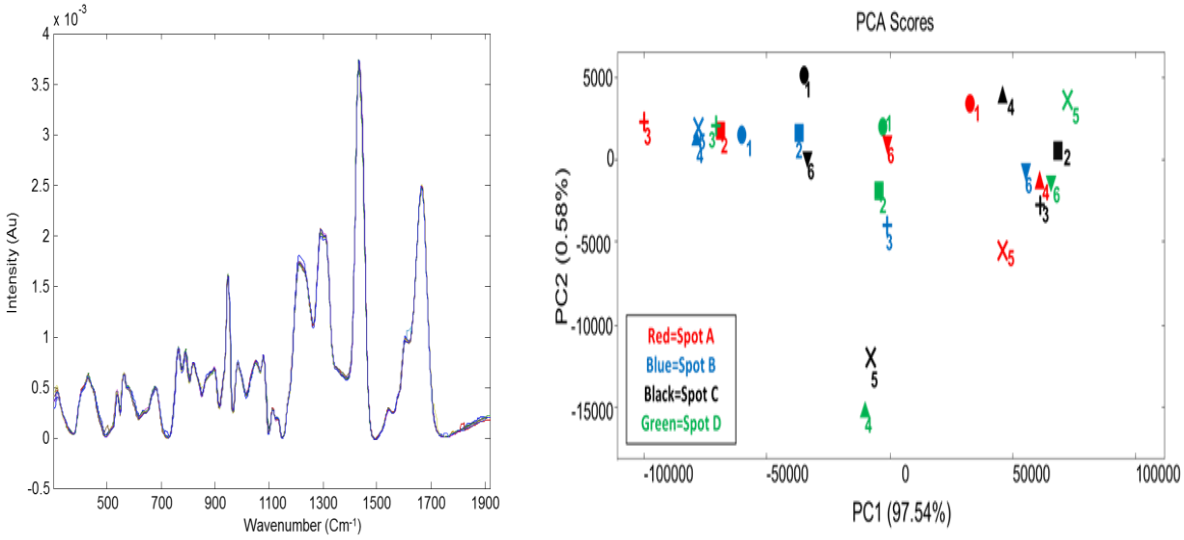

Figure 2.2: Raman Spectra and PCA Scores plot (PC1 *vs* PC2) showing the variation in the Raman spectra of RNase B recorded from six different positions (1-6) on four spots (A-D). (Raman data have been smoothed (Sav-Gol), Baseline corrected (ALS) and column mean centred).

As a further validation step, the reliability of protein spectra collected from Tienta Spectra RIM™ slides was tested by comparing the data collected previously from RNase B spotted onto a Tienta slide to data collected from the lyophilised protein power. Figure 2.3 shows the PCA scores plot for this comparison, in which there is little or no distinction between the two groups of RNase B spectra, indicating that precipitation on a Tienta slide does not affect the Raman spectra of a protein.

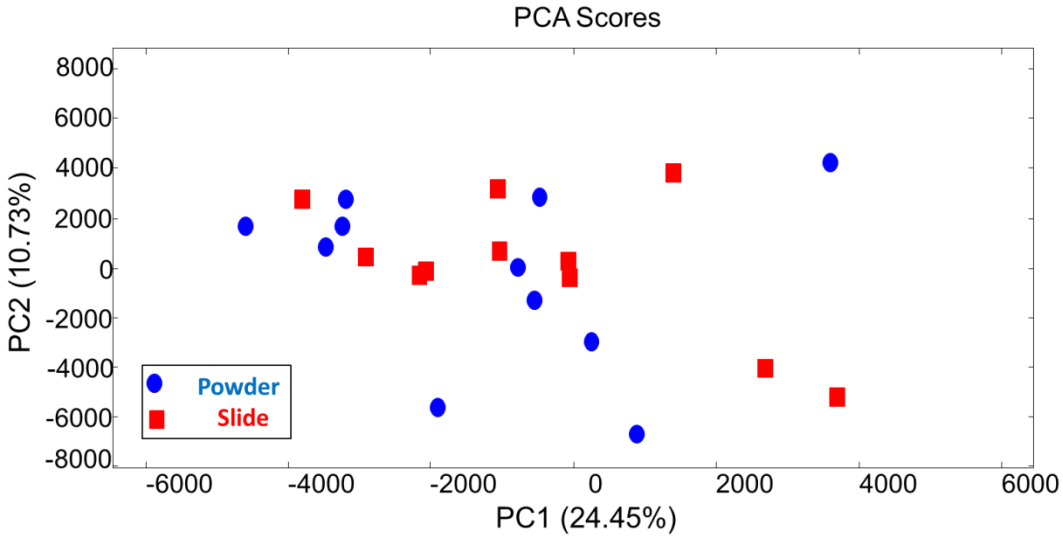

Figure 2.3: PCA Scores plot (PC1 *vs* PC2) showing the variation in the Raman spectra of RNase B recorded from lyophilised powder and a Tienta Spectra RIM™ slide.(Raman data have been smoothed (Sav-Gol), Baseline corrected (ALS) and column mean centred).

2.1.1.2 Biotools SCP ChiralRAMAN Spectrometer.

Raman data in this thesis referred to as collected on a Biotools ChiralRAMAN Spectrometer, were collected on the ROA instrument detailed below, with the spectrometer in Raman only mode. The spectrometer consists of a 532.5 nm laser, which is focussed through a prism to a quarter wave plate and two synchronised counter-rotating half wave plates. The light then passes through an incident shutter, which controls the illumination period and onto the sample solution which is contained in

a quartz cell in a sample holder. Back scattered light is divided by a beam splitter into right and left circularly polarised components and projected onto a CCD camera.

2.1.2 FT-IR Spectroscopy.

FT-IR data were recorded on a Bruker Equinox 55 infrared spectrometer (Bruker ltd, Coventry, UK) equipped with a deuterated triglycine sulfate (DTGS) detector. Opus 4 manufacturer's software operating under MS windows 2000, was used for instrument control and data capture. A microplate module HTS-XT^{TM} was used for high thoughput analysis. Initially a 'blank' spectrum was recorded from a reference well to provide a background spectrum which was used to correct the sample spectrum. Spectra were collected from each of the 96 wells over a wavenumber range of 4000-600 cm^{-1}, with a resolution of 4 cm^{-1}. For each well 64 spectra were collected, co-added and averaged, with an average collection time of 30s per sample. Data were exported from Opus software in to Microsoft Excel.

2.1.3 Avacta Optim 1000.

Optim 1000 is a multi-modal instrument designed by Avacta Analytical (Avacta Analytical, Leeds, UK) which obtains fluorescence and light scattering data simultaneously over a temperature gradient. The instrument set up, depicted in Figure 2.4, includes two lasers operating at 226 and 473 nm and a cooled CCD detector. Samples are held in a micro-cuvette array (MCA) made up of 16 quartz cuvettes which each hold 9µL of sample. A thermo electric plate is used for heating and cooling, with an added aluminium plate to help transfer heat to the samples. The instrument is controlled by Avacta's Optim Client software. Raw data can be exported directly to Excel or transferred to the Optim Analysis software which allows automated primary and secondary analysis of the data to be carried out.

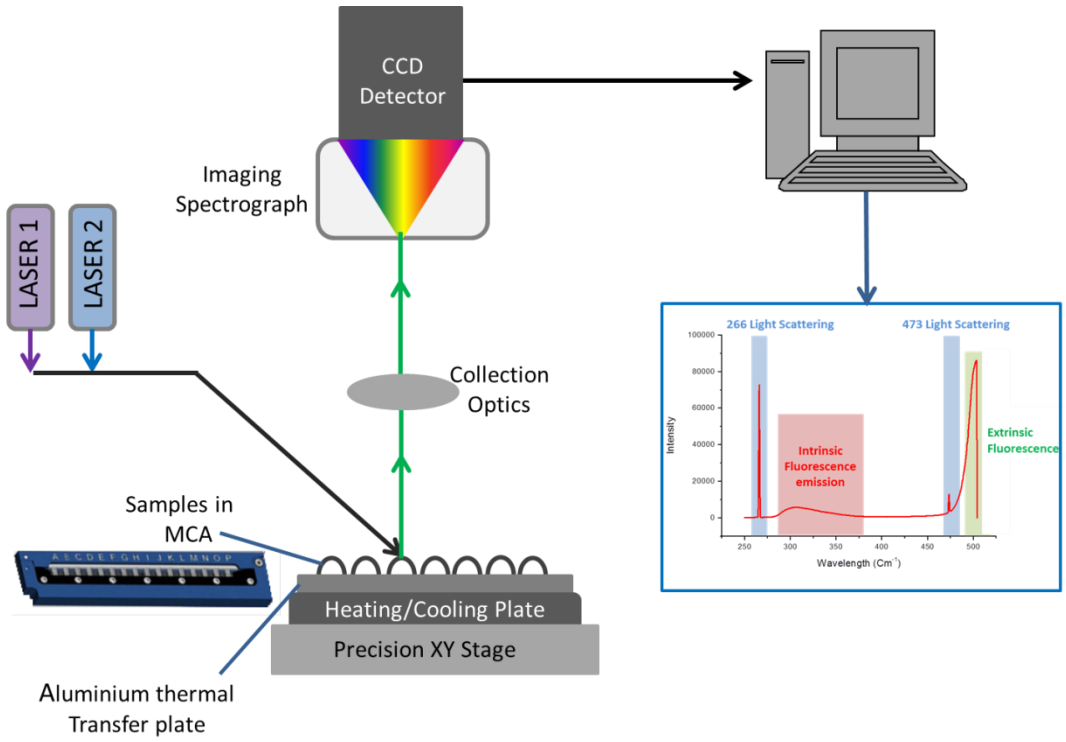

Figure 2.4: Schematic diagram showing the optical setup of the Optim 1000 and example Optim spectrum.

2.2 Data Pre-Processing.

2.2.1 Cosmic-Ray Removal.

The CCD detectors used in Raman spectrometers are susceptible to cosmic rays which pass through the detector adding charge to one or more pixels resulting in an intense sharp peak in the spectrum. For the Raman data in this thesis cosmic rays were removed manually in GRAMS Ai software (Galactic Industries Corp., 395 Main St., Salem, NH), using the 'Zap' function which removes the selected peak, replacing it with a linear trend.

2.2.2 Smoothing.

Smoothing filters are often applied to spectroscopic data in order to remove noise. Smoothing of the spectroscopic data reported in this work (Raman, IR and Optim) was carried out in Matlab R2010a (The MathWorks, Inc., Natick, MA., USA) using a Savitzky-Golay smoothing algorithm, exact details of this process are reported in the appropriate

results chapters. This works by preforming a polynomial regression on the data in order to determine a smoothed intensity value for each wavenumber (Savitzky and Golay, 1964).

2.2.3 Baseline Correction.

Fluorescence backgrounds are common artefacts in the Raman spectra of proteins; hence baseline corrections are often applied to the data in order for the spectra to be compared. For the majority of the Raman data in this thesis baseline corrections were performed in Matlab using an asymmetric least squares (ALS) algorithm, where a smoothing function is used to give a varying estimate of the baseline, in which positive deviations are weighted much higher than negative deviations (Peng et al., 2010).

2.2.4 Normalisation.

In order to reduce differences caused by shifts in spectral intensity all Raman and infrared spectra presented in this thesis were subjected to normalisation through an extended multiplicative scatter correction (EMSC) prior to multivariate data analysis. This pre-processing step was performed in Matlab for IR data and either Matlab or PyChem 3.0.5 (Jarvis et al., 2006) for Raman data. The EMSC method was originally developed to reduce the effects of light scattering in near-IR data (Naes and Isaksson, 1992). This type of normalization takes the information registered in the spectra and attempts to separate physical light-scattering effects from the actual light absorbed by molecules. EMSC also makes the spectra collapse on top of each other so that their difference in terms of spectral intensity shift is largely reduced.

2.3 Data Analysis.

2.3.1 Principal Components Analysis (PCA).

PCA is ideally suited to the interpretation of vibrational spectroscopic data due to the extremely large number of measured variables. PCA will simplify the dimensionality whilst preserving most of the variance and deriving the most important descriptors in a data set (Brereton, 2005, Jolliffe, 1986). PCA displays trends not in individual variables, but in how variables co-vary (change with respect to each other), by finding combinations of variables, factors, which describe the major trends in the data. These factors are termed principal components (PCs) and are ranked so that the first PC describes the greatest variance and the second PC the second greatest variance and so on (Jolliffe, 1986, Lindon, 2001).

The key idea behind PCA is to separate the original matrix (*X*) into two smaller matrices; the scores matrix (*T*), which contains information on how samples relate to each other, and the loadings matrix (*L*) which describes how variables relate to each other. The PCA model will then consolidate the remaining small variance factors into a residual matrix (*E*) (Otto, 1999).

Mathematically, PCA uses an eigenvector decomposition of the correlation (covariance) matrix of variables. For a data matrix, *X*, with *m* rows and *n* columns, in which each variable is a column and each sample is represented as a row, the covariance matrix (cov(*X*)) is given by:

$$\mathrm{cov}(X) = \frac{X^T X}{m-1} \qquad \textbf{Eq. 2.1}$$

PCA decomposes *X* into the vectors p_i and t_i, *where* p_i vectors are the loadings and t_i vectors are the scores. An eigenvector decomposition is then applied to the covariance matrix, where, λ_i is the eigenvalue associated with each loading vector, which is a

measure of the variance described by the loadings and scores vectors. A PCA model is therefore a combination of the scores and loadings vectors and their associated eigenvalues (Brereton, 2005, Otto, 1999, Lindon, 2001).

PCA performed on data in this work was for the most part carried out in PyChem 3.0.5 software, with the exception of chapters 9 and 10 where data was analysed using R software v. 2.9.2 (R: A Language and Environment for Statistical Computing, Vienna, Austria, 2012, http://www.R-project.org).

2.3.1.1 Multiblock PCA.

Multiblock PCA (MBPCA), also known as consensus PCA (cPCA) is a variation of PCA in which there is multiple *X* blocks. This method will produce scores and loadings bi-plots which relate to each individual block and a superscores plot which describes the covariance across the whole data set (Westerhuis et al., 1998).

2.3.1.2 Parallel Factor Analysis (PARAFAC).

PARAFAC is a generalisation of PCA which is applied to multi-way arrays. A multi-way array consists of several (3 or more) sets of categorical variables measured in a crossed fashion, were the data can be arranged in a cube with *X, Y* and *Z* dimensions, as opposed to a standard *X,Y* data matrix (Figure 2.5). PARAFAC will decompose an array into the summation over the outer product of the vectors (Bro, 2006). PARAFAC models in this work have been calculated in R.

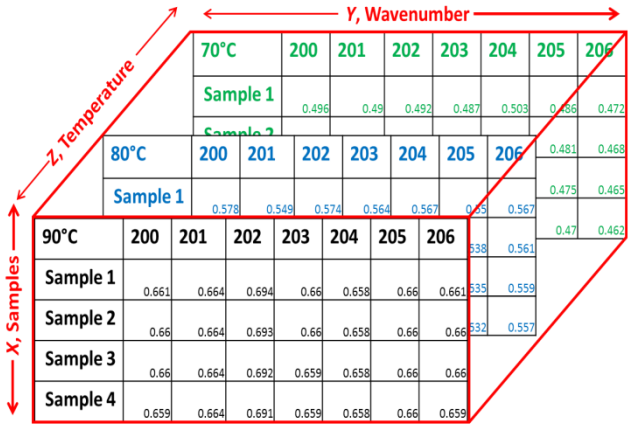

Figure 2.5: Example of Multi-Way data to be analysed by PARAFAC.

2.3.2 Discriminant Function Analysis (DFA).

Discriminant function analysis (DFA) is a supervised technique that discriminates groups using *a priori* knowledge of class membership. The algorithm works to maximize between-group variance and minimize within-group variance. The primary goal of DFA is to find linear combinations (discriminant functions) of the data variables which best discriminate between the groups (Otto, 1999). The magnitude of the absolute value of each coefficient from the discriminant function indicates the importance of the corresponding variable to the discrimination of the objects. Principal components from PCA are often used as inputs to DFA. In the present work DFA and PC-DFA were carried out in PyChem 3.0.5 software and R.

2.3.3 Partial Least Squares Regression (PLSR).

Partial least squares regression (PLSR) is a supervised multivariate analysis method which attempts to find factors, called latent variables, which can both describe the covariance and achieve correlation (Brereton, 2005). There are a number of methods for calculating PLS parameters, however the most common method, is non-iterative partial least squares (NIPALS) (DeJong et al., 2001). PLS calculates scores and loadings vectors, but also calculates an additional set of vectors known as weights (*W*), which are required to maintain orthogonal scores.

PLS will work when there is more than one predicted variable (*Y*). The *Y* matrix is split into loadings (*Q*) and scores (*U*) vectors. The PLS algorithm will attempt to find latent variables which maximise the amount of variation in *X* that is relevant for predicting *Y*, in contrast to PCA where the principal components are selected based only on the amount of variation they explain in *X*. In Brief, the PLS decomposition works by sequential calculation of the scores, weights and loadings vectors for *Y* and *X* and calculation of the inner-coefficients, which describe the relationship between *X* and *Y*. After scores and loadings have been calculated for the first latent variable *X* and *Y* residuals are calculated; the procedure is then repeated for the other latent variables (Brereton, 2005,

Otto, 1999). Once again a combination of PyChem and R software have been used in the application of PLSR to the data in this thesis.

2.3.3.1 Validation.

Supervised multivariate data analysis strategies such as PLSR need to be validated as both *X* and *Y* data are used in the model formation, and therefore the results may be subject to bias.

2.3.3.1.1 Bootstrap Cross Validation

Bootstrap is a re-sampling technique that has been applied as cross-validation to estimate the prediction performance of PLSR models in this thesis. The basic idea of this method is to select randomly, with replacement, *N* samples from a set containing exactly *N* samples. All selected samples, including the repetitions, are then used as training set and the non-selected samples are used as test set. This process is repeated a number of times, usually 100 to 1000 times, to try to approximate the real distribution of samples in the global population of cases (Efron and Tibshirani, 1994).

2.3.3.1.2 Permutation Testing

To validate the PLS models further and to confirm that the predictions are not occurring "just by chance" we have also applied a set of permutation tests to the bootstrap cross-validated models. In a permutation test the original class labels, or *Y* values, are randomly swapped and this allows the generation of a null or random model. A prediction model is then built on these permutated data, and this process is repeated several times. The accuracy value of the permutated models and is then compared to the accuracy value of the models with the original class labels. If the accuracy of the original models is significantly higher than the permutated models, then the original models are valid and not based on chance (Welch, 1990).

2.3.4 2D Correlation Analysis.

2D correlation analysis is a method of visualising a set of spectra in the form of contour maps, developed by Noda in the late 1980's (Noda, 1993, Noda, 1989, Noda and Ozaki, 2004). 2D correlation analysis can be applied to any system where a sample is subjected to external perturbation, undergoing changes which are measured by a detection method (Figure 2.6).

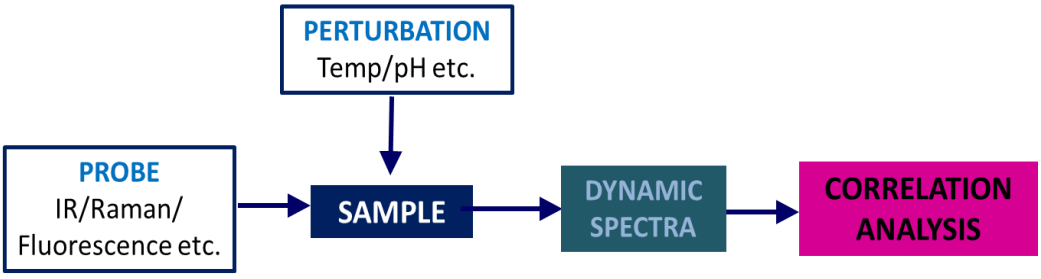

Figure 2.6: Flow diagram showing the general scheme for obtaining perturbation Induced 2D Correlation Analysis.

In order to calculate 2D correlations, data are first converted into dynamic spectra, most commonly by subtracting a perturbation-averaged spectrum from each of the spectra. 2D correlations are then calculated by applying a cross-correlation analysis to the data set, giving two orthogonal components: synchronous and asynchronous data. These data are then plotted in the form of contour plots, from which it is possible to identify bands which are changing within the data set and also probe sequential changes across a perturbation. The synchronous plot will display relative similarities, whereas the asynchronous will highlight relative differences.

An extension to 2D correlation analysis, moving windows, has also been used (Ashton and Blanch, 2008). Moving windows partitions the data into small sets, or "windows", in order to locate key transition points. This analysis is then plotted as a contour plot, which relates the spectral changes to the perturbation.

2D correlation calculations were performed using 2D Shige freeware (http://science.kwansei.ac.jp/~ozaki/index-e.html) and moving window contour plots were plotted in Matlab.

Chapter 3: Monitoring the Glycosylation Status of Ribonuclease Proteins Using Raman Spectroscopy.

Work presented in this chapter has been adapted from work published in; Brewster V.L., Ashton L. and Goodacre R. (2011). "Monitoring the Glycosylation Status of Proteins Using Raman Spectroscopy." Analytical Chemistry **83**(15): 6074-6081. L. Ashton gave advice and assistance with data interpretation and analysis, in particular 2D correlation analysis.

3.1 Introduction.

Protein-based biopharmaceuticals are becoming increasingly popular therapeutic agents, with over 50 recombinant protein products approved for use and hundreds more under development (Greer, 2008). Over one-third of these therapeutic proteins are glycosylated, in which short chain carbohydrates are covalently linked to the peptide chain of a protein following translation (Apweiler et al., 1999). The glycosylation status of a protein drug is of great importance because it can affect the stability, pharmacokinetics and perhaps most importantly immunogenicity (Greer, 2007). Consequently it is necessary to characterise the glycosylation status of a biopharmaceutical. This means determining not only whether a protein is glycosylated or not, but also that the correct glycan has been linked to the correct amino acid.

Raman spectroscopy has been used previously to characterise and quantify various carbohydrates (Oleinikov et al., 1998, Zhu et al., 2005a, Arboleda and Loppnow, 2000, Mrozek et al., 2004), as well as providing structural information about glycoproteins; in particular, glycoprotein-C of the herpes simplex virus (Kikuchi et al., 1987) and α_1-acid glycoprotein from blood plasma (Kopecky et al., 2003). However, despite the fact the Raman spectroscopy has an extensive history in protein and glycoprotein analysis, it is relatively under utilised in the monitoring of PTMs. Past work on glycoproteins centres on structural interpretations of the protein or glycan and not the differentiation of the glycosylated form from the native protein. This is essential now that the frequency of

glycoprotein use in therapy has increased significantly, and is expected to continue to increase over the next decade.

In this study we aim to develop Raman spectroscopy as a rapid approach to characterise the glycosylation status of proteins. Bovine pancreatic Ribonuclease proteins, RNase A and B, were chosen for the initial investigation as they provide a simple model system. While both proteins have an identical amino acid sequence and reported secondary and tertiary structure, RNase B is glycosylated with one N-linked high mannose glycan at asparagine residue 34 (Figure 3.1) (Taylor, 2006). Although RNase B is only glycosylated at one site, unlike the majority of therapeutic glycoproteins (which have numerous and more complex arrangements of glycans), the availability of a native non-glycosylated form makes it an ideal model glycoprotein for initial investigation.

By directly comparing RNase A and B spectra, and chemical and enzymatic deglycosylations of RNase B, we have been able to demonstrate the potential of Raman spectroscopy for characterising the glycosylation status of this protein. Furthermore through the implementation of chemometric approaches, we have established a method of quantifying levels of glycosylation in a mixture of glycosylated and non-glycosylated protein.

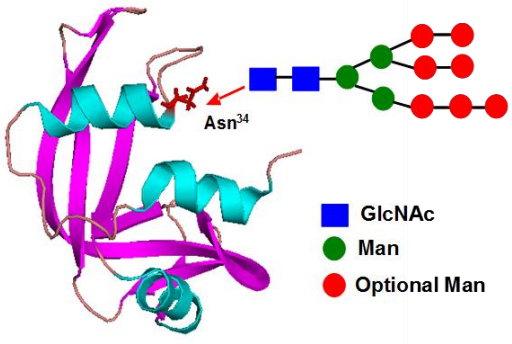

Figure 3.1: Cartoon representation of the native state of bovine RNase drawn from atomic coordinates in the PDB file (5RSA) using PyMOL; showing the Asn^{34} residue and the RNase B glycan. Optional mannose (indicated by red circles) refers to the variation in number and possible arrangements mannose residues which occurs naturally in the glycoforms of RNaseB

3.2 Materials and Methods.

3.2.1 Materials.

Ribonucleases A and B, monosaccharaides, PNGase F enzymes, trifluoromethanesulfonic acid, and other deglycosylation reagents were obtained from Sigma-Aldrich (Dorset, U.K.). MALDI matrix and calibration standards were also obtained from Sigma-Aldrich. Spectra RIM slides were purchased from Tienta Sciences Inc (Tienta Sciences Inc. Indianapolis, IN, USA).

3.2.2 Raman Spectroscopy.

Raman data were collected using the Renishaw 2000 Raman microscope described in Chapter 2. All spectra were single accumulations, extended scans between 200 and 2000 cm^{-1}, with an exposure time of between 120 s. Samples were prepared for analysis as follows; 2 µL aliquots of 2mg/ml protein solutions were spotted onto a hydrophobic SpectraRIM™ slides, detailed in section 2.1.1.1, and allowed to dry out at room temperature for approximately 1 h. Each reported spectrum is an average of 6 spectra collected from different positions within each sample spot, as depicted in 2.1.1.1.

3.2.3 Mass Spectrometry.

MALDI-MS was performed on an Axima CFR™*plus* MALDI-TOF mass spectrometer (Shimadzu Biotech, Manchester, UK), equipped with a nitrogen pulsed UV laser (337 nm), in the positive ion mode. The instrument was calibrated before each use using apomyoglobin, aldolase and albumin as calibration standards. Intact protein samples were analysed in linear time-of-flight (TOF) mode whereas protein digests were analysed in reflectron TOF mode. A total of 10 shots were recorded per profile and 1000 profiles were averaged per sample. Data were collected over a mass-to-charge (*m/z*) range of 5000-20000 with a typical laser power of 125 mW for proteins and 1-3000, laser power 75 mW for protein digests. 1 µL of sample was spotted onto a MALDI target plate and allowed to dry at room temperature, 1 µL of 10 mg/mL matrix (sinapinic acid for proteins

and α-cyano-4-hydroxycinnamic acid for tryptic digests) was then spotted on top of each sample and dried at room temperature prior to analysis.

3.2.4 Deglycosylation Methods.

Deglycosylation methods and protein recovery protocols have been adapted from those found in the literature (Edge et al., 1981, Edge, 2003, Tarentino et al., 1985, Hansen et al., 2010). Although there are established methods for both deglycosylation techniques the required amount of deglycosylating agent and incubation time vary depending on the protein, the glycan and the type of glycosylation. Therefore some degree of trial and error was involved in initial experiments in order to determine optimum conditions for ensuring complete deglycosylation. Optimised methods for both chemical and enzymatic deglycosylation of RNase B are detailed below.

3.2.4.1 Chemical Deglycosylation Method.

Trifluromethansulfonic acid (TFMS) is an established deglycosylating agent which will remove both O- and N- linked glycans by solvolytic cleavage. Glycosidic links are sensitive to cleavage by TFMS, whereas peptide bonds are stable even after prolonged exposure to the acid; hence sugars can be completely removed whilst the protein backbone remains intact. The reaction is performed in the presence of anhydrous anisole, which acts as a scavenger to protect amino acid side chains from acidic degradation.

Pre-cooled TFMS and anisole were mixed to form a solution of 10% anisole in TFMS (15 µL anisole in 140 µL TFMS). 150 µL of anisole/TFMS solution was then added to 1 mg of pre-cooled lyophilised RNase B in a reaction vial and shaken until all the protein had dissolved. The sample reaction vial was incubated on ice for 3 h, with occasional shaking. 4 µL of 0.4% bromophenol blue was then added as an indicator dye, the colour of the solution turned deep red. The sample reaction vial and a 60% pyridine solution were then cooled to -15 °C in a methanol dry ice bath. The cooled pyridine solution was then added drop wise to the sample, with mixing and cooling between drops, until the colour changed

to yellow and then to blue. A total of ~ 300 µL of pyridine solution was added. A 10-fold excess of diethyl ether with 10% hexane was then added to the reaction mixture, mixed and left to stand at -80 °C for 1 h. The sample vial was centrifuged at 8765 *x g* for 5 min and the supernatant containing the pyridinium salts was removed. The deglycosylated protein was then recovered by precipitation with ethanol, 500 µL of ethanol was added to the reaction vial, mixed and stored at -20 °C for 1 h, centrifuged at 10956 *x g* for 15 min and supernatant removed. The resulting protein pellet was re-suspended in water. Control samples were created by subjecting RNase B samples to the same experimental conditions as the deglycosylated protein with the exception of adding the deglycosylation agent.

3.2.4.2 Enzymatic Deglycosylation Method.

Peptide N-glycosidase F (PNGase F) is one of the most widely used endoglycosidase enzymes for the removal of N-linked glycans. As with the chemical method, deglyosylation with PNGase F leaves the protein intact, with the exception of the deamination of the asparagine at the glycosylation site to aspartic acid.

RNase B was prepared as a 1 mg/mL solution using 20 mM ammonium bicarbonate reaction buffer. 90 µL of glycoprotein solution was then added to a reaction vial, along with 5 µL of denaturant solution (2% octyl β-D-glucopyranoside with 100 mM 2-mercaptoethanol) and the vial was placed in a heating block at 100 °C for 10 min. The vial was allowed to cool to room temperature and the 5 µL of reaction buffer was added and the vial spun briefly, at 503 *x g* for 15 s, in a micro-centrifuge. 10 µL of PNGase F enzymes (500 unit/mL) was then added, mixed, spun and incubated at 37 °C for 24 h. The reaction was stopped by heating to 100 °C for 10 min. PNGase F enzymes were removed by precipitation (Hansen et al., 2010). Deglycosylated RNase B was recovered using the ethanol precipitation method described in the chemical deglycosylation method. RNase B control samples were created, subjecting the protein to identical conditions, minus the addition of PNGase F enzymes.

3.2.4.3 Tryptic Digests.

Protein digests were performed on deglycosylated proteins using the enzyme trypsin. 15 µL of digestion buffer (50 mM ammonium bicarbonate) and 1.5 µL of reducing buffer (100 mM DDT) were added to 10 µL of protein and incubated at 95 °C for 5 min. 3 µL of alkylation buffer (100 mM idoaceamide) was then added and incubated at room temperature for 20 min. 1 µL of trypsin (0.1 µL/µL) was then added and incubated at 37 °C for 3 h. Finally an additional 1 µL of enzyme was added and the sample incubated at 30 °C over night (Pierce-Biotechnology, 2010).

3.2.5 Data Analysis.

Raman spectroscopic data were exported into Matlab for pre-processing. PyChem was employed for PCA and PLSR. 2D correlation calculations were performed using 2D shige freeware. Spectral figures were plotted in GRAMS Ai.

3.3 Results and Discussion.

3.3.1 Detecting Glycosylation.

The ability of Raman spectroscopy to distinguish between glycosylated and non-glycosylated proteins was tested using RNase A and B. Figure 3.2 shows the average Raman spectra of RNase A and B, along with the spectra of the sugars which make up the RNase B glycan: mannose and N-acetylglucosamine (GlcNAc). A full table of band assignments for these spectra is given in Table 3.1.

The spectra show one of the main differences to be in the amide III region centred around ~1245 cm^{-1}, where a broad doublet appears in RNase A, and a more intense single peak in RNase B. The amide III band at ~1256 cm^{-1} in RNase A is a region which has previously been assigned to less-ordered proteins structure, and could be assigned to the disordered loops of the RNase protein (Ellepola et al., 2006, Ashton et al., 2007). The loss of this band in RNase B could be due to stabilisation of these structures, which are in close proximity to the glycosylation site, by the carbohydrate moiety.

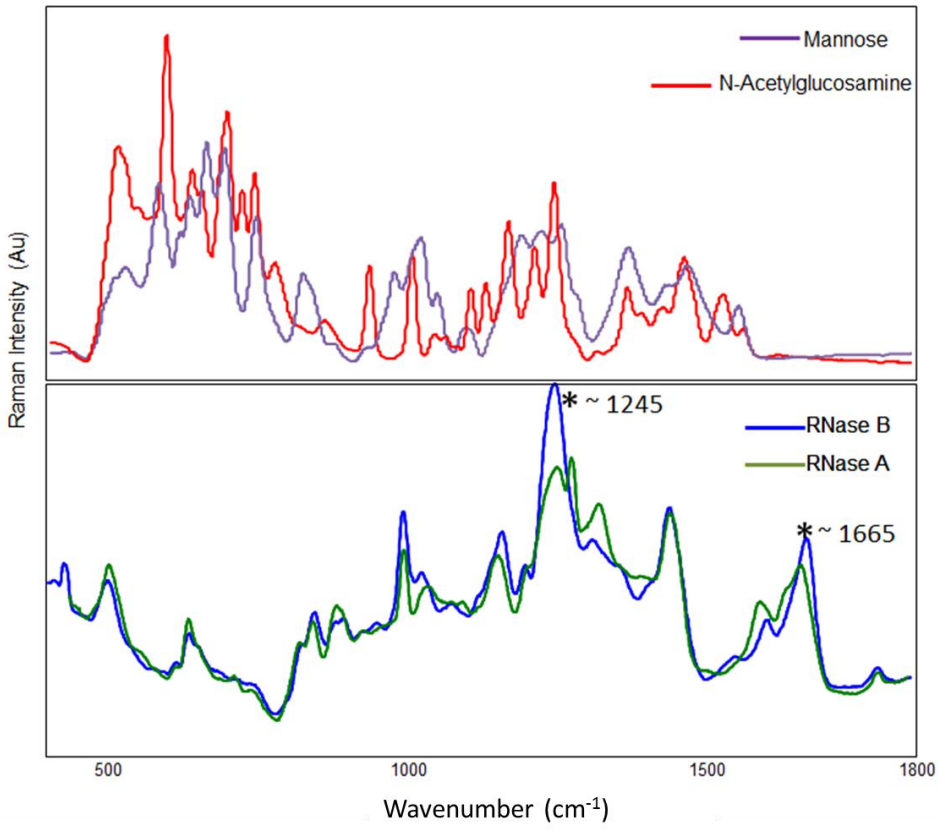

Figure 3.2: Average Raman spectra of RNase A and B, mannose and GlcNAc, astrix indicate bands indicted by PCA loadings as being important in detecting glycosylation. (Spectra have been smoothed, baseline corrected and normalised).

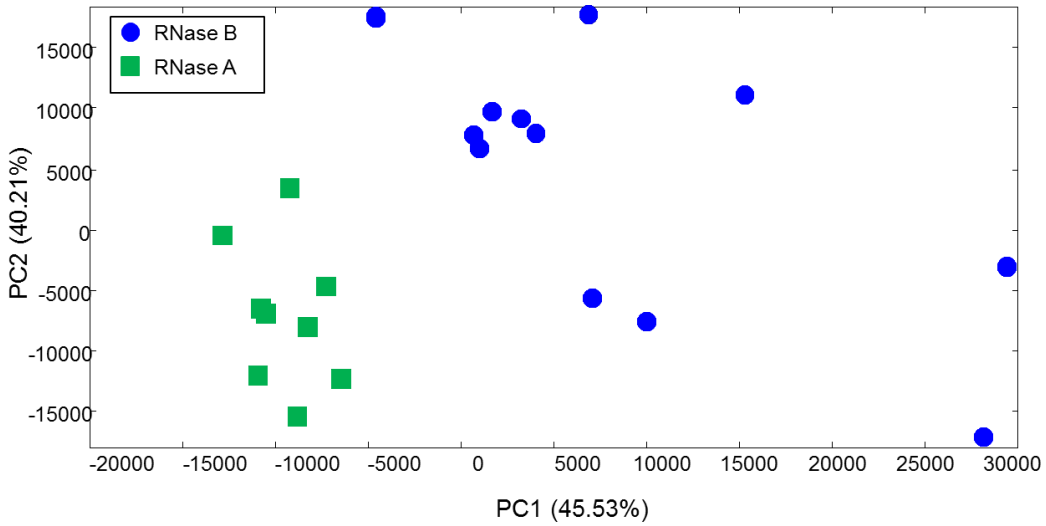

Figure 3.3: PCA scores plot (PC1 *vs* PC2) of RNase data showing RNase A and B spectra resolved into separate clusters.

Table 3.1: Raman band assignments for RNase A, GlcNAc and Mannose (Siamwiza et al., 1975, Socrates, 2001, Tuma, 2005, Oleinikov et al., 1998, Arboleda and Loppnow, 2000, Mrozek et al., 2004) .	
~Wavenumber (cm^{-1})	**Proposed Assignment**
RNase A	
511	S-S stretch (disulfide bonds)
640	C-S stretch
829	Hydroxylphenyl ring deformation (tyrosine)
850	
901	C-C stretch (peptide backbone)
935	Symmetric CH_3 stretch (alanine)
998	Phenyl ring breathing (phenylalanine)
1025	Phenyl ring vibration (phenylalanine)
1206	Phenyl ring vibration (tyrosine)
1254	Amide III vibrations (C-N stretch, N-H bend)
1318	Ring deformation (tryptophan)
1446	Asymmetric CH_3 stretch (alanine)
1608	Phenyl ring vibration (tyrosine)
1667	Amide I vibrations (C=O stretching)
GlcNAc	
413	C-C-C vibration
786	Symmetric ring breathing
863	Symmetric C-O-C stretch
927	C-C stretch
998	C-O-C Glycosidic ring vibrations
1011	
1034	
1083	
1120	Asymmetric ring breathing
1253	NH_2 twist
1642	C=O stretch
Mannose	
453	C-C ring deformation
663	Symmetric skeletal stretch
826	Ring vibration
876	Symmetric C-O-C stretch
963	C-C stretches
1004	C-O-C glycosidic ring breathing
1096	
1135	
1163	Ring breathing
1259	Ring stretch
1894	C=O stretch

Alternatively, this difference could be attributed to bands arising from the sugar molecules which mask the protein signal. Both the sugars in the glycan of RNase B exhibit bands in this region: NH_2 twisting in GlcNAc at ~1253 cm^{-1} and ring stretching in mannose at ~1259 cm^{-1} (Arboleda and Loppnow, 2000).

Also highlighted in figure 3.2 is the amide I band, which exhibits a small shift; from ~1665 cm^{-1} in RNase A to ~1676 cm^{-1} in RNase B (Figure 3.4). Amide I features at ~1665 cm^{-1} are traditionally assigned to β-sheet structure (Ashton et al., 2007), whereas bands occurring at ~1675 cm^{-1} have been associated with turn structure (Takekiyo et al., 2006). Consequently, the upward shift in peak position could be attributed to conformational changes in tertiary structure and, specifically, the turn structure of RNase brought about by the addition of a carbohydrate group.

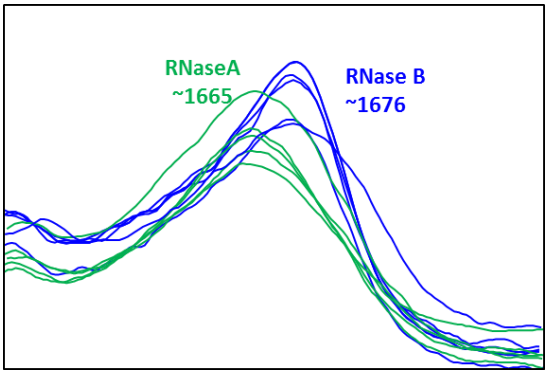

Figure 3.4: Amide I region of RNase A and B spectra, displaying an upward shift in the glycosylated protein.

By employing PCA (Figure 3.3), we were able to separate these data easily into two distinct clusters of glycoprotein and protein. The separation was largely accounted for by the first principal component score (PC1), that accounts for 45.5%of the total explained variance. Inspection of the PCA loadings matrix from PC1 revealed two major area of significance: the amide I and amide III regions as highlighted in the Raman spectra in Figure 3.2. It can be seen from the plot that the glycoprotein data does not cluster as well as the protein data. This could be due to the heterogeneous nature of the glycan attached to RNase B, but is more likely caused by orientation effects which are likely more significant for sugars than proteins.

3.3.2 Deglycosylated RNase B.

The ability of Raman spectroscopy to differentiate between deglycosylated and native glycosylated proteins was investigated using both chemical and enzymatic

deglycosylation methods. Following deglycosylation, RNase B samples were first analysed by MALDI-MS to confirm that deglycosylation had been successful. Figure 3.5 shows the mass spectra of a chemically deglycosylated protein and an enzymatically deglycosylated protein. Theoretically, RNase A has a mass of 11.7 KDa and RNase B has a mass of approximately 13 KDa, depending on the number of mannose residues in the glycan. The spectra show that in both cases deglycosylation was successful, as an average *m/z* shift of ~1460 Da can be observed (average from 10 measuremts; which equates to six mannose sugars in the glycan plus two GlcNAc). Additional conformation that this mass difference was, in fact, due to the loss of a sugar from Asn[34] was gained by performing tryptic digests on the control and deglycosylated protein samples. As trypsin enzymes cleave RNase at positions 33 and 36, a fragment containing the amino acids arginine, asparagine and leucine with a *m/z* of approximately 450 Da can be observed in the deglycosylated protein, whereas in the control protein this fragment has a much higher average *m/z* of 2130 Da, due to the glycan at position 34 (data not shown).

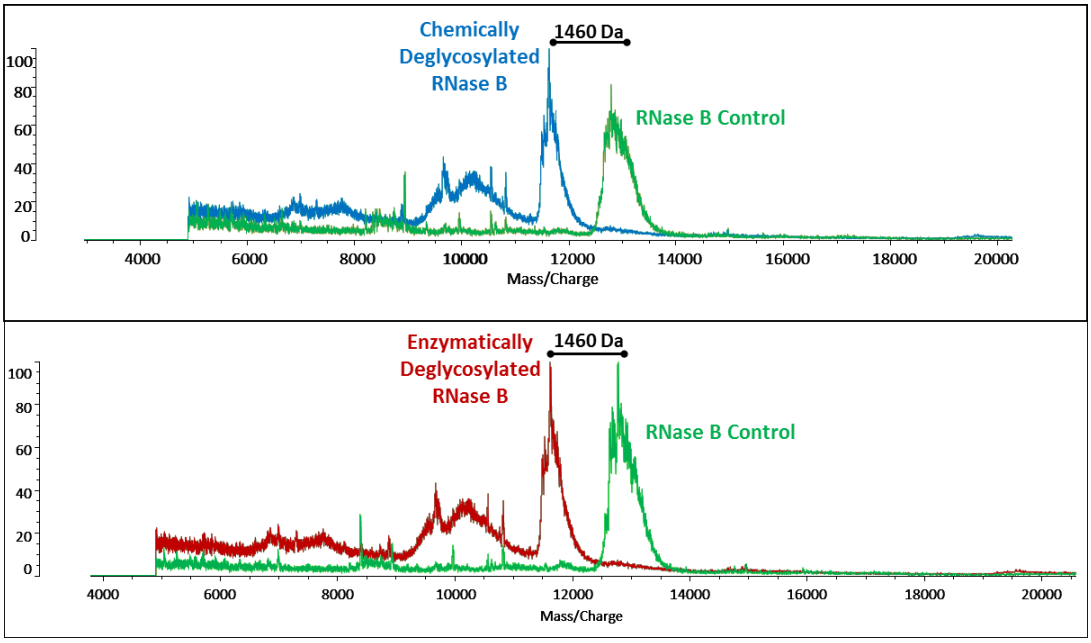

Figure 3.5: MALDI-MS spectrum of RNase B and degylcosylated RNase B, showing an average *m/z* difference of 1460 Da confirming that the protein has been successfully deglycosylated.

Raman spectra were recorded for both controls (RNase A and B) and for deglycosylated RNase B samples. The Raman spectra of deglycosylated RNase B in Figure 3.6 show that deglycosylated RNase B follows the same trends in amide I and amide III bands as observed in RNase A spectra.

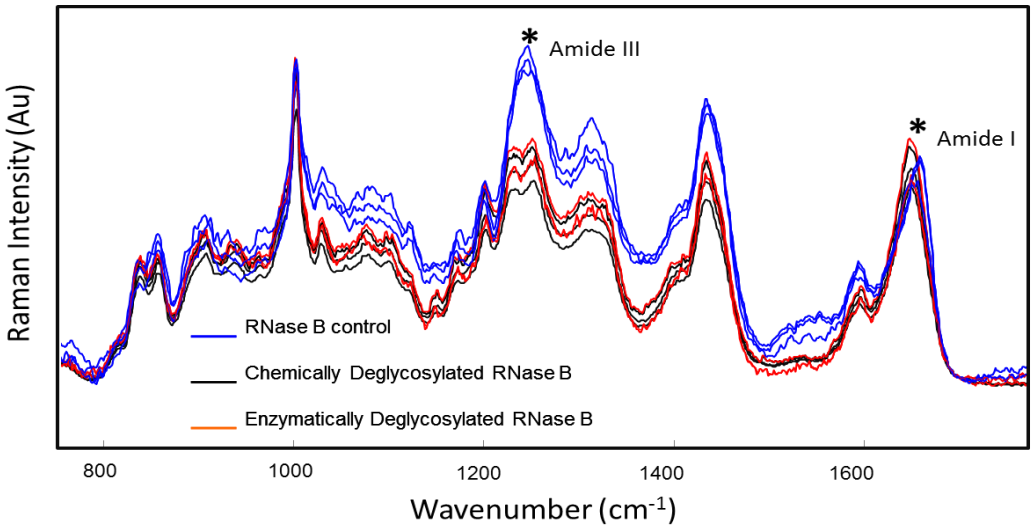

Figure 3.6: Raman spectra of control RNase B and deglycosylated RNase B, astrix indicate the amide I and III bands referred to in text. (Spectra have been smoothed, baseline corrected and normalised).

To assess the changes in the Raman spectra from the deglycosylated proteins further, PCA was performed. Figure 3.7 shows the scores plot for both chemically and enzymatically deglycosylated protein spectra along with the two controls (RNase A and RNase B). This scores biplot shows glycosylated and non-glycosylated proteins separated across PC1. The spectra from deglycosylated RNase B fall very close to the RNase A spectra, suggesting that the differences observed are, indeed, due to the removal of sugars rather than minor changes in tertiary structure. On closer inspection of this PCA scores plot, it is clear that enzyme treated samples more closely match RNase A than the chemically deglycosylated proteins. This may be due to changes in the protein conformation induced by the extreme pH needed to perform chemical deglycosylation; this is likely to be structural because MALDI-MS showed no degradation of the intact deglycosylated RNase protein when analysed in linear TOF mode. A number of samples appear to fall in the middle of the RNase A and B groups, it could be that in these

samples the deglycosylation reaction did not go to completion, hence some protein molecules may still have glycans attached.

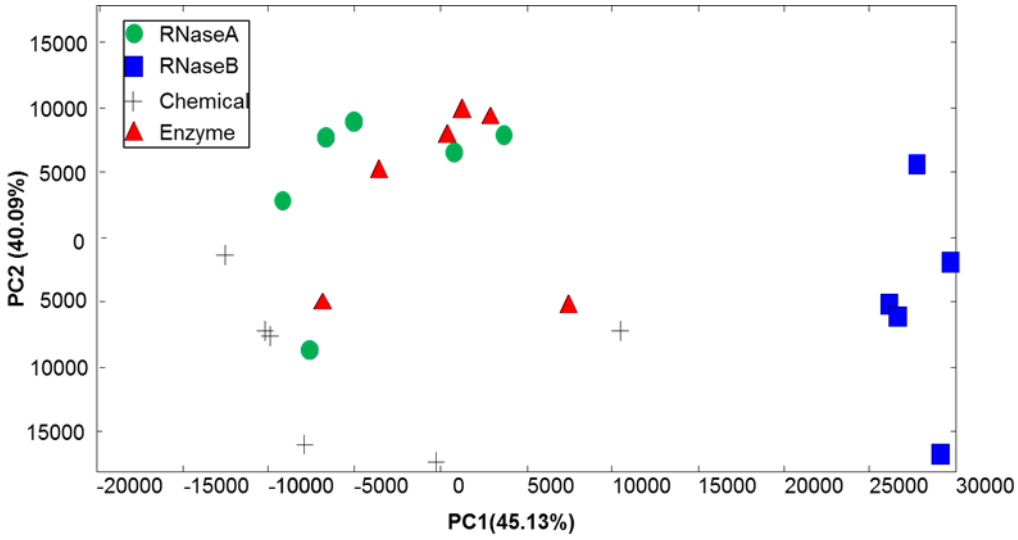

Figure 3.7: PCA scores plot (PC1 *vs* PC2) of Raman data from control RNase A and B and chemically and enzymatically deglycosylated RNase B.

3.3.3 Quantifying Glycosylation.

The next step in this study was to determine whether it was possible to predict the extent of glycosylation using Raman spectroscopy. In order to test this hypothesis, RNase A and B mixtures were used to create a model system.

Raman data were acquired randomly from 21 different mixtures of RNase A and B at 5% concentration intervals. The total protein concentration in each sample was kept constant at 1mg/ml (~70 nM). Five replicates of each mixture were analysed.

Partial least squares regression (PLSR) was applied to the data using PyChem software. Because PLSR is a supervised learning method that uses both X-data (Raman spectra) and Y-data (RNase B percentage), alternate samples were used for training and cross-validation or testing; that is to say, 0,10,20…..,90, 100% were used for training and 5,15,25%, etc. for cross validation and testing. The test data used the same series as the

cross-validation data, where the first two replicates are used for cross validation and the remaining three for the test data.

Initially data were pre-processed the same as previous RNase spectra (baseline corrected, smoothed and normalised). Figure 3.8 shows the PLSR predictions for this data, where an obvious correlation between Raman spectra and glycoprotein concentration is evident. However the test error for this model was determined to be relatively high, at 13.75%.

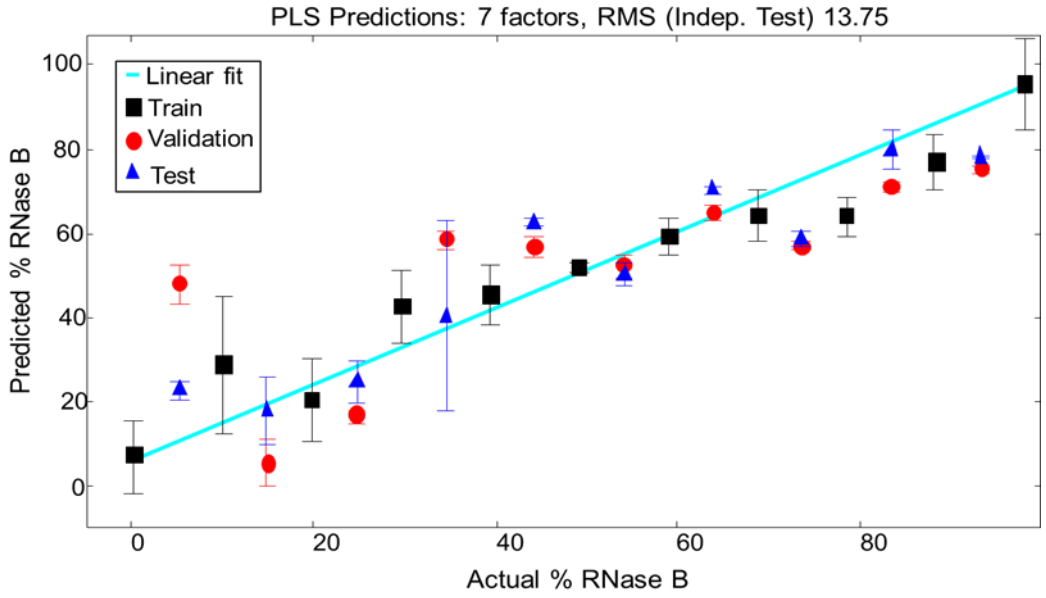

Figure 3.8: PLSR predictions from Raman data of RNase mixtures, mean predictions of five measurements are plotted with standard error bars. (Data pre-processing: Sav-Gol smoothing, ALS baseline correction and EMSC).

3.3.3.1 Pre-processing Development.

In order to improve the accuracy of the PLSR model, many different spectral pre-processing methods were tested in different combinations, and the RMS error from the test data compared. Tables 3.1 and 3.2 show the results of this testing. The optimum combination of spectral pre-processing was determined to be Smoothing (Sav-Gol (filter width 14) followed by ALS baseline correction, then scaling (min 0, max 1) and finally EMSC (polynomial order 10).

Table 3.2: Results (PLSR RMS test errors) from investigation of the most effective pre-processing methods.		
Pre-process class	**Method**	**PLS model test errors**
Raw data		22.20
Baseline correction	Weighted Least Squares	23.25
Baseline correction	Asymm Least Squares	21.90
Smoothing + BL	Savitzky-Golay + ALS	21.45
Smoothing + BL	Fourier Smoothing + ALS	22.50
Smoothing +BL	Un-weighted sliding average smooth +ALS	21.60
Normalisation	EMSC	11.50
Normalisation + BL +Smth	EMSC + ALS + Sav-Gol	9.85
Normalisation + BL +Smth	Total signal +1 + ALS + Sav-Gol	15.35
Normalisation + BL +Smth	Most intensive bin +1 +ALS + Sav-Gol	14.00
BL +Smth+ Scaling	Min 0 max+1 +ALS + sav-Gol	15.05
BL+ Norm+Scaling	ALS+EMSC + Min 0 Max +1	10.75
BL +Smth + Scaling+ Norm	Min 0 max+1 +ALS + sav-Gol +EMSC	8.35
BL+Smth+Norm+ scaling	Min 0 max+1 +ALS + sav-Gol +EMSC	9.00
Smth+ BL+ Scaling + Norm	**Min 0 max+1 +ALS + sav-Gol +EMSC**	**6.1**

Table 3.3: Results (PLS error) from investigation of smoothing filter widths and EMSC polynomial order.

Sav-Gol Filter width

EMSC Polynomial order	5	10	25	20
5	6.60	8.45	8.95	8.85
10	6.35	6.10	5.90	6.45
15	8.15	8.75	7.25	6.35
20	7.85	6.80	7.90	6.65

Sav-Gol Filter width

EMSC Polynomial order	13	14	15	16	17
13	7.10	7.10	7.70	7.82	8.55
14	6.60	6.55	7.15	5.90	5.90
15	5.60	5.56	5.91	6.50	6.95
16	5.95	5.95	7.05	6.25	6.20
17	6.30	6.20	6.90	6.45	6.4

Figure 3.9 shows the PLSR plot of predicted *vs*. actual concentrations of RNase B created from the data pre-processed as described above. This clearly shows that Raman spectroscopic data can accurately quantify relative concentrations of protein and glycoprotein, with an RMS test error in this case of only 5.56%.

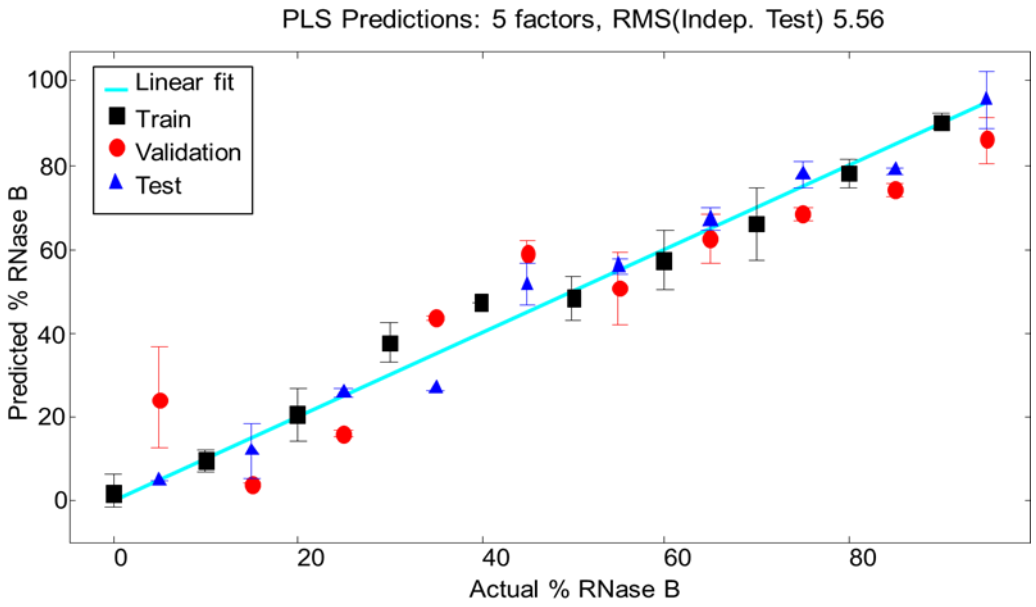

Figure 3.9: PLSR predictions from Raman data of RNase mixtures after pre-processing method development, mean predictions are plotted with standard error bars.

3.3.3.2 Model Validation.

The PLS loadings from the first two latent variables (LVs), which account for the majority of the variance in the model are plotted against each other in Figure 3.10. In this depiction each point represents a different wavenumber with each symbol coding for a different spectral region. The outer circle indicates 95% confidence level and only points outside this 95% confidence boundary have been plotted. The spectral regions that have been indicated by the loadings plot as being of importance to this model have been highlighted on the spectra in Figure 3.11 (a reproduction of Figure 3.2 with added indicators of the regions highlighted by the PLS loadings). It is clear from this representation that each of the six major areas of importance correspond to visible differences in the Raman spectra of RNase A and B, including band broadening and

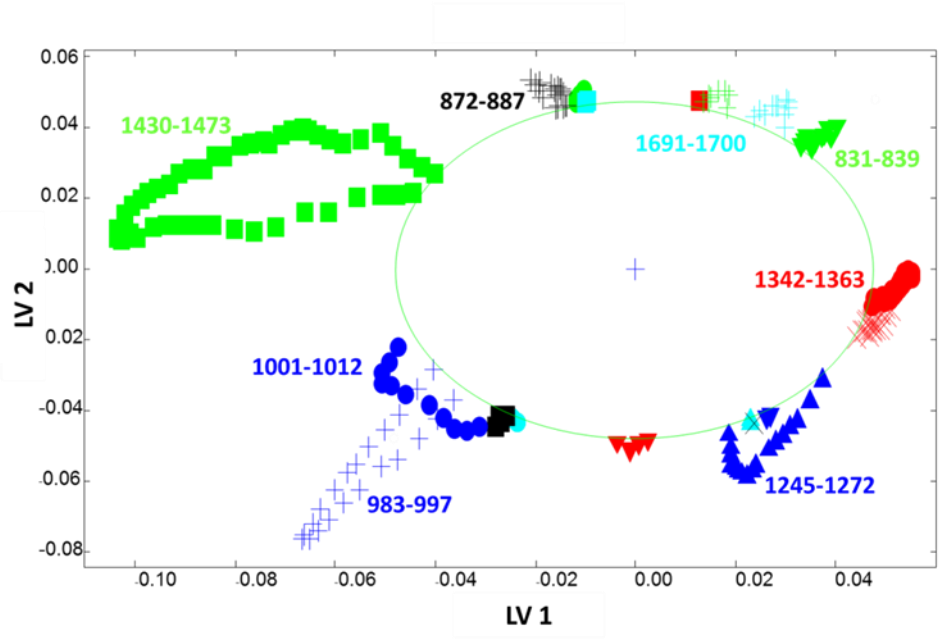

Figure 3.10: PLS loadings plot of the first two latent variables. The green circle indicates 95% confidence.

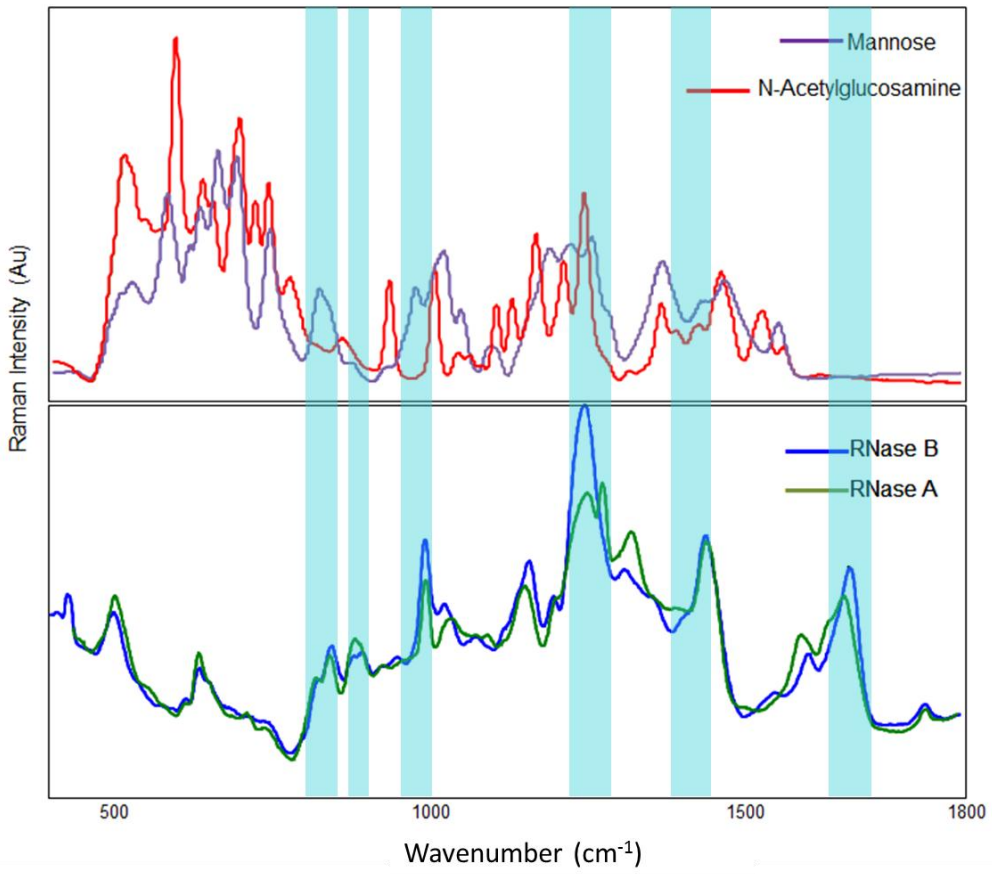

Figure 3.11: Average Raman spectra of RNase A and B, mannose and GlcNAc. Bands indicted by PLS loadings as being important are highlighted.

changes in peak intensities and position. In addition, the spectra of both sugars that form the glycan (GlcNAc and mannose) show features in the majority of these regions.

Assignments for the changes occurring in the amide I and amide III regions of the spectra have been discussed previously. Changes in the other regions (~830, 880, 1000 and 1450 cm^{-1}) could be attributed to conformational protein changes, such as changes in the local solvent environment of the aromatic amino acid side chains. However, because there are no aromatic residues close to the glycosylation site, it is more likely that these spectral differences are brought about by the presence of sugar bands in the spectra of RNase B.

The loadings plot in 3.10 is of particular value because it not only informs us which vibrational modes are the most selective for determining glycosylation status, but also assists in confirming that the quantification model is based on real spectral features, as opposed to artefacts in the baseline or nose.

Further confirmation of the importance of these regions was gained by preforming a Gaussian curve fit on the pre-processed data, using GRAMS Ai software, in order to calculate peak areas and locate peak centres. Positive correlations were found between peak parameters and RNase B concentration in five out of the six spectral regions highlighted as important by the PLSR; two examples of this are shown in Figures 3.12 and the assignments and correlation values are detailed in Table 3.4.

Table 3.4: Summary of correlations found between peak parameters and RNase B Concentration.

Wavenumber (cm^{-1})	Protein Assignment	Sugar Assignment	Peak parameter	Correlation found? (Y/N)	R^2
830	Tyr ring	glycosidic ring	Area	N	0.403
880	C-C backbone	C-O-C stretch	Area	Y	0.872
1000	Phe ring	glycosidic ring	Area	Y	0.924
1350	amide III	NH_2 twist	Area	Y	0.894
1450	alanine CH_3	glycosidic ring	Area	Y	0.732
1690	amide I	n/a	Centre	Y	0.910

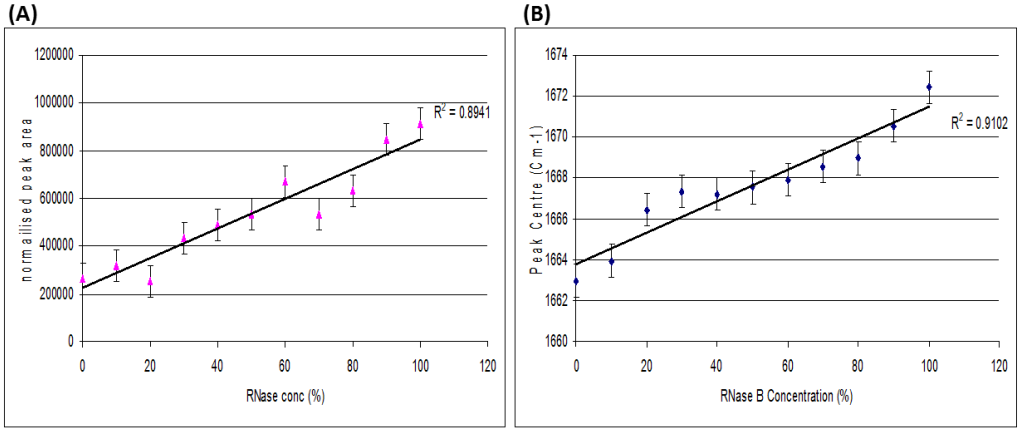

Figure 3.12 : (A) A Graph to show the correlation between peak area of the amide III band and RNase B concentration. (areas are the mean of five measurements with standard error bars shown) and **(B)** A Graph to show the correlation between peak centre of the amide I band and RNase B concentration. (plotted values are the mean of five measurements with standard error bars shown).

To investigate these correlations further, a variation of 2D correlation analysis, using moving windows, was applied as an alternative method of displaying the changes that occur within the data. Although this is not strictly a quantitative technique, it has been performed to compare the regions in which the most changes occur to those indicated by the PLS loading plot (Figure 3.10).

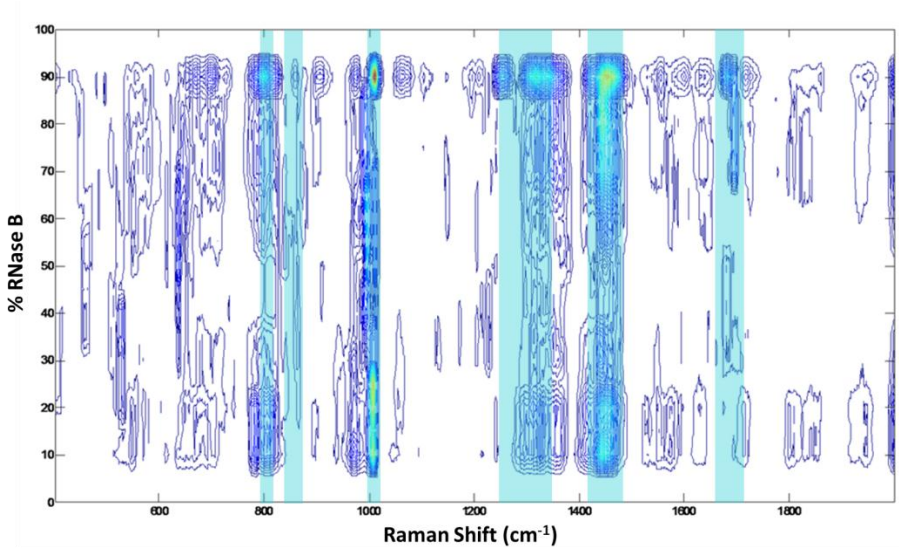

Figure 3.13: 2D-correlation moving windows contour plot as a function of spectral wavenumber and average translating window concentration of the RNase data.

A 2D moving windows contour plot for the Raman data from the RNase A and B mixtures (pre-processed in the same manner as the PLS data) is shown in Figure 3.13. The spectral regions where the most change occurs across the full percentage range, indicated by the largest number of contours, are highlighted. When compared to the spectra in Figure 3.11 and the PLSR loadings plot, it is clear that the results from both data analysis methods correlate as to which vibrational modes are the most important. It should also be noted that in the moving windows plot, contours can be observed to form two distinctive groupings at 0-30% and 80-100% RNase, which may be simply due to the detection limits of each species or errors in sampeling occurring due to issues with mixing of samples. This suggests that the majority of spectral changes are not continuous, with many occurring in two stages; 0-30% and 80-100%.

The final step in this stage of analysis was to challenge the PLS model by adding the spectra from the earlier deglycosylation experiments (Figure 3.14). The model was able to predict the control and deglycosylated spectra correctly. However, there was a much larger error margin for the whole model (10.50%) than with the standard mixtures alone (5.56%). This increase in error is likely to be due to the error associated with the deglycosylated samples, which could be caused by incomplete deglycosylation in some of the samples. Another possible cause of the higher error in this model is that many of the pre-processing methods used are global algorithms, which are influenced by the whole data set, rather than treating each spectrum separately.

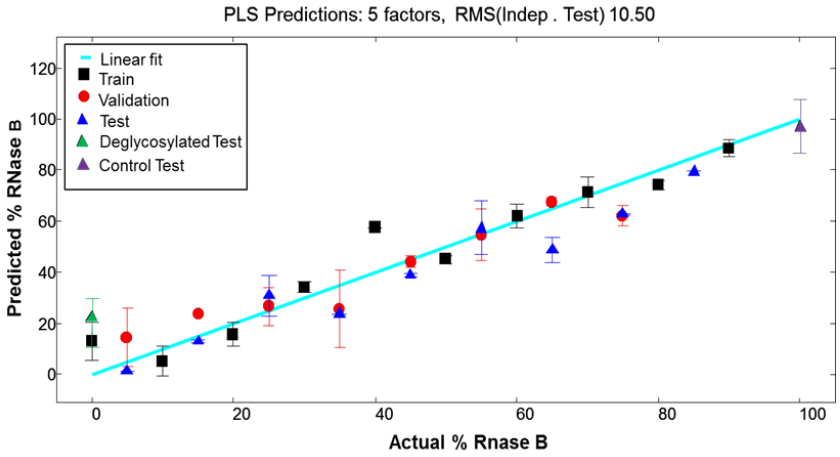

Figure 3.14: PLSR predictions from Raman data of RNase mixtures with control and deglycosylated spectra added as test data. Mean predictions are plotted with standard error bars.

3.3.3.3 Intra-Instrument Calibration Transfer.

In order to test this model for glycosylation quantification further, we investigated how transferable the model was to other instruments. Data from RNase A and B mixes was recollected on a different Renishaw 2000 microscope, with an identical optical set up and collection parameters. The data were pre-processed as described previously and then used to challenge the PLS model. In this analysis, the previously collected data were used for training and cross-validation and the newly acquired data were used to test the model. We used only 5 PLS factors, as this was the number of factors used in the previous models.

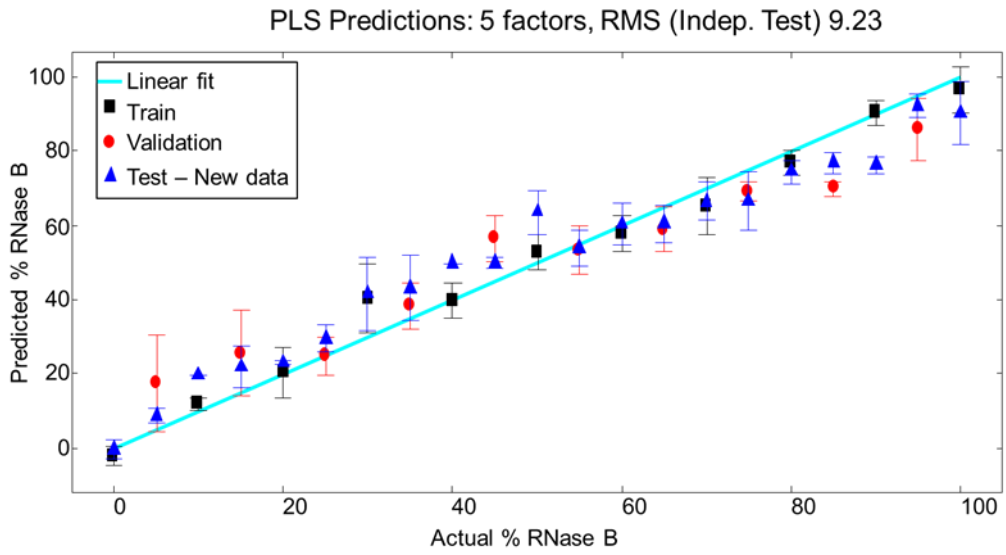

Figure 3.15: PLSR predictions from Raman data of RNase mixtures collect from two different instruments. Original data has been used for training and validation and new data from a second instrument has been used for testing. Mean predictions are plotted with standard error bars.

The PLSR predictions from this experiment are shown in Figure 3.15. The plot clearly shows a correlation between the Raman spectra in the test data and relative concentrations of protein and glycoprotein. These results show that it is possible to train the PLSR model to predict concentrations of RNase B using data collected from one Raman spectrometer, and then used this model to predict glycosylation levels accurately from data collected from a different instrument. However it should be noted that the RMS test error increased nearly two fold, to 9.23%.

This is an important result as it demonstrates the possibilities for Raman spectroscopy to be used in the biopharmaceutical industry as a technique for detecting glycosylation. In theory these results show that it possible to use data collected in one laboratory to build a quantification model and then use the same model to test data collected in another laboratory. By removing the need to build a new model for each analyses, this will increase the throughput and accessibility of Raman spectroscopy as a tool for monitoring glycosylation in therapeutic proteins.

3.4 Conclusions.

The results in this chapter clearly show the potential for Raman spectroscopy to be developed as a technique for monitoring glycosylation and deglycosylation. We have demonstrated that Raman spectroscopy with appropriate chemometric strategies is capable of distinguishing between the glycoprotein RNase B and the non-glycosylated equivalent RNase A, and also deglycosylated forms of RNase B. This work has also illustrated the potential for Raman spectroscopic data to be combined with multivariate analysis methods for the successful quantification of the glycosylation status of target proteins. Through the use of these chemometric techniques we have also identified the most selective vibrational modes for the detection and quantification of glycosylation in this system.

Although the chemometric models discussed in this chapter are specific to RNase proteins, the identification of the most informative vibrational modes can be applicable to other Raman glycoprotein investigations. By further developing these methods and building up a knowledge base of glycan standards, it should be possible to adapt the chemometirc models for this simple system to facilitate the use of Raman spectroscopy for the characterisation of glycosylation in the far more complicated glycoproteins produced by the biopharmaceutical industry. In fact since the publication of the work in this chapter, Raman spectroscopy and PLSR have been demonstrated successful in the quantification of glycosylation in two medically relevant examples; haemoglobin and albumin (Dingari et al., 2012, Barman et al., 2012).

Chapter 4: Characterising Glycosylation, Stability and Aggregation in Transferrin Using Optical Spectroscopies.

4.1 Introduction.

The use of Raman spectroscopy coupled with multivariate data analysis strategies to detect and quantify glycosylation in a simple model system was described in Chapter 3. Following on from this, we have in the present work, applied these techniques to a more complex system. As one of the major application areas for Raman spectroscopy in protein and PTM characterisation is in the biotechnology industry, we focus here on the detection of glycosylation in a real biopharmaceutical recombinant protein product. For this study a number of variants of transferrin proteins were kindly supplied by Novozymes Biopharma.

Transferrin is an iron binding globular protein which is responsible for iron transport. It occurs naturally in two forms: with iron (holo) and without iron (apo). The structures of these proteins depicted in Figure 4.1 show transferrin to be in two domains, comprised predominately of α-helical secondary structure with some turn structure and a small contribution of β-sheet structures. The iron binding site lies between these two globular lobes. As can be seen from the structures in Figure 4.1, when iron is removed to form apotransferrin the two domains are held together less tightly, and hence apotransferrin has a much more open structure than its iron containing counterpart.

Transferrin is produced as a biopharmaceutical product, not as a therapeutic protein, but as a fusion protein. Its role is to bind to other therapeutics which have short half-lives and improve the pharmacokinetics of these drugs by extending their activity. Although it is not being used as the active ingredient in a product, it is part of a drug formulation and hence these proteins still need to be robustly characterised. Complete characterisation of the

product will not only maintain the quality of the product, but more importantly help to prevent any harmful immunogenic affects.

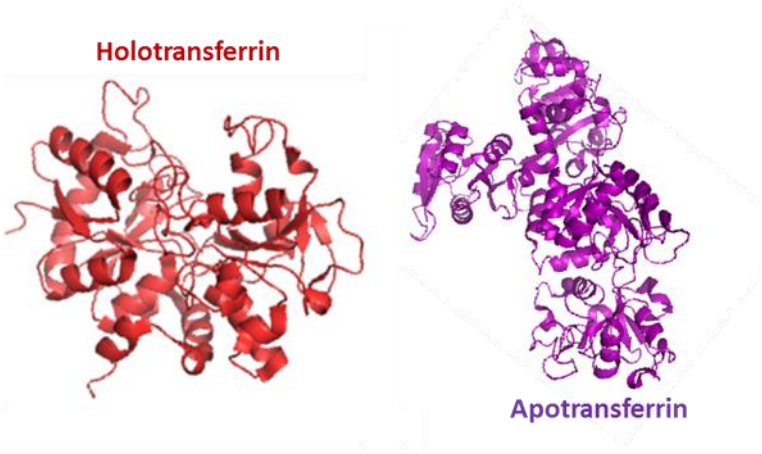

Figure 4.1: Cartoon representation of apotransferrin and holotransferrin drawn from atomic coordinates in the PDB files (2HAU and 1H76) using PyMOL.

The samples received from Novozymes are recombinant transferrins produced in yeast. The un-purified product will contain a mixture of non-glycosylated and glycosylated proteins. As these proteins are produced in yeast cell lines, all of the glycans will be of the high mannose type. The product can therefore be purified using a Con A (Concanavalin A) column which selectively binds high mannose glycans.This experiment yields three different samples for analysis: Con A flow through, containing un-mannosylated proteins, Con A retarded fraction which consists of mono-mannosylated transferrin and finally the fractions which binds to the Con A containing mainly oligo-mannosylated transferrin. After this purification step, some of the samples have been apoised to give the 'iron-less' form of the protein.

Initial work in this chapter centred on detecting glycosylation in transferrin. However as it is widely accepted that glycosylation can cause changes to the conformation of a protein, and we observed these for RNase B earlier (Chapter 3), we have in the present work investigated the structural changes brought about by glycosylation further. To do this we have employed the previously described (2.1.3) Avacta Optim 1000 instrument to profile how unfolding and aggregation profiles change with glycosylation status.

4.2 Materials and Methods.

4.2.1 Samples.

Transferrin samples were received from Dr Malcolm Saxton at Novozymes Biopharma. Samples of glycosylated variants of both apo- and holo-transferrin were received, as detailed in Table 4.1. All samples were 1mg/ml (~12 mM) and were supplied in standard PBS buffer.

Table 4.1: Summary of Transferrin Samples.

Sample Name	Apo/ Holo	Glycosylation Status	Abbreviation used in this work
Transferrin	**Holo**	**Un-mannosylated**	**Tf**
Mono-mannosylated transferrin	**Holo**	**Mono-mannosylated**	**mmTf**
Oligo-mannosylated transferrin	**Holo**	**Oligo-mannosylated**	**omTf**
Apo Transferrin	**Apo**	**Un-mannosylated**	**ApoTf**
Mono-mannosylated apotransferrin	**Apo**	**Mono-mannosylated**	**mmApoTf**
Oligo-mannosylated apotransferrin	**Apo**	**Oligo-mannosylated**	**omApoTf**
Product	**Apo**	**Unknown mixture**	**P**

4.2.2 Raman Spectroscopy.

Raman data were collected using a Renishaw 2000 Raman microscope described in Chapter 2. All spectra were single accumulation, extended scans between 400 and 1800 cm^{-1}, with an exposure time of 60 s. 2 µL of sample were spotted onto a hydrophobic SpectraRIM™ slides, detailed in section 2.1.1.1, and allowed to dry out at room temperature for approximately 1 h. Each reported spectrum is an average of 6 spectra collected from different positions within each sample spot, as depicted in 2.1.1.1. For the wavelength comparison study in 4.3.1.1., the previously described Renishaw Raman spectrometer was used operating both 785 and 633 nm excitation wavelengths. For the spectra collected at 532 nm a Biotools chiral Raman spectrometer (detailed in 2.1.1.2) was employed. Data were pre-processed (smoothed, baseline correction and normalisation) according to the method optimised in Chapter 3 (3.3.3.1).

4.2.3 FT-IR spectroscopy.

FT-IR spectra were collected on a Bruker FT-IR instrument described in 2.1.2. 4 µL of each sample was spotted onto a 96 well silicon plate and allowed to dry at room temperature. Spectra were recorded over 4000-600 wavenumbers, with 64 accumulations per sample.

4.2.4 Optim 1000.

The Avacta Optim 1000 is a multi-modal platform which simultaneously collects fluorescence emission and light scattering data. The instrument set up is described fully in 2.1.3.

4.2.4.1 Optim Thermal Ramp Experiments.

9 µL of three replicates of each sample were loaded into a micro cuvette array (MCA). A temperature ramp from 30 to 85 °C was applied to the samples with a temperature tolerance of 0.3 °C. Spectra were recorded at 1 °C intervals with a 60 s hold time at each temperature. Spectra were collected with 1 s exposure time with the slit width set to 100 µm. Each run was performed in triplicate, with three analytical replicates of each sample per run.

4.2.4.2 Optim Isothermal Experiments.

9 µL of three replicates of each sample were loaded into an MCA. Samples were rapidly heated to, and held at, a set temperature, chosen from observing the results of the previous thermal ramp experiments. In this case three separate experiments were performed at 45, 55 and 67 °C. Samples were held at these temperature, with a tolerance of 0.5 °C and spectra were recorded at 60 s intervals for 200 min. Spectra were collected with 10 s exposure time with the slit width set to 100 µm. Each run was performed in duplicate, with three analytical replicates of each sample per run.

4.2.5 Data Analysis.

Vibrational spectroscopic data were exported into Matlab for pre-processing. PyChem was employed for PCA and PLSR. PLSR models with bootstrapping and permutation testing were performed in R. Spectral figures were plotted in GRAMS Ai. Optim data were imported into Optim Analysis software for preliminary analysis. Data were then exported into Origin for further analysis and for plotting figures. 2D correlation calculations were performed using 2D shige freeware.

4.3 Results and Discussion.

4.3.1 Vibrational Spectroscopy.

4.3.1.1 Wavelength Selection.

The initial step for this study was to determine which of the available laser excitation wavelengths was most suitable for Raman spectroscopic analysis of transferrin proteins. Spectra of Holotransferrin (un-mannosylated) were recorded at 785, 633 and 532 nm. Typical spectra from each wavelength can be viewed in supplementary information (Figure S4.1). The spectrum recorded at 532 nm, shows little or no bands which can be assigned to protein structure. We believe this is due to the iron in holotransferrin being resonant at this wavelength, causing the majority of bands in the spectra to arise from the iron group. The Raman spectra recorded at 633 nm exhibited a large amount of fluorescence background. Although protein bands were observed in the spectrum, the levels of background signal encountered were problematic and the resolution of bands was poor. By contrast, the spectrum recorded at 785 nm exhibits minimal background interference and intense, well resolved protein bands. These results make a 785 nm excitation wavelength the obvious choice for future analysis of transferrin samples.

4.3.1.2 Comparing Holo- and Apo-Transferrin.

Before analysing glycosylated transferrin samples, the Raman spectra from un-mannosylated holo- and apo-transferrin were compared to see what effect the presence

of the iron group has on the Raman spectra of transferrin. The PCA scores plot in Figure 4.2 shows that the two forms of transferrin are easily distinguishable by their Raman spectra. The loadings from PC 1, which accounts for 75.44% of the total explained variance (TEV), correlate well with bands that can be seen to be changing in the Raman spectra in Figure 4.3. The PCA scores plot also appears to separate the holotransferrin into two groups across PC2, however inspection of PC2 loadings (not shown), show no particular features and are mainly comprised of noise. It should be noted that all spectra were acquired on the same day and sample collection order was randomised.

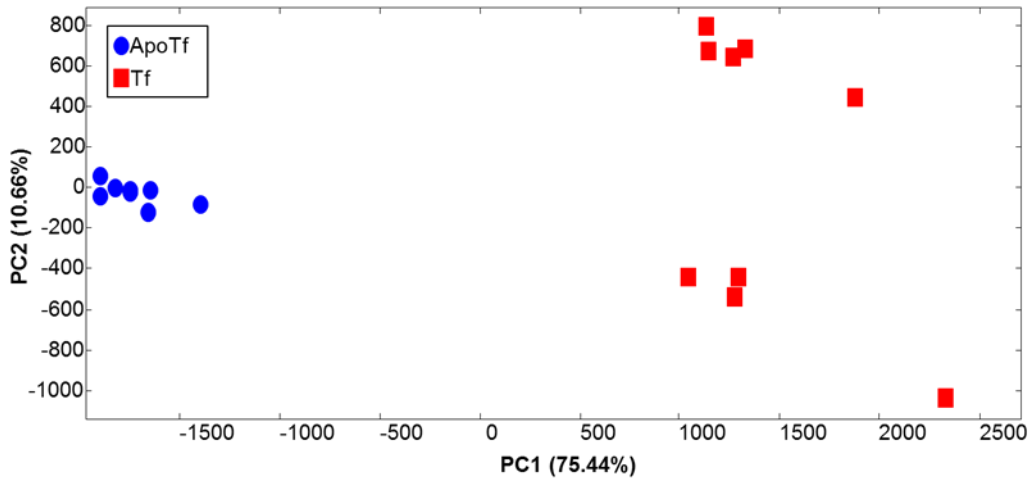

Figure 4.2: PCA scores plot (PC1 *vs* PC2) of Raman data from holotransferrin and apotransferrin, showing the data clearly resolved into two distinct clusters.

The Raman spectra in Figure 4.3 show numerous changes which can be attributed to structural differences which occur between apotransferrin and holotransferrin. The bands highlighted in the spectra are those which are indicted by the PCA loadings plot (PC1) as being significant. The features which are highlighted in red font refer to those indicated by the positive loadings, hence relating to the positive side of the PCA plot where Tf falls. The blue font indicates bands which are changing in ApoTf as indicated by the negative loadings.

The spectra show changes in the amide I and amide III regions, which is to be expected with such an alteration in protein conformation. There are also changes in the intensities

of a number of bands that can be assigned to individual amino acids, which increase or decrease in intensities as the residues become more or less exposed to the surrounding environment by the changes in tertiary structure. Finally the band indicted by the green asterisk, at ~500 cm^{-1}, which appears only in the holotransferrin spectrum can be specifically assigned to the iron-ligand vibrations, arising from both Fe-O and Fe-NO complexes (Villar et al., 2006, Soldatova et al., 2010).

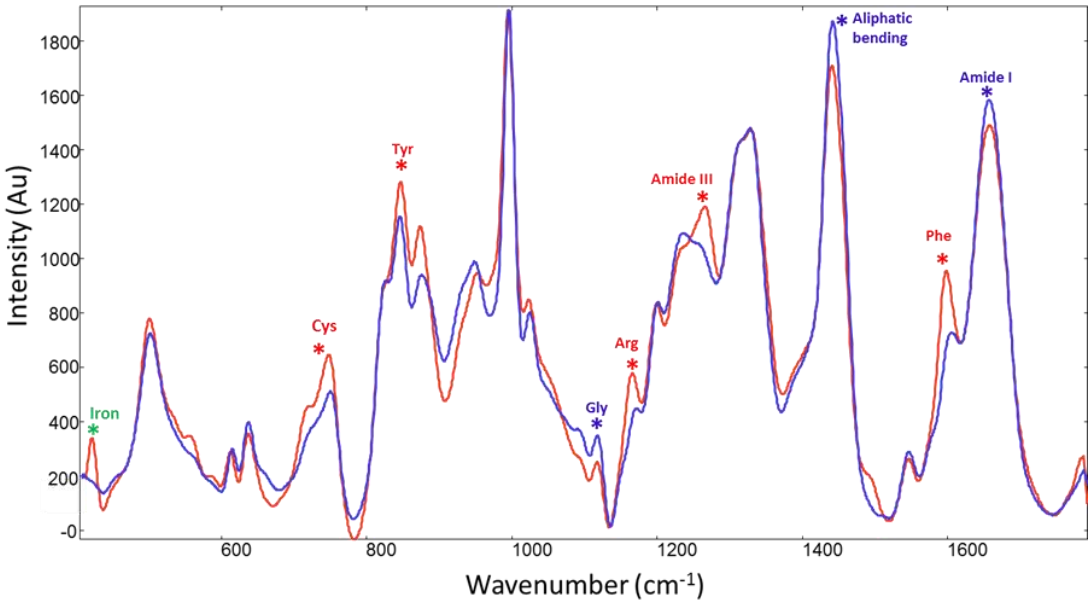

Figure 4.3: Average Raman spectra of holotransferrin (Tf) in red and apotransferrin (ApoTf) in blue. Asterisks indicate the bands highlighted by the PCA loadings. (Spectra have been smoothed and baseline corrected).

4.3.1.3 Detecting Glycosylation in Transferrin.

4.3.1.3.1 Detecting Glycosylation in Holo-transferrin.

We began investigating the use of Raman spectroscopy for the detection of glycosylation in this system by comparing the spectra of the un-mannosylated and oligo-mannosylated variants of holotransferrin (Tf and omTf). It is clearly evident from the PCA scores plot displayed in Figure 4.4A that we are easily able to distinguish between the glycosylated and non-glycosylated forms of transferrin based on their Raman spectra. The figure shows clearly defined clusters of Tf and omTf separated across PC1, which describes

58.78% of the variance in the data. An interesting point to note is that there is much more variation in the glycoprotein spectra than in the protein spectra. This is a trend we have seen previously with the RNase study, and was ascribed to the poor reproducibility in the bands arising from the vibrations of the glycan component of the molecules, which may be due to orientation effects when the proteins are dried on the SpectraRIM^{TM} slides.

The average Raman spectra of Tf and omTf are shown in Figure 4.5, along with the loadings for PC1 in Figure 4.6. The area of the loadings plot shaded blue highlights the negative loadings, which correspond to bands which can be seen to be increasing in intensity in the spectra of Tf (shown in blue). The assignments for these bands are given in Table 4.2. The majority of these features are bands which are indicative of changes in tertiary structure, showing that as with the RNase system, changes in the conformation of transferrin occur upon glycosylation and these changes are visible in the Raman spectra.

The positive bands from the loadings (shaded red) agree well with bands observed to be increasing in the Raman spectra of omTf. The changes in these bands can be attributed to contributions from glycosidic vibrations. It can be seen from the spectra of mannose shown in Figure 4.5 that the sugar exhibits bands in each of the regions highlighted in the spectra and loadings plot. However as these bands could also be assigned to bands arising from the amino acids of the protein, an alternative explanation could be that the spectral differences are due to changes that occur in the structures of the two proteins.

Table 4.2: Raman band assignments for bands highlighted in the Raman spectra of Tf and omTf (Fig 4.5) and the PCA loading plot (Fig 4.6). (Siamwiza et al., 1975, Socrates, 2001, Tuma, 2005, Oleinikov et al., 1998, Arboleda and Loppnow, 2000, Mrozek et al., 2004) .

~Wavenumber (cm^{-1})	**Proposed Assignment**
645	Tyrosine ring
820	Glycosidic ring vibration
962	Disordered structure
1009	Glycosidic ring breathing
1250	Amide III
1320	CH_2OH side chain def (mannose)
1445	Aliphatic bending

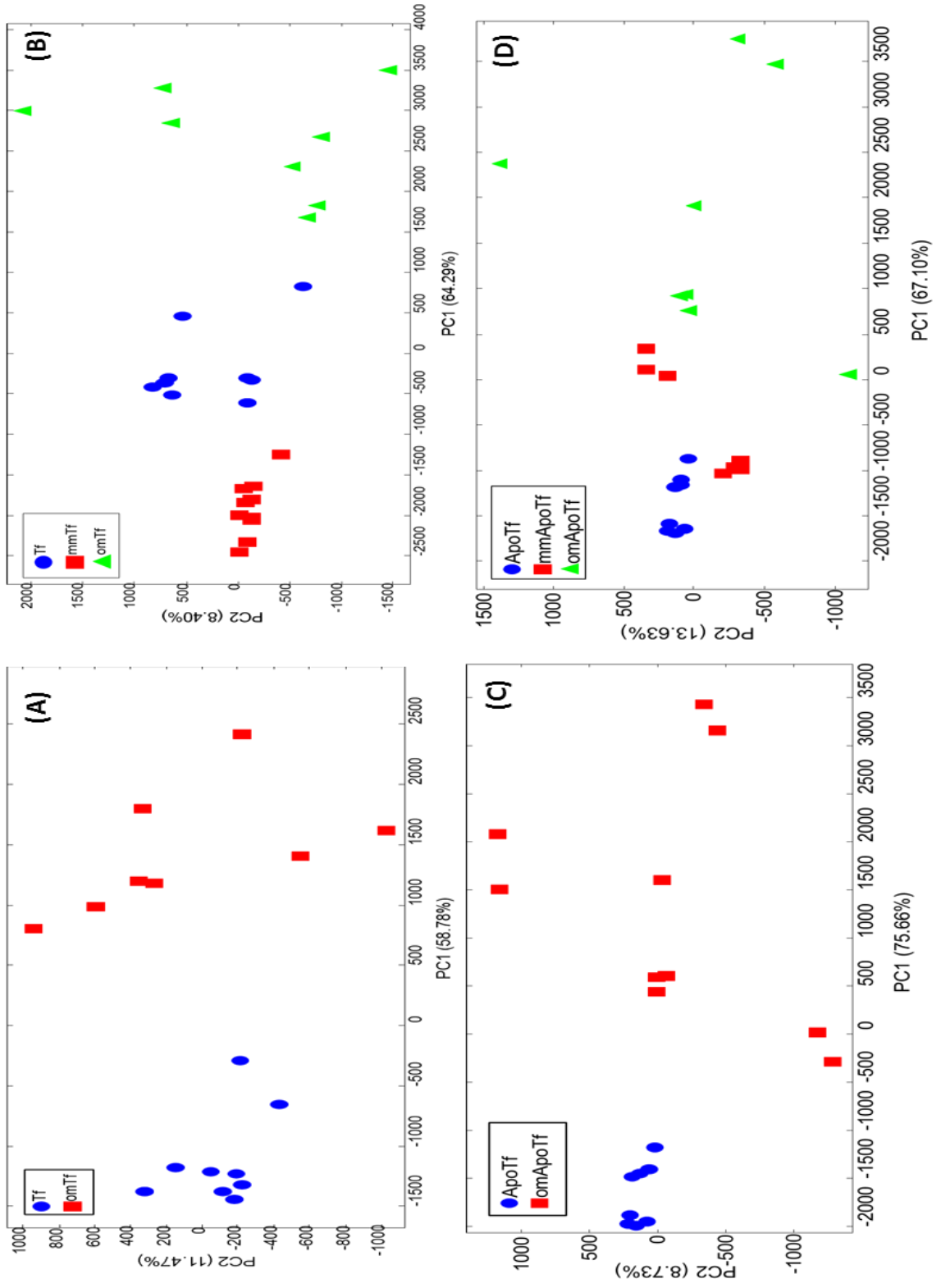

Figure 4.4: PCA scores plot (PC1 *vs* PC2) of Raman data from **(A)** un-mannosylated holotransferrin (Tf) and oligo-mannosylated holotransferrin (omTf),**(B)** un-mannosylated holotransferrin (Tf) and oligo-mannosylated holotransferrin (omTf) and mono-manosylated holotransferrin (mmTf), **(C)** un-mannosylated apotransferrin (ApoTf) and oligo-mannosylated apotransferrin (omApoTf) and **(D)** un-mannosylated apotransferrin (apoTf) and oligo-mannosylated apotransferrin (omApoTf) and mono-manosylated apotransferrin (mmApoTf).

Finally, although no changes can be observed by eye in the Raman spectra, the PCA loadings show the amide I region (~1630-1690 cm^{-1}) as being important in the separation of the two different forms of transferrin. This again indicated a notable change in protein conformation occurring upon glycosylation.

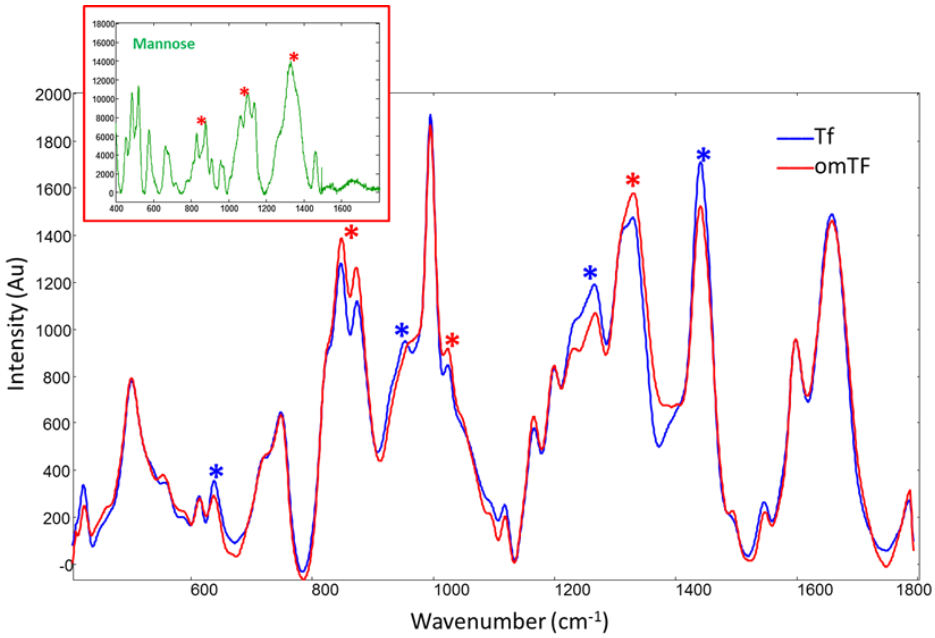

Figure 4.5: Average Raman spectra of transferrin (Tf) and oligo mannosylated transferrin (omTf) and mannose (inset). Asterisks indicate the bands highlighted by the PCA loadings. (Spectra have been smoothed and baseline corrected).

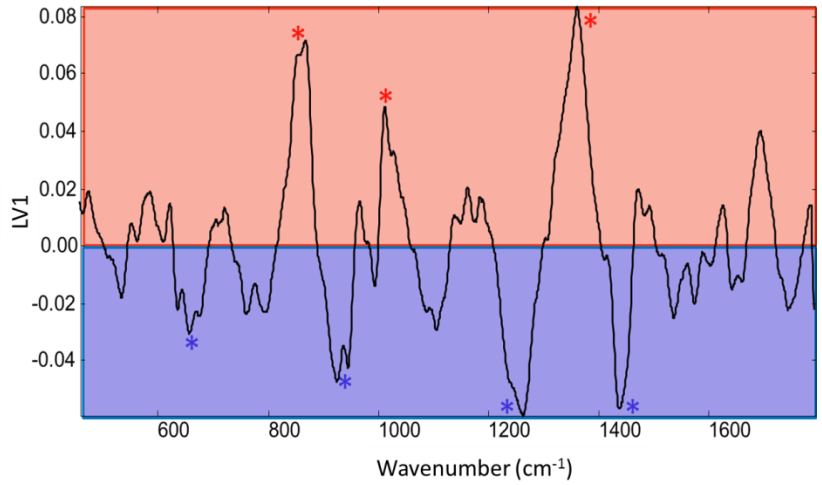

Figure 4.6: PCA loadings plot for PC1 from Raman data of un-mannosylated holotransferrin (Tf) and oligo-mannosylated holotransferrin (omTf).

The next step was to add the data from the mono-mannosylated protein (mmTf) into the PCA plot (Figure 4.4B). It is clearly evident from this plot that it is possible to distinguish mmTf from Tf and omTf, however this occurs in a rather unexpected manner with the two glycosylated variants falling at opposite sides of the of the non-glycosylated transferrin.

In order to confirm this result was a true trend (i.e., not caused by any instrumental artefacts in the spectra) we performed FT-IR analysis on the same samples. A PCA scores plot of this FT-IR data (seen in supplementary information Figure S4.2) showed a similar trend to the Raman data with mmTf and omTf separating in opposite directions.

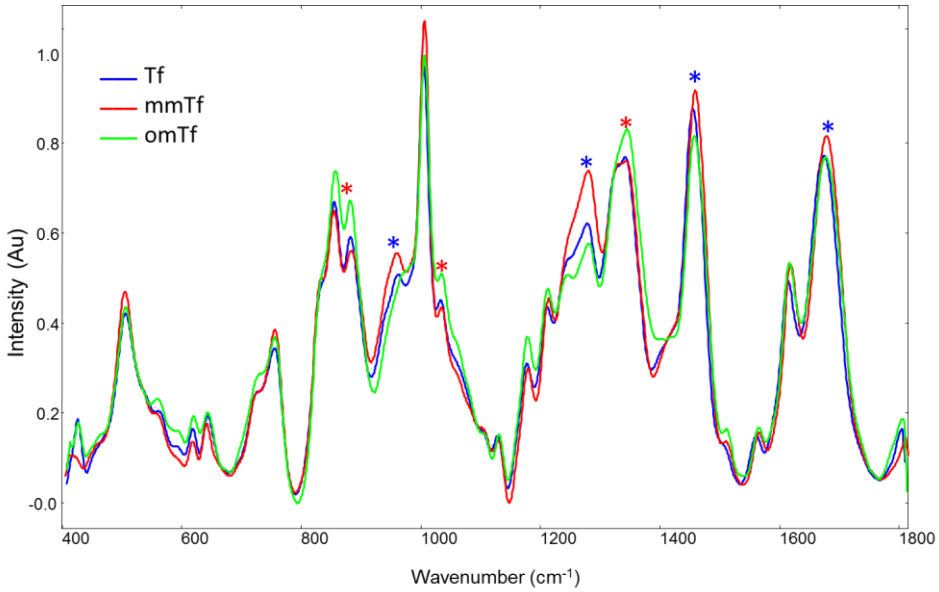

Figure 4.7: Average Raman spectra of transferrin (Tf) and oligo-mannosylated transferrin (omTf) and mono-mannosylated transferrin. Asterisks indicate the bands highlighted by the PCA loadings (Figure 4.6; positive loadings in red and negative loadings in blue). (Spectra have been smoothed and baseline corrected).

Inspection of the Raman spectra (Figure 4.7) show that bands previously seen to be increasing in omTf, which were thought to be due to glycosidic vibrations, are not increasing in the mmTf spectrum. In addition, many of the bands which are increasing in mmTf are features which we have previously assigned to protein structural changes. Therefore our hypothesis is that separation in the positive direction is based on a combination of sugar and protein bands, whereas separation in the negative direction is purely structural. The loadings from PC1 (not shown) appear to confirm this as the

negative loadings indicate structural bands and the positive loadings are very similar to the bands seen in the separation of omTf and Tf in Figure 4.6.

4.3.1.3.2 Detecting Glycosylation in Apo-transferrin.

The un-glycosylated and oligo-glycosylated variants of apotransferrin (ApoTf and omApoTf) were then compared to each other, and again showed two distinct clusters in the PCA scores plot (Figure 4.4C). ApoTf and omApoTf show very similar results to their holo- counterparts: good separation across PC1, a large amount of variation in glycosylated spectra, and PC1 loadings again reveal a combination of glycosidic and protein conformational bands.

When adding the mono-mannosylated transferrin (mmApoTf) spectra into the data matrix, the PCA scores plot shows a more expected trend than that seen with the holotransferrin data: with the mono-mannosylated samples falling in the middle of the un-glycosylated and the oligo-mannosylated samples (Figure 4.4D). The loadings for this PCA (not shown) are almost identical to the loadings for the PCA of omTf and Tf (Figure 4.4A and Figure 4.6) and the PCA of omApoTf and ApoTf (Figure 4.4C). A possible reason for this inconsistency in the behaviour of apo- and holo- glycosylated variants, could be that the differences in tertiary structure of the two transferrins leave the glycan more accessible in the apotransferrin than in the holotransferrin.

Finally, we have combined all the data from apo- and holo-transferrin, and also added the spectra from the pre-Con A purification product samples and performed PCA. The PCA scores plot for all transferrin data (Figure 4.8) displays how Raman spectroscopy can be used to distinguish between glycosylated variants of proteins (across PC1) and also between iron containing and apoised proteins (across PC2). The data from the product samples can be seen to overlay the un-mannosylated data which possibly suggests that the product is largely un-glycosylated transferrin.

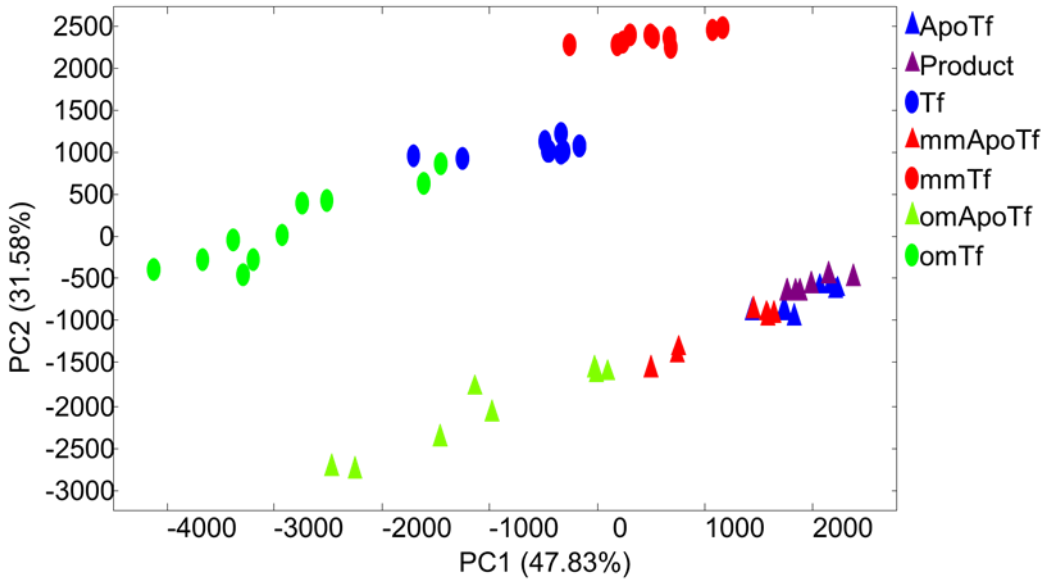

Figure 4.8: PCA scores plot (PC1 *vs* PC2) of Raman data from all transferrin samples.

4.3.1.4 Quantifying Glycosylation in Transferrin.

In order to develop the application of Raman spectroscopy for the analysis of glycosylated transferrin further, spectra were recorded from mixtures of glycosylated and non-glycosylated protein in order to attempt to quantify glycosylation from the Raman spectra. Mixtures of ApoTf and omApoTf were made, with increasing concentrations of omApoTf at 5% intervals, keeping the total protein concentration the same for each sample (1mg/ml). Three repeat measurements were recorded from each of the 21 samples and collection of all spectra was randomised. Prior to analysis data was pre-processed using the method optimised in the previous chapter (3.3.3.1).

PLSR was applied to the data, initially using PyChem software. Alternate samples were used for training and testing the PLSR model (i.e. 0%, 10%, 20% … 100% for training and model calibration and 5%, 15%, 25% etc. for test). Figure 4.9 shows the PLSR predictions, demonstrating a good correlation between predicted and actual concentrations of glycosylated proteins. However the RMS test error for this model is fairly high at 15.54%.

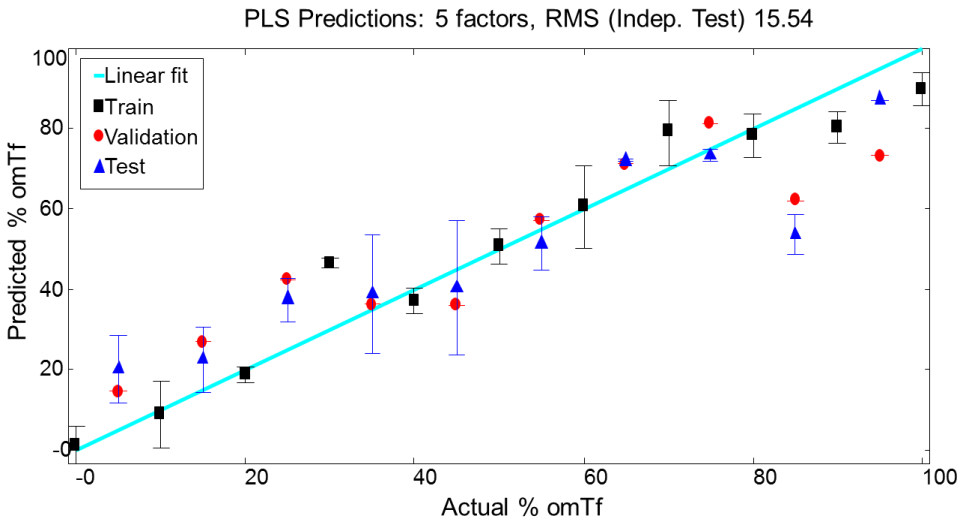

Figure 4.9: PLSR predictions from Raman data of transferrin mixtures, mean predictions of three measurements are plotted with standard error bars. (Data pre-processing: Sav-Gol smoothing, ALS baseline correction and EMSC).

In order to refine the PLSR model and to improve the error we have applied bootstrap cross validations to the data set. In this analysis we have randomly chosen which data are used for test and training and then used this selection to build the PLSR model. We have chosen to compare two bootstrapping methods for this analysis: In the first method we have made different samples of the same level of glycosylation available for use in training and testing data sets (i.e., one sample of 5% omTf could be used in training and the another sample of 5% omTf could be used for test). For the second method we have kept all replicate samples of the same glycosylated protein concentration exclusively in either the test or training sets (i.e., all 5% omTf samples were used in training and all 10% omTf samples were used for test). This process has been repeated 1000 times and the typical model achieved by each method is shown in Figure 4.10. It should be noted that in for these models additional data for lower concentrations of omTf have been used; spectra recorded at 1, 2 ,3 ,4 … 9% omTf have now been introduced to the data set.

The PLSR predictions shown in Figure 4.10 show that models produced from both of these methods outperform the original model, which was calculated without bootstrap cross validation (Figure 4.9). The range of R^2 values for each of the 1000 models and null models (permutations; where the target concentrations are randomised) were calculated

and are plotted in the inset as box and whisker plots; these show that the error is significantly reduced from the initial model. It is clear from Figure 4.10 that the validation method in which samples were free for use in both testing and training (Figure 4.10A) preforms much better than the alternative method, with an average R^2 of 0.85 in method A and 0.6 in method B. It could be said that the model shown in Figure 4.10B is a much more valid demonstration of the capability of this method to predict the concentration of glycosylated protein and that method A is perhaps an overly optimistic representation. However the loadings from model A (Figure S4.3), are very similar to the PCA loadings for ApoTf and omApoTf (Figure 4.6), and show a number of features which can be assigned to both sugar and protein vibrations. This provides conformation that these PLSR models are based on real spectral differences between the two variants of transferrin which are changing in the Raman spectra as the relative concentrations of each species vary.

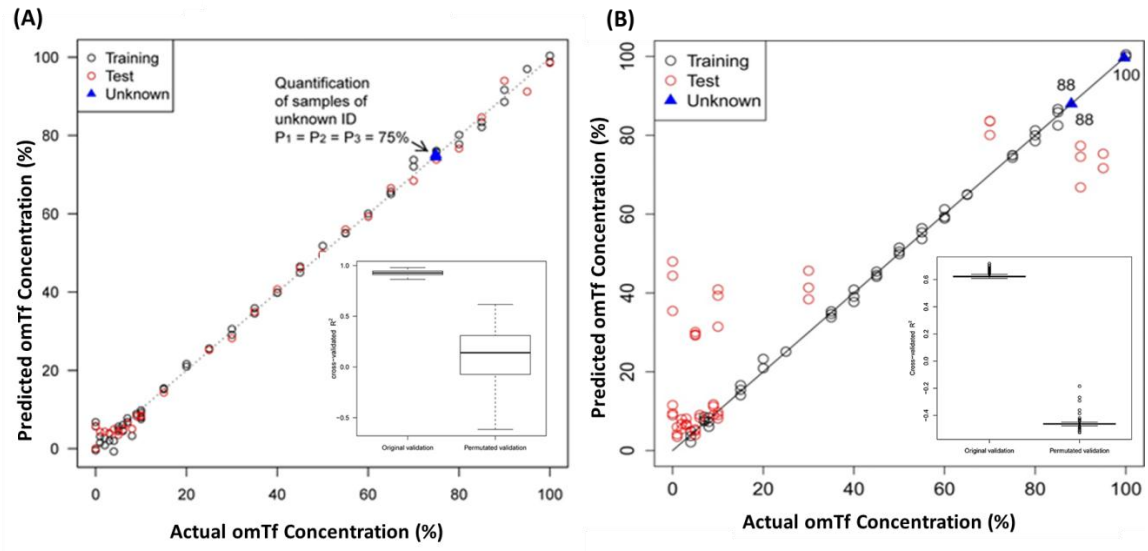

Figure 4.10: Typical PLSR predictions from Raman data of transferrin mixtures over 1000 bootstrap cross validations with **(A)** samples free for use in both training and test data sets and **(B)** Samples of same concentration kept together in either the training or test set.(Data pre-processing: Sav-Gol smoothing, ALS baseline correction and EMSC). INSET: Box and whisker plot showing R^2 values for the original and permuted models.

In order to test the validity of these PLSR models further, permutation testing has also been applied to the data (see insets in Figure 4.10), showing that the model does not perform well when all the class labels are randomly swapped. This indicates that both our PLSR models are true results which have not occurred by chance.

The PLSR model was then used to predict the level of glycosylated transferrin in the product sample and this prediction was 75%. Correlation analysis of the product spectra and the spectra of 75% omApoTf appears to confirm this (Figure S4.4). However the value supplied by Novozymes was ~50% concentration of oligo-mannosylated protein. As it possible that the product is a three component mixture of all three variants, this discrepancy could be due to an unknown concentration of mono-mannosylated transferrin in the product. Data on levels of mono-mannosylated protein in the product sample were not supplied by Novozymes.

4.3.2 Optim 1000 Analysis.

The Raman data presented in this chapter shows clear evidence of structural differences between the glycosylated variants of transferrin proteins. It was therefore of interest to determine how these structural changes could affect the stability of the recombinant protein product. This was achieved using temperature ramps and temperature holds of proteins in an Avacta Optim 1000 instrument and measuring fluorescence at light scattering in real time. This method has been developed by our industrial collaborators to probe protein stability in terms of unfolding and aggregation (Webster, 2010).

4.3.2.1 Optim Spectra of Transferrin.

The Optim spectra of all transferrin variants are shown in Figure 4.11. These spectra show apotransferrin to have a much larger intrinsic fluorescence emission band than holotransferrin, which is maybe to be expected as the apo-protein has a more open structure which leaves tryptophan and tyrosine residues more solvent exposed.

It can also be seen from the Optim spectra that there is much more variation in the intrinsic fluorescence emission of the different variants of holotransferrin than there is with the apotransferrin. The oligo-mannosylated holotransferrin has a much smaller fluorescence emission band than the non-glycosylated transferrin, which is indicative of a change in higher order structure which causes the aromatic amino acids to be less exposed to the environment.

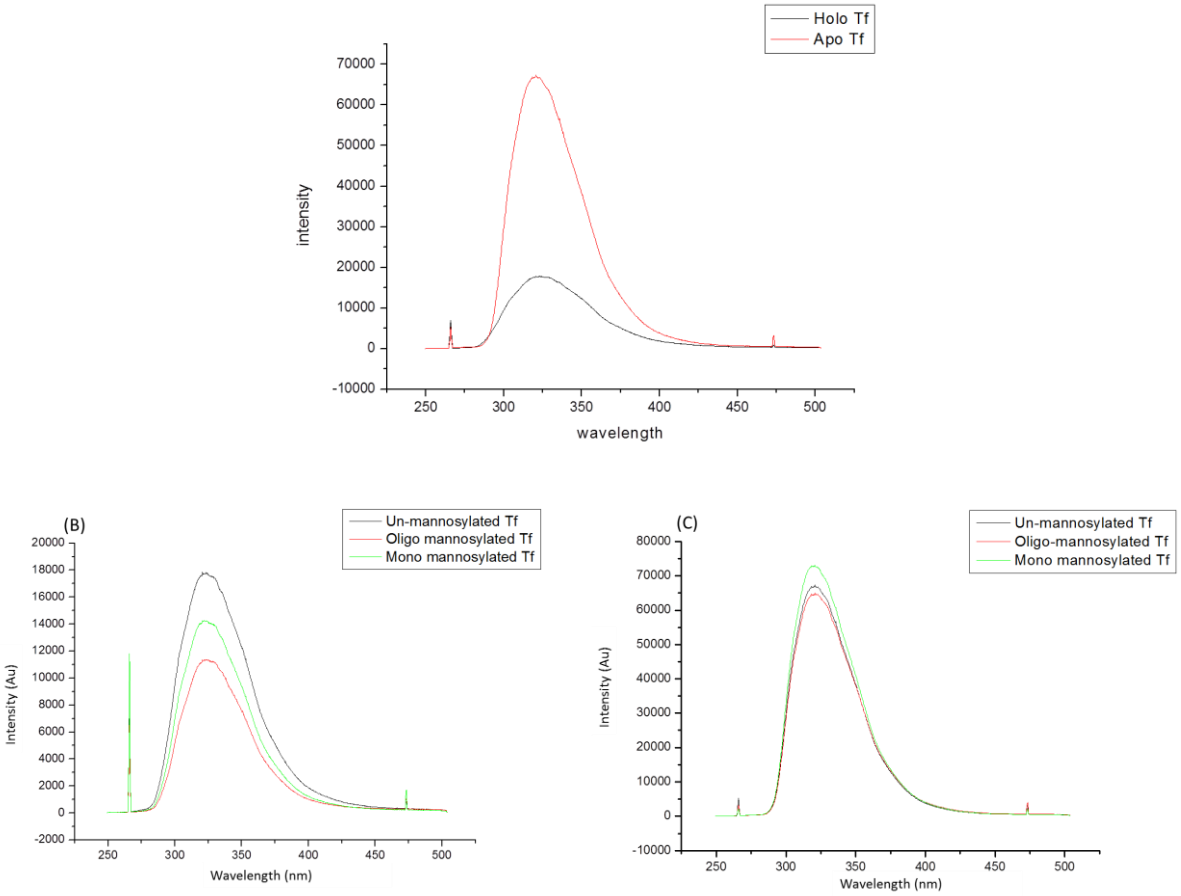

Figure 4.11: Optim 1000 spectra of **(A)** holo- *vs*. apo-transferrin, **(B)** glycosylated variants of holotransferrin and **(C)** glycosylated variants of apotransferrin.

4.3.2.2 Optim Thermal Ramp Experiments.

4.3.2.2.1 Profiling Stability.

As described in 4.2.4.1, a temperature ramp from 30 ºC to 85 ºC was applied to the samples, with spectra recorded at 1 ºC intervals. Optim spectra were then imported into the Optim analysis software for primary analysis. This allows unfolding curves to be drawn from many different spectral parameters of the Optim data (maximum fluorescence intensity, integrated peak areas, spectral centre of mass etc.) plotted against temperature. For this study we have chosen to use the ratio of fluorescence intensity at 350 nm to the fluorescence intensity at 330 nm (ratio 350:330), as this will allow us to track changes in band shape as well as intensity.

Figure 4.12 shows the unfolding curves for all transferrin samples generated from the Optim spectra, where it is clearly evident that the unfolding profiles of apotransferrins and holotransferrins are vastly different. Apotransferrin has two sharp cooperative transitions, whereas holotransferrin has two (or maybe more) broad overlapping transitions. In addition, the initial higher ratio values in holotransferrin suggest that it has more solvent exposed tryptophan.

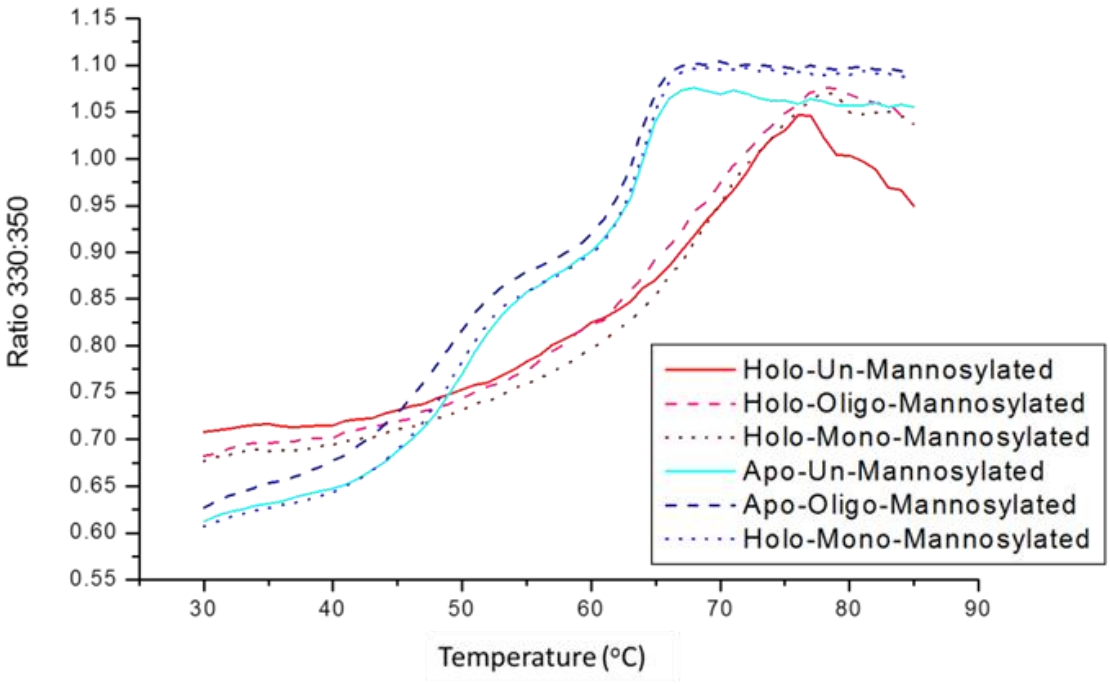

Figure 4.12: Graph to show the ratio of fluorescence intensity at 330 and 350 nm as a function of temperature. Plotted values are the mean of three independent runs, each of which contained three replicates.

There is little or no variation between the different glycosylated variants of transferrin that can be observed from these curves. Therefore more information has been extracted from these data by plotting the first derivative of the curves in order to calculate the transition midpoints; the temperature at which approximately half of the protein molecules are unfolded. The transition midpoints (T_m) for all transferrin samples are summarised in Table 4.3.

The first thing to be noted from Table 4.3 is that, overall, the calculated T_m values are higher for holotransferrin than for apotransferrin. This increase in stability in the iron containing protein is most likely due to the more closed structure it adopts, which is held in place by the iron.

Table 4.3: T_m values calculated from the first derivative of the unfolding curves. Each value is the mean nine measurements over three independent runs. (SD=standard deviation)

Sample Name	1st Transition (ºC)		2nd Transition (ºC)	
	Mean	SD	Mean	SD
Tf	51.53	0.34	70.58	0.88
mmTf	55.14	1.02	66.63	0.89
omTf	53.14	0.58	69.17	1.23
ApoTf	50.09	0.11	63.55	0.08
mmApoTf	49.69	0.17	63.21	0.28
omApoTf	49.42	0.06	63.66	0.10

Glycosylation can be seen to have only small effects on the stability of transferrin, particularly in the apotransferrin, where the variation in T_m between the different glycosylation states is only 0.4 ºC. The T_m calculations for holotransferrin show much larger differences in stability, with the oligo-mannosylated form being the most stable (highest T_m) and the un-mannosylated protein the least stable (lowest T_m), with mono-mannosylated transferrin falling between the two.

In order to try to highlight the differences between glycosylated variants further, 2D correlation analysis was applied to the data set. When 2D correlation calculations were performed on the whole Optim spectra, the resulting contour plots were found to be

dominated by the light scattering bands at 266 and 473 nm. Therefore plots shown here have been calculated from only the intrinsic fluorescence region of the spectrum (280-400 nm). The synchronous 2D correlation plots for all samples, which display the relative similarities in the spectra, can be seen in supplementary information (Figure S4.6). These plots corroborate the conclusions from the T_m calculations showing that all of the apotransferrins have very similar unfolding profiles, whereas differences can be observed in unfolding behaviour between the different glycosylation states of holotransferrin.

A variation of 2D correlation analysis, moving windows, was also been applied to the data set. Figure 4.13 shows the moving windows contour plots for all six transferrin proteins. The first thing to be noted from these plots is that the regions of maximum change (where there are the most contours, shown in red) correlate well with the transitions regions seen in the unfolding curves. Although the centers of these red contour regions do not match T_m calculations exactly, the T_m values do fall within these regions of maxium change. In addition the trends displayed here in stability between the glycosylated variants of transferrin are consistent with those observed in the unfolding curves and T_m calculations. calculations. However, unlike the unfolding curves, the moving windows analysis shows at least two clear transitions for both apo- and holo-transferrin. In addition, the regions of maximum change in the holotransferrin samples appear to be higher in the glycosylated proteins (omTf and mmTf) than in the non-glycosyated from; this further confirms that glycosylation can increase stability in transferrin, but only for the holo- form of the protein.

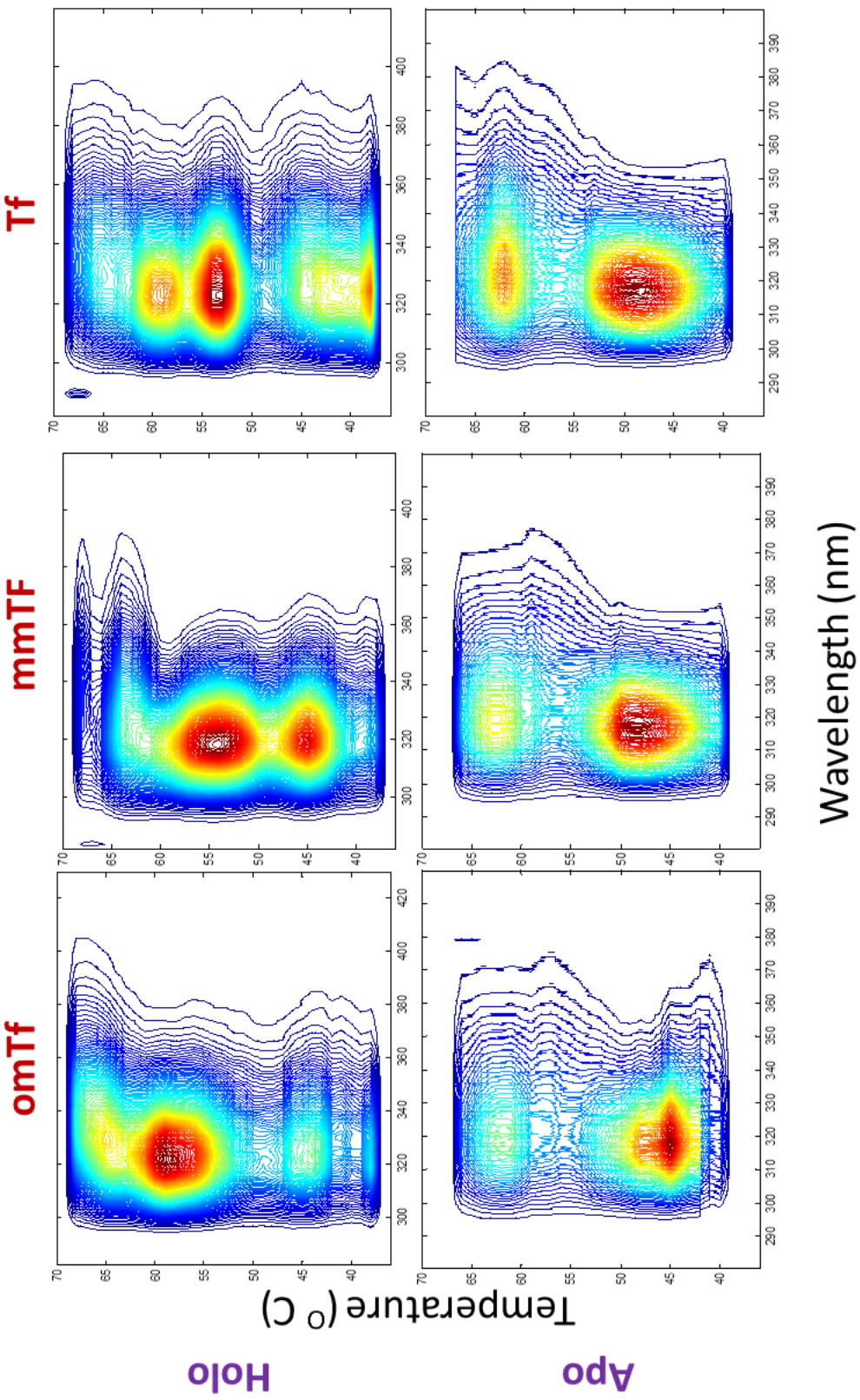

Figure 4.13: 2D-correlation moving windows contour plots as a function of spectral wavelength and average translating window temperatures for Optim transferrin data.

4.3.2.2.2 Profiling Aggregation.

In addition to unfolding profiles, Optim spectra allow us to monitor aggregation behaviour at different temperatures by tracking changes in the light scattering bands at 266 and 473 nm. The light scattering data collected from transferrin proteins suggests that glycosylation has huge effects on the aggregation propensity of the samples; this can be seen in Figure 4.14, where the intensity of the 266 nm band is plotted as a function of temperature.

The data for holotransferrin shows that initially aggregation is slow, but then there is a sharp increase in aggregation rates at ~50-60 ºC, which is associated with the unfolding transitions. The drop in light scattering intensity at ~75 ºC indicates that the protein is precipitating out of the solution, and this was confirmed visibly (data not shown). There is considerable increase in the temperature at which aggregation begins with increasing number of glycans and also a small increase in the temperature at which precipitation begins.

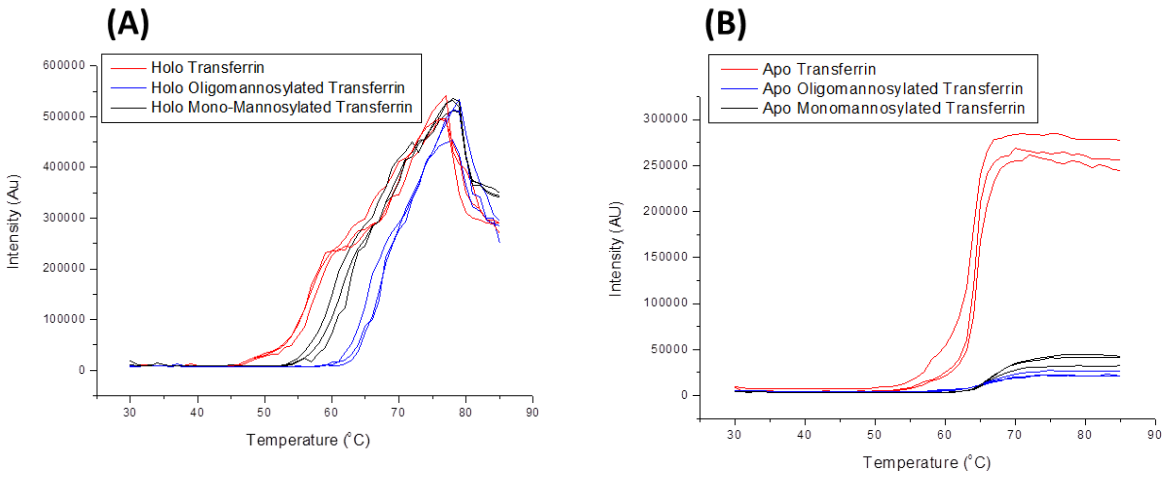

Figure 4.14: Graphs to show the intensity of light scattering at 266 nm as a function of temperature for **(A)** holotransferrin and **(B)** apotransferrin. Each trace is the mean of three repeat measurements.

Apotransferrin data again shows that the onset of aggregation coincides with the beginning of the unfolding transition. It can easily be seen from Figure 4.14B that

glycosylation dramatically reduces the total amount of aggregated protein and also slightly increases the temperature at which aggregation begins to occur.

4.3.2.3 Optim Isothermal Experiments.

In order to probe the effects of glycosylation on stability and aggregation further a series of isothermal experiments were performed. In this type of experiment samples were held at a specified temperature and spectral changes were monitored as a function of time. In this case three separate experiments were carried out at 45, 55 and 67 ºC, with spectra recorded every 60 s.

The isothermal data for unfolding display how glycosylation increases stability much more clearly than with the data collected in the thermal ramp experiments, particularly at 55 and 67 ºC. Figure 4.15 shows how the maximum fluorescence intensity of the holotransferrin proteins changes over time when held at 67 ºC. These data clearly show that oligo-mannosylated holotransferrin is much more stable at 67 ºC than its un-mannosylated equivalent. In addition, we can also see that mono-mannosylated protein appears to have a slightly slower rate of unfolding compared to the non-glycosylated transferrin. As with previous results apotransferrin samples show much less variation, with only very small differences in the rates of unfolding (data not shown).

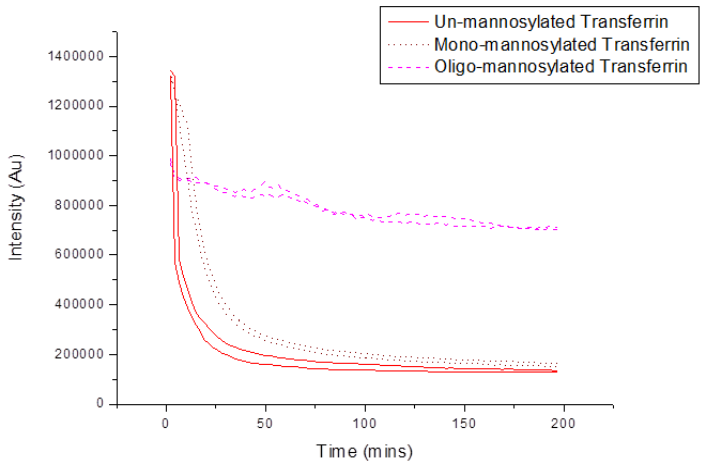

Figure 4.15: Graph to show the maximum fluorescence intensity of holotransferrin held at 67 °C as a function of time. Each trace is the mean of three replicate measurements.

Examination of the light scattering data from the isothermal experiments confirms that glycosylation is very effective at suppressing aggregation in transferrin proteins. At all three of the temperatures tested glycosylated proteins exhibit slower rates of aggregation and decreased total amount of aggregate in both holo- and apo-transferrin. In addition, these results show that apoised transferrin is much less prone to aggregation than holotransferrin, particularly at 45 and 55 ºC. It should be noted that in Figure 4.16B it may appear that the glycosylated transferrins aggregate more than the non-glycosylated holotransferrin, however this sample has a very fast initial rate of aggregation and then appears to be precipitating after ~8 min, causing a significant decrease in the intensity of light scattering observed.

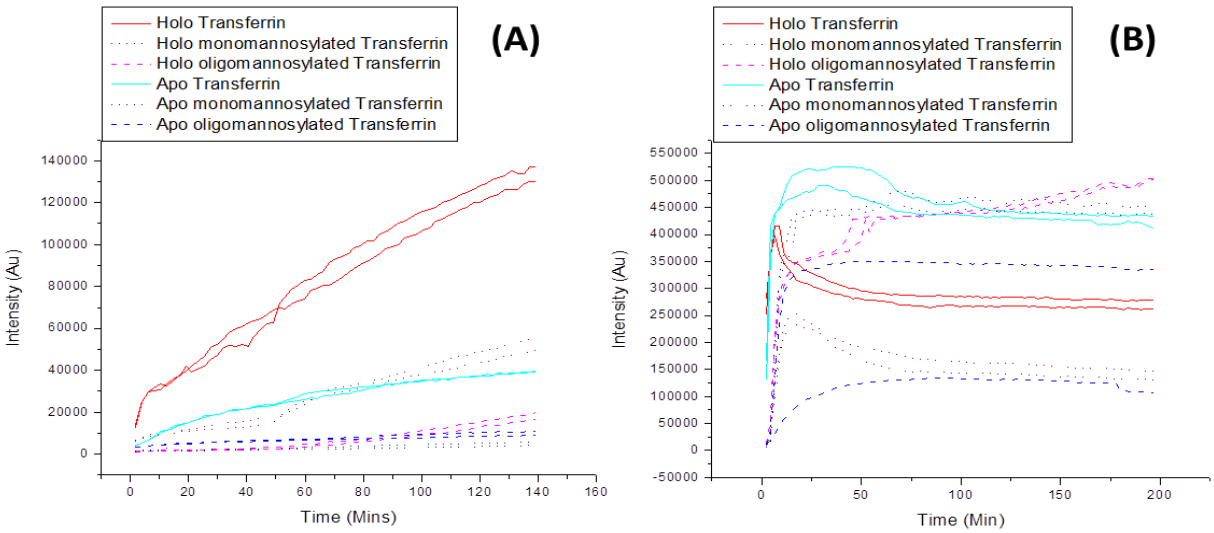

Figure 4.16: Graphs to show the intensity of light scattering at 473 nm as a function of time at **(A)** 45 °C and **(B)** 55 °C. Each trace is the mean of three replicate measurements

4.4 Conclusions.

We have shown in this study that the Raman spectroscopic methods developed in Chapter 3 for detecting glycosylation in a simple model system are transferable to more complex biopharmaceutical samples. Trends that were observed in the Raman data collected from transferrin were consistent with those seen previously with the RNase proteins, in particular which vibrational modes were used for discrimination and

quantification. We have extended the utility of Raman spectroscopy in this area further by demonstrating the ability to differentiate between different glycosylated variants of the same protein (omTf and mmTf), as well as apo- and holo-forms of these proteins. This was an important progression for this project as many cell lines express a variety of glycoforms, of which only one is the desired biopharmaceutical product.

The conformational changes which occur upon glycosylation are again clearly evident in the Raman data. For this reason we went on to characterise the implications of these structural changes with respect to stability and aggregation; two factors which need to be robustly characterised and controlled in any therapeutic protein product. Both thermal ramp and isothermal experiments show that in holotransferrin glycosylation can enhance the stability of the product, particularly in the oligo-mannosylated form. In addition, for both holo- and apo- proteins it has been found that glycosylation can greatly reduce the aggregation propensity of transferrin. These results may be useful to Novozymes in the design of a better product, especially with regards to the shelf-life of the protein.

4.5 Supplementary Information

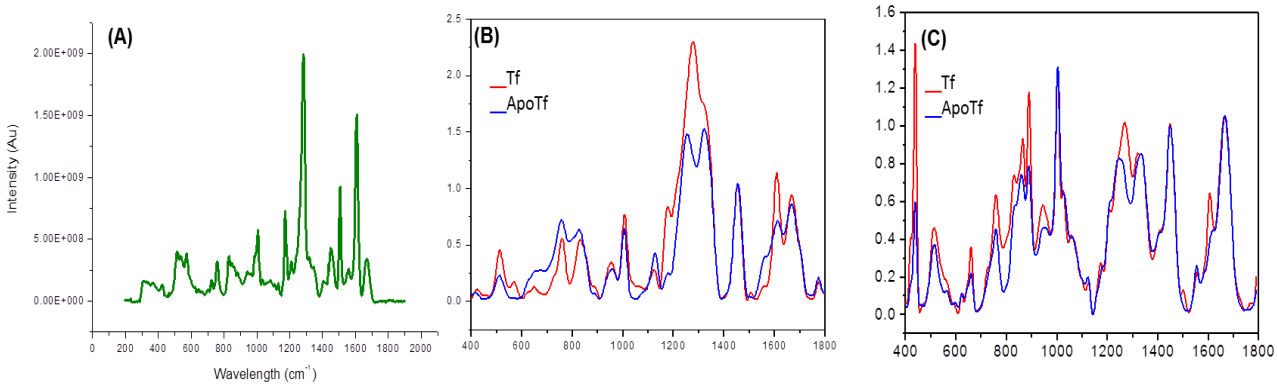

Figure S4.1: Raman spectrum of Holotransferrin recorded with **(A)** a 532 nm excitation wavelength, **(B)** a 633 nm excitation wavelength and **(C)** a 785 nm excitation wavelength. (Data have been baseline corrected (ALS)

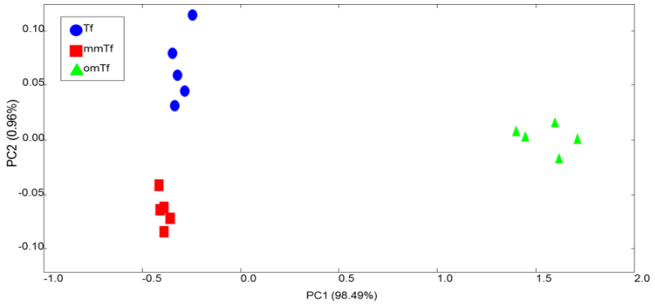

Figure S4.2: PCA scores plot (PC1 vs PC2) of FT-IR data from un-mannosylated holotransferrin (Tf) and oligo-mannosylated holotransferrin (omTf) and mono-manosylated holotransferrin (mmTf).

N.B. it should be noted that although Tf and mmTf look to separate across PC2, the total explained variance by PC2 is negligible compared the amount of variance explained in PC1. Therefore discounting the separation in the PC2 direction this plot looks similar to the corresponding Raman plot in Figure 4.4(B).

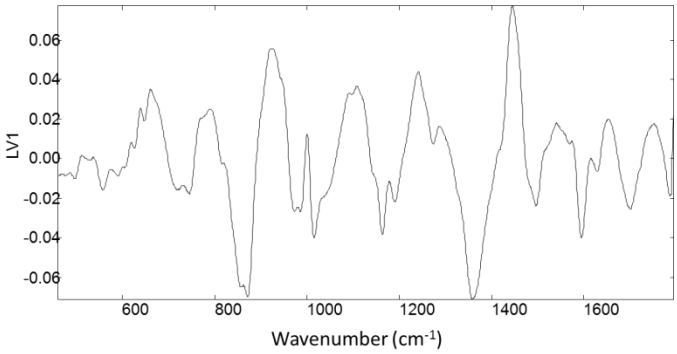

Figure S4.3: Loadings from the first LV from the PLSR model for the quantification of glycosylation in transferrin.

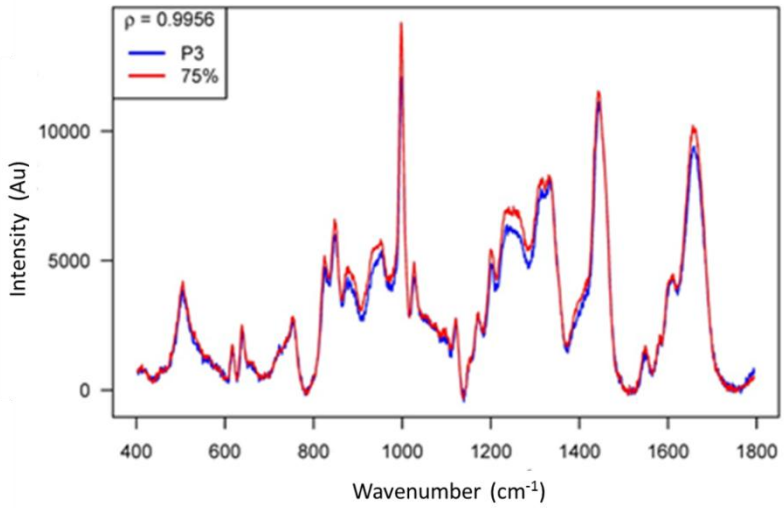

Figure S4.4: Correlation analysis of product samples with unknown levels of glycosylation and 75% glycosylated transferrin.

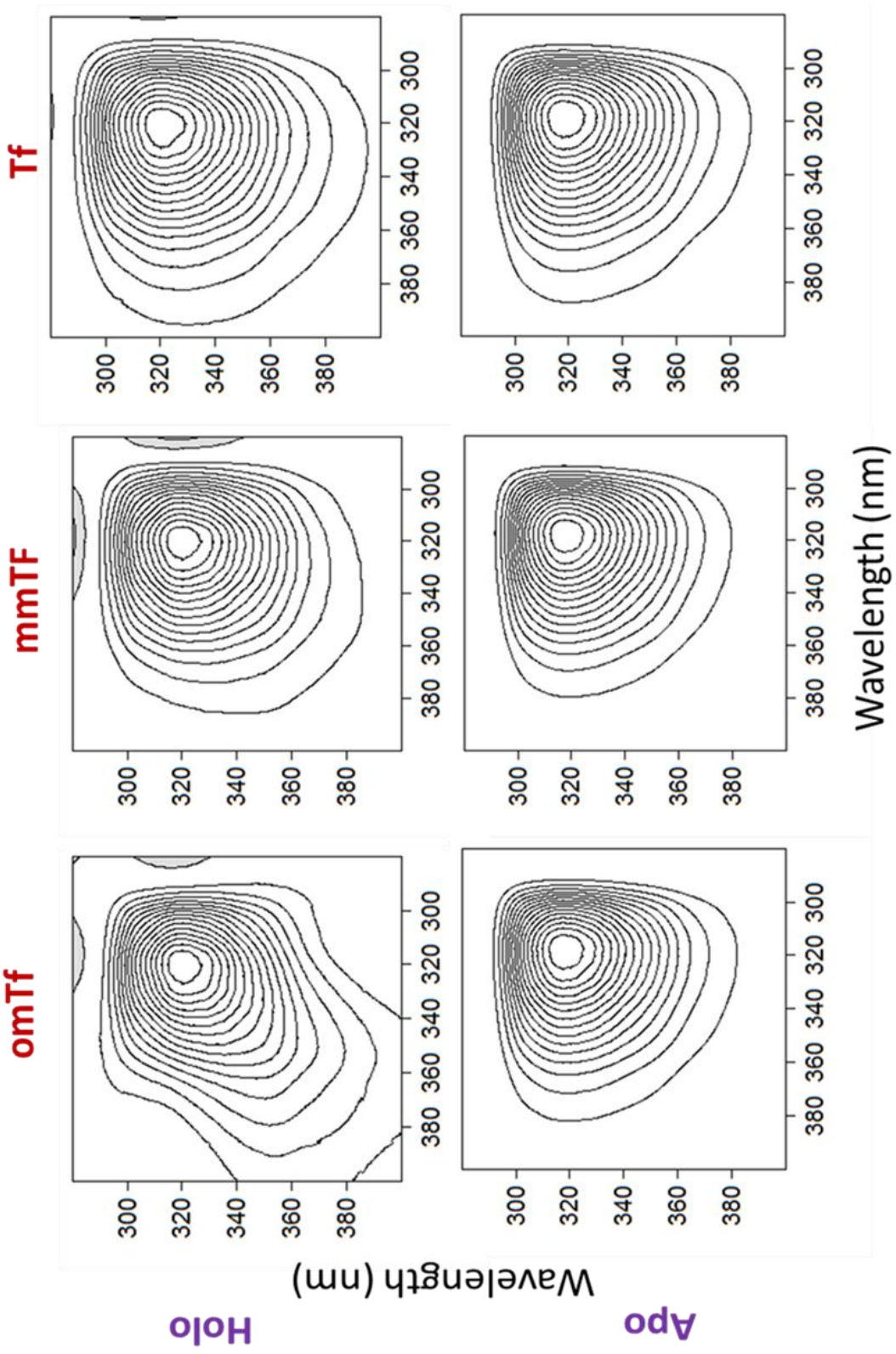

Figure S4.5: 2D-correlation synchronous contour plots of temperature dependent variations in the Optim spectra of transferrin proteins

Chapter 5: Characterising Different Variants of GFP Using Vibrational Spectroscopy and the Optim 1000.

5.1 Introduction.

Following on from the previous work, we now aim to characterise a set of green fluorescent protein (GFP) mutants which have been artificially glycosylated with various sugars at different glycosylation sites, using both vibrational spectroscopy and an Optim 1000 spectrometer.

GFP is comprised of at β-barrel type structure, with an α-helix running through the middle of the barrel (Figure 5.1). The α-helix has a chromophore covalently bound, which is responsible for the green fluorescence emission of GFP (centred ~510-550nm) (Tsien, 1998). Fluorescence emission is made possible due to interactions between oxygen atoms in the chromophore and neighbouring basic amino acid residues. Therefore when GFP is denatured, disruption to the network of hydrogen bonds causes a loss of fluorescence emission, making green fluorescence a suitable indicator of changes to tertiary structure (Tsien, 1998, Cubitt et al., 1995).

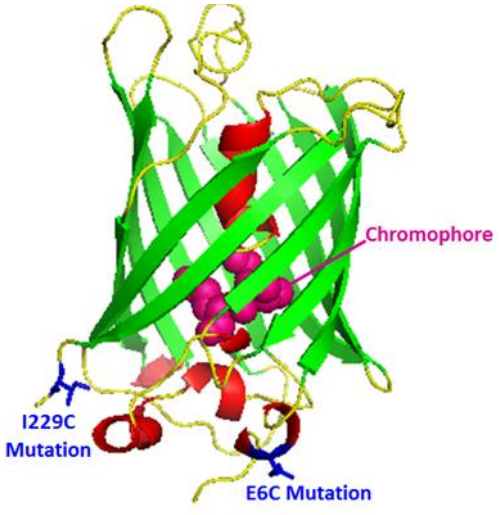

Figure 5.1: Cartoon diagram of GFP showing the positions of cysteine mutations and the Chromophore.

GFP samples received for analysis in this study had undergone point mutations to incorporate extra free cysteine residues at various positions. The proteins have then been artificially glycosylated, with either

glucose or mannose, through the thiol on the cysteine. This has been achieved by attaching an amino ethyl linker to the sugar and reacting the linker with the thiol group.

First, we aim to determine if it is still possible to detect glycosylation is systems such as this where the protein is larger (twice the size of RNase) and the glycan is much smaller than with the previously studies system. In addition, we also aim to characterise the effects of these mutations and modifications on the conformation and stability of the protein, in particular whether artificial glycosylations of this type still bring about the same type of structural changes as seen in natural glycosylation.

5.2 Materials and Methods.

5.2.1 Samples.

GFP samples were made by Andrew Martin and Sabine Flitsch at the University of Manchester. Four different mutants of GFP were received: A wild type, E6C mutant, which has a point mutation of glutamate 6 to cysteine, I229C mutant, which has a point mutation of isoleucine 229 to cysteine and a 'double' mutant which, has both of the 6 and 229 point mutations. The positions of these mutations are shown on the protein structure in Figure 5.1. In addition to the point mutation the I229 mutant has another mutation which clipped five amino acids from the end of the sequence.

Three different chemically modified variants of both E6C and I229C mutants were supplied: two glycosylated forms of the protein; one with a single mannose residue attached through the mutated cysteine, and one with a single glucose. We also received E6C, I229C and double mutants with only the amino ethyl linker which attaches the sugar. Samples are summarised in Table 5.1.

Table 5.1: Summary of GFP Samples.		
Mutant Type	**Modification**	**Abbreviation used in this work**
Wild Type	**None**	**WT**
E6C	**None**	**E6C**
E6C	**Glycosylated- Glucose**	**E6CG**
E6C	**Glycosylated-Mannose**	**E6CM**
E6C	**Linker Only**	**E6CL**
I229C	**None**	**I229C**
I229C	**Glycosylated- Glucose**	**I229G**
I229C	**Glycosylated-Mannose**	**I229M**
I229C	**Linker Only**	**I229L**
Double (E6C & I229C)	**None**	**D**
Double (E6C & I229C)	**Linker Only**	**DL**

5.2.2 Raman Spectroscopy.

Raman data were collected using the Renishaw 2000 Raman microscope described in Chapter 2. All Spectra were all were single accumulation, extended scans between 400 and 1800 cm^{-1}, with an exposure time of 120 s. 2 µL of sample were spotted onto a hydrophobic SpectraRIM™ slides, detailed in section 2.1.1.1, and allowed to dry out at room temperature for approximately 1 h. Each reported spectrum is an average of 6 spectra collected from different positions within each sample spot, as depicted in 2.1.1.1. Data were pre-processed (smoothing, baseline correction and normalisation) according to the method optimised in Chapter 3 (3.3.3.1).

5.2.3 FT-IR spectroscopy.

FT-IR spectra were collected on a Bruker FT-IR instrument described in 2.1.2. 4 µL of each sample was spotted onto a 96 well silicon plate and allowed to dry at room temperature. Spectra were recorded over 4000-600 wavenumbers, with 64 accumulations per sample.

5.2.4 Optim 1000.

5.2.4.1 Optim Thermal Ramp Experiments.

9 µL of three replicates of each sample were loaded into a micro cuvette array (MCA). A temperature ramp from 35 to 95 °C was applied to the samples with a temperature tolerance of 0.3 °C. Spectra were recorded at 1 °C intervals with a 60 s hold time at each temperature. Spectra were collected with 1 s exposure time with the slit width set to 100 µm. Each run was performed in triplicate, with three analytical replicates of each sample per run. Spectra reported in the application note in Appendix 2B, were recorded with and extended grating in order to capture the full green fluorescence peak, in this experiment data were collected with a slit width of 25 µm and a 10 ms exposure time.

5.2.4.2 Optim Isothermal Experiments.

9 µL of three replicates of each sample were loaded into an MCA. Samples were rapidly heated to, and held at, a set temperature, chosen from observing the results of the previous thermal ramp experiments. In this case samples were held at 70 °C with a tolerance of 0.5 °C and spectra were recorded at 60 second intervals for 200 min. Spectra were collected with 10 s exposure time with the slit width set to 100 µm. Each run was performed in triplicate, with three analytical replicates of each sample per run.

5.2.5 Microscopy.

Microscope images and Fluorescence correlation spectroscopy (FCS) measurements were recorded on a Zeiss LSM 51 ConfoCor 2 setup (Zeiss, Jena, Germany), equipped with an Argon laser and a 40 x objective lens. 400 µL of sample was analysed in a Lab-Tek-Nunc® eight-well chamber slide (Fisher Scientific, Leicestershire, UK). FCS measurements were performed at 90 runs each of 10 s duration. Single-component fits were applied to the FCS data and diffusion times were calculated by the instrument software.

5.2.6 Data analysis.

Vibrational spectroscopic data were exported into Matlab for pre-processing and PyChem was employed for PCA. Spectral figures were plotted in GRAMS Ai. Optim data were imported into Optim Analysis software for preliminary analysis. Data were then exported into Origin for further analysis and for plotting figures. 2D correlation calculations were performed using 2D shige freeware. PARAFAC of Optim data was performed in R.

5.3 Results and Discussion.

5.3.1 Raman Spectroscopy.

The initial aim of this section of work was to detect artificial glycosylation in E6C and I229 GFP using Raman spectroscopy. However as it became apparent that each mutant, and the different variants of each mutant, had much more conformational variation than previously expected, we also began to focus on structural interpretations of the Raman data.

5.3.1.1 Raman Spectroscopy of the I229C Mutant.

To begin with comparisons were made between non-glycosylated I229C and the protein glycosylated with glucose (I229G). PCA of this Raman data shows, once again, that differentiation between native and glycosylated forms is easily achieved. Figure 5.2A shows the two samples to be separated into two distinctive clusters across PC1, with 80.69% of the variance explained by this PC.

The Raman spectra of these samples are displayed in Figure 5.3, with the major bands from the PCA loadings indicated by the red asterisks. The most notable difference in the spectra is the band appearing at ~933 cm^{-1} in the glycosylated spectra. This band could be attributed the ring deformation modes from the sugar residue (Oleinikov et al., 1998). In addition we see a shoulder band increasing in intensity in the glycosylated GFP spectra at ~1046 cm^{-1}, again this band could be assigned to vibrations arising from the

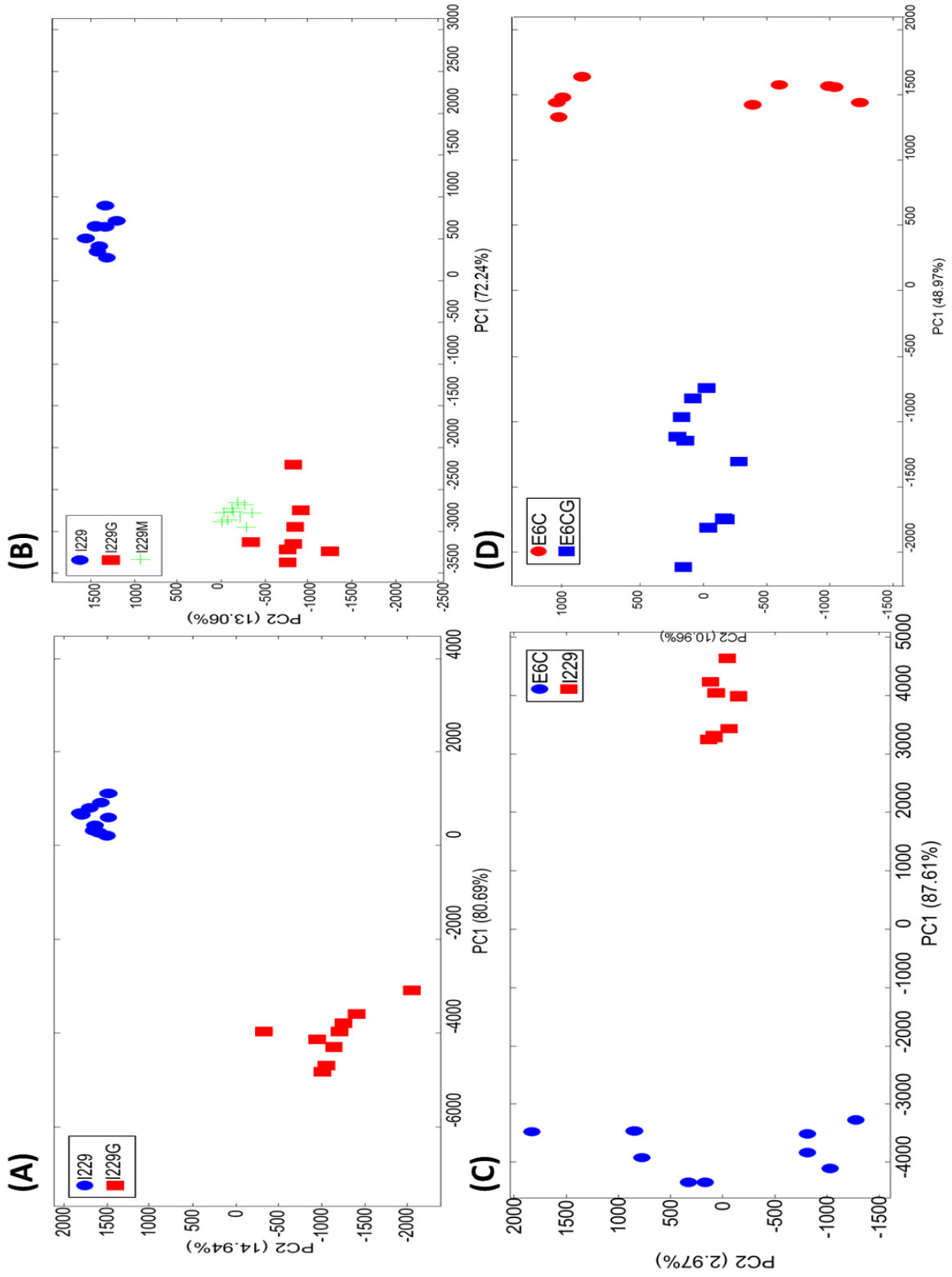

Figure 5.2: PCA scores plots (PC1 *vs* PC2) of Raman data from **(A)** I229C GFP and I229C GFP glycosylated with glucose, **(B)** I229C GFP and I229C GFP glycosylated with either glucose or mannose, **(C)** I229C GFP and E6C GFP and **(D)** E6C GFP and E6C GFP with glucose.

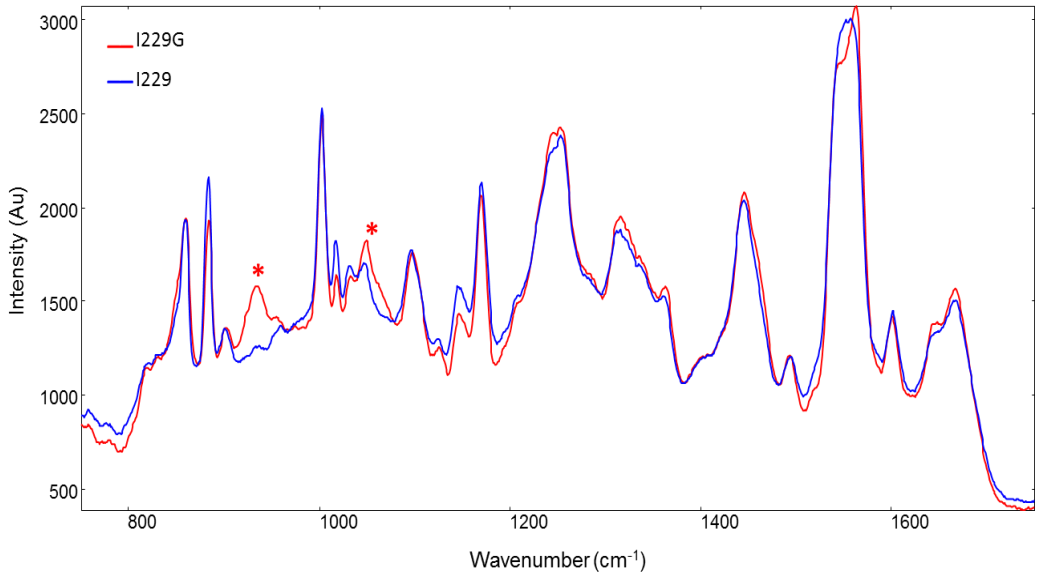

Figure 5.3: Average Raman spectra of I229C GFP and I229C GFP with glucose. Asterisks indicate the bands highlighted by the PCA loadings. (Spectra have been smoothed and baseline corrected).

glucose molecules, in this case glycosidic ring breathing (De Gelder et al., 2007). Furthermore we see a number of bands changing in the spectra which can be assigned to conformational differences between the two proteins; most notably in the band centred at ~1550 cm^{-1} arising from tryptophan residues (Barron et al., 2002). We also see variations in the amide I and amide III bands, comparable with the trends observed in these regions in the RNase system.

Following on from this we compared the Raman spectra from mannosylated I229C to the previous data. The PCA plot in Figure 5.2B shows the ability to distinguish between un-modified and glycosylated variants of I229C GFP across PC1. It could also be said that in this plot the two different glycans (mannose or glucose) appear to form separate clusters; suggesting that Raman spectroscopy has the ability to distinguish between different glycoforms of the same protein. Inspection of the Raman spectrum of mannosylated I229C (not shown) showed the same ring vibrations at ~933 and ~1046 cm^{-1} as observed in the I229G variant. This provided further proof that these vibrational modes are indeed glycosidic based.

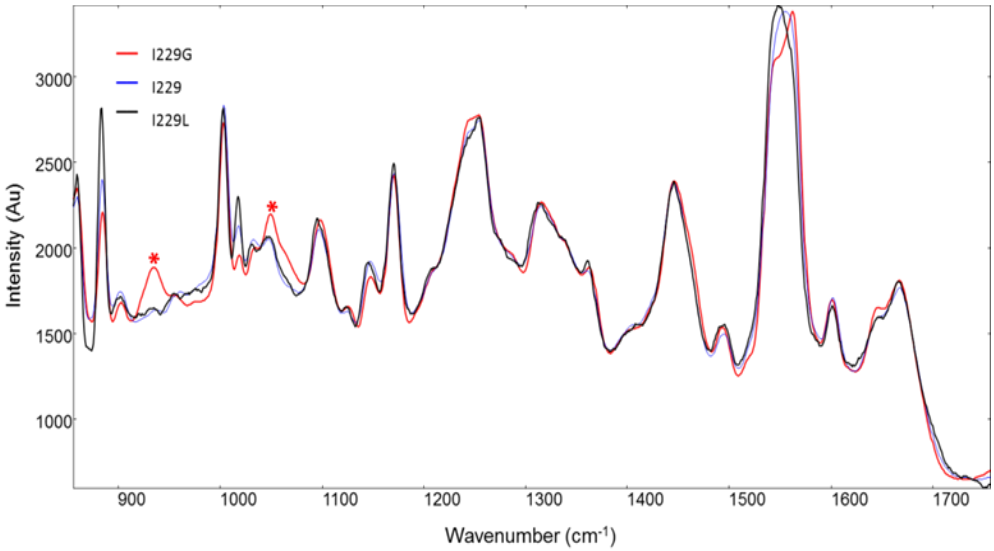

Figure 5.4: Average Raman spectra of I229C GFP and I229C GFP with glucose and I229C GFP with linker. Asterisks indicate the bands discussed in the text. (Spectra have been smoothed and baseline corrected).

In order to verify that these bands were due to the sugar moiety specifically, and are not just due to conformational differences in the proteins or bands arising from the amino-ethyl linker itself, we then examined the spectra of the I229C mutant with only the linker attached (I229L). The spectra, shown in Figure 5.4, display none of the features assigned to the glycosidic ring, and exhibit similar spectra to the un-modified protein in these regions. However it is still possible to distinguish between I229C and I229L based on differences attributed to structural changes in the proteins.

All of the I229C spectra suggest conformational differences between the proteins, the most notable difference being in the tryptophan band at ~1550 cm^{-1}, specifically due to the indole ring breathing mode of the amino acid side chain (Figure 5.5), which is indicative of changes in the higher order structure of the protein. GFP has only one tryptophan residue, which is located on the central

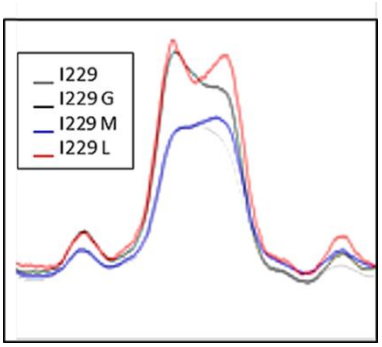

Figure 5.5: Raman spectra of all I229C GFP mutants focussed on the Indole ring breathing mode (~1550 cm^{-1}).

alpha helix. The changes in this band suggest that the different variants have differences in tertiary structure which alter the solvent exposure of the tryptophan residue.

5.3.1.2 Raman Spectroscopy of the E6C Mutant.

First we compared the Raman spectra of unmodified E6C to the spectra recoded from the I229C mutant. The two mutants were certainly distinguishable by PCA, with 87.61% of the variance described by PC1 (Figure 5.2C). This strongly suggest a notable variation in tertiary structure between the two mutants as a single point mutation alone is unlikely to cause this much variance in the Raman spectra.

This difference in higher order structure brought about by the mutations is confirmed by examining the average Raman spectra of E6C and I229C and the PCA loadings for PC1. Once again the most notable difference can be ascribed to a change in the environment around the tryptophan residue. This is corroborated by the increase in intensity of the ~1630 cm^{-1} region of the amide I band in the I229C spectra, which suggests that α-helix structures, which is where the tryptophan is, are becoming more solvated (Takekiyo et al., 2006). Furthermore, the amide III region exhibits a shoulder at ~1240 cm^{-1} which is more intense in the I229C mutant. This band can be specifically assigned to ordered β-sheet structure (Huang et al., 2006, Liang et al., 2006), suggesting disruptions to the β-barrel structure in one of the mutants.

Spectra recorded from glycosylated E6C, shown in Figure 5.6, do not show the same trends as seen in the I229C samples. The glycosidic bands seen in I229G and I229M at ~933 and ~1046 cm^{-1} are not present in the spectra of E6C glycosylated with either mannose or glucose. Nonetheless it is still possible to distinguish between glycosylated and un-modified E6C variants using PCA, as displayed in Figure 5.2D. PCA loadings show that this separation is based primarily on the structural features discussed previously: amide I, amide III and tryptophan bands.

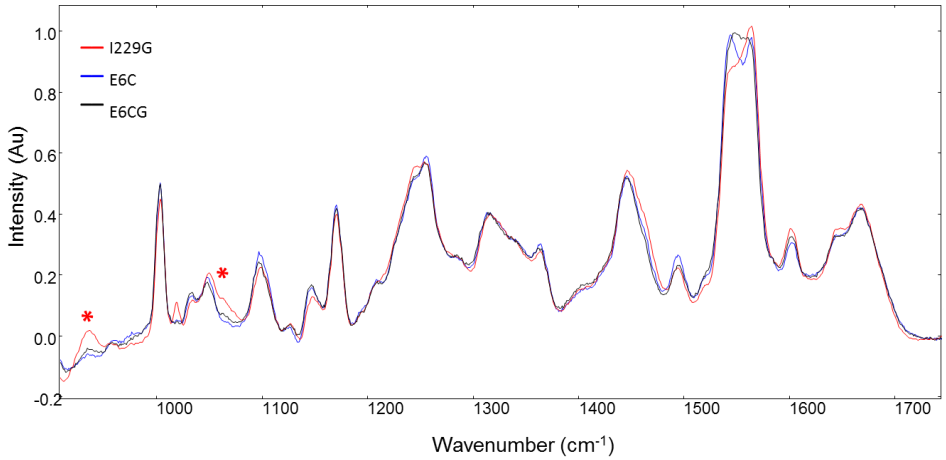

Figure 5.6: Average Raman spectra of E6C GFP and E6C GFP with glucose and I229C GFP with glucose. Asterisks indicate the bands discussed in the text. (Spectra have been smoothed and baseline corrected).

The final step in this analysis was to compare the data collected from all E6C and I229C variants. This comparison displaying the trends between all samples is shown in Figure 5.7. There is a clear separation across PC1 of the modified (glycosylated and linker) and the un-modified mutants, with sub-clusters of glycosylated variants and protein with only the linker attached. It could also be argued that there is a separation between the proteins glycosylated with mannose and the proteins glycosylated with glucose. Moreover the E6C and I229 mutants separate based on structural differences across PC2.

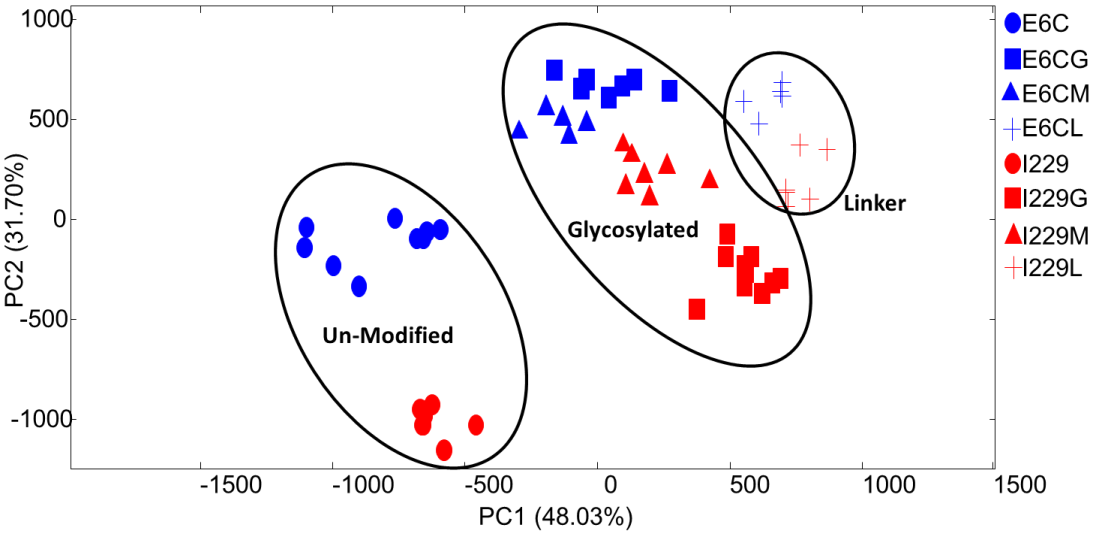

Figure 5.7: PCA scores plot (PC1 *vs* PC2) of Raman data from all I229C variants and all E6C variants.

In conclusion, Raman data clearly shows structural differences between the I229C and E6C mutants. The major difference in the Raman spectra of these mutants was found to be in the indole ring breathing mode of the tryptophan residue; which could be due to a change in higher order structure which leaves the tryptophan more exposed. A change in conformation of this magnitude is unlikely to be caused by a single point mutation alone, but could be due to the additional mutation which deleted 5 amino acids from the end of the I229C sequence (Figure 5.8). An alternative hypothesis for the variation in the data collected from these GFP samples is that the free cysteine residues which are present in the un-modified mutants make the molecules more prone to aggregation and therefore altering the structural bands which arise from the proteins.

We have also shown that we are able to distinguish between glycosylated and non-glycosylated proteins in this artificially produced system. In the I229C mutant we have identified bands in the spectra which are solely due to the glycan vibrations. However these bands were not detected in the spectra of the glycosylated E6C mutants, and distinction in this case was based on the structural differences brought about by glycosylation. This leads to the question: why can we see these sugar vibrations in one mutant and not in the other? One theory is that the glycan is simply more accessible in the I229C mutant. Although the glycosylation sites appear to be equally exposed in the cartoon shown in Figure 5.1, a closer look at the amino acid sequences of these particular mutants show us that the addition of a polyhistidine-tag (to aid in purification) at the start of the sequence and the clipping of five amino acids at the end, will leave the I229C mutation much closer to the end of the sequence than the E6C glycosylation.

5.3.2 FT-IR Spectroscopy.

FT-IR spectroscopy was performed in order to corroborate the structural differences alluded to by the Raman data. For FT-IR spectra and PCA plots see supplementary information (Figure S5.1). As with the Raman data it was possible to distinguish between E6C and I229 mutants based on their IR spectra; again the PCA loadings indicate a band and ~1550 cm^{-1} to be instrumental in this separation (Figure S5.1A and C). FT-IR spectra

I229C Mutant

MHHHHHHSSGLVPRGSGMKETAAAKF
ERQHMDSPDLGTDDDDKAMADIGSEF
MSKGEELFTGVVPILVELDGDVNGHKFS
VSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTFSYGVQCFSRYPDHMKRHDF
FKSAMPEGYVQERTISFKDDGNYKTRA
EVKFEGDTLVNRIELKGIDFKEDGNILGH
KLEYNYNSHNVYITADKQKNGIKANFKI
RHNIEDGSVQLADHYQQNTPIGDGPVL
LPDNHYLSTQSALSKDPNEKRDHMVLL
EFVTAAGCTHGMG

E6C Mutant

MHHHHHHSSGLVPRGSGMKETAAAKF
ERQHMDSPDLGTDDDDKAMADIGSEF
MSKGECLFTGVVPILVELDGDVNGHKFS
VSGEGEGDATYGKLTLKFICTTGKLPVPW
PTLVTTFSYGVQCFSRYPDHMKRHDFFK
SAMPEGYVQERTISFKDDGNYKTRAEVK
FEGDTLVNRIELKGIDFKEDGNILGHKLEY
NYNSHNVYITADKQKNGIKANFKIRHNI
EDGSVQLADHYQQNTPIGDGPVLLPDN
HYLSTQSALSKDPNEKRDHMVLLEFVTA
GITHGMDELYK

Figure 5.8: Amino acid sequences of I229C and E6C mutants. Red font indicates the his tag sequence, green font indicates sites of cysteine mutations and blue font highlights the amino acids which are clipped from the end of the sequence in I229C.

of I229C and glycosylated I229C can be easily separated by PCA, with loadings indicating that separation across PC1 is largely due to a band at ~1064 cm^{-1}, possibly arising from glycosidic vibrations (Figure S5.1B and D). Glycosylated E6C spectra do not exhibit this band. The PCA plot drawn from FT-IR data of all samples displays a similar trend to the Raman data with separation of modified and unmodified mutants across PC1, although distinction between mutants is less clear.

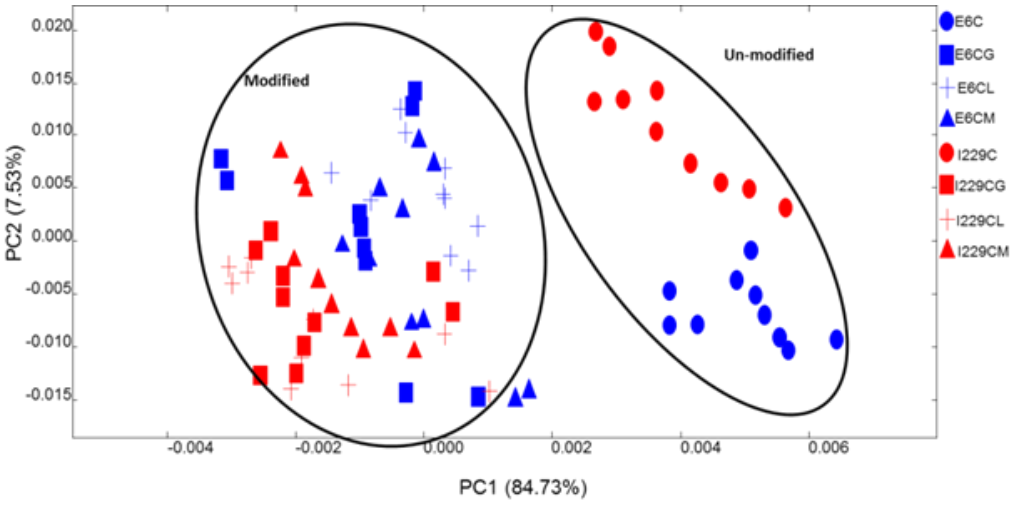

Figure 5.9: PCA scores plot (PC1 *vs* PC2) FT-IR data from all I229C variants and all E6C variants.

5.3.3 Optim 1000 Analysis.

Both thermal ramp and isothermal Optim experiments were performed on the GFP mutants, with the aim of confirming or disproving our hypotheses as to why there is so much variance displayed in the vibrational spectroscopy data. Furthermore, we wished to explore the effects of these conformational differences on the stability of the proteins. In addition to the I229C and E6C mutants have also compared here the profiles from the wild type and double mutants.

5.3.3.1 Optim Spectra of GFP.

The initial spectra at 30 °C recorded from each mutant are shown in Figure 5.10. We can observe from these spectra that the I229C sample has a substantially bigger fluorescence emission band when compared to other three mutants; indicating a change in the environment around the tryptophan residue. In addition we observe a red shift in the peak centre of the I229C mutant, ~323 nm in I229C compared with an average of ~305 nm in the other three spectra. This shift is indicative of more solvent exposed tryptophan in this sample, which was also suggested by the Raman data.

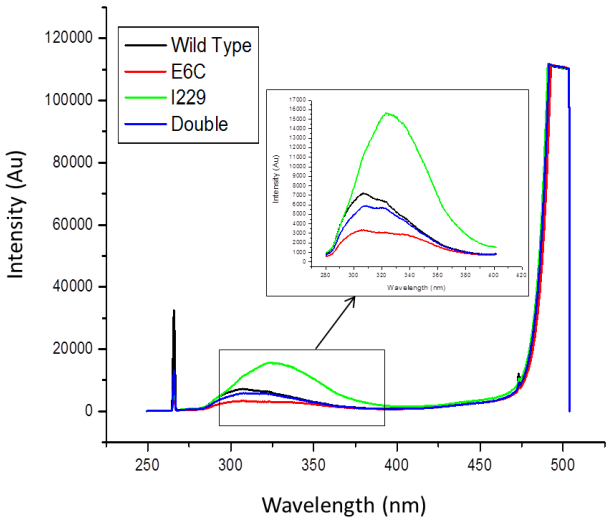

Figure 5.10: Optim 1000 spectra of all four GFP mutants showing a zoom of the intrinsic fluorescence region.

5.3.3.2 Optim Thermal Ramp Experiments.

A temperature ramp from 35 ºC to 95 ºC was applied to the samples, with spectra recorded at 1 ºC intervals. By plotting the Optim spectra collected at each temperature (Figure 5.11), we can observe a decrease in the natural GFP fluorescence (~500 nm) with increasing temperature and also changes in the intensity and shape of the intrinsic fluorescence band (~280-360 nm). Furthermore we can see increases in the intensity of both 266 and 473 nm light scattering bands above 75 °C.

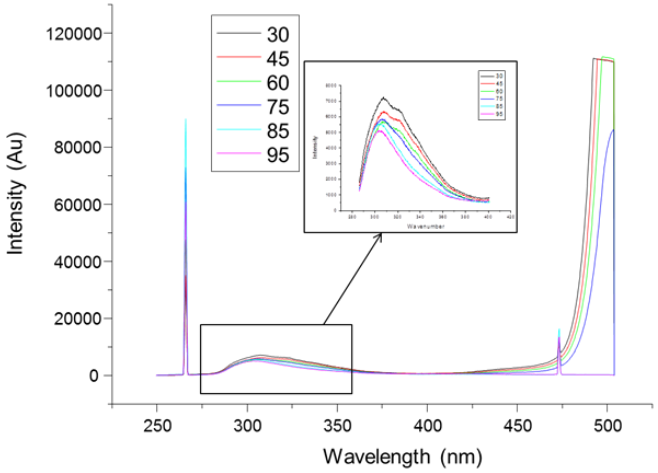

Figure 5.11: Optim 1000 spectra of Wild Type GFP over a temperature range showing a zoom of the intrinsic fluorescence region.

Optim spectra were then imported into the Optim analysis software for primary analysis, where unfolding curves were drawn from many different spectral parameters of the Optim data (maximum fluorescence intensity, integrated peak areas, spectral centre of mass etc.) plotted against temperature. A traditional unfolding curve was drawn by plotting the maximum intrinsic fluorescence intensity as a function of temperature, Figure 5.12A. This graph shows no particular unfolding transitions, which was also true for graphs drawn from the integrated peak area and the ratio of intensities at 350:330 nm (not shown), suggesting that it is not possible to track unfolding in GFP by monitoring changes in intrinsic fluorescence emission.

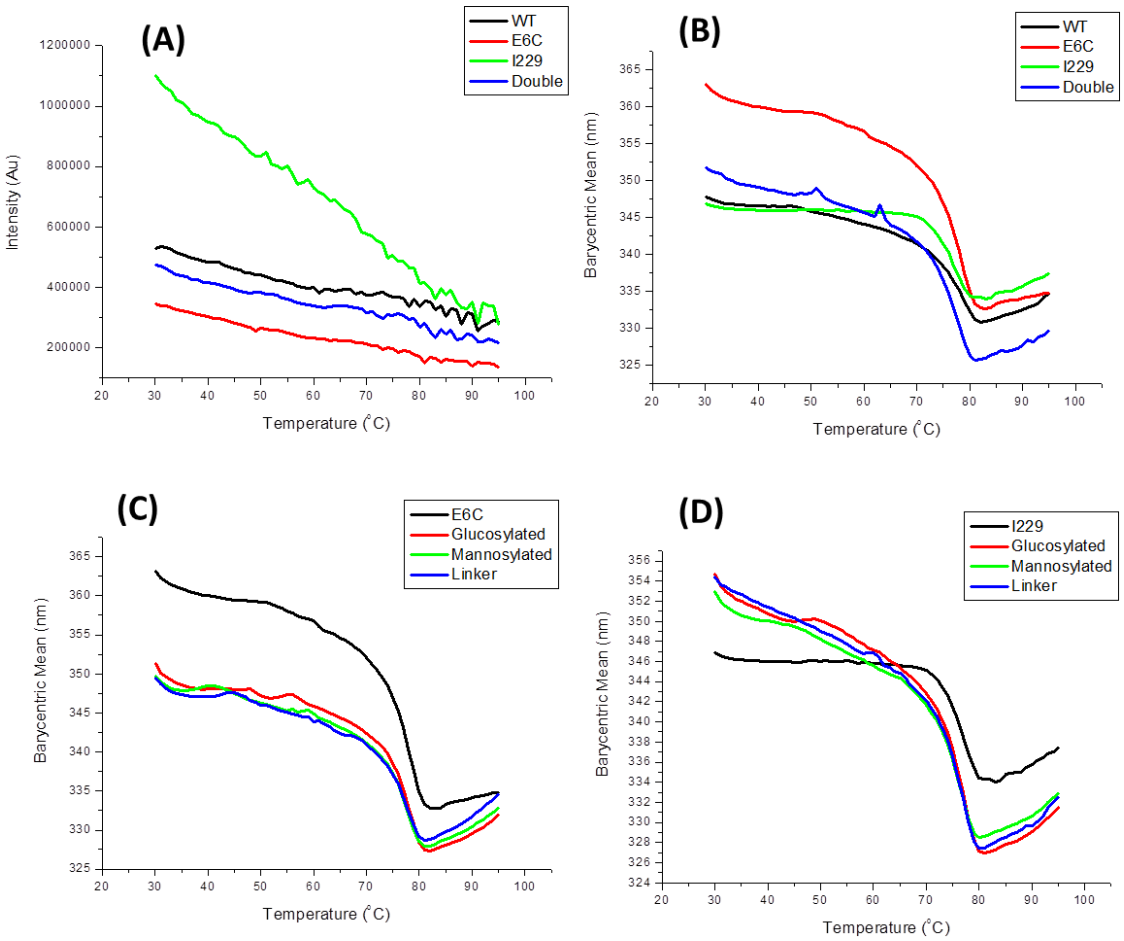

Figure 5.12: Graphs to show: **(A)** the maximum intrinsic fluorescence intensity as a function of temperature for unmodified mutants, **(B)** the barycentric mean of Optim spectra of unmodified mutants as a function of temperature, **(C) t**he barycentric mean of Optim spectra of all I229C variants as a function of temperature and **(D)** the barycentric mean of Optim spectra of all E6C variants as a function of temperature.

However, Figure 5.12B appears to indicate that we are able to monitor unfolding in GFP by plotting the barycetric mean (the spectral centre of mass) as a function of temperature. Although this particular spectral feature is heavily influenced by the changes in the natural green fluorescence of GFP, hence we are technically measuring disruption to the chromphore rather than unfolding of the protein. Work was carried out to investigate the accuracy of using of green fluorescence to monitor GFP denaturation, which lead to the production an application note for Avacta Analytical, which can be viewed in the appendix.

The unfolding profiles calculated from the barycentric mean display slightly different unfolding behaviour for each mutant. T_m values, calculated by potting the derivatives of the curves, are summarised in Table 5.2. T_m calculations confirm that there are small variations in the stability of each mutant; with I229C being the least stable and the wild type the most stable. Furthermore when comparing the profiles for modified I229C and E6C variants to the un-modified proteins (Figures 5.12A,C and D), we can see that in both mutants all of the modified proteins display similar behaviour which is distinctly different from the un-modified samples. In addition all of the GFP mutants investigated in this study exhibit slightly lower T_m values than those quoted in the literature for other GFP mutants which range between 83 to 85 ^{o}C (Alkaabi et al., 2005, Melnik et al., 2011).

Table 5.2: Summary T_m values calculated from the first derivative of the unfolding curves. Calculations were made from three independent measurements; T_m1, T_m 2 & T_m 3.

Sample Name	Tm1	Tm2	Tm3	Mean Tm
WT	77.99	77.99	77.06	77.68
E6C	76.99	76.96	75.93	76.62
I229C	74.98	75.35	75.37	75.23
D	75.42	75.97	75.04	75.48

5.3.3.2.1 Data analysis strategies for Optim 1000 data.

Various data analysis strategies were then investigated for use in extracting more useful information from the Optim data. As the curve drawn using the barycentric mean of the Optim spectra was largely based on the loss of green fluorescence and not changes in intrinsic fluorescence, we aim to use a variety of chemometric approaches to extract more meaningful information from the intrinsic fluorescence emission spectra. Even though the curve drawn fluorescence intensity showed no unfolding transitions, we were hopeful that data analysis can provide valuable information on the variations in intrinsic emission between mutants, as the raw spectra shown in Figures 5.10 and 5.11 display obvious differences in this region both between mutants and across a temperature range.

2D correlation analysis was applied to Optim spectra, however due to interference from the light scattering and green fluorescence bands, calculations were performed on the intrinsic fluorescence region only (280-400nm). 2D correlation contour plots, Figure 5.13, show that there is a large amount of variation in the unfolding profiles of the I229C samples, but that the E6C proteins all have very similar unfolding behaviour. 2D correlation spectroscopy was investigated in more detail for use in the analysis of Optim 1000 data which show little or no transitions in the Optim primary analysis, using the GFP data set. This work has been compiled as an application note for Avacta Analytical which can be found in the appendix.

Principal components analysis was also applied to the data set to determine if any more information about the variations in unfolding profiles could be extracted, and also to aid in the visualisation of differences and similarities between the samples. Although the unfolding curves in Figure 5.12 showed only small variations in unfolding behaviour, in the PCA plot of the Optim data from all GFP samples the un-modified I229C protein shows a very different trend to all the other samples, and the plot is dominated by the I229C profile (Figure 5.14A). When re-plotting the PCA with the I229C data removed we can easily see differences between all of the samples (Figure 5.12B).

Figure 5.15C shows the PCA of only the E6C samples, unlike with the 2D correlation analysis we are now able to see obvious differences in the unfolding behaviour of each variant. An interesting point to note here is that all of the modified mutants appear to be grouped together, a validation of the conformational differences between modified and un-modified proteins alluded to by the Raman spectra. Furthermore if we remove the sample labels and instead label each point with the temperature at which each spectrum was recorded we are able to observe the transition region for each sample, which correlates well with the T_m calculations in Table 5.2 (Figure 5.15D)

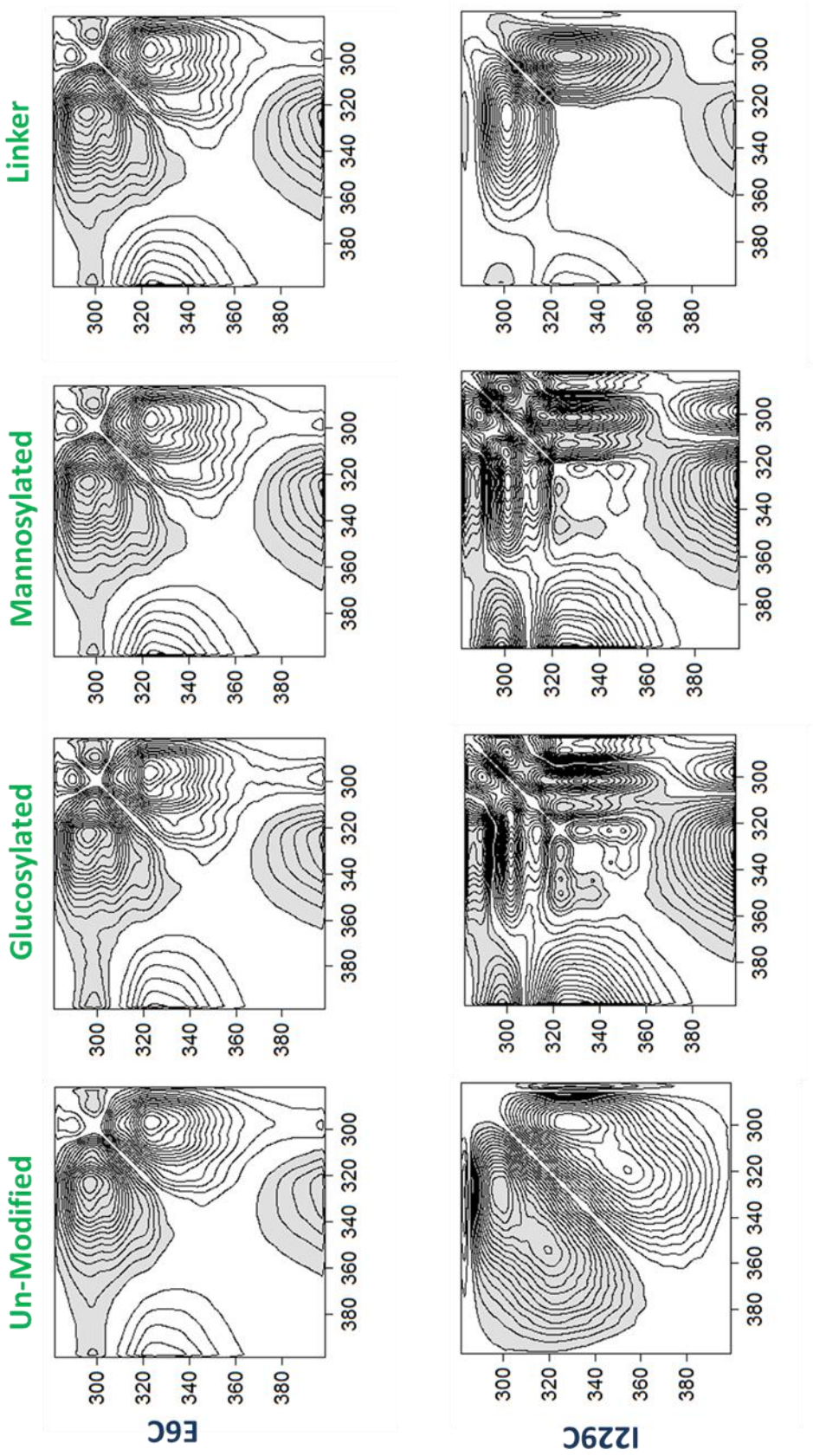

Figure 5.13: 2D-correlation asynchronous contour plots of temperature dependant variations in the intrinsic region (208-400 nm) of the Optim spectra of the E6C and I229C mutants.

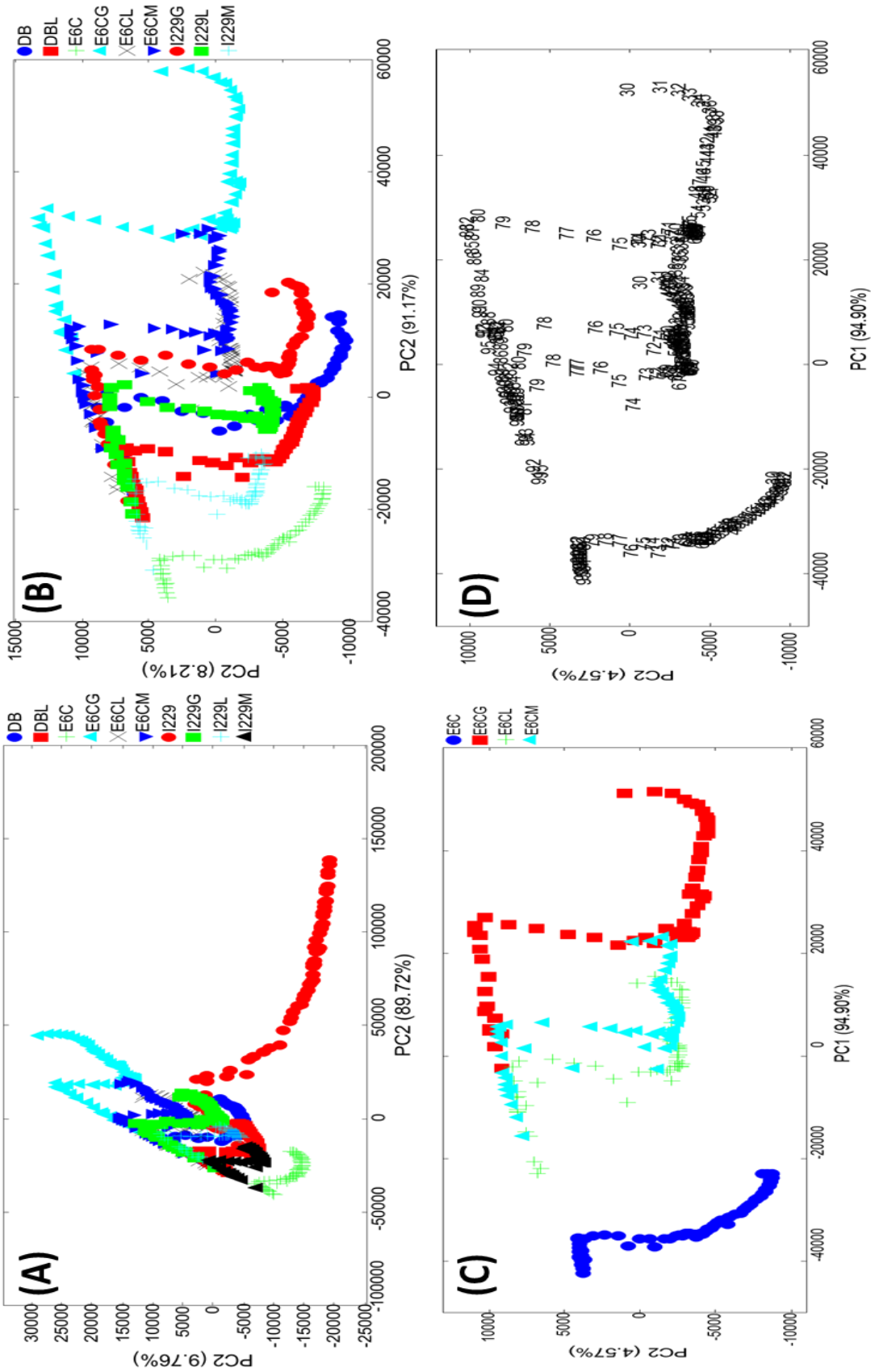

Figure 5.14: PCA Scores plots (PC1 *vs* PC2) of Optim data from (A) all mutants, (B) all mutants without I229C, (C) E6C mutant and (D) E6C mutant with samples labelled by temperature.

In order to simplify visualisation of the similarities between GFP samples further we have used parallel factor analysis (PARAFAC). The scores plot, seen in Figure 5.15 confirms that the un-modified I229C is the most different and also shows a separation between the modified and unmodified proteins.

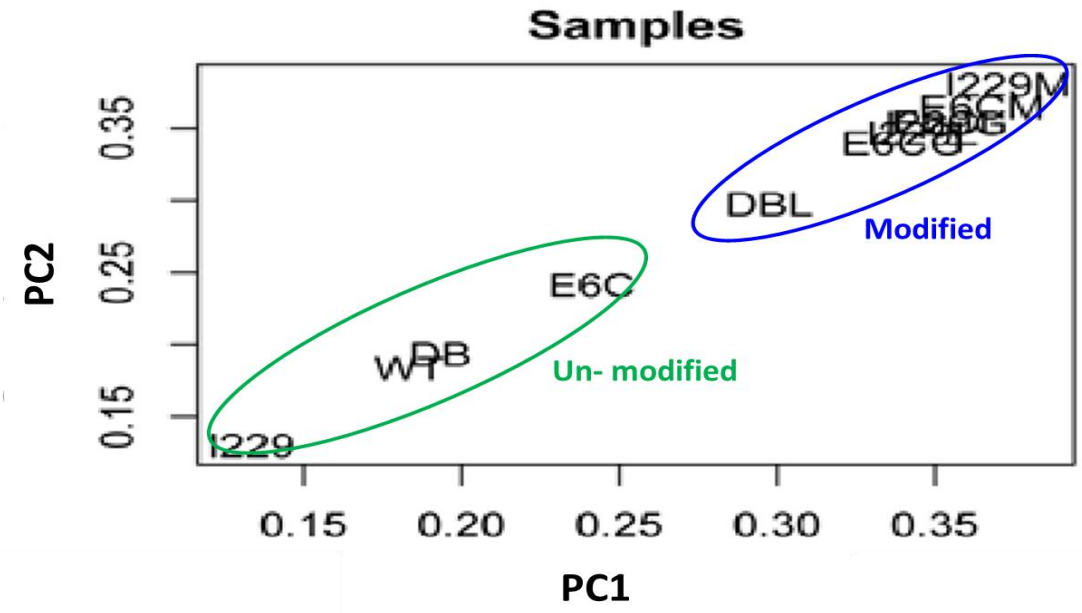

Figure 5.15: PARAFAC sample scores plot (PC1 *vs* PC2) of intrinsic region of Optim spectra.

A third application note detailing how PCA and PARAFAC can be used to highlight transitions which cannot be seen in a traditional curve, and also simplify visualisation of similarities and differences between unfolding profiles was produced and can be viewed in the appendix.

5.3.3.2.2 Optim Light Scattering Data.

Finally we examined the light scattering data collected in Optim thermal ramp experiments. The intensity of the light scattering band at 266 nm for all four of the GFP mutants is shown as a function of temperature in Figure 5.16A, where the I229C mutants can be seen to have a vastly different profile from all the other samples. The Wild type and Double mutants both begin to aggregate at around ~74 °C which coincides with the

beginning of unfolding transition, whereas in the I229C mutant has a much higher aggregation propensity with aggregation beginning much earlier (~35 °C). The E6C mutants appears much less prone to aggregation, giving credence to the hypothesis that variation in the Raman data of E6C and I229C mutants could be in part due to differences in aggregation behaviour. The levels of aggregation seen in I229C could be due to the I229C already having a more open structure, which was hinted to by the red shift in the initial Optim spectrum and the tryptophan band in the Raman spectrum. However, the increase in aggregation is more likely to be due to the unreacted cysteine on the outside of the protein, a theory which is confirmed by the aggregation profiles of the modified I229C (Figure 5.16B); which show a much lower aggregation propensity when the free thiol group is bound to the amino ethyl linker.

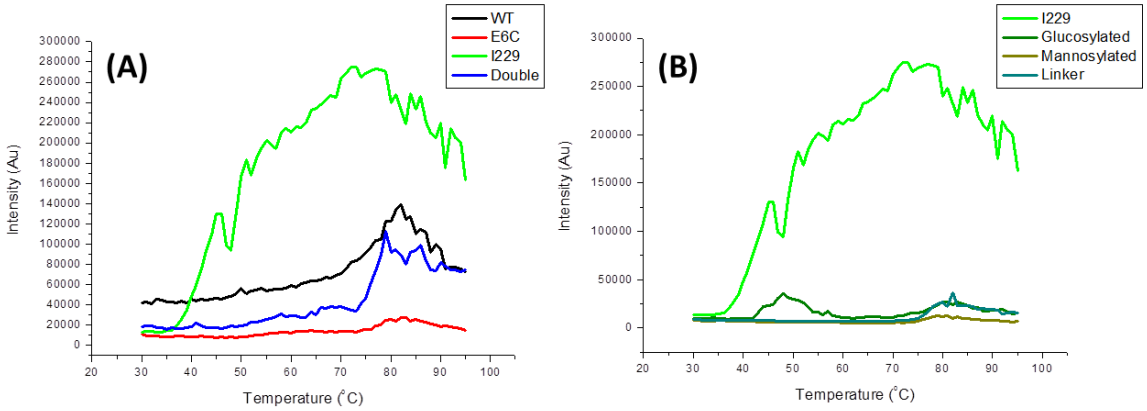

Figure 5.16: Graphs to show the intensity of light scattering data at 226 nm as a function of temperature for **(A)** all GFP mutants and **(B)** all I229C variants.

5.3.3.3 Optim Isothermal Experiments.

In order to probe the variations in the thermal unfolding of these samples further an isothermal experiment was performed. Samples were held at 70 °C with spectra recorded every 60 s. As with the thermal ramp experiments, these data indicate that the I229C mutant has a very different profile, however in these experiments this is clearly evident from the curves without further data analysis. Figure 5.17A shows I229C to be much less stable than the other mutants at 70 °C, with a very fast rate of unfolding recorded in the

first 20 min. All three of the other mutants appear to be stable at 70 °C. Once again this trend was shown to be unique to only the unmodified I229C protein (Figure 5.17B).

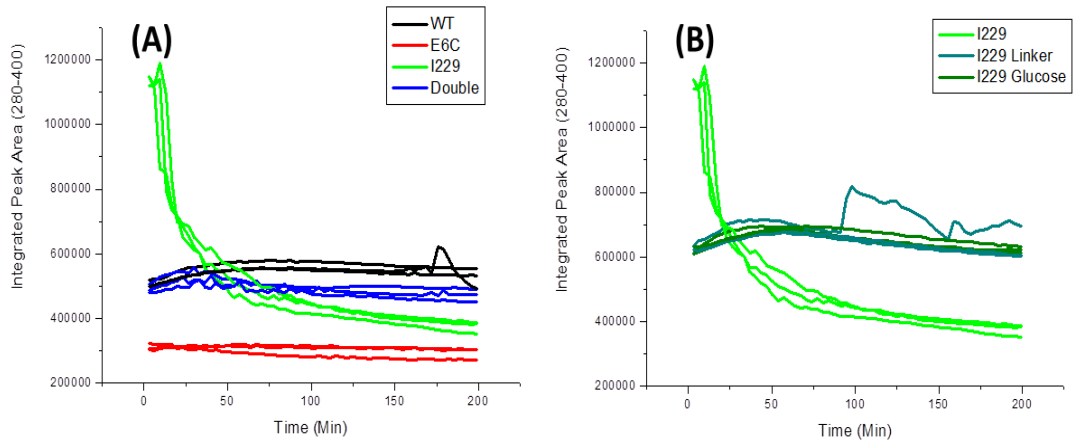

Figure 5.17: Graphs to show: **(A)** the integrated area of intrinsic fluorescence in all GFP mutants held at 70 °C as a function of time and **(B)** the integrated area of intrinsic fluorescence in all I229C variants held at 70 °C as a function of time.

`

5.3.4 Investigating Aggregation Further- Microscopy.

To understand the variations in the aggregation propensity of the GFP samples indicated by the Optim data better, further investigations were carried out on the un-modified mutants using optical microscopy. 100 images were collected from a well containing 400 µL of sample. The images collected from the wild type sample showed little or no signs of aggregates (Figure 5.18A), whereas both the E6C and double mutants exhibited aggregates between 2-4 µm in size, which occurred more frequently in the images collected from the double mutant. The I229C sample was found to have very large aggregates (~10 µm) which were observed in nearly every frame (Figure 5.18B). A larger collection of microscope images can be viewed in supplementary information (Figure S5.2). In addition images were also collected from the bottom of the well plate, Figure 5.18C and D. These images indicated that both the I229C and double GFP mutants had a large amount of protein precipitated on the bottom of the well; in the I229C images these aggregates are as big as 100 µm and appear to be fibril like in nature.

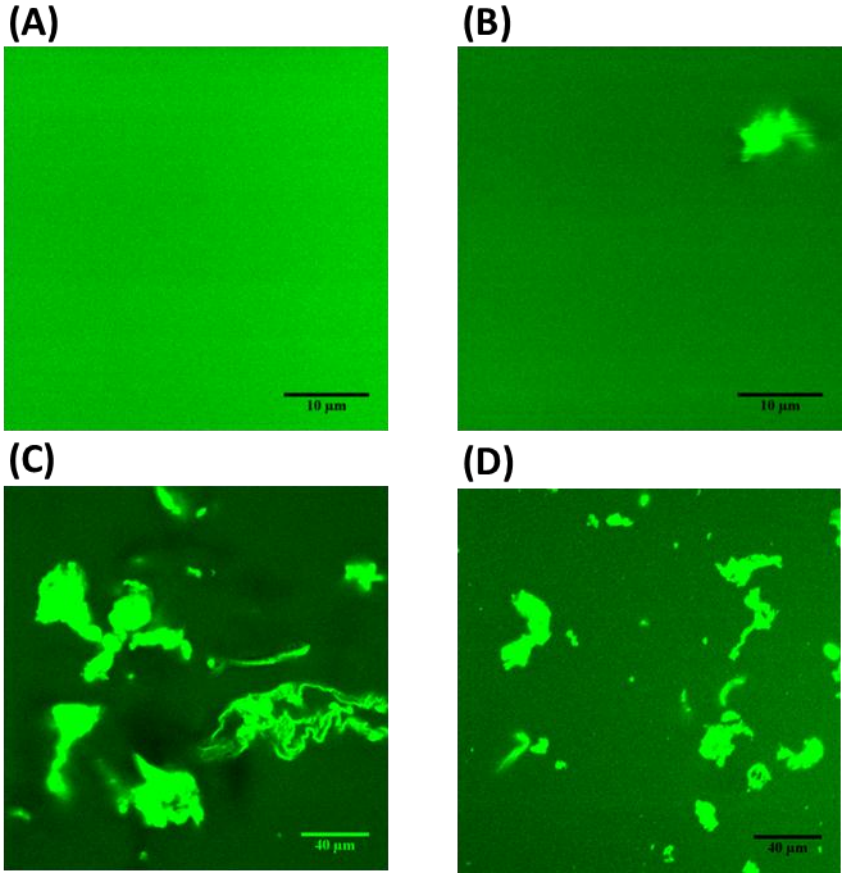

Figure 5.18: (A) Microscope Image from Wild Type GFP, **(B)** Microscope Image from I229 GFP, **(C)** Microscope Image from bottom of sample well for the I229C GFP mutant and **(D)** Microscope Image from bottom of sample well for the double GFP mutant.

Further proof of these observations was gained by performing FCS on the microscope images. FCS is a technique that calculates a correlation analysis of the fluctuations in fluorescence intensity within a sample, which are caused by the Brownian motion of particles. Using this method it is possible to calculate the diffusion time for each sample, which is proportional to the size of the particles in the sample, i.e. larger particles will have longer diffusion times.

The results from this analysis are summarised in the box and whisker plot in Figure 5.19 and the average diffusion time for each sample (mean of 90 measurements) is given in Table 5.3. Wild type GFP has the smallest average diffusion time and also the smallest deviation between the measurements, confirming that the wild type GFP is the least

prone to aggregation of the four mutants tested. The E6C mutant has a slightly higher average partial size than the wild type sample, which verifies the observations from the microscope images, but contradicts the previous Optim data which shows the E6C mutant to be the most resistant to aggregation at elevated temperatures (Figure 5.16A).

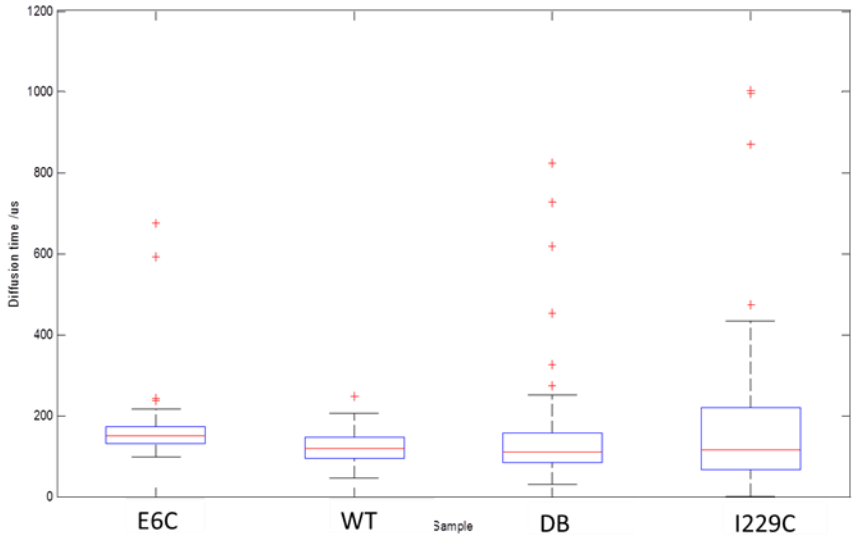

Figure 5.19: Box and whisker plot displaying diffusion times for GFP mutants calculated from FCS measurements.

Finally, both the I229C and double mutants have similar average diffusion times, with I229C being slightly higher. In addition Figure 5.19 shows the I229C sample to have a much higher frequency of larger particles than the other three samples. This confirms the results from the Optim experiments which show the I229C to be much more prone to aggregation. If, as suggested earlier, this increase in the aggregation propensity in this particular GFP is in fact due to the un-reacted cysteine residue in the protein, then we must ask ourselves, why does this only occur in the I229C sample when the E6C mutant also has a free thiol. The most plausible explanation for this is that the cysteine mutation at position 229 is more available than the one at position 6; due to the clipping of five amino acids from the end of the chain after the 229 cysteine and the addition of a His-tag at the opposite end of the sequence (see Figure 5.8). The

Table 5.3: Average diffusion timed for GFP mutants calculated from FCS results. (values are the mean of 90 measurements)

Sample Name	Mean Diffusion Time (µs)
WT	120.468
E6C	148.376
I229C	169.089
D	162.416

hypothesis that the changes in aggregation profiles is due specifically to the cysteine mutations is substantiated by the fact that the wild type sample, which has no external free cysteine residues has been shown by microscopy and FCS to be significantly less prone to aggregation that all of the other samples.

5.4 Conclusions.

Raman spectroscopy has for a third time in this thesis, been shown to be capable of distinguishing between glycosylated and non-glycosylated proteins; in this case a series of synthetic glycosylations in GFP mutants. Data collected from these proteins have highlighted that the position of the glycan is an important factor for the detection of glycosidic bands. Moreover, It has been shown that in cases where sugar vibrations are not observed in the Raman spectrum of a glycoprotein we are still able to easily differentiate between the glycoprotein and it's non-glycosylated equivalent based on structural differences between the two proteins.

Raman data has also been proved able to detect differences in the vibrations corresponding to higher order structure, indicating conformational variations between the E6C and I229C mutants. Further investigations by light scattering and fluorescence correlation spectroscopy indicate that there are vast differences in the aggregation propensity of these two GFP mutants; which could account for the unusual variations observed in vibrational spectroscopic data.

In addition, fluorescence emission spectra of the intrinsic protein fluorescence suggests that the I229C sample has more solvent exposed tryptophan than the E6C GFP, corroborating the variations in the indole ring breathing mode observed in the Raman and IR spectra. This points towards additional structural changes between the mutants. Optim thermal ramp and isothermal experiments suggests that these structural differences decrease the stability of the I229C GFP.

5.5 Supplementary Information.

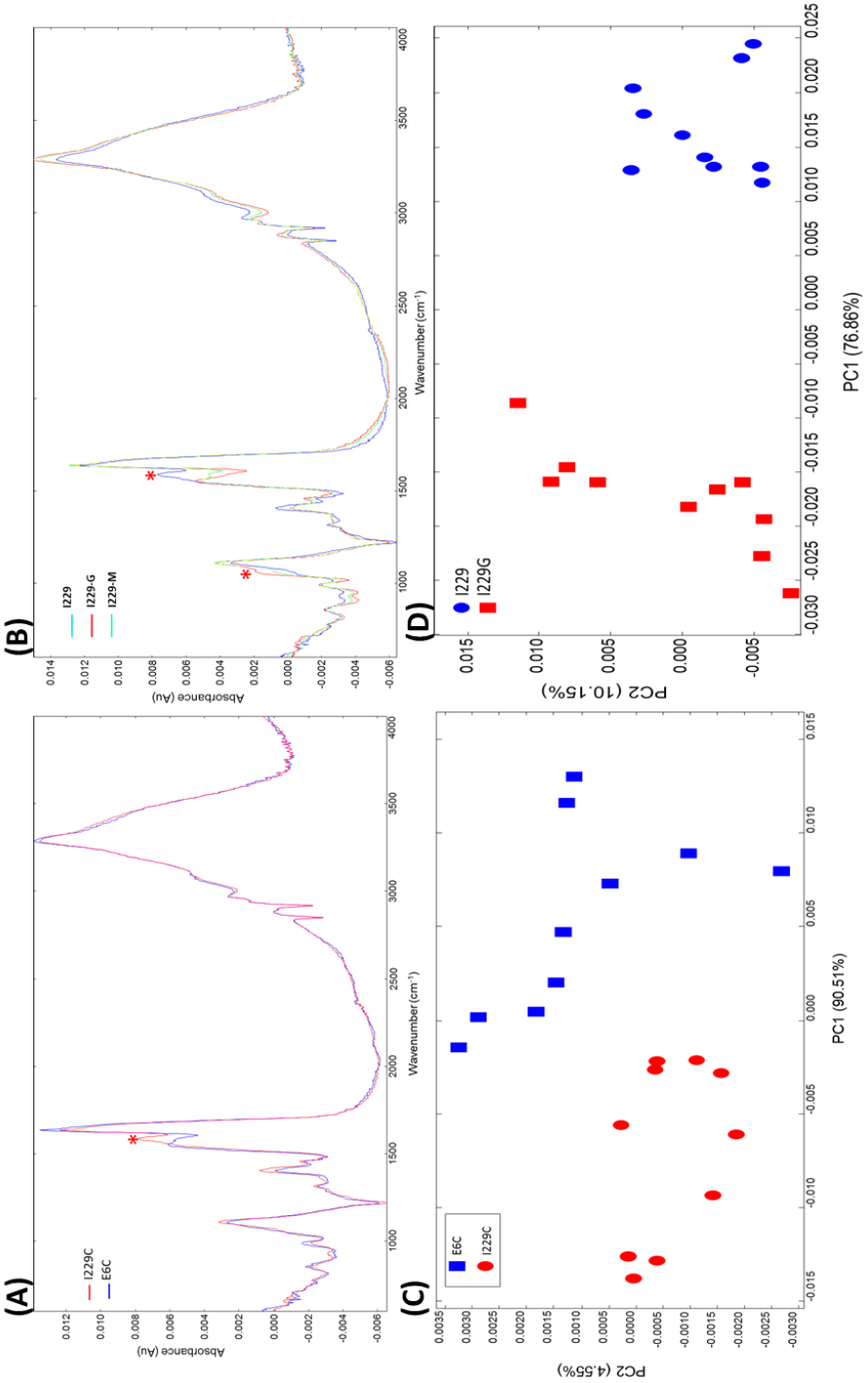

Figure S5.1: (A) FT-IR Spectra of I229C mutant and E6C mutant, **(B)** FT-IR Spectra of I229C mutant and I229C mutant glycosylated with glucose and mannose, **(C)** PCA scores plot (PC1 *vs* PC2) of FT-IR data from E6C and I229C and **(D)** PCA scores plot (PC1 *vs* PC2) of FT-IR data from I229C and I229G

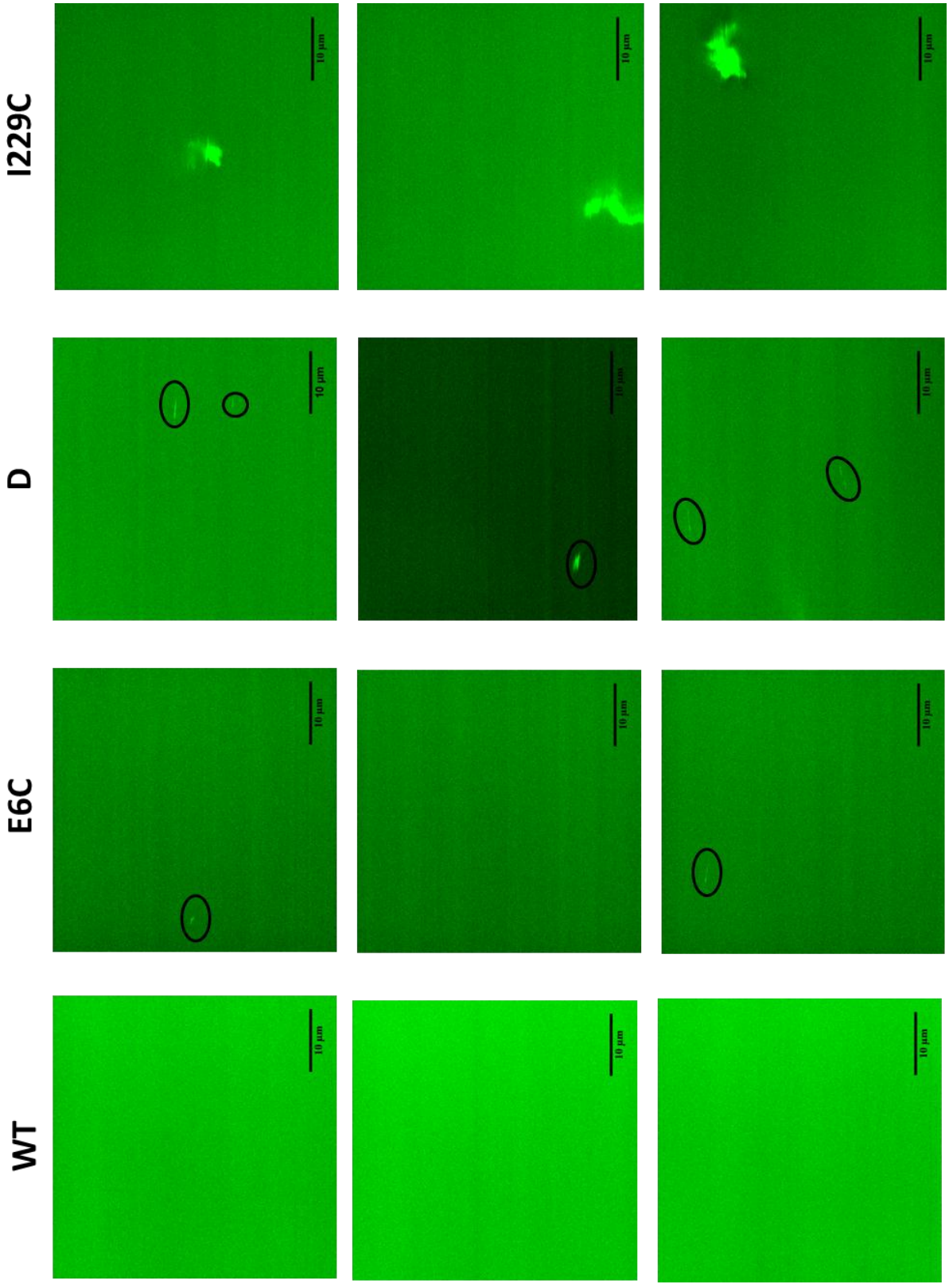

Figure S5.2: Microscope Images From GFP.

Chapter 6: Detection of the Sickle Cell Mutation in Haemoglobin Using Raman Spectroscopy.

Portions of the data in this chapter we collected by Sarah Newton as part of a MChem project under my supervision.

6.1 Introduction.

This portion of work was concerned with the use of Raman spectroscopy for the discrimination and quantification of the sickle cell mutation in haemoglobin. In these preliminary investigations we aimed to assess the suitability of Raman spectroscopy as an approach that could be used as a point-of-care diagnostic tool for the detection of sickle cell anaemia and the sickle cell trait from biological samples.

Haemoglobin (HbA) is the main constituent of red blood cells (erythrocytes) and is responsible for the transport of oxygen around the body (Silverstein and Nunn, 1997). Haemoglobin is a globular protein comprised of four chains and four haem groups (Figure 6.1). The four chains are organised into two pairs, two α-subunits and two β-subunits, in a tetrahedral arrangement (Stryer et al., 2002, Nienhuis and Bunn, 1974).

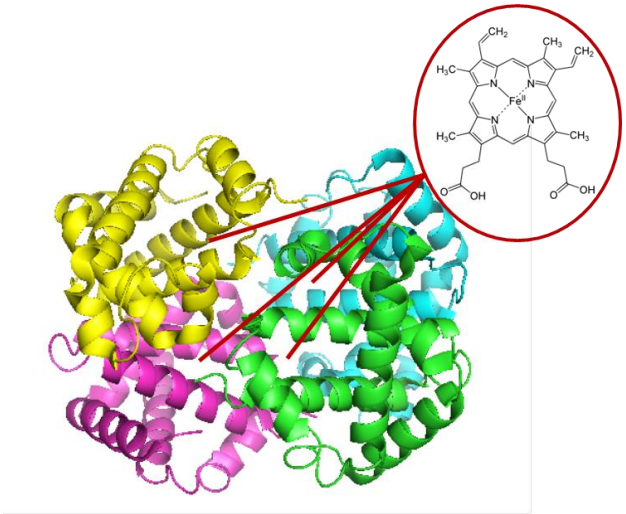

Figure 6.1: Cartoon diagram of HbA showing the position and structure of haem groups. Drawn in PyMol from PDB file 2HHB.

Sickle cell haemoglobin (HbS) is one of approximately 500 possible genetic mutations of HbA, most of which are point mutations (Stryer et al., 2002). The mutation in HbS, shown in Figure 6.2A, is a point mutation that results

in the single amino acid substitution of a glutamic acid for a valine at position 6 on the β-chain (Silverstein and Nunn, 1997). The side chain of valine is neutral, whereas the glutamic acid has a negative charge; this leads to a 'sticky' hydrophobic point on the surface of the β-chain which causes increased levels of aggregation in HbS (Weiss et al., 2009, Pumphrey and Steinhar.J, 1973, Murayama, 1972).

Sickle cell anaemia is a hereditary blood disorder which will affect one in every four hundred African-American children every year. Children of other ethnic origins are also affected but with a much lower probability (Aygun and Odame, 2012). A homozygous child, with two copies of the sickle cell gene will suffer from sickle cell anaemia, whereas a heterozygous individual, with one normal HbA gene and one HbS gene, will be a carrier of the sickle cell trait, but will not suffer the symptoms of the disease (Silverstein and Nunn, 1997). In a sufferer of sickle cell anaemia approximately 97% of haemoglobin will be HbS, compared to around 40% HbS in a carrier of the trait (Silverstein and Nunn, 1997, Pumphrey and Steinhar.J, 1973).

In a patient suffering from sickle cell anaemia, the HbS molecules are able to bind to oxygen just as efficiently as HbA, the problematic properties of HbS are encountered only after deoxygenation of haemoglobin (McCavit, 2012). The hydrophobic point on HbS can cause molecules to polymerise into stiff, rod-like, fibril structures after deoxygenation (Figure 6.2B). The formation of these fibrils will cause erythrocytes to adopt a sickle shape; these sickle shaped erythrocytes

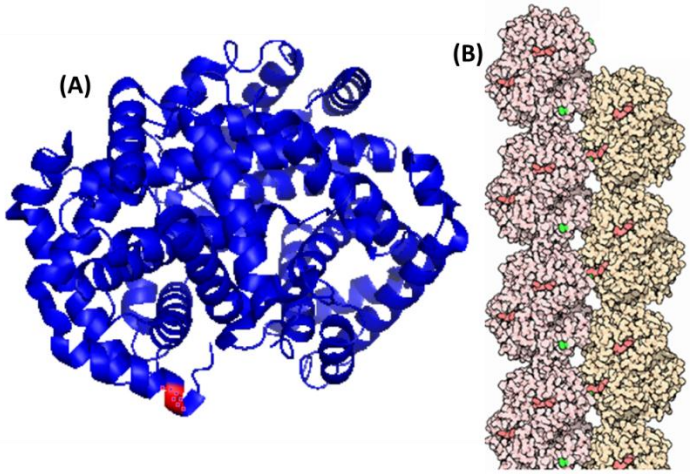

Figure 6.2: (A) Cartoon diagram of HbS showing the position of the glutamic acid substitution in red. Drawn in PyMol from PDB file 2HHB and **(B)** Cartoon depiction of HbS fibrillation.

will have difficulty passing through capillaries and can 'stack-up' in blood vessels, causing blockages which result in blood deprivation to the affected tissues (Silverstein and Nunn, 1997, Meier and Miller, 2012, Chien et al., 1970). In addition, sickle cell erythrocytes will have a shorter life cycle than normal erythrocytes and can rupture and die after 20 days, resulting in a reduced red blood cell count, which can lead to anaemia (Silverstein and Nunn, 1997). A carrier of the sickle cell trait will have none of the symptoms described here, as the majority (~60%) of their haemoglobin will be HbA (McCavit, 2012, Meier and Miller, 2012).

The most routine method of diagnosing sickle cell disease is the sodium metabisulfate test; in which a sample of patient's blood is mixed with sodium metabisulfate which causes defective erythrocytes to sickle. The sample is placed on a glass slide and sickled cells are identified using microscopy. However, this method cannot distinguish between samples from patients with sickle cell anaemia and samples from patients with the sickle cell trait, and therefore has a high rate of false positives (Aygun and Odame, 2012, Silverstein and Nunn, 1997). The Sikledex method is another chemical test which uses haemolysis of erythrocytes. Once cells are lysed HbA will dissolve readily in solution but HbS will aggregate; therefore a cloudy solution will indicate the presence of HbS. However this method can be inaccurate in new born babies, and again cannot differentiate sickle cell anaemia from the sickle cell trait (Meier and Miller, 2012, Nalbandi et al., 1971). The most reliable approaches, which do offer diagnosis of both sickle cell anaemia and the sickle cell trait, are electrophoresis-based tests (Cotton and Gulbis, 2013, Meier and Miller, 2012). Although these methods are reliable and relatively inexpensive, a Raman spectroscopy based approach would be a less laborious, higher throughput alternative, with the potential for portable *in situ* diagnosis.

The suitability of Raman spectroscopy as a tool for the diagnostic analysis of haemoglobin, erythrocytes and whole blood has been well documented in the literature. Many previous studies have centred on the imaging of erythrocytes using Raman microscopy (Wood et al., 2011). This approach was also used to investigate the

aggregation of HbS inside blood cells and also to detect haemozoin in blood for the diagnosis of malaria (Webster et al., 2008, Wood et al., 2005). More recently, the ability of Raman spectroscopy to detect glycated haemoglobin (Hb1Ac) was investigated for use in the assessment of glycaemic control in patients with diabetes mellitus (Barman et al., 2012). Raman spectra taken from whole blood have also allowed the quantification of blood glucose concentrations (Shao et al., 2012).

In this study we have investigated the use of Raman spectroscopy as a method that may be able to discriminate between HbA and HbS in both pure protein samples and mock biological samples, in which proteins were spiked into human plasma stock. In addition, we have attempted to quantify relative concentrations of HbA and HbS in a sample, in order to distinguish between patients with sickle cell anaemia and carriers of the sickle cell trait. Finally, we have studied aggregation profiles of HbS and HbA using light scattering data in order to compare this method to the results obtained using Raman spectroscopy.

6.2 Materials and Methods.

6.2.1 Materials.

Haemoglobin A (HbA) and haemoglobin S (HbS) were purchased as lyophilized powders from Sigma-Aldrich (Dorset, U.K.). Human plasma stock was also purchased from Sigma-Aldrich and was prepared using a protein A chromatography cartridge to remove plasma proteins.

6.2.2 Raman Spectroscopy.

Raman data were collected using a Renishaw 2000 Raman microscope described in Chapter 2. All spectra were single accumulation, extended scans between 400 and 1800 cm^{-1}, with an exposure time of 120 s. 2 µL of sample were spotted onto a hydrophobic Tienta SpectraRIM^{TM} slides (section 2.1.1.1) and allowed to air dry at room temperature for approximately 1 h. Each reported spectrum is an average of 6 spectra collected from different positions within each sample spot. For the wavelength comparison study in

6.3.1, the previously described Renishaw Raman spectrometer was used, operating either 785 nm or 633 nm excitation wavelengths. For the spectra collected at 532 nm a Biotools chiral Raman spectrometer (detailed in 2.1.1.2) was employed.

A Delta Nu® Advantage 200A portable Raman probe equipped with a diode laser operating at 785 nm was used in the preliminary investigations (Delta Nu Inc, Laramie, WY, USA). The spectral range was 200 to 2000 cm^{-1} with a spectral resolution of 8 cm^{-1}. The output of the laser is ~60 mW at source and ~30 mW at sample. Daily calibration of the instrument was achived by obtaining a spectrum of polystyrene using the calibration routine built into the software. The spectrometer was controlled using Delta Nu, Nu Spec™ software.

6.2.3 Optim 1000.

An Avacta Optim 1000 (described in 2.1.3) was used to collect light scattering data. 9 µL of three replicates of each sample (1 mg/ml haemoglobin in ultra-pure water) were loaded into a multi cuvette array (MCA). A temperature ramp from 25 to 80 °C was applied to the samples with a temperature tolerance of 0.3 °C. Spectra were recorded at 1 °C intervals with a 60 s hold time at each temperature. Spectra were collected with 1 s exposure time with the slit width set to 100 µm. Each run was performed in triplicate, with three analytical replicates of each sample per run.

6.2.4 Data Analysis.

Raman data were pre-processed in Matlab (smoothing, baseline correction and normalisation) according to the method optimised in Chapter 3 (3.3.3.1). PyChem was employed for PCA and PLSR. Spectral figures were plotted in GRAMS Ai. Optim data were imported into Optim Analysis software for preliminary analysis. Data were then exported into Origin for further analysis and for plotting figures.

6.3 Results and Discussion.

6.3.1 Preliminary Investigations.

Prior to further experiments, we investigated the most suitable excitation wavelength for Raman analysis of haemoglobin and spectra were recorded at 532, 633 and 785 nm (data not shown). Due to the large amount of fluorescence background present in both the 532 and 633 data, we determined that 785 nm was the most suitable available wavelength for this application.

Initial experiments were conducted to investigate the feasibility of a Raman based method to detect haemoglobin in biological samples at relevant physiological concentrations (2.2-2.4 mM or ~35% of total blood content) (Barman et al., 2012). HbA was spiked into human plasma stock at 5% wt/vol increments from 0-100%. For these preliminary studies protein depleted plasma was used in order to identify which bands are due to haemoglobin, without the complication of bands arising from other plasma proteins. Samples were analysed on Tienta Spectra RIM^{TM} slides using a Renishaw Raman microscope with 785 nm excitation.

The Raman spectra for 0, 50 and 100% HbA shown in Figure 6.3 display increases in the intensities of Raman bands with increasing HbA concentrations. This data set was then subjected to analysis by PLSR, where alternate concentrations were used for training and test sets. The PLSR predictions in Figure 6.4 indicate a strong correlation between HbA concentration and the Raman data, with a RMS error of 8.5% for the test data. Predictions were accurate at the lowest concentration tested of 5% HbA (or ~3.5 µM); significantly lower than the concentration range of haemoglobin which would be encountered in biological samples. Loadings from the PLSR analysis (not shown) indicate that the bands rising from amide I, amide III and phenylalanine vibrations, at ~1650, ~1230, and ~998 cm^{-1}, respectively, were the most selective for predicting HbA concentrations (Tuma, 2005, Lord and Yu, 1970).

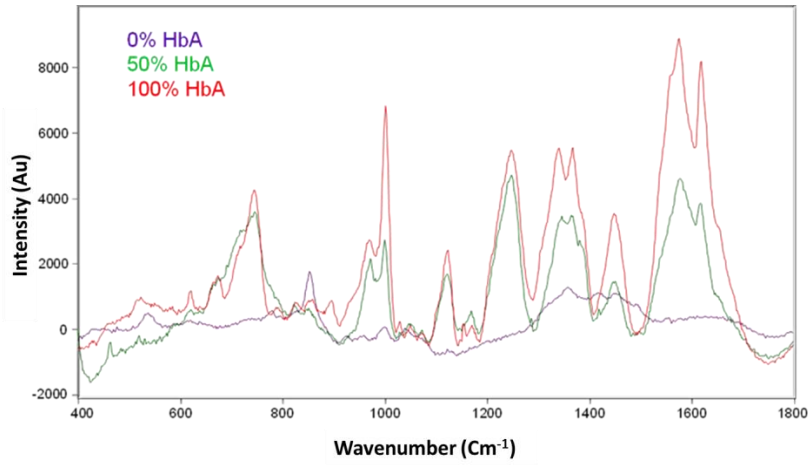

Figure 6.3: Raman Spectra of protein depleted plasma (0% HbA), Plasma with 50% Haemoglobin (50% HbA) and pure Haemoglobin (100% HbA) collected on Renishaw Raman Microscope. (Spectra have been Baseline corrected (ALS) and normalised (EMSC).

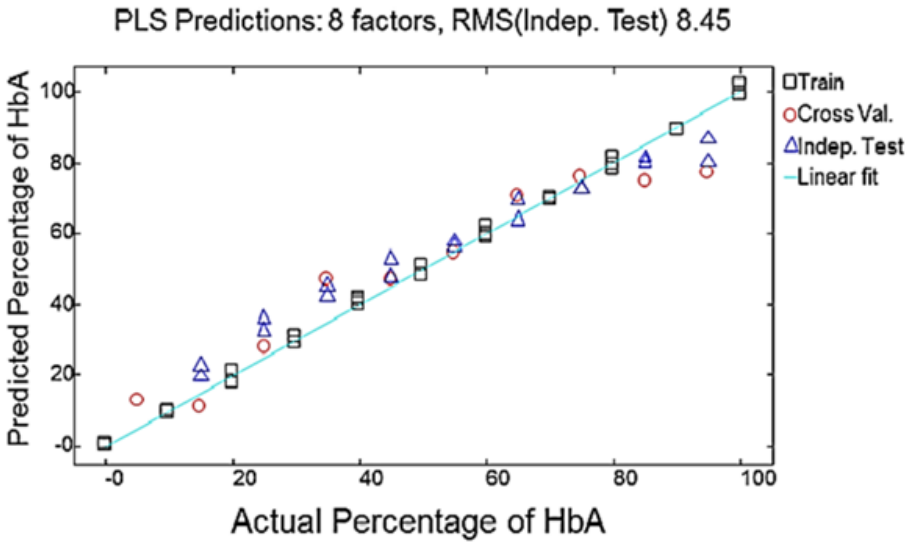

Figure 6.4:.PLSR predictions from Raman data of HbA in plasma collected on Renishaw Raman Microscope.

We also carried out this same experiment on a portable Raman probe (Delta Nu advantage 200A) in order to simulate an *in situ* diagnostic test. Although the raw spectra (Figure S6.1A) were heavily plagued by fluorescence with poor peak resolution, PLSR predictions (Figure S6.1B), still show an ability to predict HbA concentrations from the Raman spectra, but with a higher RMS test error than that achieved with the microscope data. The loadings plot for this model (Figure S6.1C) exhibits many bands which can be attributed to proteins, most notably the hydroxyphenyl ring deformation mode of tyrosine

at ~830 cm^{-1} (Siamwiza et al., 1975), which confirms that this model is based on real spectral features rather than differences observed in the background or baseline.

6.3.2 Detecting Sickle Cell Haemoglobin using Raman Spectrscopy.

Spectra were recorded from HbA and HbS on a Raman microscope. The average Raman spectra for each protein are displayed in Figure 6.5. Key features in the spectra are highlighted with asterisks and assignments for these bands are detailed in Table 6.1. Without the need for further chemometrics, visual inspection of the data shows that it is easy to discriminate between the two haemoglobin variants based on their Raman spectra.

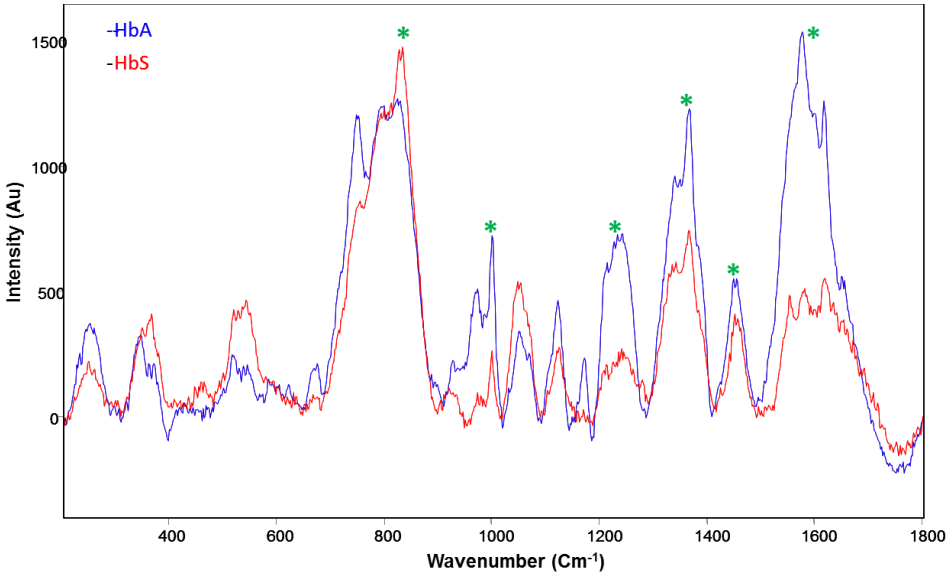

Figure 6.5: Average Raman Spectra of HbA and HbS. Asterisk indicate key features which are assigned in Table 6.1.
(Spectra have been Baseline corrected (ALS) and normalised (EMSC).

The band in the spectra at ~820 cm^{-1} has been attributed to the amino acid tyrosine; this band is known to increase in intensity as tyrosine residues become less exposed (De Gelder et al., 2007, Prevelige et al., 1993). Therefore, the sharp increase in the intensity of this peak observed in the spectrum of HbS suggests that HbS has an altered conformation which leaves tyrosine residues more buried than in HbA. This change in

higher order structure could be brought about by the glutamic acid mutation which increases the aggregation propensity of HbS. Aggregation of HbS molecules may leave internal tyrosine residues less exposed to the external environment and therefore cause an increase the band observed at ~820 cm^{-1}. These conformational differences are also reflected by the variations detected in the amide III (~1230 cm^{-1}) and amide I (~1650 cm^{-1}) regions of the spectra.

Table 6.1: Raman band assignments for bands highlighted in the Raman spectra of HbA and HbS (Fig 6.5) (Ashton et al., 2007, Tuma, 2005, Chen and Lord, 1980, Lord and Yu, 1970, De Gelder et al., 2007).

~Wavenumber (cm^{-1})	Proposed Assignment
820	Tyrosine ring
1050	Phenylalanine Ring
1220	Amide III
1380	Glutamic Acid
1420	CH_2/CH_3 from amino acid side chains
1600	Amide I

Most interestingly, there is a sharp peak at ~1380 cm^{-1} in the spectrum of HbA which can be specifically assigned to the ionised carboxyl group of glutamic acid (Ashton et al., 2007, De Gelder et al., 2007). This difference corresponds to the substitution of a glutamic acid residue on the outer surface of the β-chain of HbS. The spectral variations described here correlate well to the bands described in studies by Wood and colleagues, which monitored the aggregation of HbS and HbA inside red blood cells (Wood et al., 2005).

PCA was applied to this data set and the PCA scores (Figure 6.6A) show an excellent discrimination between HbA and HbS across PC1 with 78.71% TEV. The PCA loadings for this separation, shown in Figure 6.6B, correlate well with the bands which can be seen to be changing in the Raman data (Figure 6.5). Positive loadings, relating to HbA, show a good agreement with the spectrum of HbA, likewise the negative loadings correspond to the bands highlighted in the spectrum of HbS, most notably the previously discussed tyrosine vibrations at ~820 cm^{-1}.

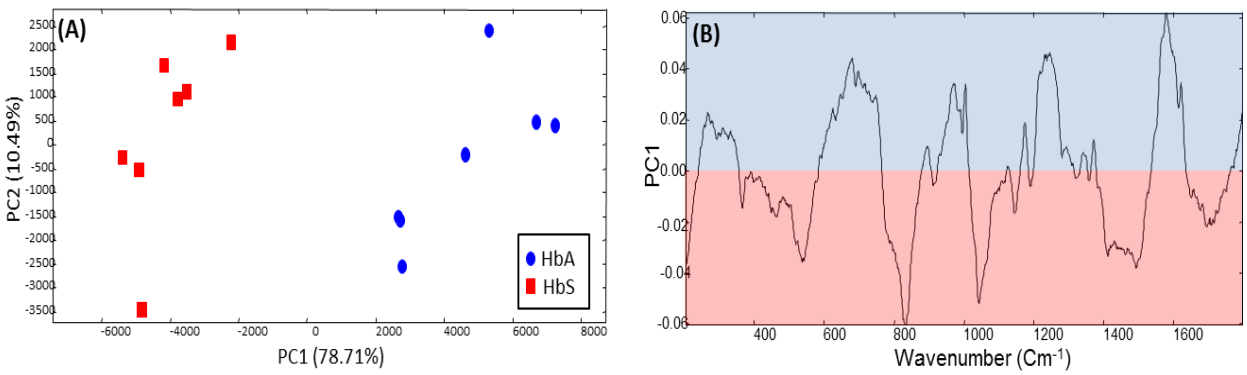

Figure 6.6: (A) PCA scores plot (PC1 *vs.* PC2) for the discrimination of HbA and HbS and **(B)** PCA loadings from PC1

The next step in this investigation was to spike HbA and HbS into human plasma stock in order to ascertain if these differences can still be identified in a complex 'mock' mixture. For this experiment we used plasma stock prior to protein A purification (i.e. with plasma proteins), in order to simulate more accurately a biological sample. Solutions were prepared at relevant, physiological concentrations of ~35% haemoglobin (equivalent to ~ 2 mM), by dissolving 129 mg of lyophilised haemoglobin into 1 mL of plasma. PCA was then applied to this data and PCA scores in Figure 6.7 show a clear separation of HbA and HbS across PC1, with the loadings for PC1 (not shown) being comparable with those show in in Figure 6.6B. This result confirms that it is possible to differentiate between HbA and HbS in a mock biological sample.

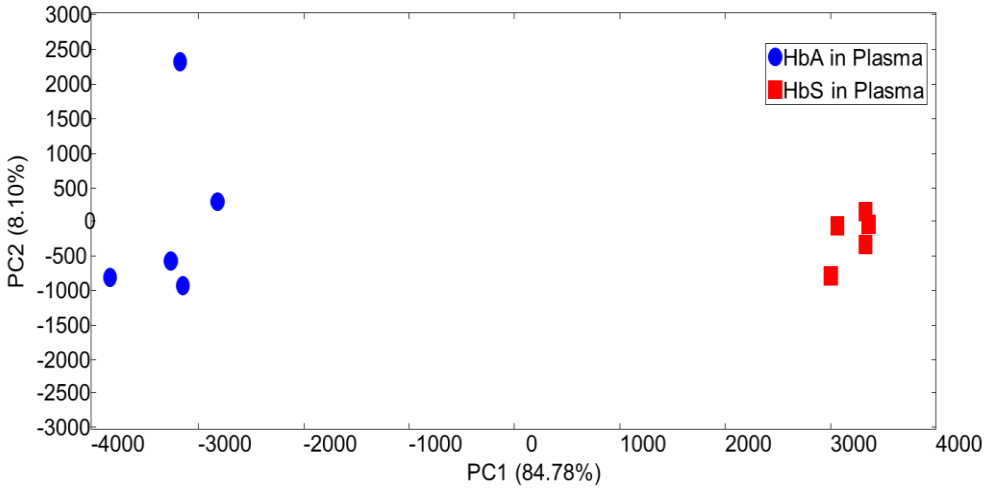

Figure 6.7: PCA scores plot (PC1 *vs* PC2) for the discrimination of ~ 2mM HbA and HbS in plasma.

6.3.3 Detecting the Sickle Cell Trait using Raman Spectroscopy.

Having previously demonstrated the ability of Raman spectroscopy to differentiate between mock samples from patients with sickle cell anaemia (HbS) and mock samples from healthy patients (HbA), we next attempt to quantify relative concentrations of HbA and HbS in a sample in order to discriminate between sickle cell anaemia and the sickle cell trait.

Initially, spectra were recorded from a mixture of HbA and HbS at the medically relevant ratio of 40% HbS and 60% HbA, in order to simulate a sample from a carrier of the sickle cell trait. These data were then compared, through the use of PCA, to the spectra recorded previously from pure HbA and HbS. This experiment was performed twice; once with the proteins in aqueous solutions and once with the proteins spiked into human plasma stock.

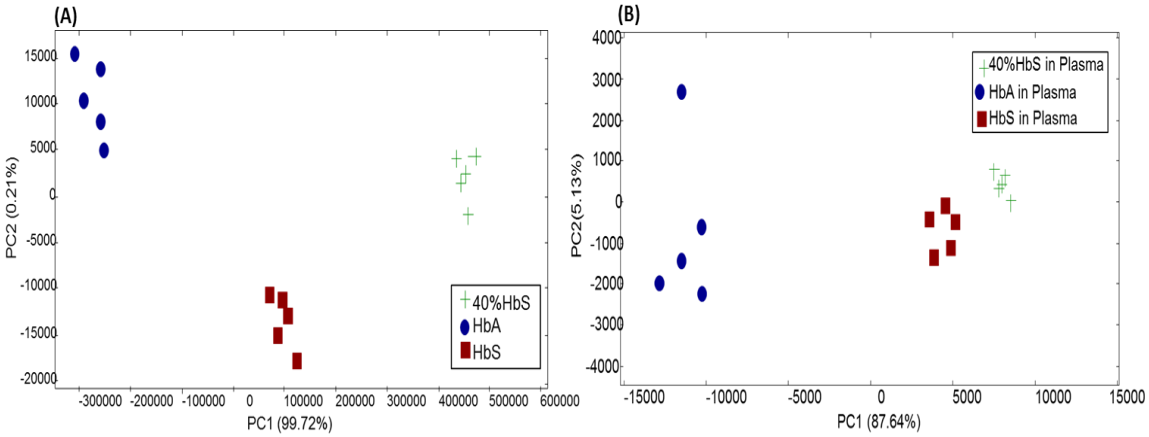

Figure 6.8: PCA scores plot for the discrimination of **(A)** HbA, HbS and a mixture of 40% HbS and 60% HbA and **(B)** HbA, HbS and a mixture of 40% HbS and 60% HbA spiked in human plasma stock.

PCA scores plots for both analysis of protein samples and analysis of mock biological samples are shown in Figure 6.8A and 6.8B, respectively. Results from both experiments display clearly that we are able to differentiate 100% HbA and 100% HbS from the mixture of 40% HbS and 60% HbA. This suggests that it may be possible to diagnose not only sickle cell anaemia but also the sickle cell trait using Raman spectroscopy. However,

it should be noted that, unexpectedly, the spectra from pure samples lie closer together than the spectra collected from the mixtures, and it was thought that this could be due to protein-protein interactions between HbA and HbS. For this reason we have chosen to further investigate HbS and HbA mixtures by analysing a wider range of HbS concentrations; from 0-100% at 10% increments. Due to the success of previous experiments carried out on mock biological samples, we carried out this analysis solely on samples of haemoglobin spiked into human plasma.

Raman spectra collected from all concentrations of HbS in HbA show changes in a number of peak intensities with varying HbS concentration (data not shown). To determine if these trends are linear, we focussed on the tyrosine band at ~820 cm^{-1}. As this band was previously seen to increase in intensity in the Raman spectrum of HbS (Figure 6.5), we would expect this band to increase in intensity as the concentration of HbS increases. Using GRAMS Ai software, a peak fitting function was applied to this band in order to calculate the peak area at each concentration. Normalised peak areas were then plotted as a function of HbS concentration, in Figure 6.9A, and a linear trend was observed. This indicates that it may be possible to use this band to predict relative concentrations of HbS and HbA in a sample, and hence diagnose sickle cell anaemia and the sickle cell trait. There are a number of anomalous results in the graph, particularly at 40 and 60% HbS. This could possibly be due to drying effects and orientation effects introduced by using drop-coating method on the Tienta Spectra RIM™ slides.

Data were then subjected to interrogation by PCA. The PCA scores plot in Figure 6.9B indicates a clear trend across PC1 with increasing HbS concentration. There is a definite distinction between pure HbS, pure HbA and the mixed samples, which once again suggests that Raman spectroscopy is capable of diagnosing both sickle cell anaemia and the sickle cell trait. There is also a clustering of individual HbS concentrations within the mixed samples. This quantitative trend is displayed more clearly by plotting PC1 scores

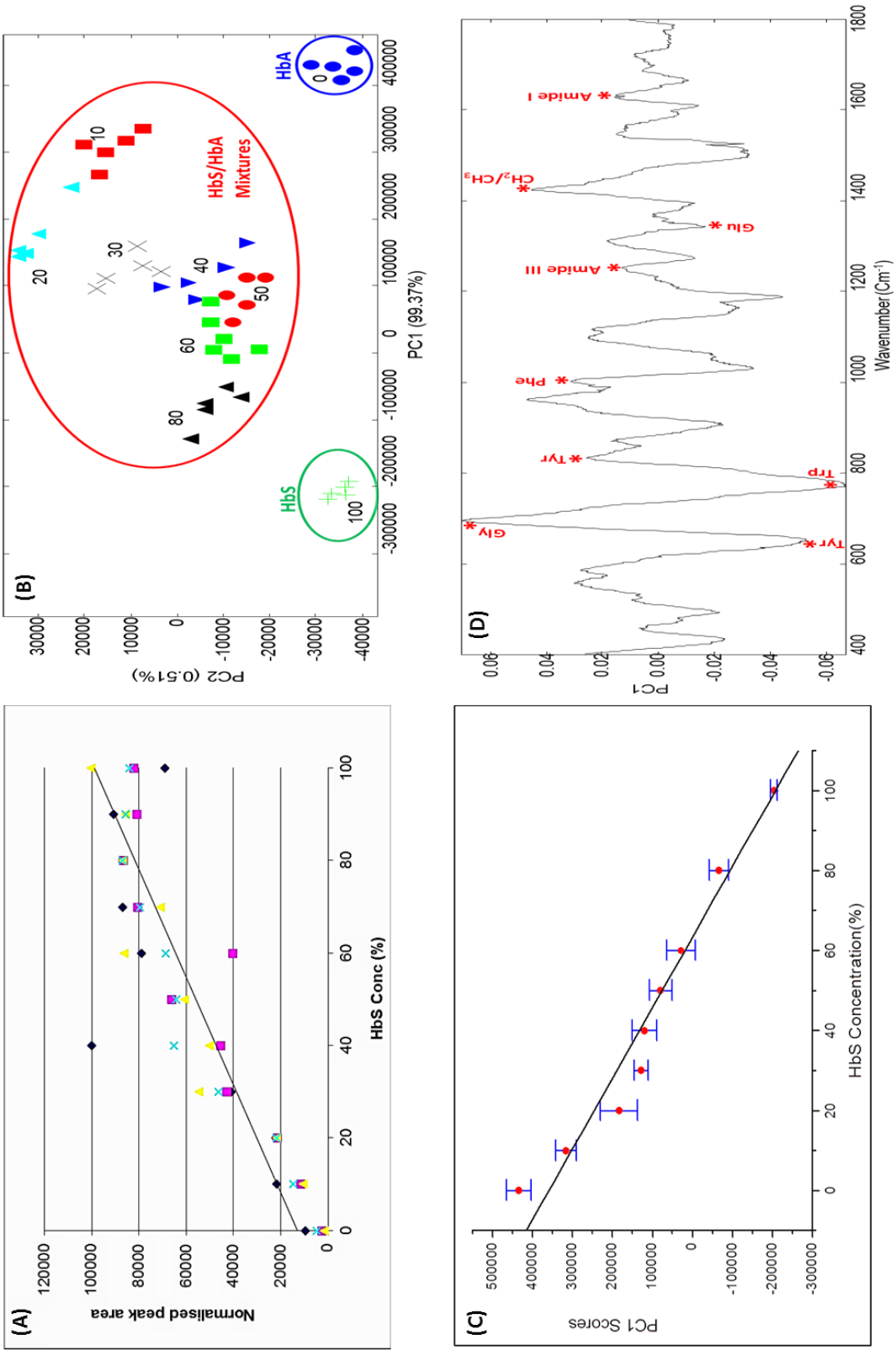

Figure 6.9: (A) Graph to show the correlation between HbS concentration and peak area of the band at ~820 cm^{-1} (3 independent measurements are shown with the mean measurement indicated by the blue cross), **(B)** PCA scores plot of Raman data from HbS and HbA mixtures, **(C)** Graph of PCA scores from PC1 plotted as a function of HbS concentration (values are the mean of 5 measurements with standard error bars) and **(D)** PCA loadings from PC1, with bands discussed in the text highlighted by the red asterisk.

as a function of HbS concentration in Figure 6.9C, where mean PC1 scores for five measurements are plotted with standard error bars. Although there is some overlap between samples of different HbS concentrations in both Figure 6.9B and 6.9C, for this diagnostic application there is no need to quantify the exact amount of HbS, only to discriminate between HbA, HbS and a 40% HbS and 60% HbA mixture; which these results clearly demonstrate is possible.

The loadings from PC1 for this analysis, shown in Figure 6.9D, show that all of the bands which were discussed previously for the discrimination of pure HbA and HbS were used in this separation, including the ~820 cm^{-1} tyrosine band, ~1000 cm^{-1} phenyl ring breathing mode, ~1230 cm^{-1} amide III vibrations, ~1380 cm^{-1} glutamic acid band, and the amide I vibrations at ~1650 cm^{-1}. In addition, there are three large peaks present at ~785 cm^{-1}, ~698 cm^{-1} and ~650 cm^{-1}, which can all be attributed to vibrations arising from the amino acid side chains, specifically; vibrational stretching of the indole ring in tryptophan (Liang et al., 2006), NH_2 bending in glycine (Kumar et al., 2005) and ring vibrations from tyrosine (Lord and Yu, 1970), respectively.

As with the previous PCA analysis in Figure 6.8, it could be said that the two pure samples are closer to each other, with separation of pure and mixed samples across PC2. However the amount of variance explained by PC2, 0.51% TEV, is negligible compared to the 99.37% TEV found in PC1. Nevertheless, we have explored the possibility of a trend occurring between pure and mixed samples due to protein-protein interactions by investigating the aggregation profiles of mixed samples using light scattering experiments.

6.3.4 Investigating HbS aggregation using light scattering.

Thermal ramp experiments were carried out in order to analyse variations in light scattering data between HbA, HbS and mixtures of HbA and HbS. In addition to investigating any interactions which maybe occurring between HbA and HbS, we also discuss the potential use of light scattering experiments carried out on a Optim instrument

as a high throughput method of diagnosing the sickle cell trait based on the aggregation profile of a sample.

6.3.4.1 Comparing HbA and HbS.

Initially, we compared the variations in light scattering intensities over a temperature range of 25-80 ºC, for both HbA and HbS. Figure 6.10 shows the intensity of light scattering at both 473 nm (Figure 6.10A) and 266 nm (Figure 6.10B) as a function of temperature. Traces shown are the average values from nine measurements across three independent runs.

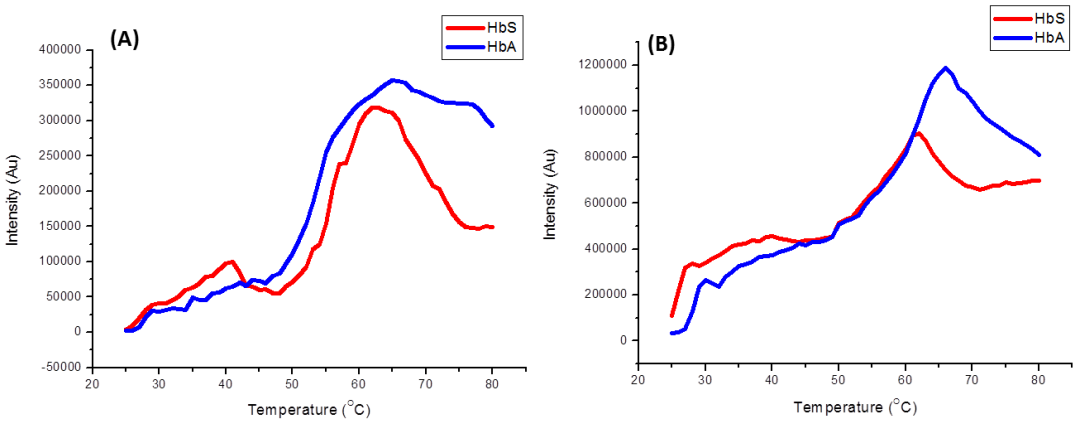

Figure 6.10: Graphs to show the intensity of light scattering a function of temperature for HbS and HbA at **(A)** 473 nm and **(B)** 266 nm. Each trace is the mean of nine repeat measurements, which overlay well with each other.

Significant differences between the curves from HbA samples and HbS samples can be observed. Both 473 and 266 nm curves indicate that aggregation begins at slightly earlier temperatures in HbS, and the 473 nm graph displays a sharper increase in aggregation of HbS at lower temperatures, which is not observed in the HbA data. The drop in light scattering intensities seen at higher temperatures (above 65 ºC) is due to the precipitation of aggregates in the solution. We can see from both Figures 6.10A and 6.10B that precipitation begins at a much lower temperature in HbS than in HbA; ~61 ºC in HbS compared with ~70 ºC in HbA. This suggests that it may be possible to use aggregation profiles monitored using light scattering data to diagnose sickle cell anaemia.

6.3.4.1 Diagnosing the Sickle Cell Trait using Light Scattering.

By comparing the light scattering intensities at 266 nm for mixtures of HbA and HbS, Figure 6.11A, we are able to observe differences in the aggregation profiles of mixed samples. A decrease in the temperature at which precipitation begins is observed from the data over increasing HbS concentrations, which suggests that we are able to differentiate between sickle cell anaemia and the sickle cell trait using light scattering data.

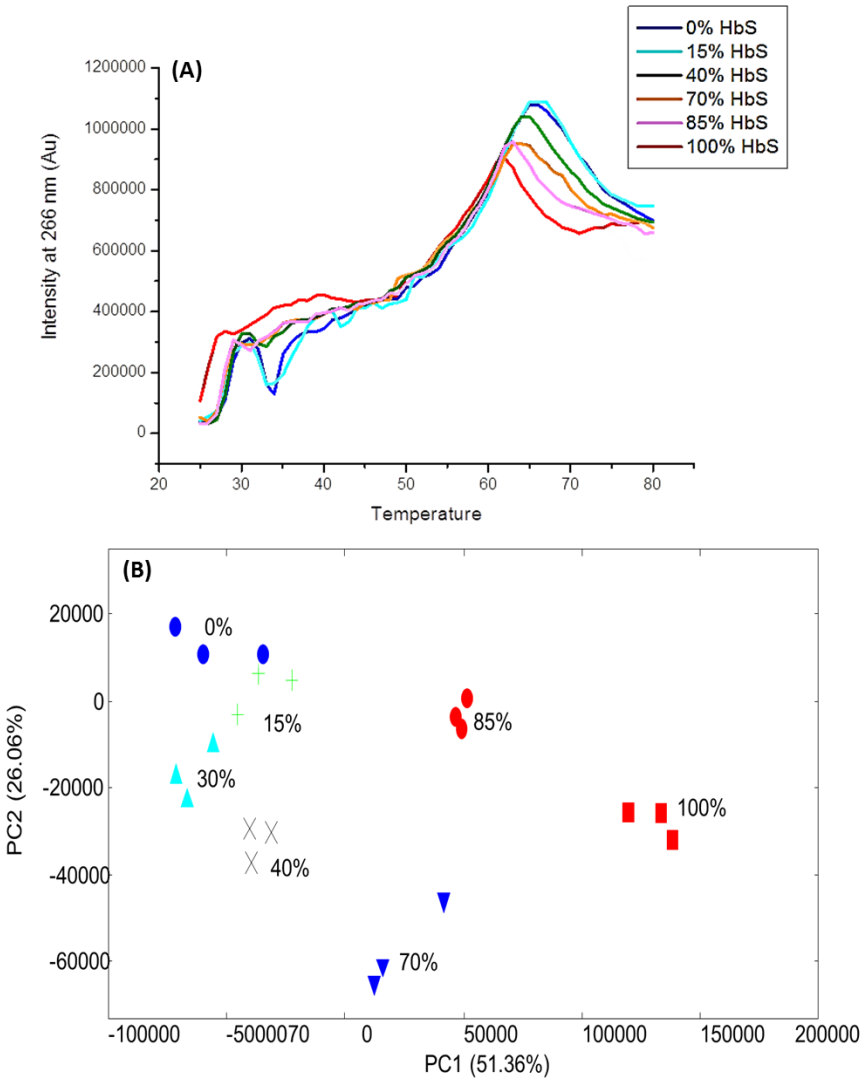

Figure 6.11: (A) Graphs to show the intensity of light scattering at 266 nm a function of temperature for HbS and HbA and mixtures of HbS and HbA. Each trace is the mean of nine repeat measurements and **(B)** PCA scores plot for the discrimination of HbS samples using light scattering data.

In order to simplify this discrimination based on aggregation profiles we used the average trace of light scattering intensity *vs.* HbS concentration for each of the triplicate runs as input variables for a PCA model. The resulting scores plot (Figure 6.11B) clearly displays that both 100% HbS and 40% HbS samples can be easily identified, hence it is possible to diagnose both sickle cell anaemia and the sickle cell trait from the light scattering data.

Furthermore, this PCA shows a trend in the scores plot with increasing HbS concentrations. Unlike previous Raman analysis (Figures 6.8 and 6.9B), these data do not show any grouping of un-mixed samples. A more expected trend is apparent, with HbS and HbA falling at opposite sides of the plot and mixed samples lying in between. This would seem to indicate that the variations observed between mixed and pure samples in the Raman data are due to instrumental artefacts in the data set rather than differences introduced by variations in protein conformation brought about by interactions between HbA and HbS.

6.4 Conclusions.

Work by Wood and colleagues clearly displayed the potential of Raman spectroscopy to detect structural information about haemoglobin molecules from within red blood cells (Wood et al., 2005). However, this study focussed on investigations into the mechanisms of aggregation in haemoglobin and sickle cell haemoglobin, rather than differentiation of the two proteins with a view to being applied in a diagnostic setting.

In these preliminary investigations we have shown Raman spectroscopy to be capable of distinguishing between HbA and HbS proteins both in pure protein samples and when spiked into human plasma. We have also demonstrated the potential of Raman analysis to quantify levels of HbS in this mock biological sample containing a mixture of HbS and HbA. This is an important result as it shows how Raman spectroscopic analysis of samples can allow diagnosis of both sickle cell disease and the sickle cell trait, which current routinely used methods such as the sodium metabisulfate test and the Sickledex

test do not allow (Silverstein and Nunn, 1997, Aygun and Odame, 2012, Nalbandi et al., 1971, Meier and Miller, 2012).

Furthermore, we have shown the feasibility of this method as a point-of-care diagnostic tool, by demonstrating that a portable Raman probe is capable of detecting haemoglobin at physiologically relevant concentrations in protein depleted plasma. These encouraging preliminary results motivate future work comparing the Raman spectra of HbA and HbS in mock biological fluids recorded on a portable Raman instrument, leading on to the analysis of HbS in a complex biological sample which will more accurately simulate a patient's blood sample.

Finally, due some unusual trends visible in the PCA scores plot drawn from Raman data of mixed HbA and HbS samples, we have investigated the possibility of interactions between the two proteins using light scattering. Results here showed no distinction between mixed and un-mixed samples, suggesting that the variations observed in the Raman data were potentially analytical artefacts, possibly introduced by drying effects on the Tienta Spectra RIM^{TM} slides. It is hoped that future experiments in solution based systems will not have this error associated. Nevertheless, light scattering data showed it to also be a suitable method for the analysis haemoglobin samples, which, as with the Raman method, allows distinction of mock samples from healthy patients, carriers of the sickle cell trait and sufferers of sickle cell anaemia.

6.5 Supplementary Information.

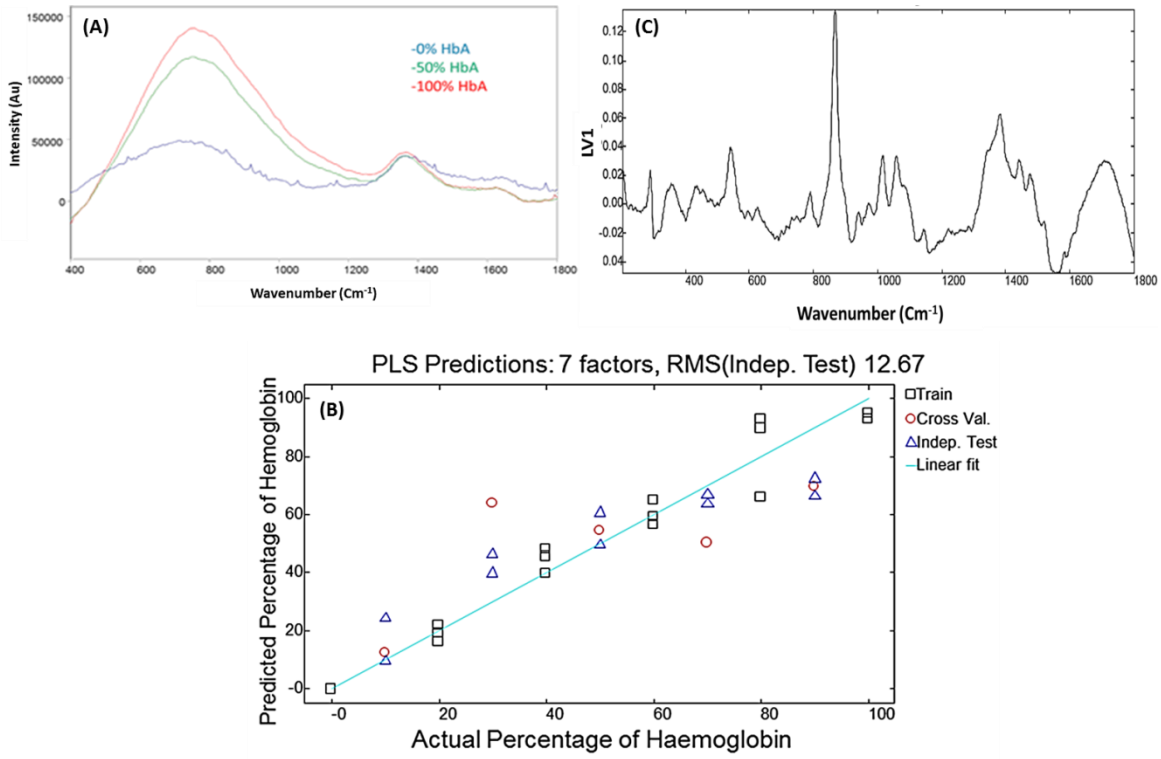

Figure S6.1: (A) Raw Raman Spectra of protein depleted plasma (0% HbA), Plasma with 50% Haemoglobin (50% HbA) and pure Haemoglobin (100% HbA) collected on Delta Nu portable Raman probe, **(B)** PLSR predictions from Raman data of HbA in plasma collected on Delta Nu portable Raman probe (Data pre-processing: Sav-Gol smoothing, ALS baseline correction and EMSC) and **(C)** Loadings from the first LV of the PLSR model for the quantification of HbA in plasma built from Raman data collected on Delta Nu probe.

Chapter 7: Monitoring Guanidinium-Induced Structural Changes in Ribonuclease Proteins Using Raman Spectroscopy and 2D Correlation Analysis.

Work presented in this chapter has been adapted from work accepted for publication in *Analytical Chemistry*; Brewster V. L., Ashton L. and Goodacre R. "Monitoring Guanidinium-Induced Structural Changes in Ribonuclease Proteins Using Raman Spectroscopy and 2D Correlation Analysis". L. Ashton gave advice on production and interpretation of 2D correlation contour plots.

7.1 Introduction.

It is widely known that changes in the tertiary, three-dimensional (3D) structure of a protein-based pharmaceutical can directly affect drug activity and may also induce protein aggregation (Goddard, 1991). Therefore it is essential that the stability of a biopharmaceutical product is well characterised. In this study we investigate Raman spectroscopy as an alternative to the current gold standard analytical methods for monitoring protein unfolding: fluorescence spectroscopy and differential scanning calorimetry (DSC) (Spink, 2008, Serrano et al., 2012).

The tertiary structure of a protein depends largely on its solvent environment. When in aqueous solutions at neutral pH, proteins will adopt an extremely ordered 'native' conformation. For most protein biopharmaceuticals this native tertiary structure is the biologically active conformation (Goddard, 1991). When the solvent environment is perturbed by extreme temperatures, pH or the addition of a chemical denaturant, the 3D protein structure will unfold into a disordered or denatured state. Guanidine hydrochloride (GuHCl) is one of the most commonly used chaotropic agents for protein unfolding

studies. However, despite its widespread use as a chemical denaturant the mechanism of action is still unclear. It is thought to involve the formation of hydrogen bonds between the denaturant and the peptide backbone and also an increase in the solubility of both polar and non-polar amino acid side chains (Konermann, 2004).

It has been reported in the literature and demonstrated in the work displayed in this thesis (Chapter 4) that the stability of a protein molecule is enhanced by the addition of a glycan group (Sola and Griebenow, 2009). Therefore, we will in this study also compare the unfolding behaviour of a protein and its glycosylated equivalent. The addition of an oligosaccharide increases the internal non-covalent forces which hold the protein in its folded form by decreasing exchange rates of the backbone amide protons and hence increasing the concentration of denaturant required (Taylor, 2006). The Ribonuclease (RNase) proteins, RNase A and its glycosylated equivalent RNase B, previously used in this thesis for glycosylation studies (Chapter 3), were again chosen as a model system. This was largely due to their similarity in secondary and tertiary structure, but also because these proteins are well characterised in literature in terms of the structure and stability of both proteins (Taylor, 2006, Naidu and Prabhu, 2011, Scheraga, 2011).

Much of the previous work investigating the perturbation of protein molecules by Raman spectroscopy takes advantage of the sensitivity of the amide III region to conformational changes, in particular monitoring changes which occur during acid induced unfolding of proteins by UV Resonance Raman. (Tuma, 2005, Tuma et al., 1995, Chi and Asher, 1998, Zheng et al., 2004). Other studies have focussed on changes to the tryptophan modes as the hydrophobic amino acids become more or less exposed (Liang et al., 2006, Chen and Lord, 1980). The thermal unfolding of RNase A has been measured previously by Raman spectroscopy, however this approach centred on analysis of the C-H stretching modes in the 3000 cm^{-1} region (Verma and Wallach, 1977). Raman spectroscopy has also been used to characterise various glycoproteins, including RNase B, but these studies did not investigate unfolding but rather the detection of the glycan (Barman et al., 2012, Dingari et al., 2012, Brewster et al., 2011).

In this chapter we report the use of Raman spectroscopy to monitor the unfolding of Ribonuclease proteins in the presence of GuHCl. Through the use of unfolding curves and 2D correlation analysis, we compare the results derived from Raman spectroscopy to those obtained by a conventional fluorescence unfolding experiment. We also acquired Raman and fluorescence data from the glycosylated equivalent of RNase A, RNase B in order to evaluate how the addition of a complex glycan affects the stability and unfolding of RNase proteins.

7.2 Materials and Methods.

7.2.1 Materials.

Ribonuclease A (RNase A), Ribonuclease B (RNase B), guanidine hydrochloride (GuHCl) and phosphate buffered saline (PBS) tablets were all of analytical grade and purchased from Sigma Aldrich (Dorset, U.K).

7.2.2 Method.

For fluorescence spectroscopy 10 mg/mL solutions of RNase A and RNase B were prepared in PBS solution (0.01 M phosphate buffer, 0.0027 M KCl, 0.137 M NaCl; pH 7.4). A 7 M stock solution of GuHCl in PBS was made and subsequently diluted with the RNase solutions into three sub-stocks of 2, 4 and 6 M GuHCl. Sub-stocks were then diluted into the desired GuHCl range (between 0-6 M at 0.2 M intervals), where the final protein concentration was 0.1 mg/mL (~7 μM) and the final sample volume 1 mL. Samples were then incubated at 37 °C overnight before analysis. Each unfolding experiment was performed in triplicate. For Raman spectroscopy initial solutions of 400 mg/mL RNase A and B were prepared in PBS. The above method was then employed so that the final protein concentration was 10 mg/mL (~700 μM) within a sample volume of 400 μL.

7.2.3 Fluorescence Spectroscopy.

Fluorescence spectra were obtained on a Shimadzu RF-5301PC spectrofluorophotometer (Shimadzu Biotech, Manchester, UK) equipped with a 150 W xenon lamp and a holographic grating with 1300 grooves/mm. The excitation wavelength range was 220-990 nm and the measurement range was 220-750 nm. Shimadzu Pop Up Scan software was used for instrument control. For this study the excitation wavelength was set at 280 nm and the excitation slit was set at 3 μm. The emission slit was set at 5 μm with an emission wavelength range of 220-550 nm. Samples were analysed in a quartz cell with a fast scanning speed and sensitivity set to high.

7.2.4 Raman Spectroscopy.

Raman data were collected using the Renishaw 2000 Raman microscope described in Chapter 2. 100 μL of each sample was pipetted into a 96 well plate, and a 20x objective lens was focussed on the top of the solution. Spectra were collected over a spectral range on 200-2000 cm^{-1}, over five accumulations with 120 s exposure time. Raman data were pre-processed by smoothing (Sav-Gol, Filter width: 7) followed by a baseline correction (ALS), both performed in Matlab.

7.2.5 Data Analysis.

Raman spectroscopic data were exported from the instrument software into Matlab where data pre-processing was performed, in order to allow direct comparison of the data. 2D correlation calculations were performed using 2D Shige freeware and moving window contour plots were plotted in Matlab. Spectral subtractions and peak fitting were calculated in GRAMS Ai software. GRAMS Ai was also used to plot Raman spectral figures and fluorescence data were plotted in Excel. Unfolding curves were drawn using Origin software.

7.3 Results and Discussion.

7.3.1 Fluorescence Spectroscopy.

The fluorescence spectra of RNase A (7 µM) at a number of different guanidine hydrochloride concentrations are shown in Figure 7.1. It can be seen from these spectra that fluorescence emission increases as the concentration of the denaturant increases, this is due to fluorescent, hydrophobic groups such as tryptophan and tyrosine being exposed as the protein unfolds into a less ordered state. The intensity of the fluorescence emission at the approximate peak centre (345 nm) was calculated for each GuHCl concentration and plotted as an equilibrium curve in Figure 7.2.

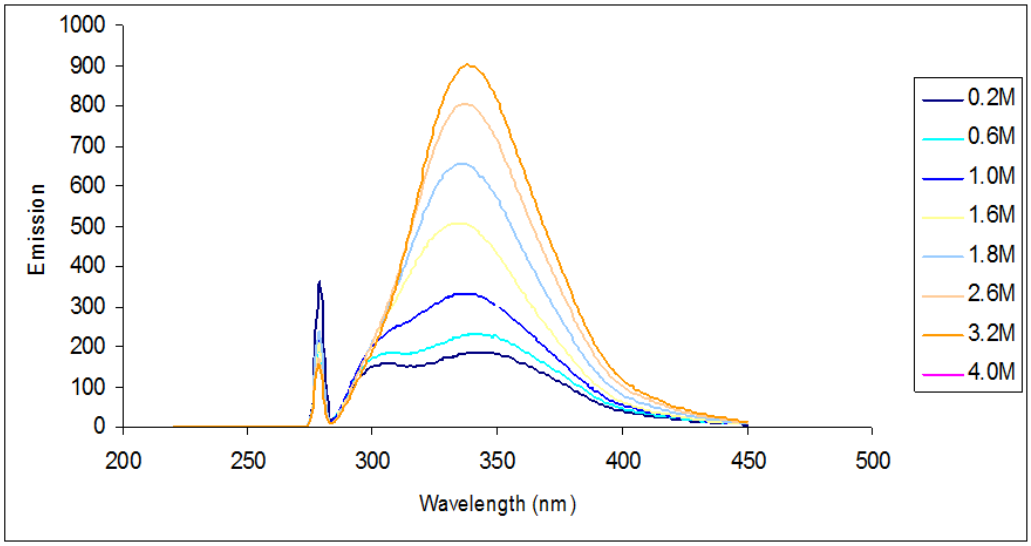

Figure 7.1: Fluorescence spectra of 7µM RNase A at different GuHCl concentrations.

It is possible from this plot to calculate the fractions of folded (*ff*) and unfolded (*fu*) protein molecules at each GuHCl concentration using equations 7.1A and B, in which X is the observed signal, X_N is the signal observed from the native protein and X_U is the signal observed from the unfolded protein.

$$fu = \frac{X - X_N}{X_U - X_N} \qquad \textbf{Eq. 7.1 (A)}$$

$$ff = 1 - fu \qquad \textbf{Eq. 7.1 (B)}$$

From these calculations it is possible to estimate the concentration of denaturant which is needed to unfold half of the protein molecules, the $[D]_{50.}$ The $[D]_{50}$ calculated for RNase A in this study was 3.1 M, which is consistent with $[D]_{50}$ quoted in the literature for RNase A as 3.1-3.2 M (Arnold and Ulbrich-Hofmann, 2000, Greene and Pace, 1974).

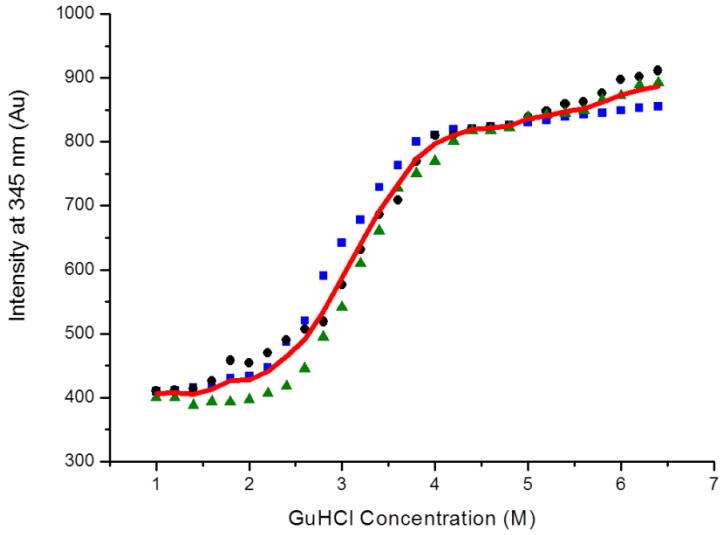

Figure 7.2: Protein unfolding curve for RNase A drawn from fluorescence data (intensity at 345nm). The red line indicates the mean of the triplicate measurements taken.

7.3.2 Raman Spectroscopy.

Raman spectra of RNase A (700 µM) over a similar concentration gradient of GuHCl were recorded along with a control spectrum of GuHCl in PBS at each concentration; the latter was used since GuHCl denaturant itself has a Raman spectrum and this needs to be compensated for. Thus prior to data analysis the control spectrum at each concentration was subtracted from the corresponding spectrum of protein and GuHCl; the resulting spectra are shown in Figure 7.3A, which focuses on the amide I region (1600-1700 cm^{-1}). A notable shift in this band as the concentration of denaturant increases can easily be observed along with a change in the band shape. In order to find the peak centre of the band a Gaussian curve fit was applied to the amide I region, using the peak fitting function in GRAMS Ai. Peak centre values were then plotted as a function of

denaturant concentration giving a traditional protein stability curve drawn from the Raman data (Figure 7.3B).

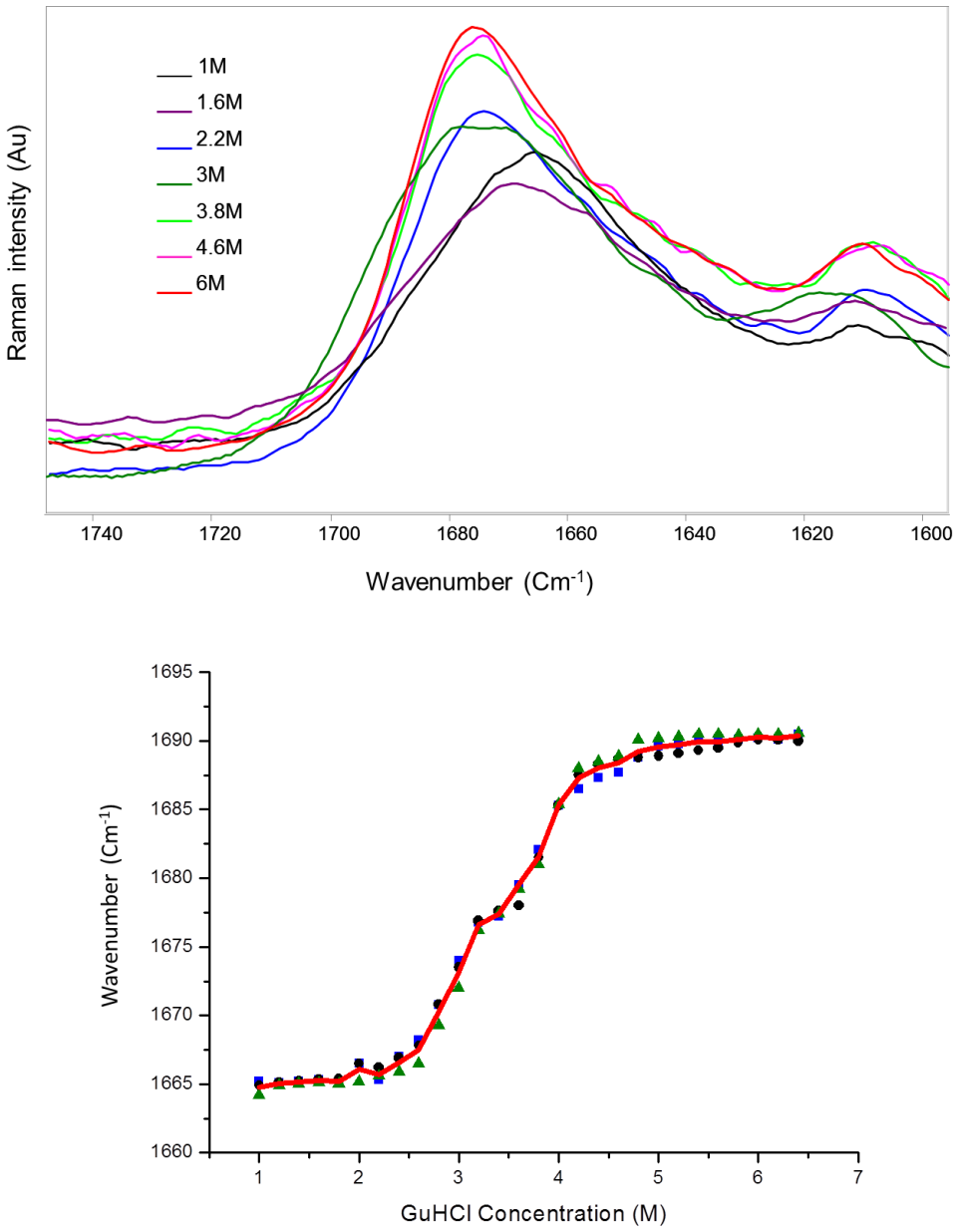

Figure 7.3: (A) Amide I region of the Raman spectra of 700 µM RNase A at various GuHCl concentrations after subtraction of the control GuHCl spectra, smoothing and baseline correction. **(B)** Protein unfolding curve for RNase A drawn from Raman Spectroscopy data (peak centre of Amide I band). The red line indicates the mean of the triplicate measurements taken**.**

7.3.3 Method Comparison.

In order to validate the use Raman spectroscopy for monitoring denaturant-induced unfolding in proteins, we compare the results gained from the Raman experiments to those obtained by 'traditional' fluorescence spectroscopy. We start by simply comparing the protein unfolding equilibrium curves drawn by each method; it can be seen easily that the curves compare favourably with each other, both indicating that the main unfolding events take place between GuHCl concentrations of 2.5-4 M. In order to compare both methods directly the data from both curves are plotted against each other in Figure 7.4. If the two methods produced identical results a linear trend would be seen in the graph and the best-fit line indicates this. The general shape of this curve (blue diamond symbols) does increase concomitantly confirming that the Raman results are somewhat comparable with the traditional fluorescence based method.

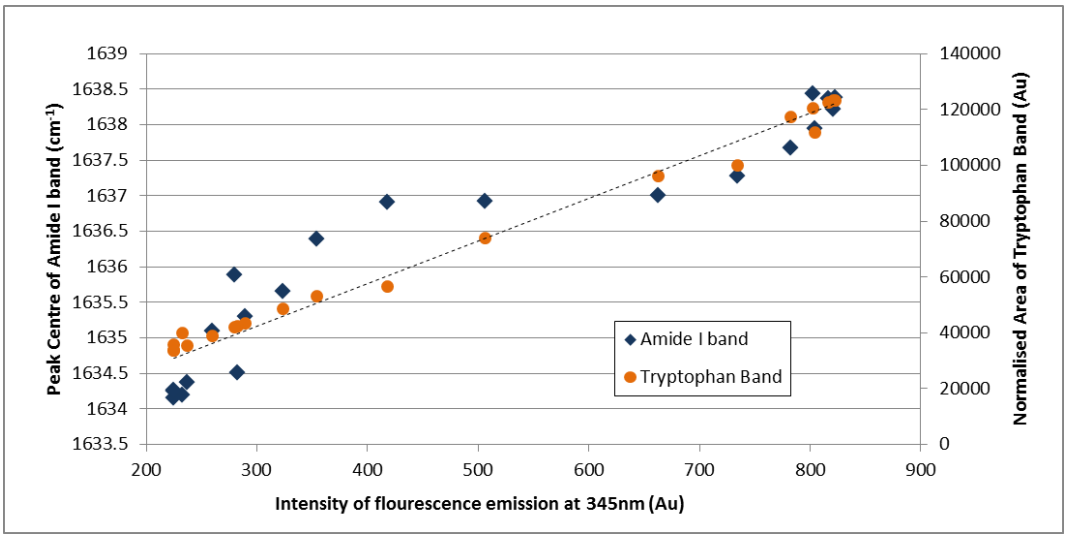

Figure 7.4: A Graph to compare RNase A unfolding data obtained by Raman and fluorescence spectroscopy.

However it is evident when comparing the stability curves in Figures 7.2 and 7.3A that the curves drawn from the Raman data appears to hint towards a two-stage transition, whereas the fluorescence data shows only one. This could be due to an inaccuracy in the Raman method, but as this the trend is observed consistently over three independent measurements it is more likely that the Raman spectroscopy data are able to highlight smaller conformational changes not observable by fluorescence spectroscopy. This

Raman method is therefore able to separate out two distinct transitions which overlap in the fluorescence data to appear as one transition, which has never before been reported for RNase A.

We have also compared our Raman method to the more traditional Raman-based approach for detecting conformational changes in proteins; this involves measuring changes in intensity of the tryptophan vibration centred at ~875 cm^{-1}; a band which will decrease in intensity as the protein unfolds and tryptophan residues become more exposed. This method, like the fluorescence method shows only one unfolding transition which occurs at ~3.1M GuHCl (Data are shown in Supplementary Information, Figure S7.1). This has been confirmed by adding this data to the method comparison graph in Figure 7.4 (orange circle symbols), where to allow ease of comparison the normalised reciprocal of the data has been plotted.

The $[D]_{50}$ values calculated from each of the triplicate measurements and the average $[D]_{50}$ from each method for RNase A are shown in Table 7.1. The values calculated from the Raman data are in very good agreement with the fluorescence results calculated from this study, and most encouragingly both values fall within the literature values for the $[D]_{50}$ of RNase A: 3.1-3.2 M GuHCl (Arnold and Ulbrich-Hofmann, 2000, Greene and Pace, 1974). The standard deviation of the triplicate measurements are also given in Table 7.1, and these show that the Raman spectroscopy to be a reproducible method of tracking protein unfolding and calculating $[D]_{50}$ values.

Furthermore, it is also possible to calculate the Gibbs free energy (ΔG) of a protein at each given denaturant concentration from the unfolding curves (Becktel and Schellman, 1987), using equation 7.2:

$$\Delta G = -RT \ln\left(\frac{fu}{ff}\right) \qquad \textbf{Eq 7.2}$$

where *R* is the ideal gas constant and *T* is the temperature in K.

The ΔG values for RNase A at the $[D]_{50}$ concentration (3.1 M) were calculated for both the Raman and fluorescence methods, also shown in Table 7.1, and were found to be in good agreement.

Table 7.1: Comparison of $[D]_{50}$ and ΔG values for RNase A and RNase B obtained from the fluorescence and Raman methods.

	$[D]_{50}$ (M)	ΔG (J mol^{-1})
RNase A		
Raman	3.14 (0.08)	-10971.47 (30)
Fluorescence	3.13 (0.04)	-11043.58 (12)
RNase B		
Raman	3.51 (0.06)	-3866.82 (41)
Fluorescence	3.56 (0.04)	-4037.50 (15)

Values are the mean of triplicate repeats and standard deviation is provided in parentheses. ΔG values quoted are the free energy at the $[D]_{50}$ concentration: for RNase A this was at 3.1, and at 3.5 for RNase B.

As an extra comparison step, and also to probe the two-step transition seen in the Raman unfolding profiles further, we have applied 2D correlation moving windows analysis to the data. The results of the 2D correlation moving windows analysis on the fluorescence and Raman data are shown in Figure 7.5A and B, respectively. Like with the stability curves, the 2D contour plots show that both methods indicate that unfolding transitions occur between guanidine concentrations of 2.8 and 4 M. The regions where the most changes are occurring in the spectra, indicated by the red contours, correspond well with the calculated $[D]_{50}$ values. Interestingly, as with the Raman equilibrium curve in Figure 7.3B, Figure 7.5B also shows the unfolding transition to be comprised of two separate events (and all three repeats show the same transitions; data not shown). Figure 7.5 shows the 2D correlation moving windows contour plot for the 860-900 cm^{-1} region of the Raman spectra, showing that the changes occurring in the tryptophan band are in one single transition. These results confirm that not only is this Raman spectroscopic method a suitable alternative to fluorescence spectroscopy for probing

protein structural changes, but that it can also provide additional information on more subtle changes which are not observed by the fluorescence method.

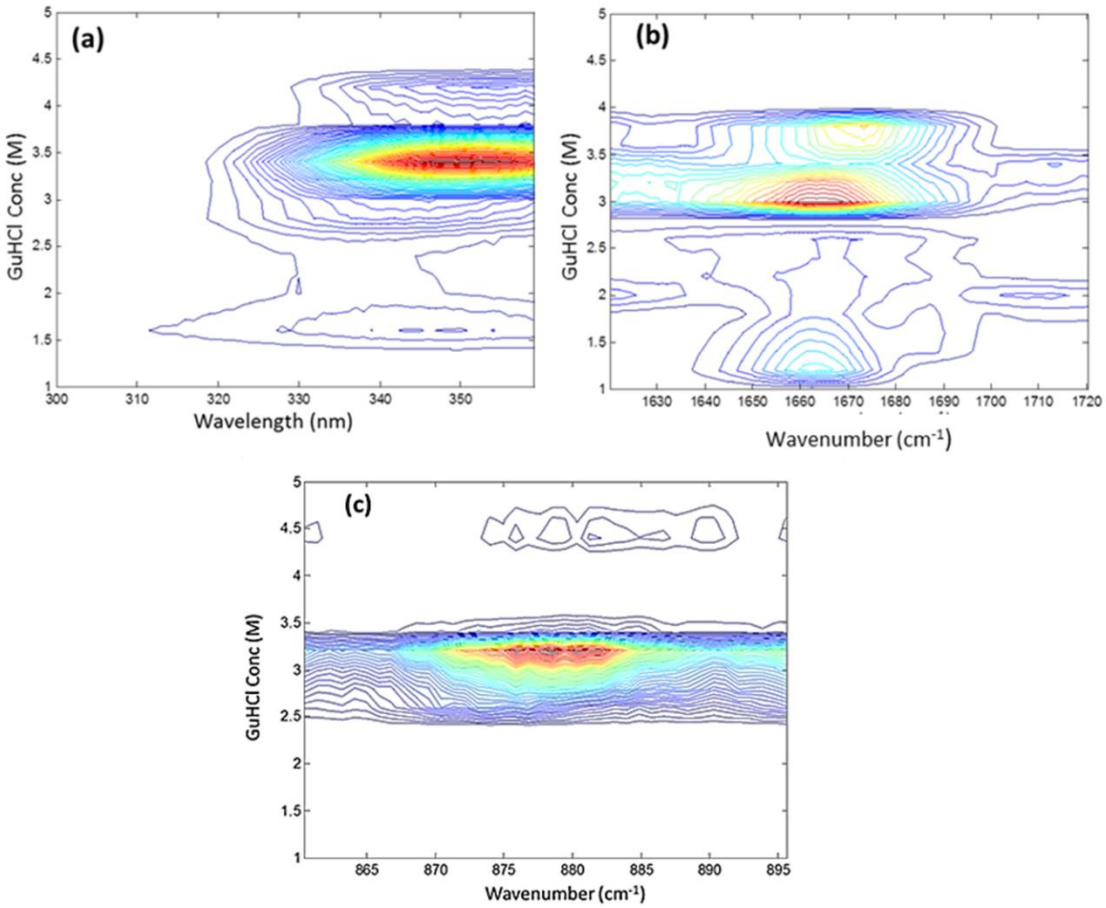

Figure 7.5: 2D Moving Window Contour Plots as a function of average translating window of GuHCl concentration from RNase A using: **(A)** fluorescence spectra; or Raman spectra from: **(B)** amide I region (1620-1720 cm^{-1}) or **(C)** tryptophan region (860-900 cm^{-1}).

7.3.4 Comparing the Stability of RNase A and B.

We used the Raman spectroscopy method detailed and tested above to monitor the unfolding of RNase B, the glycosylated form of RNase A. Both proteins have identical primary and secondary structure and very similar tertiary structures, with the differences that do occur being due to the addition of the glycan group. Therefore an additional aim of this work was to investigate whether Raman spectroscopy can be used to detect differences in protein stability which are brought about by glycosylation.

The Raman spectra of RNase B at various GuHCl concentrations show, as with the RNase A data, an upward shift in the position of the amide I band as the denaturant concentration increases (Figure 7.6A). Using the method described previously, protein unfolding curves were generated from triplicate measurements and the average curve from RNase B is shown in Figure 7.6B, compared to the average RNase A curve. It is clear in these stability curves that the concentration of GuHCl needed to denature RNase A is significantly lower than that needed for RNase B, confirming that the presence of a sugar group does indeed increase stability in RNase proteins.

The equilibrium curve was used to find the $[D]_{50}$ for RNase B (Table 7.1), which was calculated as 3.5 M, compared to 3.1 M GuHCl $[D]_{50}$ for RNase A, confirming that the glycoprotein is, indeed, the more stable molecule; unfortunately there are no literature values to confirm this, but the fluorescence measurements of RNase B shown in Table 7.1 do corroborate the Raman data. In addition, we can compare the ΔG values of each protein at the same GuHCl concentration, as ΔG will decrease (become negative) as the stability of a protein decreases. Therefore, comparing ΔG at 3.0 M GuHCl gives further proof that RNase B is the more stable system as ΔG for RNase B is -1002.7 J mol^{-1} compared with -9905.5 J mol^{-1} for RNase A.

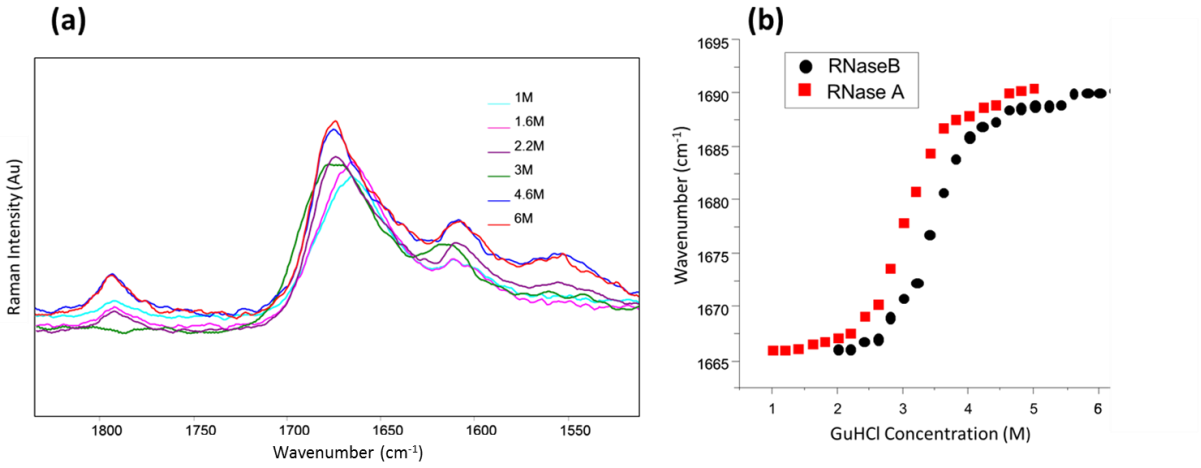

Figure 7.6: (A) Amide I region of the Raman spectra of RNase B at various GuHCl concentrations after subtraction of the control GuHCl spectra, smoothing and baseline correction. **(B)** Protein unfolding curves for RNase A and RNase B drawn from Raman Spectroscopy data (peak centre of Amide I band). Plotted values are the mean of three measurements.

7.4 Conclusions.

We have described a novel Raman spectroscopy-based method for monitoring the unfolding of proteins in the presence of a chemical denaturant, and have shown the potential of this technique through comparisons with fluorescence spectroscopy. Through the use of unfolding curves and $[D]_{50}$ calculations we have shown that the results obtained from Raman spectroscopy are very comparable to those obtained by a conventional fluorescence unfolding experiment. By employing 2D correlation moving windows analysis, we have been able to demonstrate that Raman spectroscopy is more sensitive to smaller conformational changes than fluorescence emission data. Finally, using RNase A and B as model proteins, we have shown that this Raman method is capable of evaluating increases that occur in stability when this protein is glycosylated.

7.5 Supplementary Information .

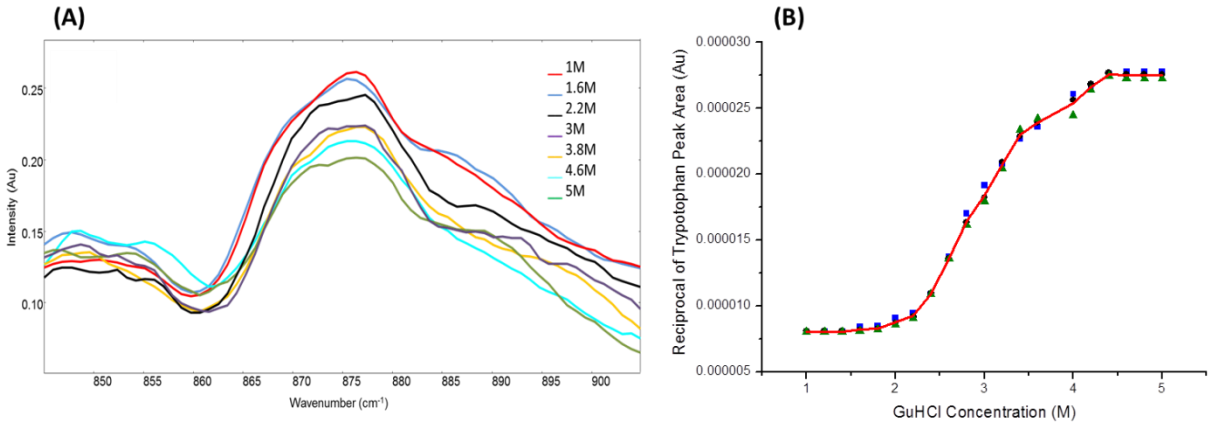

Figure S7.1: (A) Tryptophan region of the Raman spectra of RNase A at various GuHCl concentrations after subtraction of the control GuHCl spectra, smoothing and baseline correction and **(B)** Protein unfolding curve for RNase A drawn from Raman Spectroscopy data (peak area of tryptophan band). The red line indicates the mean of the triplicate measurements taken, reciprocal values (1/N) have been plotted to make curves easily comparable with Figures 7.2 and 7.3 B.

Chapter 8: Detecting Foreign Protein Contamination in Protein Samples Using High Throughput FT-IR Spectroscopy and Multivariate Analysis.

Work presented in this chapter has been adapted from the following paper submitted for publication; Correa E., Brewster V.L., and Goodacre R. "Fast Detection of Low Levels of Protein Contamination Using FT-IR Spectroscopy Coupled to Chemometric Analysis".
E. Correa performed multivariate analysis on the data and assisted in interpretation of the results.

8.1 Introduction.

As protein based therapeutics begin to dominate the pharmaceutical market, the purification and characterisation of these products continue to pose many analytical challenges. One of these is the separation of the protein product from any other proteins that the bioprocess yields, and the validation of this step (Greer, 2008). Removal of host cell protein (HCP) based contaminants is a vital downstream processing step, as these "foreign" proteins can give the product undesirable immunogenic effects (Goddard, 1991). This study was concerned with the development of a high throughput method that uses vibrational spectroscopy coupled to chemometric models to detect protein contaminants in a biopharmaceutical product.

Lysozyme and cytochrome *c* (cyt *c*) were used to mimic a therapeutic protein contaminated with a single foreign protein and to establish a model. Ribonuclease (RNase) A and B were also used to simulate a protein contaminated with a glycosylated equivalent. FT-IR spectroscopy was chosen as the analytical technique for this application because it allows high throughput (typically 30s per sample) and automated

data acquisition, an essential criteria for biopharmaceutical process analytical technology (PAT).

FT-IR spectroscopy has a long history as a versatile, non-destructive, qualitative and quantitative method of monitoring proteins. The Infrared spectra of proteins are information rich and contain many bands which are highly sensitive to secondary and tertiary conformational changes (Manning, 2005). Infrared spectra of biological molecules contain large numbers of variables, and are often complex containing many overlapping bands, therefore, changes in the FT-IR spectra can be difficult to identify by visual examination of the dataset alone, and hence the application of multivariate analysis (MVA) becomes necessary.

In this work we demonstrate the ability of FT-IR spectroscopy coupled with MVA techniques to detect low level (1-5%) contamination of proteins in two model systems. Several well-known chemometric methods were employed for data analysis, all analyses were rigorously tested using resampling methods, and the results are discussed and compared against each other.

8.2 Materials and Methods.

8.2.1 Materials.

Initially, well characterised proteins were chosen to establish a model. Lysozyme and cytochrome *c* were chosen to mimic a protein product contaminated with a foreign protein. These proteins were chosen because lysozyme contains both α-helix and β-sheet structures, whereas cyt *c* is mainly helical as shown in Figure 8.1. RNase A and RNase B were used as a mock system where a protein is contaminated with a glycosylated equivalent. Bovine pancreatic Ribonuclease A and B, lysozyme from chicken egg white and cytochrome *c* from equine heart were all purchased from Sigma-Aldrich (Dorset, UK).

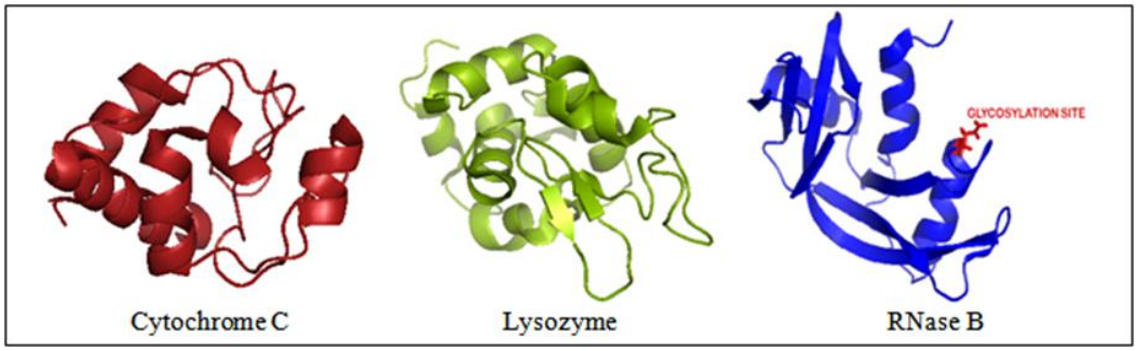

Figure 8.1: Cartoon representations of the three-dimensional structures of cytochrome *c* (left), lysozyme (middle) and RNase B (right). Drawn from PDB files (1CRC, 3IJV and 5RSA) using PyMOL

8.2.2 Method.

Lyophilised proteins (cyt *c*, lysozyme, RNase A and RNase B) were dissolved in ultra-pure water at a concentration of 5mg/mL. The spiked samples were then created by mixing the appropriate amount (5% v/v and 1% v/v) of lysozyme into cyt *c* and RNase B into RNase A. 5 µL of each sample was loaded into a 96 well silicon plate, which had been pre-washed in methanol and deionised water. The plate was then left to air dry at room temperature for 60 min, until completely dry. Eight different 96 well silicon plates were prepared as described in Table 8.1 and to minimise instrumental drift being incorporated into the measurements the samples were semi-randomised by being positioned in alternate rows on each plate.

Table 8.1: Summary of samples spotted onto each of the 96 well silicon plates.

	Sample and Number of wells
Plate 1	Cyt *c* (94)
Plate 2	Cyt *c* (47) Lysozyme (47)
Plate 3	Cyt *c* (47) Cyt *c* with 5% Lysozyme (47)
Plate 4	Cyt *c* (47) Cyt *c* with 1% Lysozyme (47)
Plate 5	RNase A (94)
Plate 6	RNase A (47) RNase B (47)
Plate 7	RNase A (47) RNase A with 5% RNase B (47)
Plate 8	RNase A (47) RNase A with 1% RNase B (47)

8.2.3 FT-IR Spectroscopy.

FT-IR spectra were collected on a Bruker FT-IR instrument, as described in 2.1.2. Spectra were collected from each of the 96 wells over a wavenumber range of 4000-600 cm^{-1}, with a spectral resolution of 4 cm^{-1}. For each well 64 accumulations were collected and co-added to improve the signal-to-noise ratio. A total of 752 spectra were collected over 8 plates with an average collection time of 30s per sample. Data were exported from the instrument manufacture's Opus software to Excel spread sheets, which were later analysed in R.

8.2.4 Data Analysis.

Prior to multivariate analysis data were pre-processed in R. Spectra were normalised using EMSC (polynomial order 3), and auto-scaled so that each data column, corresponding to a wavenumber variable, had a mean equal to zero and a standard deviation equal to one. Data were then analysed using R to perform PCA, DFA, PLSR, PLS-DA, support vector machine (SVM) and artificial neural networks (ANNs). As these are supervised multivariate data analysis methods (with the exception of PCA), results need to be validated as both *X* and *Y* data is used in the model formation, and therefore the results may be subject to bias. The bootstrap cross validation method described previously in 2.3.3.1.1 was used, with results calculated over 1000 bootstrap cross validations.

8.3 Results and Discussion.

8.3.1 FT-IR Spectra.

FT-IR spectra were recorded of pure cytochrome *c*, pure lysozyme and cytochrome *c* spiked with lysozyme at 5 and 1%. This was repeated for RNase A, RNase B and RNase A spiked with RNase B. Data are shown in Figure 8.2. The need for chemometrics is highlighted here as little or no changes are visible by eye in the spectra of the spiked and pure samples, particularly in the RNase A/B system.

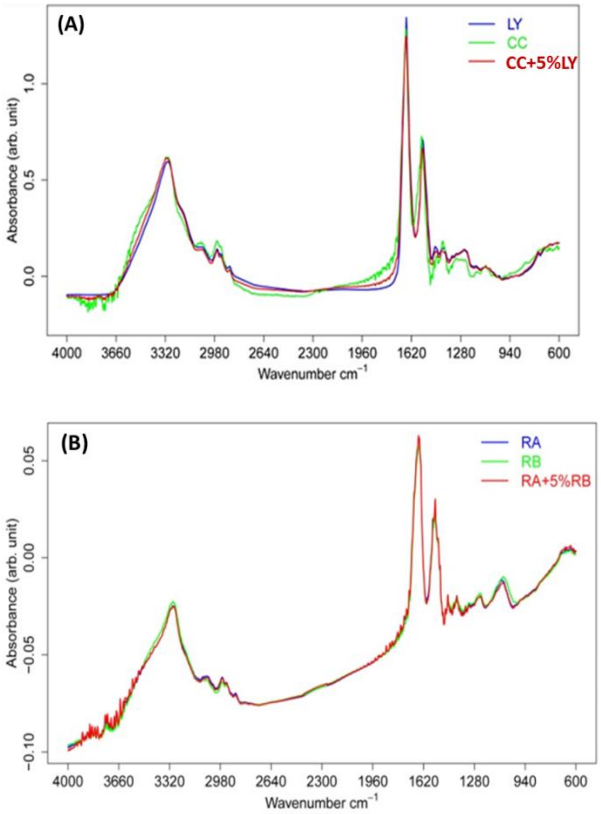

Figure 8.2: Average FT-IR spectra of **(A)** lysozyme (LY), cytochrome *c* (CC) and CC contaminated with 5% LY (CC5%LY) and **(B)** RNase A (RA), RNase B (RB) and RA contaminated with 5% RB (RA+5%RB).

8.3.2 Unsupervised Clustering - PCA.

First we applied PCA to both of the model systems tested; scores plots for this analysis are shown in Figure 8.3. We can see from Figures 8.3A and 8.3B that with both the cyt *c* and RNase samples we are unable to differentiate between pure samples and those spiked with 1% contamination. When the concentration of spiked lysozyme and RNase B is increased to 5%, the PCA scores plots (Figure 8.3C and 8.3D) show a slight improvement in the ability to distinguish between the FT-IR spectra of pure and contaminated samples. However there is still a large amount of overlap for the two groups, particularly in the cyt *c*-lysozyme system. As PCA plots show significant overlap of spiked and pure samples for both of the data sets, we have then applied more sophisticated supervised MVA methods to aid in discrimination of the samples.

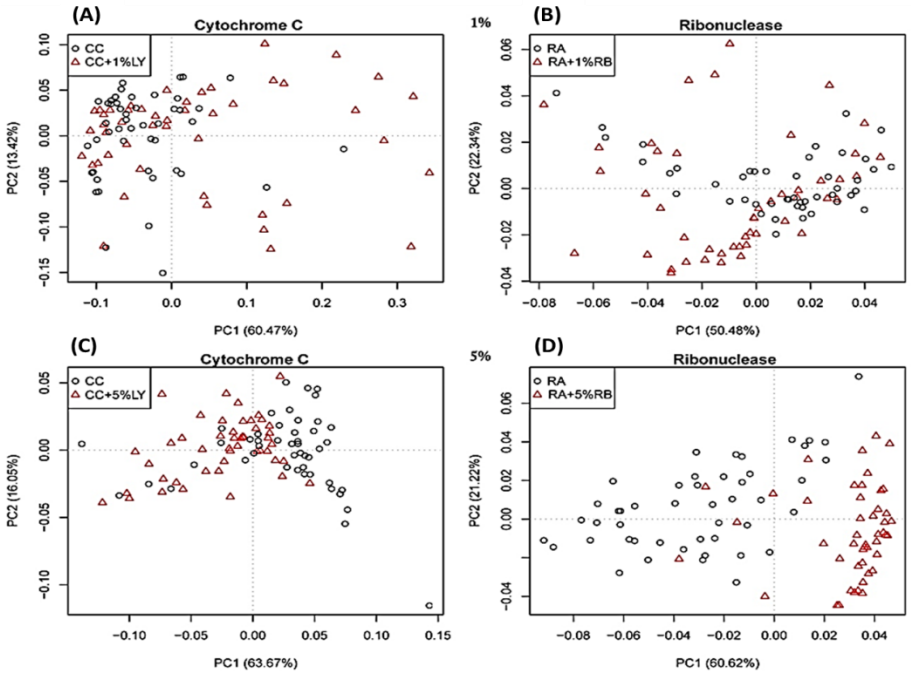

Figure 8.3: PCA scores plots of **(A)** pure *vs.* 1% contaminated cyt *c*, **(B)** pure *vs.* 1% contaminated RNase A, **(C)** pure *vs.* 5% contaminated cyt *c* and **(D)** pure *vs.* 5% contaminated RNase A.

8.3.3 Supervised Clustering.

8.3.3.1 PC-DFA.

In an attempt to improve discrimination, and also to find out how much variance among the proteins the FT-IR analysis has detected, we have applied DFA to the data. In this case we used the first 10 PCs from PCA as the input variables, as opposed to the full FT-IR spectra.

Figure 8.4 shows the resulting DFA scores plot for all samples. As expected, the analysis detected more similarities in the proteins which have similar secondary and tertiary structures; RNase A and B. As with the PCA results, this method clearly discriminates between pure cyt *c* and lysozyme, but there is still a large amount of overlap between pure and spiked cyt *c* samples.

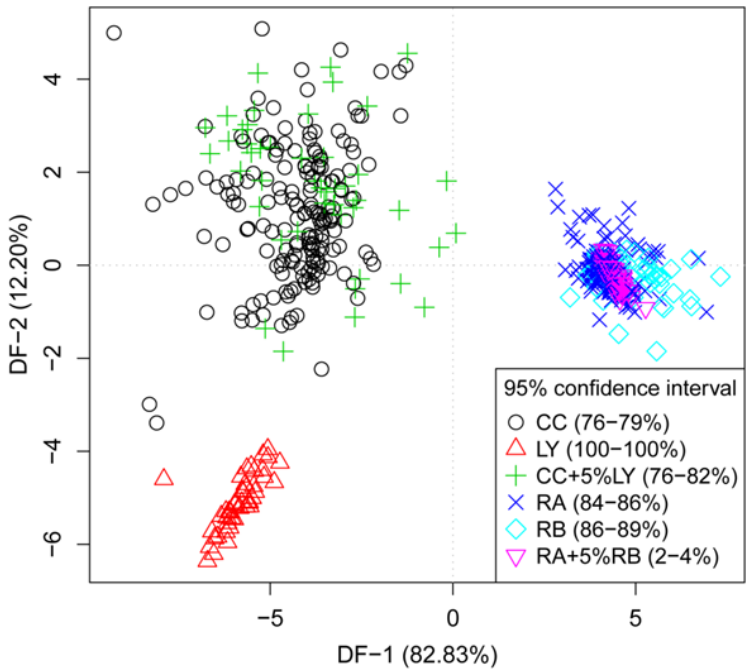

Figure 8.4: PC-DFA applied to the full dataset. 10 PCs (TVE = 99.2%) were used as input variables for DFA. The inset reports 95% confidence intervals for correct classifications estimated over 1000 bootstrap cross-validations.

The numbers in the parenthesis the figure ledgend report the percentage of samples which fall within the 95% confidence interval for each class, which has been estimated over 1000 bootstrap cross validations. The statistics here indicate a good discrimination across all protein classes, with the exception of RNase A spiked with 5% contamination, where the model performs particularly poorly with only 2-4% of samples correctly classified.

From these results we propose that PC-DFA could be used as a filtering step for these samples; proteins analysed by PC-DFA would be either classified as (A) unequivocally identified as lysosome and no further action is needed, (B) recognized as being cyt *c* type (either pure or spiked) where further MVA is required, or (C) identified as being RNase proteins, where again subsequent MVA is needed for further discrimination. Therefore additional supervised methods have been employed to resolve the identity of samples classified in group B or C, where each group has been treated as a separated data set.

8.3.3.2 PLSR.

Once samples were separated into sub-classes of RNase and cyt *c* PLSR models were applied to classify samples as pure or contaminated. We aimed to use these models to quantify contamination; i.e., distinguish between 0%, 1% and 5% contamination.

Figures 8.5A and 8.5B show the distribution of PLSR predictions over 1000 bootstrap cross validations for the discrimination of pure samples and samples spiked with 1% contamination. For comparison, we also show the distribution of predicted values when we try to differentiate pure samples from pure samples, i.e. a null distribution (Figures 8.5C and 8.5D). For RNase samples we see that in the null distribution the curves lie on top of each other; indicating, as one would expect that the model can not discriminate between these identical samples. Figure 8.5A shows that the curves for RNase A and RNase A spiked with 1% RNase B lie offset from each other indicating that in the majority of cases the model can discriminate between pure and contaminated RNase samples. Figure 8.5B shows significantly poorer results for the cyt *c* system, showing that the

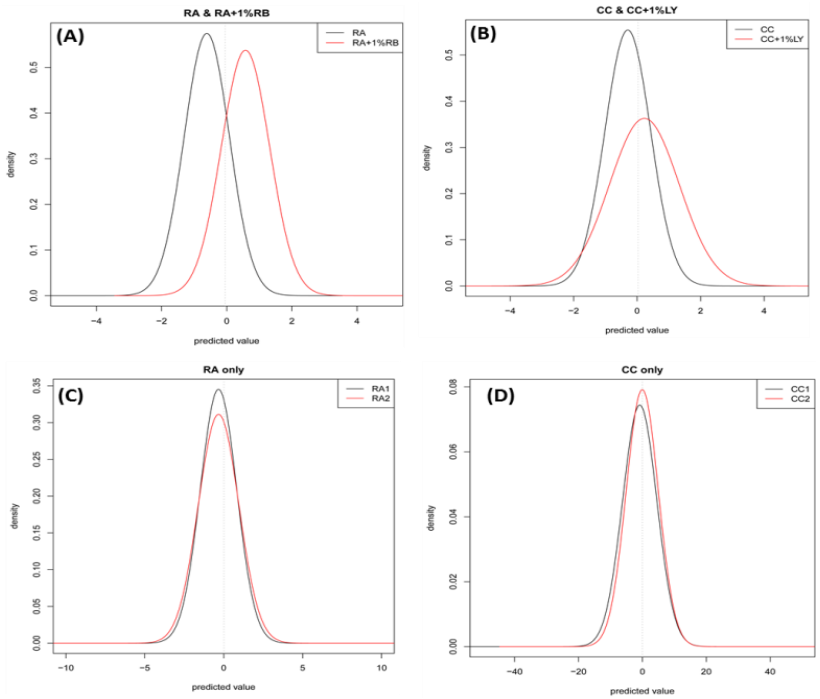

Figure 8.5: Graphs to show **(A)** the distribution of predicted values for RNase A and RNase A spiked with 1% RNase B, **(B)** the distribution of predicted values for cyt *c* and cyt *c* spiked with 1% lysozyme, **(C)** the null distribution of predicted values for RNase A and RNase A and **(D)** the null distribution of predicted values for cyt *c* and cyt *c*.

model can only correctly classify a small portion of the samples. Whilst this contradicts the results for the PC-DFA, which discriminated poorly between pure and spiked RNase A, but performed much better for the cyt *c* samples, this is likely to be because we are now treating each protein set separately.

The R^2 values for the PLSR models over the 1000 bootstrap cross validations are shown in Figure 8.6. The PLSR model for the cyt *c* data, Figure 8.6A, shows a mean R^2 of 0.99. This very high prediction result indicates that FT-IR spectroscopy detected significant differences between pure and contaminated samples. By contrast, the permuted model shows poor prediction results, with a cross validated mean R^2 of –1.1, confirming that the success of the original PLSR model did not occur by chance.

Figure 8.6B shows the equivalent results for the RNase A and B system, where original bootstrap models suggest a successful quantification of contamination levels; with a mean R^2 of 0.87. Once more, results of the permutation testing confirm the validity of the models built from the correctly labelled data. Figure 8.6 also reports the p-values from *t*-tests performed on each respective bootstrap and permutation test result; in both cases p-values are ~10^{-6}, providing additional confirmation that the PLSR results correctly reflect the patterns within the data rather than results which are occurring by chance.

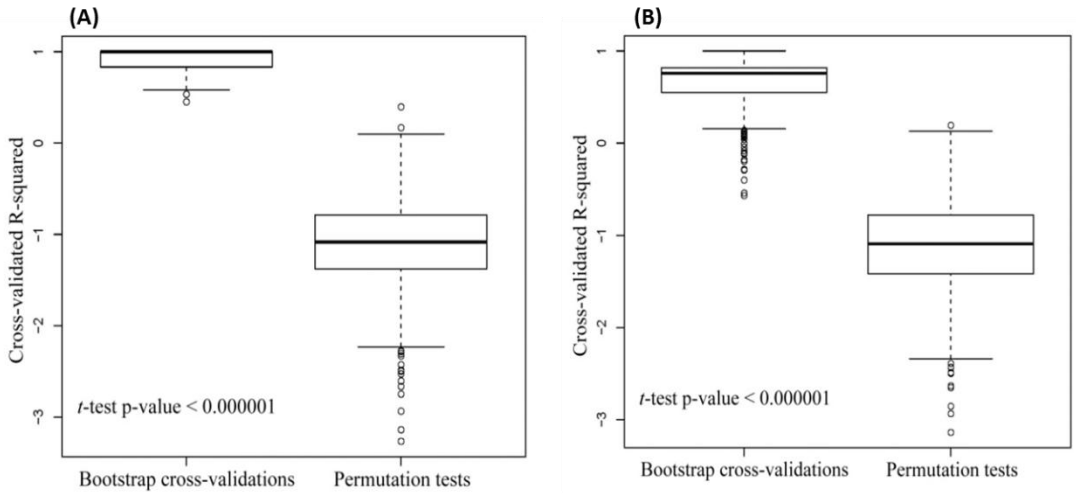

Figure 8.6: PLS model validation**. (A)** reports 1000 cross-validations for the PLS models applied to pure and contaminated cyt *c***. (B)** reports 1000 cross-validations for the PLS models applied to pure and contaminated RNase A.

8.3.4 Optimising discrimination of pure and contaminated proteins.

In order to assess the best chemometrics approach for discrimination from the FT-IR data, we compared three additional supervised classification algorithms: partial least squares-discriminant analysis (PLS-DA), support vector machines (SVMs) and artificial neural networks (ANNs). PLS-DA is a particular case of PLSR which finds a linear function which best defines or separates the data (Wold et al., 2001). In PLS-DA the dependant variable is a binary or dummy variable. A dummy variable is a transformed variable that encodes the presence or absence of a characteristic or property, in the present work the dummy variable encoded the following characteristics: “pure protein” (absence of contamination encoded as 0) and “contaminated protein” (presence of contamination encoded as 1). SVM is an alternative supervised algorithm, which uses data points which are near the border of the two groups to create a support vector (Burges, 1998). These vectors are used to form separating hyperplanes which are applied to define the boundaries between groups. ANNs are non-linear methods which are designed to simulate the way brain neurons communicate (Burges, 1998). A neural network consists of groups of interconnecting nodes which receive numerical inputs and process them into multiple outputs. The nodes are organised into layers, and each node of one layer is connected to all nodes of the next layer. The network used in this work was the most common ANN model: a single hidden layer feed-forward back-propagation network, which consists of one input layer (comprising 1764 inputs), one hidden layer (consisting of 150 nodes) and a final single node output unit.

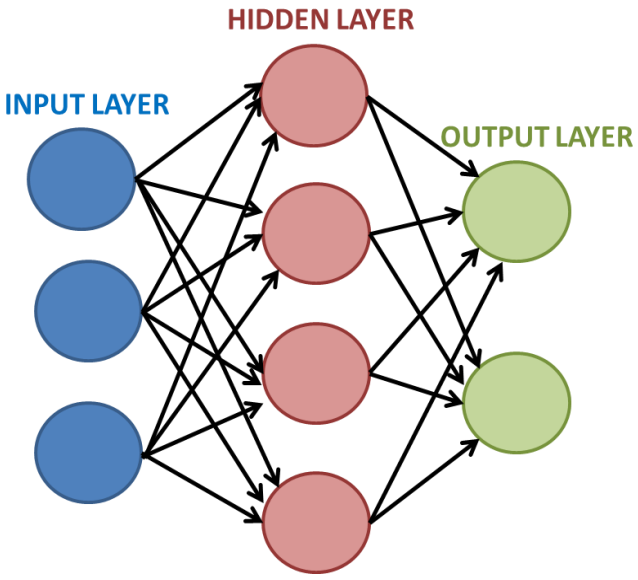

Figure 8.7: Schematic diagram representing an artificial neural network with 3-4-2 topology, in this study the topology was 1764-150-1.

For this analysis the data set from each model system was partitioned into the sub-sets as described in Table 8.2, thus giving results for discrimination between pure and pure samples, pure and 1% contaminated sample and pure and 5% contaminated samples.

Table 8.2: Summary of data partitions into sub-sets for MVA.

DATA SET	NAME	DISCRIPTION
CC 1	CC *vs* cc	94 pure cyt *c* samples, 47 of which are intentionally mis-labelled as spiked.
CC 2	CC *vs* CC1%LY	47 pure cyt *c* and 47 cyt *c* with 1% lysozyme
CC 3	CC *vs* CC5%LY	47 pure cyt *c* and 47 cyt *c* with 5% lysozyme
RA 1	RA *vs* ra	94 pure RNase A samples, 47 of which are intentionally miss-labelled as spiked.
RA 2	RA *vs* RV1%RB	47 pure RNase A and 47 RNase A with 1% lysozyme
RA 3	RA *vs* RV5%RB	47 pure RNase A and 47 RNase A with 5% lysozyme

The results from all three classification algorithms for all three data sub-sets were calculated over 1000 bootstrap cross-validations and the hold out sets are summarised in the classification matrix displayed in Table 8.3. As expected, for the 'null models' CC *vs.* cc and RA *vs.* Ra, all algorithms predicted an average of 50% of the samples correctly. For discrimination between pure and spiked samples PLS-DA on average outperformed the other methods. The PLS-DA model performed particularly well with the RNase samples; correctly identifying over 80% of the 1% samples and 94% of the 5% samples. For the cyt *c* system, PLS-DA was able to correctly identify 98% of samples with 5% contamination, but only 57% of the samples contaminated with 1% lysozyme.

The loadings from PLS-DA for both RNase and cyt C data were examined to allow identification of the bands in the IR spectra which contribute significantly to the discrimination between pure and spiked samples. Figure 8.8 shows the loadings from the model which discriminated between pure cyt *c* and cyt *c* spiked with 5% lysozyme, with assignments given in Table 8.4. The band at ~1665 cm^{-1} arises from the amide I vibrations (C=O), and can be specifically assigned to β-sheet contributions (Tuma, 2005). This difference most likely arises because lysozyme has ~10% β-sheet content and cyt *c*

Table 8.3: Classification matrix for ANN, PLS-DA and SVM applied to discriminate between pure and contaminated protein samples. Data shown are from the test sets only.

			Predicted class					
			ANN		PLSDA		SVM	
			CC	Cc	CC	Cc	CC	Cc
Real class	Cytochrome C	**CC**	**53%**	**47%**	**51%**	**49%**	**54%**	**46%**
		cc	45%	55%	56%	44%	48%	52%
			ANN		PLSDA		SVM	
			CC	CC1%LY	CC	CC1%LY	CC	CC1%LY
		CC	63%	37%	69%	31%	66%	34%
		CC1%LY	42%	58%	43%	57%	37%	63%
			ANN		PLSDA		SVM	
			CC	CC5%LY	CC	CC5%LY	CC	CC5%LY
		CC	95%	5%	100%	0%	99%	1%
		CC5%LY	5%	95%	2%	98%	2%	98%
			ANN		PLSDA		SVM	
			RA	Ra	RA	Ra	RA	Ra
	Ribonuclease A	RA	51%	49%	57%	43%	59%	41%
		ra	50%	50%	54%	46%	51%	49%
			ANN		PLSDA		SVM	
			RA	RA1%RB	RA	RA1%RB	RA	RA1%RB
		RA	75%	25%	84%	16%	82%	18%
		RA1%RB	27%	73%	20%	80%	23%	77%
			ANN		PLSDA		SVM	
			RA	RA5%RB	RA	RA5%RB	RA	RA5%RB
		RA	87%	13%	96%	4%	92%	8%
		RA5%RB	11%	89%	6%	94%	8%	92%

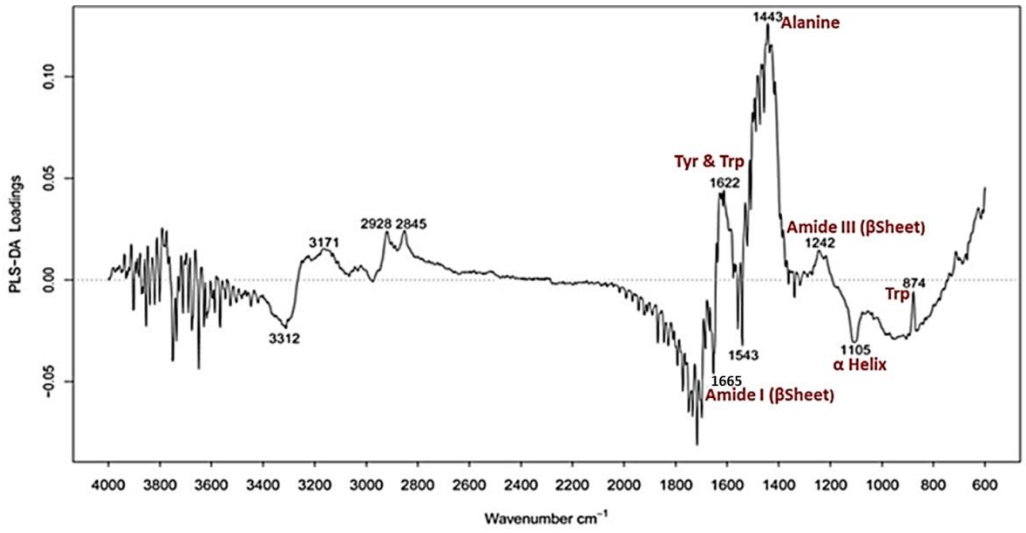

Figure 8.8: PLS-DA loadings plot for the discrimination between pure cyt *c* and cyt *c* + 5% LY.

only ~1% β-sheet content, so the addition of lysozyme to cyt *c* increases the percentage of β-sheet structures in the sample. A similar explanation can be attributed to the ~1242 cm^{-1} amide III band (C-N and N-H), which can also be assigned to β-sheet structure (Socrates, 2001). Other features in the loadings can be assigned to the contributions from the side chains of amino acids which increase in number when lysozyme is added. For instance, the bands at ~1622 cm^{-1} and ~874 cm^{-1} can both be assigned to the aromatic side chain of tryptophan (Liang et al., 2006). Variations in this band can be explained by the fact that cyt *c* has only 1 tryptophan residue but lysozyme has 6, therefore spiked samples will have a higher concentration of tryptophan, and therefore these bands are likely to increase in intensity in spiked samples.

Table 8.4: Assignment and Discussion of FT-IR bands identified as relevant for the discrimination between pure cyt *c* and cyt *c* + 5%LY.

~Wavenumber (cm^{-1})	Assignment	Justification
1665	Amide I – β-sheet	Lysozyme is 10% β-sheet, cyt *c* only 1%
1622	Tyrosine and Tryptophan	Cyt *c* has 4 Tyr and 1 Trp, Lysozyme has 3 Tyr and 6 Trp
1443	Alanine	Increase in number of alanine, cyt *c* has 6 and lysozyme 12
1242	Amide III – β-sheet	Lysozyme is 10% β-sheet, cyt *c* only 1%
1105	A-helix	Cyt *c* is mainly α-helix and lysozyme is α-helix and β-sheet
874	Tryptophan	Lysozyme as 6, cyt *c* 1; Trp is more buried in cyt *c*.

Figure 8.9 presents the loadings for the discrimination between pure RNase A and RNase A + 5% RNase B. As with the loadings from the cyt *c* data, the RNase loadings show contributions from amide I and III bands. This could be because the RNase B spectra exhibited a change in these bands as a result of the changes to the higher order structure of the protein transformed by the addition of a glycan. Other bands can be assigned to glycosidic vibrations arising from the sugar component of RNase B. Full assignments for the bands highlighted in these loadings are discussed in Table 8.5.

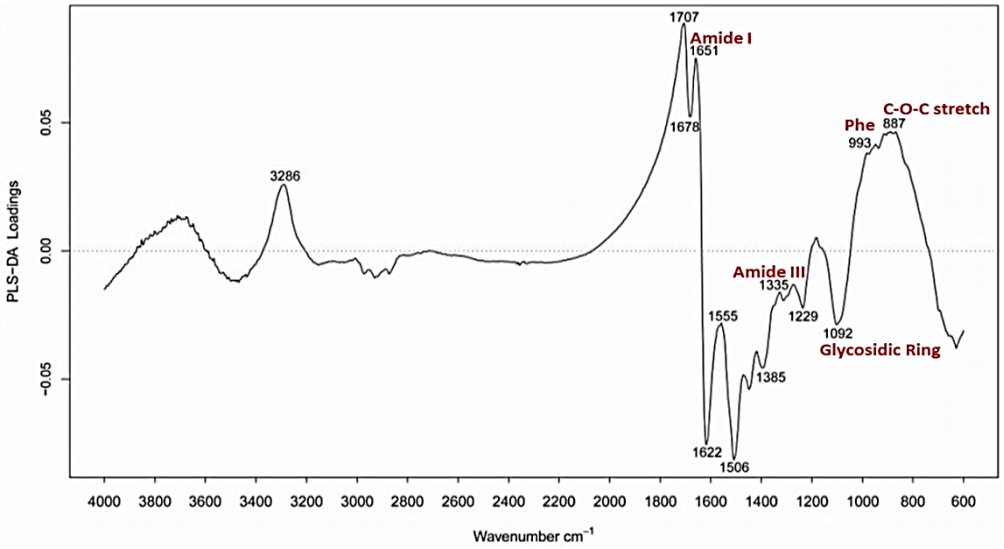

Figure 8.9: PLS-DA loadings plot for the discrimination between pure RNase A and RNaseA + 5% RNase B.

Table 8.5: Assignment and Discussion of FT-IR bands identified as relevant for the discrimination between pure RNase A and RNase A +5% RNase B.

~Wavenumber (Cm^{-1})	Assignment	Justification
1678/1651	Amide I	Conformational changes brought about by addition of glycan
1335	Amide III/ NH_2 twist	NH_2 twisting mode of GlcNAc in glycan appears as a shoulder of amide III
1092	Glycosidic ring	Glycosidic ring deformation modes of the glycan component of RNase B
993	Phenylalanine	Change in local environment due to conformational changes
887	C-O-C stretch	Glycan based vibrations from RNase B

8.3.4.1 Probability of Correct Classification.

In this final step we have converted the PLS-DA results from Table 8.3 into numerical values which describe the precision of this method; this is a measure which assesses the percentage of samples identified as contaminated that actually are contaminated. The precision of our chemometric models was calculated using equation 8.1; where *TP* is the rate of true positives and *FP* is the false positive rate. This value is essentially a Bayesian

or conditional probability that a sample identified by FT-IR spectroscopy as contaminated is in fact a real contaminated sample.

$$Prescision = \frac{TP}{TP+FP} \qquad \textbf{Eq. 8.1}$$

Table 8.6 reports the calculated probabilities of correct classification for each model. This table illustrates that this method is a precise method of detecting contamination in RNase proteins; with the probability a sample is correctly identified as contaminated being 0.963 for 5% contamination, along with 0.863 for 1% RNase B. The precision is equally high for detection of cyt *c* spiked with 5% lysozyme, 0.998, but falls significantly to 0.649 when the contaminant concentration is lowered to 1%.

Table 8.6: Model precision: probability of correct identification of contaminated samples for spiked cytochrome *c* and ribonuclease A samples.

Precision for spiked samples				
Pr(X) = Pr(spiked|predict_as_spiked) = probability that a sample is in fact spiked given that the algorithm identified the sample as spiked				
Algorithm	**Comparison**	**Pr(X)**	**Comparison**	**Pr(X)**
PLSDA	(100%CC) vs (99%CC+1%LY)	0.649	(100%CC) vs (95%CC+5%LY)	0.998
	(100%RA) vs (99%RA+1%RB)	0.830	(100%RA) vs (95%RA+5%RB)	0.963
ANN	(100%CC) vs (99%CC+1%LY)	0.608	(100%CC) vs (95%CC+5%LY)	0.954
	(100%RA) vs (99%RA+1%RB)	0.746	(100%RA) vs (95%RA+5%RB)	0.868
SVM	(100%CC) vs (99%CC+1%LY)	0.647	(100%CC) vs (95%CC+5%LY)	0.994
	(100%RA) vs (99%RA+1%RB)	0.815	(100%RA) vs (95%RA+5%RB)	0.924
Precision for non-spiked (pure) samples				
Pr(Y) = Pr(pure|predict_as_pure) = probability that a sample is in fact pure given that the algorithm identified the sample as pure				
Algorithm	**Comparison**	**Pr(Y)**	**Comparison**	**Pr(Y)**
PLSDA	(100%CC) vs (99%CC+1%LY)	0.694	(100%CC) vs (95%CC+5%LY)	0.998
	(100%RA) vs (99%RA+1%RB)	0.836	(100%RA) vs (95%RA+5%RB)	0.963
ANN	(100%CC) vs (99%CC+1%LY)	0.627	(100%CC) vs (95%CC+5%LY)	0.953
	(100%RA) vs (99%RA+1%RB)	0.750	(100%RA) vs (95%RA+5%RB)	0.868
SVM	(100%CC) vs (99%CC+1%LY)	0.661	(100%CC) vs (95%CC+5%LY)	0.994
	(100%RA) vs (99%RA+1%RB)	0.824	(100%RA) vs (95%RA+5%RB)	0.924

8.3 Conclusions.

Successful identification of low levels of protein contamination has been achieved using FT-IR spectroscopy and chemometric analysis. Initial inspection of FT-IR data and PCA scores plots highlighted the need for a sophisticated chemometric approach, as only small amounts of variation were observed in the spectra of pure and contaminated proteins. DFA results demonstrated that the analytical technique chosen recorded spectral information from the samples that matched the expected protein clustering according to prior biological knowledge of their respective structures. The DFA approach can be applied as a filter step to narrow down possibilities and the final identity of the samples resolved by other MVA methods. Low levels of protein contamination were consistently quantified correctly by PLSR regression models.

We tested three supervised classification methods, all of which demonstrated a high level of discrimination between pure and contaminated proteins based of their FT-IR spectra. On average, PLS-DA produced the best discrimination results for the dataset analysed, and PLS-DA loadings indicated bands that could be assigned to secondary structure features or the side chains of amino acids, which increase in quantity in the protein contaminant, and therefore in the spiked sample.

These results confirmed that this approach of FT-IR spectroscopy coupled to chemometric analysis can be used for fast and reliable detection of low levels of protein contamination and this has several possible applications for validation of the recovery and purification of biosynthetic products. The positive results from this study motivate and suggest future work using even lower concentrations of protein contamination and the inclusion of other protein types closer to those that would be encountered in the biopharmaceutical industry.

Chapter 9: Detecting Protein Contamination Using FT-IR Spectroscopy and Chemometrics: A Biopharmaceutical Example.

9.1 Introduction.

We have previously demonstrated the ability for FT-IR spectroscopy coupled with multivariate analysis to discriminate successfully between pure and contaminated proteins in two models systems. Although this approach is not as sensitive as the currently favoured immunoassay based approaches, it is much less laborious and less prone to human error; making it an ideal high-throughput screening method, with the potential for automated 'at-line' use. Following on from the work described in Chapter 8, we have investigated the detection of contamination in a more biopharmaceutically relevant system; Immunoglobulin G (IgG) spiked with transferrin. In addition, in this study we have explored a much wider range of contaminant concentrations: 0.25-60%.

IgG was chosen as a mock protein product due to the fact that over 60% of the products produced by biotechnology are antibodies or antibody derived (e.g. antibody fragments) (Redwan, 2007). Transferrin was chosen to use as a protein contaminant as it a commonly encountered HCP in the mammalian cell lines which are used to produce IgGs.

In this work we once more demonstrate the ability of FT-IR spectroscopy combined with MVA techniques to detect low level contamination of proteins, in this case we have successfully detected contamination levels as low as 0.25%.

9.2 Materials and Methods.

9.2.1 Materials.

Polyclonal human IgG and human apotransferrin were both purchased from Sigma-Aldrich (Dorset, UK). Phosphate buffered saline (PBS) tablets for the light scattering studies were also purchased from Sigma-Aldrich.

9.2.2 Method.

Lyophilised IgG and transferrin were dissolved in ultra-pure water at a concentration of 5mg/mL. The spiked samples were then created in triplicate by mixing the appropriate amount of transferrin into IgG (0.25, 0.5, 0.75, 1, 2, 3, …10, 20 … 60% v/v). 5 µL of each sample was loaded into a 96 well silicon plate, which had been pre-washed in methanol and deionised water. The plate was then left to air dry at room temperature for 60 min, until completely dry. To minimise instrumental drift being incorporated into the measurements samples were positioned in alternate rows on each plate. 52 analytical replicates of each sample were analysed with replicates coming from 3 independent mixtures and spread over multiple plates and measurement days.

9.2.3 FT-IR Spectroscopy.

FT-IR spectra were collected on a Bruker FT-IR instrument, as described in 2.1.2. Spectra were collected from each of the 96 wells over a wavenumber range of 4000-600 cm^{-1}, with a spectral resolution of 4 cm^{-1}. For each well 64 accumulations were collected, co-added and averaged to improve the signal-to-noise ratio. Data were exported from the Opus software to Excel spread sheets, which were later analysed in R.

9.2.4 Raman Spectroscopy.

Raman data were collected using a Renishaw 2000 Raman microscope described in Chapter 2. All spectra were all were single accumulation, extended scans between 400 and 1800 cm^{-1}, with an exposure time of 60 s. 2 µL of sample were spotted onto a hydrophobic SpectraRIM™ slides (section 2.1.1.1) and allowed to air dry out at room

temperature for ~1 h. Each reported spectrum is an average of 6 spectra collected from different positions within each sample spot. Data were pre-processed using an ALS baseline collection and EMSC normalisation (Polynomial order, 9).

9.25 Light Scattering.

Light scattering experiments were performed by employing composition-gradient multi-angle light scattering (CG-MALS). A Wyatt Calypso II (Wyatt Technology Corp., Santa Barbara, CA, USA) was used for sample injection, and a Wyatt miniDAWN TREOS was used as a light scattering detector, in which detection angles were; 49°, 90° and 131°. The setup also included a Waters 2487 dual wavelength UV detector operating at 280 and 254 nm (Waters Corp., MA, USA). Calypso 1.2.9.1 software was used for instrument control and data capture, data were then exported into Excel for further analysis. The method used in this analysis is depicted in supplementary information, Figure S9.1A.

9.2.6 Data Analysis.

Prior to multivariate analysis FT-IR data were pre-processed in R. Spectra were normalised using EMSC (polynomial order 3), and auto-scaled so that each data column, corresponding to a wavenumber variable, had a mean equal to zero and a standard deviation equal to one. Data were then analysed using R to perform PCA, DFA, PLSR, PLS-DA, SVMs and ANNs. Raman data were pre-processed in Matlab prior to analysis in R.

9.3 Results and Discussion.

9.3.1 FT-IR Spectra.

FT-IR spectra were recorded of pure IgG, pure transferrin (Tf) and samples of IgG spiked with varying amounts of Tf. Spectra of both pure proteins and IgG with 1% and 5% Tf contamination are shown in Figure 9.1, once again the need for further analysis is highlighted here as little or no variations between the spectra of pure and spiked IgG can

be observed by eye, and only very small differences are visible in the spectra of two pure proteins.

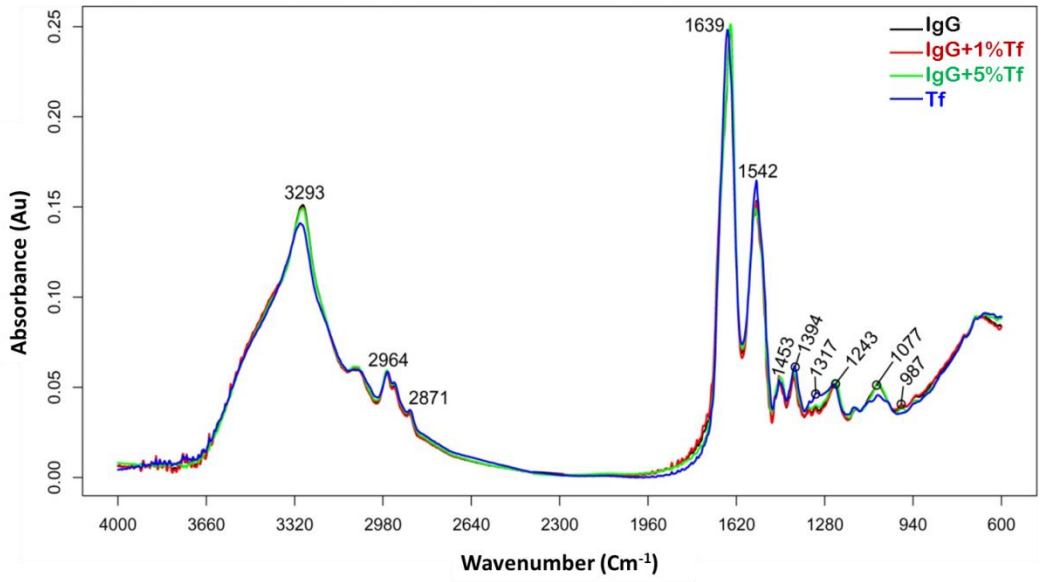

Figure 9.1: Average FT-IR spectra of IgG, transferrin (Tf), and IgG contaminated with 1 and 5% Tf.

Traditionally, the most effective way to differentiate IR spectra recorded from various proteins, without the application of MVA, is to perform a spectral de-convolution on the amide I region (~1600-1700 cm^{-1}) (Byler and Susi, 1986, Jackson and Mantsch, 1995, Sane et al., 1999, Ganim et al., 2008). This will allow us to view the component peaks underneath this band, which can then be assigned to structural features in order to discriminate between proteins based on their secondary structure. Peak fitting and de-convolution has been performed in GRAMS Ai software using the inbuilt Gaussian and Lorentzian peak fitting function.

Spectral differences in the proteins understudy in this chapter should be easily visible in the amide I region, as IgG is predominantly formed of β-sheet structures (39% as opposed to 18% in Tf), with only 5% of its structural features being classified as α-helical, compared to 31% α-helix in Tf (Wikstrom et al., 1994, Novotny et al., 1986, Zikan et al., 1985, Bailey et al., 1988, Matsuo et al., 2005). The results of this analysis on the spectra of IgG and transferrin are shown in Figure 9.2A-B. It is clear from the de-convolved spectra of this region that the spectra indicate very different structural features for the two

proteins. The spectra of IgG show large bands arising from β-sheet structures at ~1624 and 1685 cm^{-1} (Liang et al., 2006, Chen and Lord, 1980, Zheng et al., 2004). In comparison the transferrin spectra show numerous bands which can be assigned to α-helix structures (at ~1629, ~1630, ~1648 and ~1652 cm^{-1} (Takekiyo et al., 2006, Barron et al., 2002, Ellepola et al., 2006), but only very small contributions from β-sheet features. In addition the Tf spectra also exhibit a large band at ~1678 cm^{-1} which can be attributed to the turn structures in the protein (Takekiyo et al., 2006).

We next investigated this approach for detecting contaminated IgG; the de-convolved amide I regions for the samples of IgG spiked with 1% and 30% Tf are displayed in Figure 9.2C-D. In both cases we can observe an increase in the intensity of bands assigned to α-helix compared to the spectra of pure IgG, this is associated with the increase in α-helix content in Tf. Furthermore there is also an increase in bands assigned to turn structure in spectra of contaminated IgG, which corresponds to the large band attributed to turn structure seen in the spectra of Tf.

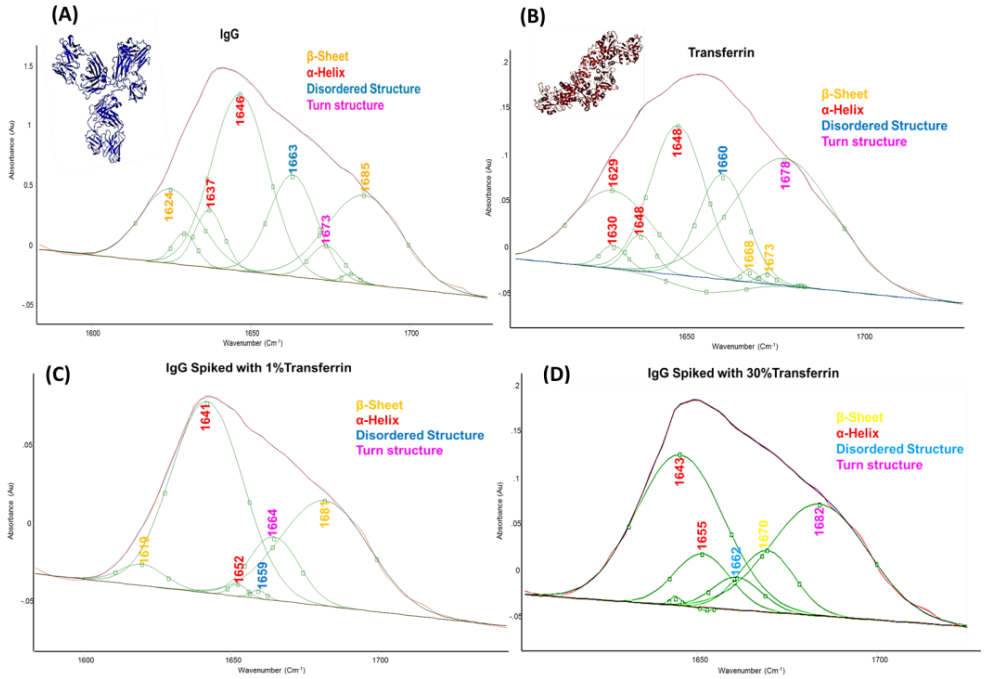

Figure 9.2: Spectral de-convolution of the amide I region of the FT-IR spectrum of **(A)**Pure IgG (with cartoon structure inset, drawn from PDB file 1HGY), **(B)** Pure Tf (with cartoon structure inset, drawn from PDB file 2HAU), **(C)** IgG spiked with 1% Tf and **(D)** IgG spiked with 30% Tf.

Although this method appears to allow us to distinguish between pure and spiked samples it is a time consuming approach which requires a certain level of expertise in order to make correct assignments. It also could be argued that this method is a subjective approach as results may vary depending on the proposed assignments which could differ between analysts. Therefore there still remains a need for an objective chemometrics based method to facilitate rapid and reliable detection of contamination.

9.3.2 Unsupervised Clustering - PCA.

We next applied PCA to this data set and scores plots for the analysis of 1% and 5% contamination levels are shown in Figure 9.3. The results show that although we are able readily to separate the two pure protein samples there is a significant overlap of spiked and pure IgG samples. As PCA alone shows little discrimination between contaminated and non-contaminated proteins supervised machine learning is needed to aid in identification of the contaminated samples.

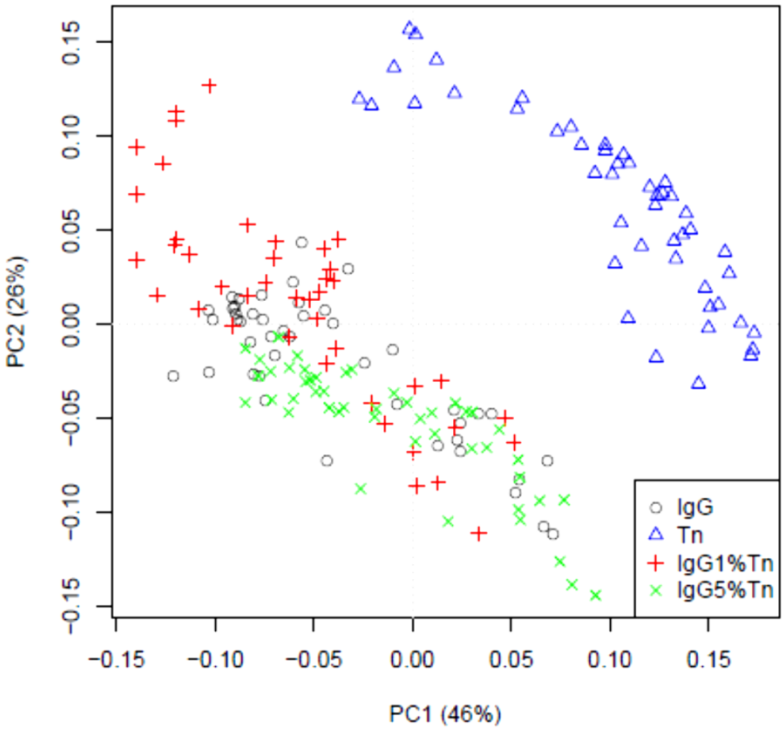

Figure 9.3: PCA scores plots (PC1 *vs.* PC2) of pure IgG, pure Tf, and IgG contaminated with 1% and 5% Tf.

9.3.3 Supervised Clustering.

9.3.3.1 PC-DFA.

DFA was first applied to the data to improve discrimination between pure and spiked samples and also to attempt to quantify how much contaminant is in a sample. We have used the PCs calculated in the previous analysis as input variables for this model (200 PCs). In our preliminary tests to determine the number of PCs that should be retained, 200 was the number of PCs (input variables to DFA) that maximized the predictive accuracy of the DFA models. One possible explanation for the high number of PCs required for DFA to discriminate so many different levels of contamination, ranging from 0 to 60%, is that the first PCs do discriminate well between very low and very high levels of contamination (between 0.5% and 60% for instance) but many more PCs are needed to clearly separate the lowest contaminated samples (0.5% and 0.75% for instance).

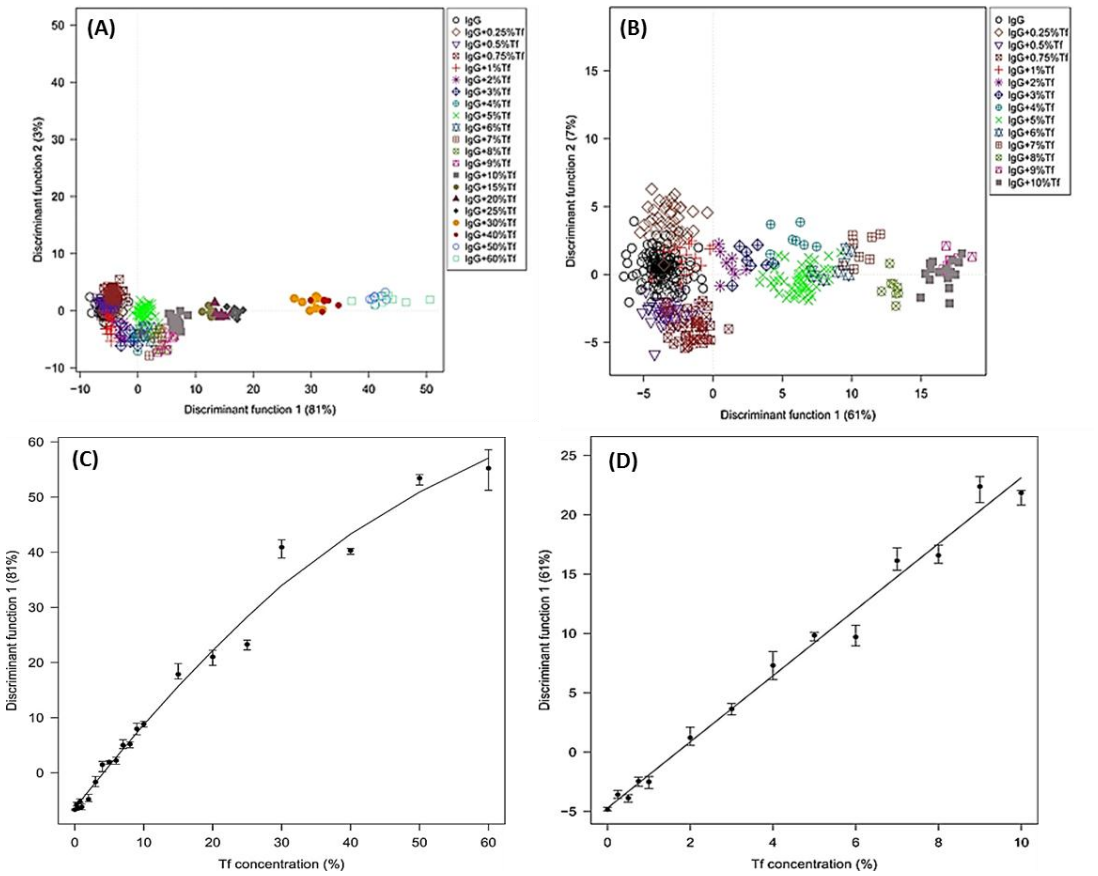

Figure 9.4: (A) PC-DFA scores from the full dataset, **(B)** PC-DFA scores from IgG samples spiked with 0-10% Tf **(C)** DF 1 Scores from A plotted as a function of Tf concentration and **(D)** DF 1 Scores from B plotted as a function of Tf concentration.

Figure 9.4A shows the resulting DFA scores plot for all IgG samples. A clear trend can be observed across DF1 with increasing Tf concentration; however there is still a large amount of overlap between pure IgG and the samples with lower levels of contamination (0.25-3%). As an alternative way to display this quantitative trend we have plotted a graph with the scores from DF1 as a function of Tf concentration (Figure 9.4C), where a strong correlation between the FT-IR spectra and contamination levels is evident. If we calculate DFA scores using only PCs from samples contaminated with between 0 and 10% Tf, we can see a marked improvement in the resolution on clusters from lower Tf concentrations (Figure 9.4B and D). When further reducing the number of samples used to build the DFA model to only those spiked with 0-1% Tf (Figure 9.5), we can easily discriminate between pure and contaminated IgG samples based on their IR spectra, even at much lower contaminant concentrations than those investigated in the previous chapter.

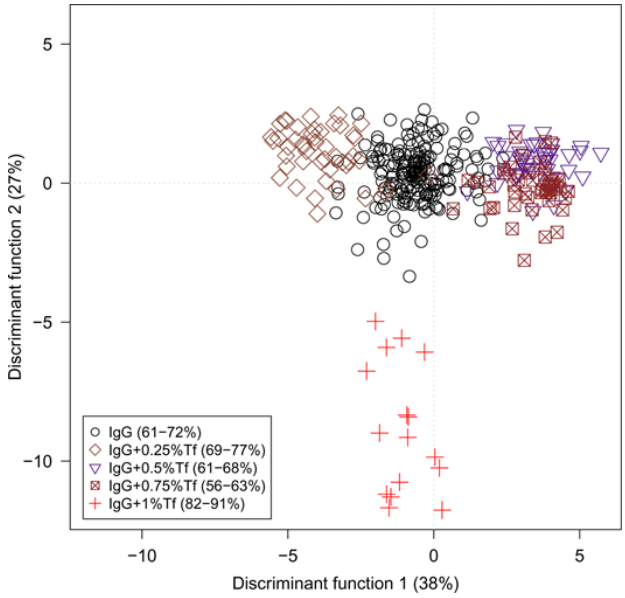

Figure 9.5: PC-DFA scores from pure IgG and IgG spiked with 0.25-1% Tf.

9.3.3.2 PLSR.

Next a PLSR model was applied to attempt to improve quantification of contamination. Figures 9.6A-B show the distribution of PLSR predictions over 1000 bootstrap cross validations for the discrimination of pure IgG and samples spiked with 5 and 1% Tf, respectively. The null distribution is also shown in Figure 9.6C, as seen previously for cyt *c* and RNase these curves lie directly on top of each other. The distribution curves for

spiked samples indicate an excellent discrimination of pure and 5% contaminated samples, with only a small amount of overlap seen between the two groups. The 1% Tf samples also show good discrimination results with the model being able to correctly classify spiked IgG in a majority of samples. Furthermore, if we compare these curves to those produced from the cyt *c* and RNase samples in the previous chapter (Figure 8.5), it can be seen that discrimination by PLSR in this biopharmaceutical example appears to be more successful than the results achieved with our model systems (Chapter 8).

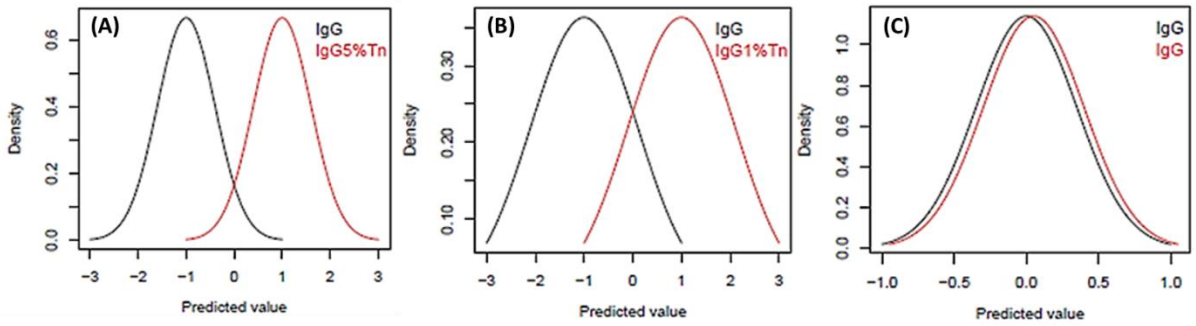

Figure 9.6: Graphs to show **(A)** the distribution of predicted values for IgG and IgG spiked with 5% Tf, **(B)** the distribution of predicted values for IgG and IgG spiked with 1% Tf and **(C)** the null distribution of predicted values for IgG and IgG.

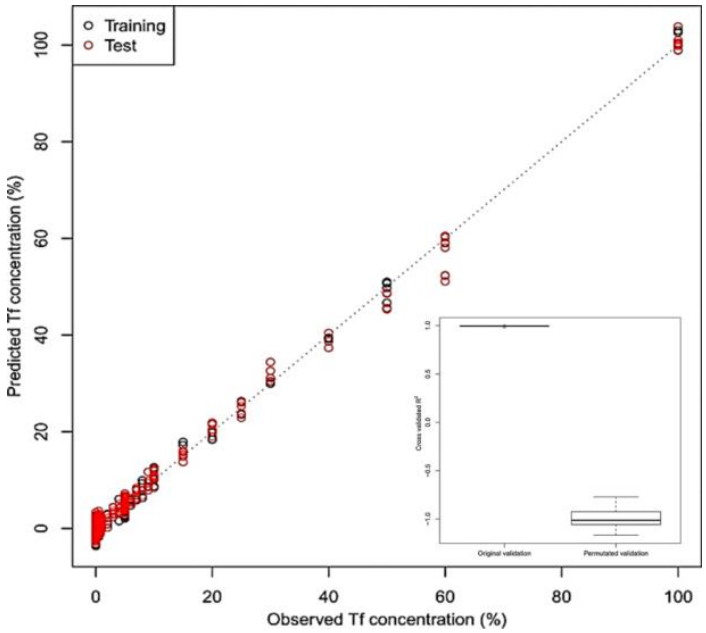

Figure 9.7: Typical PLSR predictions from FT-IR data of IgG spiked with Tf over 1000 bootstrap cross validations.
INSET: Box and whisker plot showing R^2 values for the original and permuted models.

A typical model showing the PLSR predictions from models built using all of the IgG samples, with Tf concentrations between 0 and 60% and also pure Tf (100%) is shown in Figure 9.7. It is clear from this plot that FT-IR analysis of this particular system provides an excellent method of quantifying contamination. There is a very strong correlation between the FT-IR spectra and Tf concentration, with the majority of models having R^2 values close to 1 (the R^2 values for the 1000 models are shown in the box and whisker plot inset in Figure 9.7). The box and whisker plot also shows the results obtained from permuted models, which confirm that the original model is a true result.

9.3.4 Supervised Classification Methods: PLS-DA, SVM, ANN and RF.

As with the previous work, we next compare the three different supervised algorithms discussed in the last chapter, PLS-DA, SVM and ANN, in order to find the best classification model for our FT-IR data and improve the precision of spectroscopy and MVA based detection of contamination. We have, in the present work, also investigated a fourth classification method; random forests (RFs). RF is an alternative supervised learning algorithm which uses many decision trees (i.e. a forest), where the input for each decision tree is a bootstrapped version of the original training data, and each node of the trees has a degree of randomness incorporated (Breiman, 2001).

Data for this analysis was partitioned into many subsets, allowing discrimination between pure IgG and contaminated IgG at each level of contamination. Results were calculated over 1000 bootstrap cross validations for each method and are summarised in Table 9.1. The values quoted in this table refer to the precision of the method at each concentration, which have been calculated from the classification results using the method described in Chapter 8 (8.3.4.1).

All four of the methods tested performed extremely well; however on average PLS-DA was the most successful with the highest average precision over all contaminant concentrations. PLS-DA showed excellent discrimination between pure and spiked

samples, demonstrated by a precision value of 1.00 for the majority of higher Tf concentrations and 0.97 for the lowest Tf concentration (0.25%).

Table 9.1: Precision (probability that the classification is correct) of ANN, PLS-DA, SVM and RF applied to discriminate between pure and contaminated protein samples.

	Null distribution							
	pure IgG	pure IgG	pure IgG	pure IgG	pure IgG	pure IgG	pure IgG	pure IgG
0	0.50	0.50	0.49	0.49	0.50	0.50	0.50	0.50
	ANN		PLSDA		SVM		RF	
% Tf added	pure IgG	IgG+Tf	pure IgG	IgG+Tf	pure IgG	IgG+Tf	pure IgG	IgG+Tf
0.25	1.00	1.00	0.96	1.00	0.97	0.99	1.00	1.00
0.50	0.98	0.97	1.00	1.00	0.99	1.00	0.97	1.00
0.75	0.98	0.97	0.96	0.98	0.93	0.95	0.97	0.92
1	0.85	0.93	0.92	0.97	0.89	0.96	0.81	0.84
2	0.90	0.94	0.96	0.97	0.93	0.96	0.89	0.85
3	0.91	0.96	0.96	0.99	0.94	0.97	0.96	0.90
4	0.99	1.00	1.00	1.00	1.00	1.00	0.98	0.97
5	0.74	0.78	0.85	0.87	0.73	0.74	0.72	0.74
6	0.96	1.00	0.96	1.00	0.96	0.99	0.98	0.95
7	0.98	0.99	0.98	0.99	0.97	0.99	0.97	0.95
8	0.95	0.98	0.98	1.00	0.96	0.98	0.96	0.95
9	0.97	0.98	0.98	1.00	0.97	0.98	0.96	0.95
10	0.97	1.00	0.99	1.00	0.96	0.98	0.93	0.94
15	0.97	0.99	0.98	1.00	0.97	0.98	0.93	0.98
20	0.97	1.00	0.97	1.00	0.98	0.98	0.92	0.98
25	0.98	0.99	0.99	1.00	0.99	0.99	0.95	0.98
30	0.95	1.00	0.98	1.00	0.97	1.00	0.96	0.96
40	0.98	1.00	1.00	1.00	0.96	1.00	0.93	0.97
50	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.97
60	0.96	1.00	1.00	1.00	1.00	1.00	0.94	0.98
100	1.00	1.00	1.00	1.00	0.98	1.00	0.94	0.93
Average	0.95	0.98	0.97	0.99	0.96	0.97	0.93	0.94

It is difficult to precisely compare computation time between the classification algorithms used in this work as they require different parameters to optimize performance; such as the number of input variables, the number of latent variables computed and the network structure for ANN. However, an estimation of the elapsed time that each algorithm takes to perform 100 bootstrap cross-validations using exactly 10 input variables is shown in Figure 9.8. The results are averaged over 100 independent runs (100 bootstrap cross-

validations repeated 100 times). As it can be seen from Figure 9.8 PLS-DA also shows the fastest computational time with ANN and RF being the most time consuming approaches.

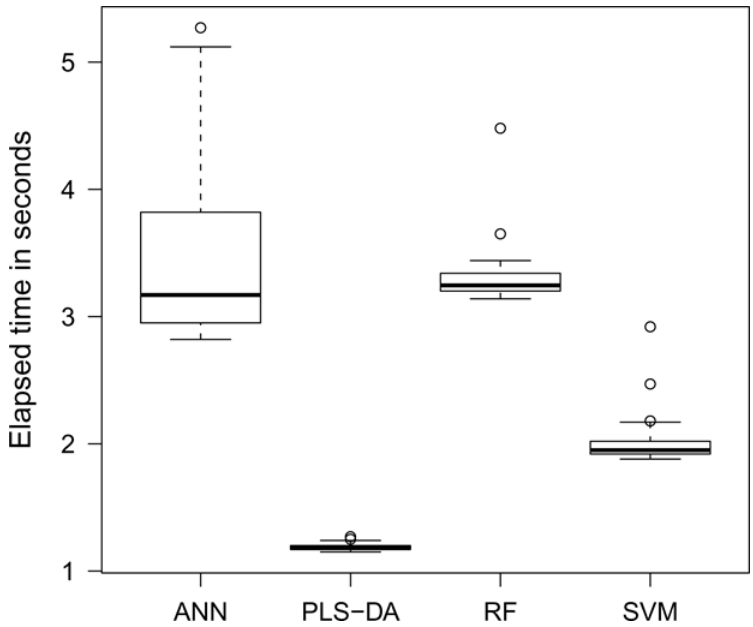

Figure 9.8: Estimation of the elapsed time that each algorithm takes to perform 100 bootstrap cross-validations using exactly 10 input variables. The results are averaged over 100 independent runs (100 bootstrap cross-validations repeated 100 times).

The loadings plot for the classification of pure IgG and IgG spiked with 1% Tf, displayed in Figure 9.9, show that a mixture of bands from secondary structure features and amino acid side chains are responsible for the discrimination. Bands at ~1636, ~1300 and ~1105 cm^{-1} can all be assigned to α-helix structure (Takekiyo et al., 2006, Prevelige et al., 1993, Barron et al., 2002), which could be attributed to the increase an in α-helix content in the sample when Tf is added, as IgG is only 5% α-helix but Tf is 31%. In addition the band at ~1456 cm^{-1} could be attributed to turn structure (Tuma, 2005) which correlates to the increase in turn structure indicated by the amide I region of the Tf spectrum displayed in Figure 9.2 B. There are also a number of bands (~1589, ~1558, and ~1032 cm^{-1}) which have been assigned to the ring breathing and ring deformation modes from the side chains of aromatic amino acids (Prevelige et al., 1993, Howell and LiChan, 1996, Lord and Yu, 1970). As these bands will vary in intensity depending on the burial or exposure of residues, this difference may be due to the tyrosine, tryptophan and phenylalanine residues being more exposed in one of these proteins.

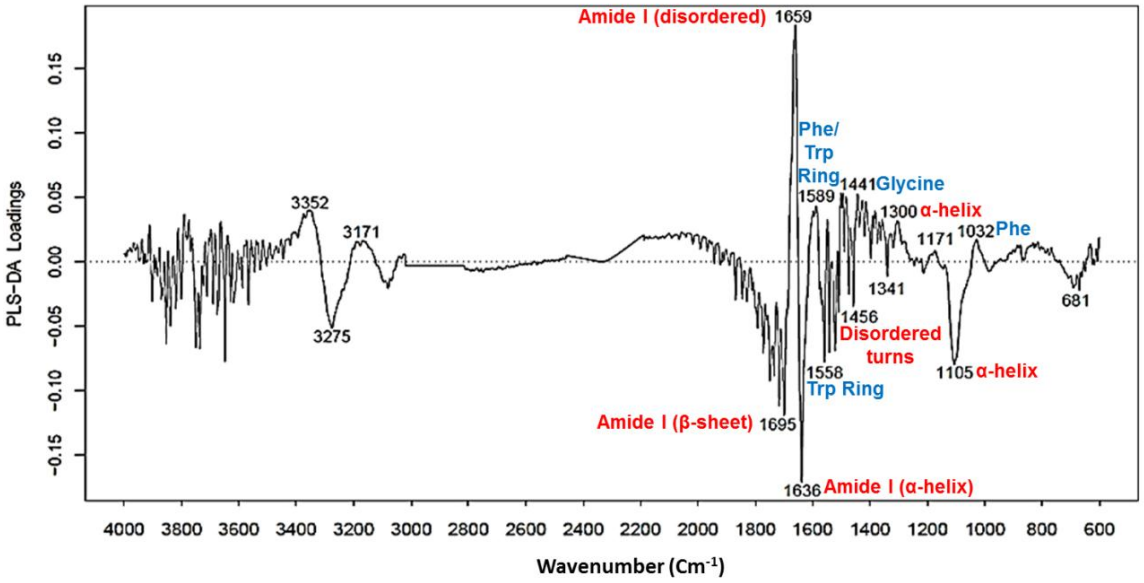

Figure 9.9: PLS-DA loadings plot for the discrimination between pure IgG and IgG + 1% Tf, with assignments relating to secondary structure given in red font and assignments for amino acid in blue font.

An important point to note from the PLS-DA results is that the discrimination between pure and 5% contaminated IgG yields strangely low results compared to all other contamination levels; this was usual and was not due to experimental error as this was repeatable (x2; data not shown). The precision for 5% contamination was calculated to be 0.73, in contrast with 0.96 for 6% Tf and 1.00 for 4% Tf. This result is consistent over the four different MVA methods tested. If we inspect the PLS-DA loading for 4,5 and 6% contamination (Figure 9.10), we can see that the 5% sample is indeed the odd one out, as the 4% and 6% loadings are not only similar to each other, but also correlate well with the previously reported loadings for IgG spiked with 1% Tf. The major bands appearing in the 5% Tf sample can be assigned to disordered turns (~1456 cm^{-1}) and β-turns (~1358 cm^{-1}) (Tuma, 2005, Barron et al., 2002, McColl et al., 2003), although the reason for these changes occurring in only the 5% samples is unclear and needs to be investigated further.

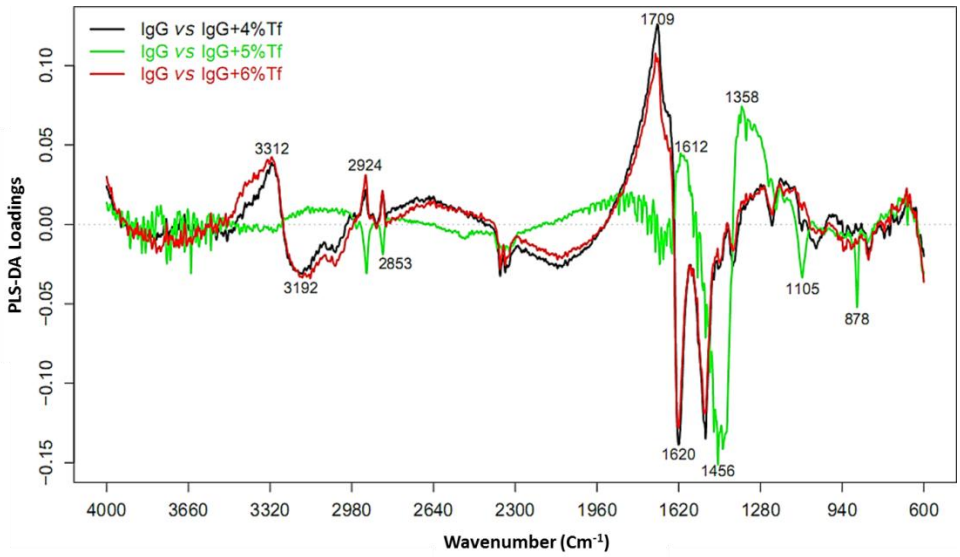

Figure 9.10: Comparison of PLS-DA loadings for the discrimination between pure IgG and IgG spiked with 4, 5 and 6% Tf.

9.3.5 Investigating IgG spiked with 5% transferrin.

Having previously detected an anomalous result for the PLS-DA analysis of IgG samples contaminated with 5% Tf, we investigated this occurrence further. If we look back to the PC-DFA results from 9.3.3.1 and re-plot the DFA scores showing just 1 and 5% Tf concentrations (Figure 9.11) we see that the 1% and 5% samples separate in opposite directions across DF2; with 1% falling below the pure IgG with negative scores and 5% sample having positive scores.

Our initial thoughts were that this change in the FT-IR spectra of IgG contaminated with 5% Tf could be due to protein-protein interactions between IgG and Tf. Although there is no known reason why this would happen only at this particular contaminant concentration, due to the fact that we know transferrin to be used as a fusion protein we investigated this theory using a light scattering experiment to detect any complex formation that may be occurring between IgG and Tf. The results of this analysis are shown in supplementary information (Figure S9.1B). From the light scattering data it was possible to calculate the weight average molecular weight of each individual species and the two proteins together. As the weight average molecular weight of IgG and Tf mixed together was much lower than the combined values calculated from the injections of the individual proteins, this suggests that no complex was being formed between IgG and Tf.

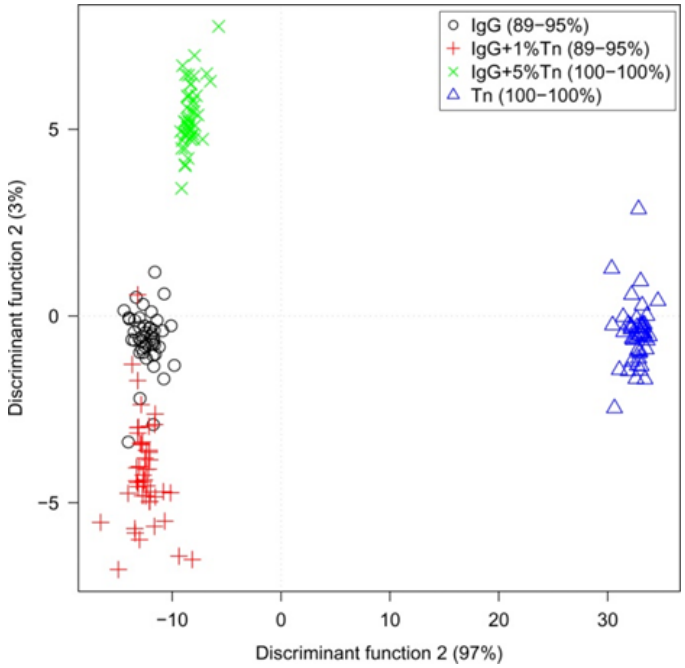

Figure 9.11: PC-DFA scores from FT-IR data of IgG, Tf and IgG spiked with 1 and 5% Tf.

We further investigated the spectra of IgG spiked with 5% Tf by applying a peak fitting function to the amide I region and comparing this to the spectra displayed in Figure 9.2. The de-convolved amide I region of IgG with 5% Tf (Figure 9.12) looks comparable to the amide I band of IgG spiked with 1% Tf, with the exception of the additional band seen at ~1700 cm^{-1}. The only assignment found in the literature for protein features in this wavenumber region was for protonated amide I bands (Ashton et al., 2007, Chi and Asher, 1998). A possible reason for this band, which was also present in the spectra of IgG with 6, 7, 10, 15 and 30% Tf (not shown), is that the sialic acid present in the glycan of IgG is effecting the environment of the proteins and hence has an effect on the IR spectrum of Tf. In addition, we have confirmed that the presence of acid groups affects the spectrum of Tf in this region by adjusting the pH of a solution of Tf; in this experiment a shoulder can be seen appearing on the amide I band at ~1720 cm^{-1} when at low pH (Figure S9.2). Although this is a potential hypothesis as to why transferrin spectra may alter when IgG is added, as this band was also observed in the spectra of IgG at higher Tf concentrations, it still does not provide an explanation for the anomalies which occur solely in the 5% samples.

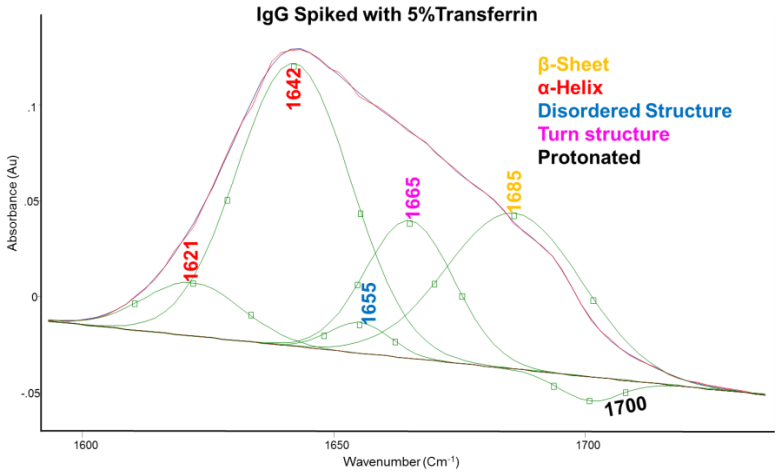

Figure 9.12: Spectral deconvolution of the amide I region of the FT-IR spectrum of IgG spiked with 5% Tf.

In an attempt to explain this trend we repeated the experiment using Raman spectroscopy with 785 nm excitation. As Raman spectra of proteins tend to be more information rich than the IR spectra it was hoped that this analysis may reveal structural bands which highlight possible conformational changes occurring in the 5% sample. The DFA scores for this Raman analysis are shown in Figure 9.13 (Raman spectra can be viewed in supplementary information, Figure S9.3), where we can observe that when using this complementary technique we see a much more expected trend in the scores plot, with separation across DF2 with increasing Tf concentration.

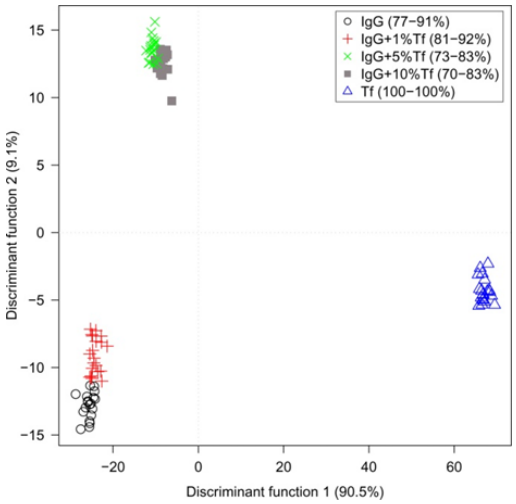

Figure 9.13: PC-DFA scores from Raman data of IgG, Tf and IgG spiked with 1 and 5% Tf.

All of the results from the exploration of 5% samples seem to suggest that the strange PLS-DA and PC-DFA results for the samples of IgG spiked with 5% Tf are likely to be caused by analytical artefacts, such as noise, instrumental drift or baseline differences, rather than structural changes in the proteins under study. However this is equally unlikely as measurements were collected across numerous 96 well plates, on different days, with replicates coming from three independent mixtures of IgG and Tf. Therefore this anomaly in the IR data must unfortunately be left unexplained for the present work.

9.4 Conclusions.

We have successfully identified low levels of Tf contamination in IgG samples using supervised MVA methods applied to FT-IR data. FT-IR data showed little variation between samples, although we were able to detect contaminated IgG samples by de-convolving the amide I region of the spectra. However, as this method is lengthy and subjective, a MVA method is still need to facilitate fast and precise discrimination.

Both PC-DFA and PLSR show an excellent ability to discriminate between pure and contaminated IgG samples at the lowest concentration of Tf tested, 0.25%. In addition both methods clearly display effective quantification of contaminant concentration. Classification by PLS-DA showed vastly improved results for this system compared to those analysed in the previous chapter, with precision being close to 1.00 for the majority of higher Tf concentrations and 0.97 for the 0.25% Tf samples. The loadings for this discrimination are largely due to variations in bands assigned to secondary structure.

The PC-DFA and PLS-DA results both indicate an anomalous result for the sample of IgG contaminated with 5% Tf. Further investigations carried out by FT-IR, Raman and CG-MALS analysis did not yield any plausible explanations for this unusual trend.

Results shown here have provided further evidence that FT-IR spectroscopy coupled to chemometric analysis, in particular PLS-DA, is a suitable method for the rapid and reliable detection of protein contamination in biopharmaceutical products. The positive

results presented in this chapter could lead to further work in which multiple proteins are spiked into a monoclonal IgG sample in order to more accurately simulate a protein therapeutic contaminated with HCPs, and ultimately analysis of a real biopharmaceutical samples.

9.5 Supplementary Information.

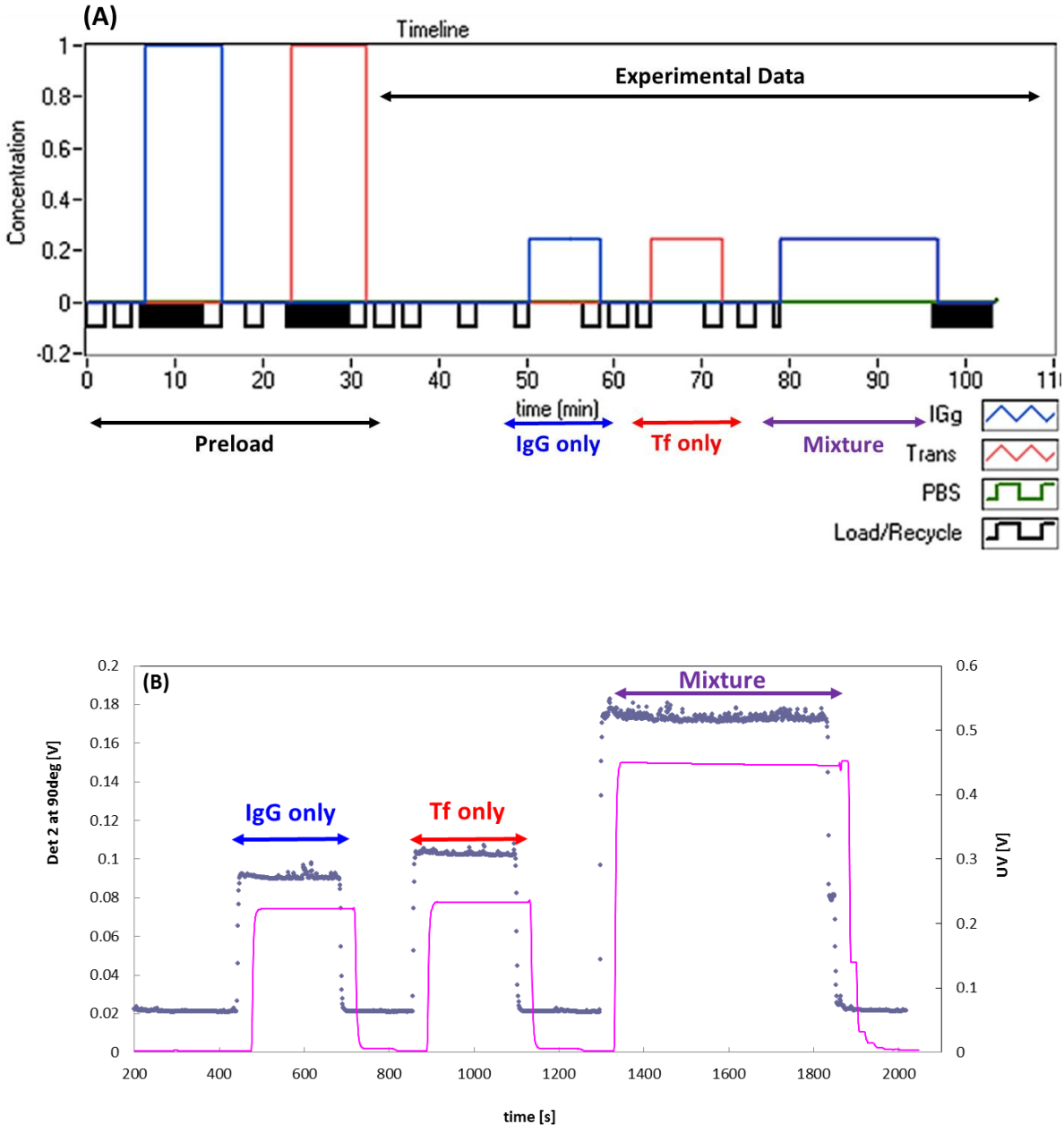

Figure S 9.1: (A) Diagram summarising the method used for CG-MALS experiment and **(B)** A graph to show the results of CG-MSLS experiment, where the Pink line indicated UV absorbance results and the blue points indicate MALS results.

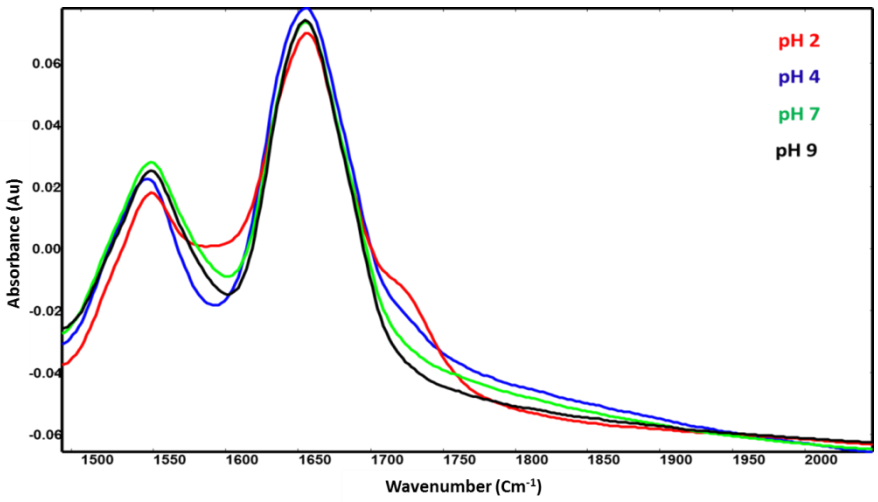

Figure S 9.2: Comparison of the amide I region of the FT-IR spectra of Tf at pH 2, 4, 7 and 9.

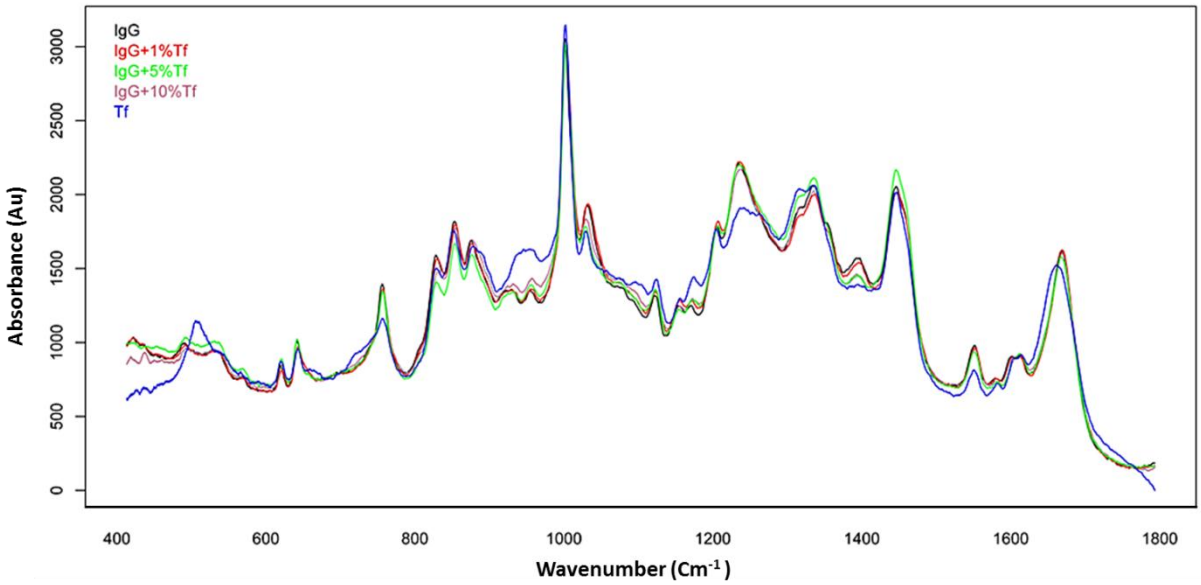

Figure S 9.3: Average Raman spectra of IgG, Tf, and IgG spiked with 1, 5 and 10% Tf (Data have been baseline corrected (ALS) and normalised (EMSC, polynomial=9)).

Chapter 10: Conclusions

It is estimated that by the year 2016, 8 out of the top 10 selling pharmaceuticals will be recombinant protein or antibody products (Redwan, 2007). With this growth in biotechnology-derived therapeutics comes an increasing need to develop reliable and high throughput process analytical technologies (PATs) for the robust characterisation of such products. Many of the current methods employed for this purpose are off-line destructive techniques (Greer, 2008). Raman spectroscopy provides an attractive alternative. The non-invasive, confocal nature of this technique coupled with the ability to obtain both quantitative and qualitative data, makes Raman spectroscopy a unique and valuable candidate to be developed for the *in situ* analysis of bioprocesses and biopharmaceuticals. The data which arise from these biopharmaceutical systems are often complex and difficult to interpret. The use of chemometrics and machine learning to convert these large data sets into meaningful information about the system under study is key to the success of vibrational spectroscopy in this arena.

10.1 Monitoring the Glycosylation Status of Proteins.

Glycosylation is the most common PTM, with over half of human gene products and one third of biological therapeutics being glycoproteins (Apweiler et al., 1999, Greer, 2007). As it is widely known that the glycosylation status of a protein drug is important for both biological activity and clinical efficiency, this modification must be fully characterised (Greer, 2007, Berman, 1985). The complexity of this characterisation challenge is greatly increased by the fact that the majority of glycoproteins are expressed in a variety of glycoforms. Consequently, complete characterisation of a glycosylated product involves not only detecting if a protein is glycosylated, but also that the correct glycan has been attached to the correct glycosylation site, and that the level of glycosylation is within acceptable limits (ICH, 1999).

The current gold standard techniques for glycoprotein analysis are chromatography and mass spectrometry based methods (Greer, 2008, Morris, 1980, Anumula, 2006). Although these techniques provide precise and reliable results, a Raman spectroscopy based method would hold the advantages of being non-destructive, involving minimal sample preparation and having the potential for at-line or on-line analysis. Despite the fact that Raman spectroscopy has a rich history of successful use in protein analysis (Tuma, 2005, Wen, 2007), it is relatively under-utilised in the study of PTMs, particularly glycosylation. Previous work carried out by Raman (and ROA) spectroscopy on glycans and glycoproteins has involved structural interpretations of proteins and carbohydrates, rather than the discrimination of glycosylated and non-glycosylated forms of a protein. Such studies include; characterising and quantifying sugar moieties (Oleinikov et al., 1998, Arboleda and Loppnow, 2000, Mrozek et al., 2004) and investigations into the interactions of glycoproteins (Fleury et al., 1999, Cui et al., 2005).

Through the work presented in this thesis we have undoubtedly demonstrated the potential for Raman spectroscopy to be developed as a tool for detecting and quantifying protein glycosylation. For the first time we have shown Raman spectroscopy to be capable of distinguishing between a glycoprotein and its non-glycosylated equivalent, and also quantifying relative concentrations of protein and glycoprotein. This was initially demonstrated in a simple model system of RNase A and B, then subsequently in a real biopharmaceutical sample (transferrin), as well as in an artificially glycosylated system (GFP).

Initial studies carried out on RNase A and its glycosylated equivalent RNase B showed that through the use of Raman spectroscopy combined with PCA the protein and the glycoprotein can be readily identified. In addition Raman spectra of chemically and enzymatically deglycosylated RNase B were found to be similar to that of RNase A, confirming that the previous discrimination between RNase A and B was indeed due to the addition of a carbohydrate moiety. Furthermore, by applying PLSR analysis to data that were acquired from mixtures of RNase A and B, we have clearly established the

potential of Raman spectroscopy to be used in predicting levels of glycosylation. Since the publication of this work (Brewster et al., 2011), other studies have proved to be successful in the detection and quantification of glycosylation in haemoglobin and albumin by Raman spectroscopy combined with PCA and PLSR (Barman et al., 2012, Dingari et al., 2012).

Following on from this, it was proven that the methods developed in Chapter 3 were transferable to more complex systems. Identification and quantification of glycosylation in a recombinant biopharmaceutical sample was demonstrated using transferrin proteins; where in addition to the discrimination between glycosylated and non-glycosylated forms we were also able to distinguish between the holo- (iron containing) and apo- (without iron) forms of the protein. Successful detection of glycosylation in a synthetic system has also be displayed in Chapter 5, where GFP mutants with sugars attached through free cysteine residues were subjected to interrogation by Raman spectroscopy and PCA. Both studies on GFP and transferrin have extended the value of Raman spectroscopy in this application by successfully identifying different glycoforms of the proteins.

Through the use of PCA and PLSR we have identified the vibrational modes which are most selective in the detection of glycosylation. The Raman bands used for the discrimination and quantification of glycosylation consisted of a mixture of features attributed to both protein vibrations and bands arising from the glycan components. It was found that these vibrational modes were relatively consistent over the three models investigated. From this we are able to deduce some general rules for the identification of proteins and glycoproteins based on variations in their Raman spectra: All three of our systems, and also results subsequently published for haemoglobin and albumin, showed variations in both the amide I and amide III regions of the spectra, at ~1650 cm^{-1} and ~1250 cm^{-1} respectively, due to conformational differences between glycosylated and non-glycosylated proteins (Ellepola et al., 2006, Ashton et al., 2007, Tuma, 2005, Barman et al., 2012, Dingari et al., 2012). In addition we also saw changes in the bands which arise from the side chains of aromatic amino acids, specifically the ~830 cm^{-1} band in

RNase assigned the tyrosine (Siamwiza et al., 1975), the band at ~1550 cm^{-1} in GFP due to the indole ring mode of tryptophan (Howell and LiChan, 1996) and the ~650 cm^{-1} tyrolsyl ring vibrations in transferrin (Lord and Yu, 1970). Furthermore, in all three cases glycosylated proteins exhibited bands between ~1000 and ~1046 cm^{-1}, which could be assigned to the ring breathing modes of the glycosidic rings in the carbohydrate moiety (Arboleda and Loppnow, 2000, Oleinikov et al., 1998). Bands from glycosidic ring stretching were also observed at ~1259 cm^{-1} in the spectrum of RNase B (Socrates, 2001) and ~820 cm^{-1} in the spectrum of mannosylated transferrin (Degen, 1997) and ring deformation modes were seen in spectrum of glycosylated GFP at ~933 cm^{-1} (Dollish et al., 1974).

Results obtained in Chapters 4 and 5 have highlighted the importance of glycan position in the observation of vibrational modes which can be attributed to the sugars. Nevertheless, in both of these cases the protein and glycoprotein were still easily identified in the absence of glycosidic bands, by examining bands in the spectra which were attributed to conformational differences in the protein molecules brought about by the addition of a glycan.

In order to ensure the success of Raman spectroscopy in this application, the ability to distinguish between different glycoforms of proteins must be developed further. This task would be greatly assisted by building up a knowledge base of glycan standards and gaining a better understanding of how sensitive glycosidic vibrations are to variations in glycan structure. Some preliminary investigations into this area, presented in the appendix, have shown Raman data to be capable of distinguishing between various monosaccharaides, glycan fragments and whole glycans; of particular importance was the fact that glycans with the same sugar residues in different spatial arrangements are able to be classified using PCA. Chemometric analysis of this data set coupled with further investigations, which compared monosaccharaides to disaccharides and complex glycans, have highlighted the potential to use bands arising from glycosidic links in the discrimination of free sugars and glycans. Future work will focus on determining if there is

a linear trend between these bands and the number of glycosidic bonds present in a sample, hence allowing quantification of the number of sugar residues in a glycan.

In order for the work presented in this thesis to be transferable to an at-line or on-line method, another avenue of further investigation would be to determine if glycosylation can still be detected by Raman spectroscopy in a more complex sample which contains free sugars as well as those attached to the protein. This is an important step as at either end of the biopharmaceutical pipeline the samples encountered will include carbohydrate molecules: these may be in the form of feeds and metabolites in cell culture media (Butler and Meneses-Acosta, 2012), or the sugars used as stabilisers at the formulation stage (Hajare et al., 2011, Ohtake and Wang, 2011). In these more complex systems the identification of glycosylated proteins may be confused by the glycosidic vibrations from the free sugars. However, as we have previously shown that the discrimination of proteins and glycoproteins is partly based on changes in bands assigned to protein conformation and also that the spectrum of a glycan differs from that of is component monosaccharaides, we are hopeful that successful detection and quantification of glycosylation will still be possible.

10.2 Monitoring the Conformation and Stability of Proteins.

We have discussed previously how the Raman data collected from glycosylated proteins has highlighted conformational differences between a protein and equivalent glycoprotein. It was therefore of interest to investigate how these structural changes affect the stability of the glycoproteins studied. The stability of a biopharmaceutical product is an important consideration, as disordered or unfolded proteins may suffer from loss of activity, increased aggregation and decreased solubility (Goddard, 1991).

Experiments to determine the unfolding and aggregation profiles of transferrin and GFP were carried out by utilising a Optim 1000 spectrometer developed by Avacta analytical, which simultaneously measures light scattering and fluorescence emission as a function of temperature. Results from these experiments have clearly displayed that glycosylation

in transferrin proteins greatly reduces aggregation propensity, with an increase in the temperature at which aggregation begins of ~15-20 ºC with increasing number of glycans. Analysis of unfolding curves drawn from transferrin and GFP data has shown that in both cases glycosylation can increase the stability of the proteins.

As well as differences between proteins of different glycosylation states, Raman data from GFP samples also alluded to structural differences between the I229C and E6C mutants. The light scattering investigations into these samples suggested that these differences may be due to increased levels of aggregation in I229C GFP and this was confirmed by microscopy and FCS data.

The interpretation of fluorescence data collected on the Optim 1000 was assisted by the development of data analysis strategies for this type of spectroscopic data. 2D correlation analysis of fluorescence data from transferrin samples was able to resolve the broad overlapping transitions seen in the unfolding curves into multiple distinct transitions. In addition, comparison of moving windows contour plots from each transferrin species allowed the increase in the temperature at which transitions occur to be more readily observed. 2D correlation moving windows analysis of the data collected from GFP samples was also able to display multiple unfolding transitions in cases where conventional instrument manufacturer's analysis methods failed to describe any transitions. Furthermore using the synchronous and asynchronous contour plots drawn from GFP data we have been able to deconvolve the intrinsic fluorescence emission spectrum into two overlapping features, determine if these bands are changing in the same or opposite directions and sequence the changes with respects to the temperature at which they occur. Multivariate data analysis methods were also applied the GFP data: Both PCA and PARAFAC were successfully used to simplify a matrix of Optim data into a visual depiction of the relative similarities and differences between samples. We were also able to pinpoint the temperatures at which the most variations were occurring in the spectra, and these transition regions were in good agreement with T_m values published for other GFP mutants (Alkaabi et al., 2005, Cubitt et al., 1995).

Due to the obvious indications of conformational differences observed in the Raman data of proteins, mainly indicated by the amide I vibrations, we investigated the use of this region of the Raman spectrum to monitor protein unfolding induced by a chemical denaturant. We have compared this method to one of the gold standard methods, fluorescence spectroscopy (Serrano et al., 2012) and also to a previously established Raman method which focusses on observing changes in the tryptophan modes as the residues become more exposed to the external environment (Prevelige et al., 1993, Liang et al., 2006, Chen and Lord, 1980). Through the use of unfolding curves, $[D]_{50}$ and ΔG calculations we have shown that results obtained through this novel Raman spectroscopic method are comparable with those obtained by a conventional fluorescence unfolding experiment. By employing 2D correlation moving windows analysis to both data sets we have demonstrated that this Raman based approach is more sensitive to smaller conformational changes than both the fluorescence method and the tryptophan region of the Raman spectrum.

10.3 Detecting Foreign Protein Contamination in Proteins.

Another commonly encountered characterisation challenge in biopharmaceutical production is the contamination of a therapeutic protein product with any other proteins produced by the bioprocess. The removal of these HCPs and the validation of this step is an important stage of downstream processing, as these foreign proteins can have undesirable immunogenic effects (Greer, 2008, Goddard, 1991). Although the sensitivity of the current immunoassay based methods of detection is unrivalled, the FT-IR method described in this thesis is a high-throughput and precise alternative for the rapid screening of samples. We have successfully demonstrated that FT-IR spectroscopy combined with supervised machine learning is capable of detecting low-level protein contamination in three separate systems.

Initial model systems, mimicking a protein product contaminated with HCP (cytochrome *c* and lysozyme) and a protein contaminated with a glycosylated equivalent (RNase A and B), both highlighted the need for sophisticated supervised learning chemometrics by the

small amount of variance seen in the spectra of pure and spiked samples and poor discrimination by PCA, an unsupervised learning approach. Results from PLS-DA showed that discrimination of 5% contaminated samples was easily achieved for both cytochrome *c* and RNase samples. Examination of the loadings plots from these models showed that this identification was based on bands that could be assigned to secondary structure features or the side chains of amino acids which increase in quantity in the protein contaminant, and therefore also increase in the spiked samples.

This method was then tested using a more biopharmaceutically relevant example of IgG spiked with transferrin, where encouraging results were achieved. PLS-DA showed an excellent ability to discriminate between pure and spiked samples at a wider contaminant concentration range of 0.25-60%; precision values of ~1.00 were achieved at higher concentrations of contaminant, and ~0.97 for the lowest concentration of Tf tested (0.25%). In addition, both PC-DFA and PLSR achieved successful quantification of contamination with low prediction errors.

Due to the success of this method of detecting protein contamination using FT-IR and multivariate analysis, future work will involve numerous proteins being spiked into a monoclonal IgG sample in order to more accurately simulate a protein therapeutic contaminated with HCPs. Following on from this it is hoped that this method can be tested on real biopharmaceutical samples.

10.4 Detecting Sickle Cell Anaemia and the Sickle Cell Trait.

Finally, we have investigated the application of Raman spectroscopy to a medical diagnostic problem: the detection of sickle cell anaemia and the sickle cell trait. Many of the currently used diagnostic tests are unable to differentiate between the disease and the trait, thus leading to a high rate of false positives (Nalbandi et al., 1971, McCavit, 2012, Meier and Miller, 2012). Therefore we have attempted to use Raman spectroscopy to not only detect sickle cell anaemia by identifying the defective form of haemoglobin (HbS), but also differentiate between samples of pure HbS and mixtures of HbA and HbS,

which would be characteristic of a blood sample from a carrier of the sickle cell trait (Silverstein and Nunn, 1997). This study will build on previously published work in which Raman microscopy was used to monitor aggregation of HbA and HbS inside erythrocytes (Wood et al., 2005).

We have shown the potential for Raman Spectroscopy in this application by initially distinguishing between HbA and HbS proteins both in pure protein samples and mock biological samples. We went on to demonstrate the potential of Raman analysis to quantify levels of HbS in mock biological samples containing a mixture of HbS and HbA, displaying how Raman spectroscopic analysis can allow diagnosis of both sickle cell disease and the sickle cell trait. Moreover, we have shown the viability of this method as a point-of-care diagnostic tool, by demonstrating that a portable Raman probe was capable of detecting haemoglobin at physiological concentrations in mock biological samples. These successful preliminary experiments will lead to future work carried out on a portable Raman instrument, beginning with the mock biological samples previously analysed by Raman microscopy and leading on to the analysis of HbS in a complex biological sample which will more accurately simulate a patient's blood sample.

10.5 Concluding Remarks.

Throughout this thesis the utility of vibrational spectroscopy in the detection of protein modifications has been undoubtedly displayed. We have investigated applications of these techniques in both biopharmaceutical characterisation and medical diagnostics. These 'proof of principle' type studies have shown Raman spectroscopy to be capable of determining the glycosylation status of a protein, detecting mutant forms of proteins and monitoring conformational changes in proteins. In addition we have shown FT-IR spectroscopy coupled with MVA to be successful in the identification of foreign protein contamination. Moving forward, since the potential of these techniques has been so clearly demonstrated in these studies, we hope to transfer these methods to real on-line and at-line analyses.

References

Alkaabi, K. M., Yafea, A. & Ashraf, S. S. 2005. Effect of Ph on thermal- and chemical-induced denaturation of GFP. *Applied Biochemistry and Biotechnology,* **126,** 149-156.

Ambrose, E. J. & Elliot, A. 1950. Infrared spectrocopic studies of globular protein structure. *Procedings of the Royal Society B: Biological Sciences,* **208,** 75-90.

Anumula, K. R. 2006. High-sensitivity and high-resolution methods for glycoprotein analysis. *Analytical Biochemistry,* **350,** 1-23.

Apweiler, R., Hermjakob, H. & Sharon, N. 1999. On the frequency of protein glycosylation, as deduced from analysis of the swiss-prot database. *Biochimica Et Biophysica Acta-General Subjects,* **1473,** 4-8.

Arakawa, T., Philo, J. S., Ejima, D., Tsumoto, K. & Arisaka, F. 2007. Aggregation analysis of therapeutic proteins, part 2: Analytical ultracentrifugation and dynamic light scattering. *BioProcess International,* **5,** 36-47.

Arboleda, P., H. & Loppnow, G., R. 2000. Raman spectroscopy as a discovery tool in carbohydrate chemistry. *Analytical Chemistry,* **72,** 2093-2098.

Arnold, U. & Ulbrich-Hofmann, R. 2000. Differences in the denaturation behavior of ribonuclease A induced by temperature and guanidine hydrochloride. *Journal of Protein Chemistry,* **19,** 345-352.

Ashton, L., Barron, L. D., Hecht, L., Hyde, J. & Blanch, E. W. 2007. Two-dimensional Raman and Raman optical activity correlation analysis of the α-helix-to-disordered transition in poly-l-glutamic acid. *Analyst,* **132,** 468-479.

Ashton, L. & Blanch, E. W. 2008. Investigation of polypeptide conformational transitions with two-dimensional Raman optical activity correlation analysis, applying autocorrelation and moving window approaches. *Applied Spectroscopy,* **62,** 469-475.

Ashton, L., Johannessen, C. & Goodacre, R. 2011. The importance of protonation in the investigation of protein phosphorylation using Raman spectroscopy and Raman optical activity. *Analytical Chemistry,* **83,** 7978-7983.

Atkins, P. W. 1998 *Atkins physical chemistry sixth edition,* Oxford, Oxford University Press.

Aygun, B. & Odame, I. 2012. A global perspective on sickle cell disease. *Pediatric Blood & Cancer,* **59,** 386-390.

Bailey, S., Evans, R. W., Garratt, R. C., Gorinsky, B., Hasnain, S., Horsburgh, C., Jhoti, H., Lindley, P. F., Mydin, A., Sarra, R. & Watson, J. L. 1988. Molecular-structure of serum transferrin at 3.3-A resolution. *Biochemistry,* **27,** 5804-5812.

Baldwin, M. A. 2005. Mass spectrometers for the analysis of biomolecules. *Biological mass spectrometry.* San Diego: Elsevier Academic Press Inc.

Banwell, C. & McCash, E. 2006. *Fundamentals of molecular spectroscopy 4th edition,* Maidenhead, McGraw-Hill College.

Barman, I., Dingari, N. C., Kang, J. W., Horowitz, G. L., Dasari, R. R. & Feld, M. S. 2012. Raman spectroscopy-based sensitive and specific detection of glycated hemoglobin. *Analytical Chemistry,* **84,** 2474-2482.

Barron, L. D. 2004. *Molecular light scattering and optical activity.,* Cambridge, Cambridge University Press.

Barron, L. D., Blanch, E. W. & Hecht, L. 2002. Unfolded proteins studied by Raman optical activity. *Unfolded Proteins,* **62,** 51-90.

Barth, A. 2007. Infrared spectroscopy of proteins. *Biochimica Et Biophysica Acta-Bioenergetics,* **1767,** 1073-1101.

Becktel, W. J. & Schellman, J. A. 1987. Protein stability curves. *Biopolymers,* **26,** 1859-1877.

Berkowitz, S. A. 2006. Role of analytical ultracentrifugation in assessing the aggregation of protein biopharmaceuticals. *AAPS Journal,* **8,** 590-605.

Berman, L. 1985. Engineering glycoproteins for use as pharmaceuticals. *Trends in Biotechnology,* **3,** 51-54.

Bhattacharya, M., Jain, N., Bhasne, K., Kumari, V. & Mukhopadhyay, S. 2011. Ph-induced conformational isomerization of bovine serum albumin studied by extrinsic and intrinsic protein fluorescence. *Journal of Fluorescence,* **21,** 1083-1090.

Breiman, L. 2001. Random forests. *Machine Learning,* **45,** 5-32.

Brereton, R. G. 2005. *Chemometrics: Data analysis for the laboratory and chemical plant,* Chichester, Wiley

Brewster, V. L., Ashton, L. & Goodacre, R. 2011. Monitoring the glycosylation status of proteins using Raman spectroscopy. *Analytical Chemistry,* **83,** 6074-6081.

Bro, R. 2006. Review on multiway analysis in chemistry - 2000-2005. *Critical Reviews in Analytical Chemistry,* **36,** 279-293.

Bruylants, G., Wouters, J. & Michaux, C. 2005. Differential scanning calorimetry in life science: Thermodynamics, stability, molecular recognition and application in drug design. *Current Medicinal Chemistry,* **12,** 2011-2020.

Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery,* **2,** 121-167.

Butler, M. & Meneses-Acosta, A. 2012. Recent advances in technology supporting biopharmaceutical production from mammalian cells. *Applied Microbiology and Biotechnology,* **96,** 885-894.

Byler, D. M. & Susi, H. 1986. Examination of the secondaty structure of proteins by deconvolved FT-IR spectra. *Biopolymers,* **25,** 469-487.

Cannon, W. R. & Jarman, K. D. 2003. Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Communications in Mass Spectrometry,* **17,** 1793-1801.

Chai, W., Piskarev, V. & Lawson, A. M. 2001. Negative-ion electrospray mass spectrometry of neutral underivatized oligosaccharidses. *Analytical Chemistry,* **73,** 651-656.

Chalmers, J. M., Griffiths, P. 2001. *Handbook of vibrational spectroscopy: Volume 1, theory and instrumentation.,* Chichester, J Wiley and Sons.

Chang, V. T., Crispin, M., Aricescu, A. R., Harvey, D. J., Nettleship, J. E., Fennelly, J. A., Yu, C., Boles, K. S., Evans, E. J., Stuart, D. I., Dwek, R. A., Jones, E. Y., Owens, R. J. & Davis, S. J. 2007. Glycoprotein structural genomics: Solving the glycosylation problem. *Structure,* **15,** 267-273.

Chen, M. C. & Lord, R. C. 1980. Laser-excited Raman spectrocopy of biomolecules:13:Conformational study of alpha-chymotrypsin and trypsin. *Journal of Raman Spectroscopy,* **9,** 304-307.

Chi, Z. H. & Asher, S. A. 1998. UV resonance Raman determination of protein acid denaturation: Selective unfolding of helical segments of horse myoglobin. *Biochemistry,* **37,** 2865-2872.

Chi, Z. H. & Asher, S. A. 1999. Ultraviolet resonance Raman examination of horse apomyoglobin acid unfolding intermediates. *Biochemistry,* **38,** 8196-8203.

Chien, S., Usami, S. & Bertles, J. F. 1970. Abnormal rheology of oxygenated blood in sickle cell aneamia. *Journal of Clinical Investigation,* **49,** 623-639.

Clarke, S. J., Littleford, R. E., Smith, W. E. & Goodacre, R. 2005. Rapid monitoring of antibiotics using Raman and surface enhanced Raman spectroscopy. *Analyst,* **130,** 1019-1026.

Clowers, B. H., Dwivedi, P., Steiner, W. E., Hill, H. H. & Bendiak, B. 2005. Seperation of sodiated isobaric disaccharides and trisaccharides using elctrospray ionization-atomic preassure ion mobility-time of flight mass spectrometry. *Journal of Amerian Society of Mass Spectrometry,* **16,** 660-669.

Cotton, F. & Gulbis, B. 2013. Separation of hemoglobin variants by capillary electrophoresis. *Methods in molecular biology,* **919,** 121-130.

Cromwell, M., Hilario, E. & Jacobson, F. 2006. Protein aggregation and bioprocess. *AAPS Journal,* **66,** 572-579.

Cubitt, A. B., Heim, R., Adams, S. R., Boyd, A. E., Gross, L. A. & Tsien, R. Y. 1995. Understanding, improving and using green fluorescent proteins. *Trends in Biochemical Sciences,* **20,** 448-455.

Cui, Y., Turner, G., Roy, U. N., Guo, M., Pan, Z., Morgan, S., Burger, A. & Yeh, Y. 2005. Raman spectroscopy shows antifreeze glycoproteins interact with highly oriented pyrolytic graphite. *Journal of Raman Spectroscopy,* **36,** 1113-1117.

Das, G., Gentile, F., Coluccio, M. L., Perri, A. M., Nicastri, A., Mecarini, F., Cojoc, G., Candeloro, P., Liberale, C., De Angelis, F. & Di Fabrizio, E. 2011. Principal component analysis based methodology to distinguish protein SERS spectra. *Journal of Molecular Structure,* **993,** 500-505.

De Gelder, J., De Gussem, K., Vandenabeele, P. & Moens, L. 2007. Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy,* **38,** 1133-1147.

Degen, I. A. 1997. *Tables of characteristic group frequencies for the interpretation of infrared and Raman spectra,* Harrow, Acolyte.

DeJong, S. B. M., Wise, B. M. & Ricker, N. L. 2001. Canonical partial least squares and continuum power regression. *Journal of Chemometrics,* **15,** 85-100.

Dell, A. & Morris, H. R. 2001. Glycoprotein structure determination mass spectrometry. *Science,* **291,** 2351-2356.

Demeule, B., Lawrence, M. J., Drake, A. F., Gurny, R. & Arvinte, T. 2007. Characterization of protein aggregation: The case of a therapeutic immunoglobulin. *Biochimica Et Biophysica Acta-Proteins and Proteomics,* **1774,** 146-153.

Dingari, N. C., Horowitz, G. L., Kang, J. W., Dasari, R. R. & Barman, I. 2012. Raman spectroscopy provides a powerful diagnostic tool for accurate determination of albumin glycation. *PLOS One,* **7,** 1-11.

Dollish, F. R., Fateley, W. G. & Bentley, F. F. 1974. *Characteristic Raman frequencies of organic compounds,* New York, Wiley-Interscience.

Edge, A. S. B. 2003. Deglycosylation of glycoproteins with trifluoromethanesulphonic acid: Elucidation of molecular structure and function. *Biochemical Journal,* **376,** 339-350.

Edge, A. S. B., Faltynek, C. R., Hof, L., Reichert, L. E. & Weber, P. 1981. Deglycosylation of glycoproteins by trifluoromethanesulfonic acid. *Analytical Biochemistry,* **118,** 131-137.

Efron, B. & Tibshirani, R. J. 1994. *An introduction to the bootstrap.,* New York, Chapman and Hall/CRC.

El-Aneed, A., Cohen, A. & Banoub, J. 2009. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews,* **44,** 210-230.

Ellepola, S. W., Choi, S.-Z., Phillips, D. L. & Ma, C.-Y. 2006. Raman spectroscopic study of rice globulin. *Journal of Cereal Science,* **43,** 85-93.

Elliot, A. & Ambrose, E. J. 1950. Structure of synthetic polypeptides. *Nature,* **165,** 921-922.

Ellis, D. I. & Goodacre, R. 2006. Metabolic fingerprinting in disease diagnosis: Biomedical applications of infrared and Raman spectroscopy. *Analyst,* **131,** 875-885.

Elortza, F., Nuhse, T. S., Foster, L. J., Stensballe, A., Peck, S. C. & Jenson, O., J. 2003. Proteomic analysis of glycosylphosphatatidylinositol- anchored membrane proteins. *Molecular & Cellular Proteomics,* **2,** 1261-1270.

Ettrich, R., Kopecky, V., Hofbauerova, K., Baumruk, V., Novak, P., Pompach, P., Man, P., Plihal, O., Kuty, M., Kulik, N., Sklenar, J., Ryslava, H., Kren, V. & Bezouska, K. 2007. Structure of the dimeric N-glycosylated form of fungal beta-n-acetylhexosaminidase revealed by computer modeling, vibrational spectroscopy, and biochemical studies. *BMC Structural Biology,* **7,** 1-14.

Fabian, H. & Mantsch, H. H. 1995. Ribonuclease-A revisited - infrared spectroscopic evidence for lack of native-like secondary structures in the thermally denatured state. *Biochemistry,* **34,** 13651-13655.

Faulds, K., Jarvis, R., Smith, W. E., Graham, D. & Goodacre, R. 2008. Multiplexed detection of labelled oligonucleotides using surface enhanced resonance Raman scattering (SERRS). *Analyst,* **133,** 1505-1512.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. 1990. Electrospray ionization-principles and practice. *Mass Spectrometry Reviews,* **9,** 37-70.

Fleury, F., Ianoul, A., Baggetto, L., Jardillier, J. C., Alix, A. J. P. & Nabiev, I. Raman, SERS, and induced circular dichroism techniques as a probe of pharmaceuticals in their interactions with the human serum albumin and p-glycoprotein. *In:* Morris, M. D., ed. Conference on Biomedical Applications of Raman Spectroscopy, Jan 25-26 1999 San Jose, Ca. Spie-Int Soc Optical Engineering, 80-89.

From, N. B. & Bowler, B. E. 1998. Urea denaturation of staphylococcal nuclease monitored by fourier transform infrared spectroscopy. *Biochemistry,* **37,** 1623-1631.

Gabrielson, J. P., Arthur, K. K., Stoner, M. R., Winn, B. C., Kendrick, B. S., Razinkov, V., Svitel, J., Jiang, Y., Voelker, P. J., Fernandes, C. A. & Ridgeway, R. 2010. Precision of protein aggregation measurements by sedimentation velocity analytical ultracentrifugation in biopharmaceutical applications. *Analytical Biochemistry,* **396,** 231-241.

Ganim, Z., Chung, H. S., Smith, A. W., Deflores, L. P., Jones, K. C. & Tokmakoff, A. 2008. Amide I two-dimensional infrared spectroscopy of proteins. *Accounts of Chemical Research,* **41,** 432-441.

Garcia-Fruitos, E., Vazquez, E., Gonzalez-Montalban, N., Ferrer-Miralles, N. & Villaverde, A. 2011. Analytical approaches for assessing aggregation of protein biopharmaceuticals. *Current Pharmaceutical Biotechnology,* **12,** 1530-1536.

Ge, Y., Rajkumar, L., Guzman, R. C., Nandi, S. C., Patton, W. F. & Agnew, B. J. 2004. Enrichment of phosphoproteins for proteomic analysis using immobolised fe(III)-affinity chromatography. *Proteomics,* **4,** 3464-3469.

Goddard, P. 1991. Therapeutic proteins - a pharmaceutical perspective. *Advanced Drug Delivery Reviews,* **6,** 103-131.

Graham, D. & Faulds, K. 2008. Quantitative SERRS for DNA sequence analysis. *Chemical Society Reviews,* **37,** 1042-1050.

Greene, R. F. & Pace, C. N. 1974. Urea and guanidine-hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *Journal of Biological Chemistry,* **249,** 5388-5393.

Greer, F. 2007. Glycosylated bioproducts- breaking down the benefits. *European BioPharmaceutical Review*.

Greer, F. 2008. Biopharmaceutical charcterisation- considering the key questions. *European BioPharmaceutical Review*.

Greer, F., Reason, A. & Rodgers, M. 2008. Post- translational modifications of biopharmaceuticals - a challenge for analytical characterisation. *European BioPharmaceutical Review*.

Grimsley, G. R., Scholtz, J. M. & Pace, C. N. 2009. A summary of the measured pk values of the ionizable groups in folded proteins. *Protein Science,* **18,** 247-251.

Hajare, A. A., More, H. N. & Pisal, S. S. 2011. Effect of sugar additives on stability of human serum albumin during vacuum foam drying and storage. *Current Drug Delivery,* **8,** 678-690.

Hames, B. D. & Hooper, N. M. 2000. *Instant notes in biochemistry,* Oxford, Bios Scientific.

Hansen, R., Dickson, A. J., Goodacre, R., Stephens, G. M. & Sellick, C. A. 2010. Rapid characterization of N-linked glycans from secreted and gel-purified monoclonal antibodies using MALDI-TOF mass spectrometry. *Biotechnology and Bioengineering,* **107,** 902-908.

Harris, D. & Bertolucci, M. 1978. *Symmetry and spectroscopy: An introduction to vibrational and electronic spectroscopy* New York, Dover Publications.

Henzel, W. J., Watanabe, C. & Stults, J. T. 2003. Protein identification: The origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry,* **14,** 931-942.

Hering, J. A., Innocent, P. R. & Haris, P. I. 2002. Automatic amide I frequency selection for rapid quantification of protein secondary structure from fourier transform infrared spectra of proteins. *Proteomics,* **2,** 839-849.

Howell, N. & LiChan, E. 1996. Elucidation of interactions of lysozyme with whey proteins by Raman spectroscopy. *International Journal of Food Science and Technology,* **31,** 439-451.

Huang, K., Maiti, N. C., Phillips, N. B., Carey, P. R. & Weiss, M. A. 2006. Structure-specific effects of protein topology on cross-beta assembly: Studies of insulin fibrillation. *Biochemistry,* **45,** 10278-10293.

ICH 1999. Q6B test procedures and acceptance criteria for biotechnological/biological products. International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use.

Jackson, M., Haris, P. I. & Chapman, D. 1989. Fourier-transform infrared spectroscopic studies of lipids, polypeptides and proteins. *Journal of Molecular Structure,* **214,** 329-355.

Jackson, M. & Mantsch, H. H. 1995. The use and misuse of FTIR spectroscopy in the determination of protein-structure. *Critical Reviews in Biochemistry and Molecular Biology,* **30,** 95-120.

Jarvis, R. M., Blanch, E. W., Golovanov, A. P., Screen, J. & Goodacre, R. 2007. Quantification of casein phosphorylation with conformational interpretation using Raman spectroscopy. *Analyst,* **132,** 1053-1060.

Jarvis, R. M., Broadhurst, D., Johnson, H., O'Boyle, N. M. & Goodacre, R. 2006. Pychem: A multivariate analysis package for python. *Bioinformatics,* **22,** 2565-2566.

Jolliffe, I. T. 1986. *Principal components analysis,* New York Springer-Verlag.

Josephs, J. L. & Sanders, M. 2004. Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadruople mass spectrometers. *Rapid Communications in Mass Spectrometry,* **18,** 743-759.

Kaufmann, J. E. B. & Fussenegger, M. 2001. Use of antibodies for detection of phosphorylated proteins seperated by two dimensional gel electrophoresis. *Proteomics,* **38,** 1757-1764.

Kealey, D., Haines, P.J. 2002. *Instant notes in analytical chemistry,* Oxford, Bios Scientific

Kikuchi, G. E., Baker, S. A., Merajver, S. D., Coligan, J. E., Levine, M., Glorioso, J. C. & Nairn, R. 1987. Purification and structural characterization of herpes-simplex virus glycoprotein-c. *Biochemistry,* **26,** 424-431.

Konermann, L. 2004. Protein unfolding and denaturants. *Encyclopedia of Life Sciences.* John Wiley and Sons.

Kong, J. & Yu, S. 2007. Fourier transform infrared spectroscopic analysis of protein secondary structures. *Acta Biochimica Et Biophysica Sinica,* **39,** 549-559.

Kopecky, V., Ettrich, R., Hofbauerova, K. & Baumruk, V. 2003. Structure of human alpha-1-acid glycoprotein and its high-affinity binding site. *Biochemical and Biophysical Research Communications,* **300,** 41-46.

Kreimer, D. I., Ben-Amotz, D., Zhang, D., Xie, Y., Ortiz, C., DeGrella, R. F., Adar, F. & Davisson, V. J. 2004. Raman, reflectance FTIR and MALDI-MS of proteins, peptides, glycoproteins and small molecules on a single tienta substrate - preliminary data and implications for protein characterization in drug discovery. *Protein Science,* **13,** 125.

Kumar, S., Rai, A. K., Singh, V. B. & Rai, S. B. 2005. Vibrational spectrum of glycine molecule. *Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy,* **61,** 2741-2746.

Lambert, J. P., Ethier, M., Smith, J. C. & Figeys, D. 2005. Proteomics: Gel based to gel free. *Analytical Chemistry,* **77,** 3771-3789.

Li, C. H., Xichdao, N., Narhi, L., Chemmalil, L., Towers, E., Muzammil, S., Gabrielson, J. & Jiang, Y. 2011. Applications of circular dichroism (CD) for structural analysis of proteins: Qualification of near- and far-UV CD for protein higher order structural analysis. *Journal of Pharmaceutical Sciences,* **100,** 4642-4654.

Liang, M., Chen, V., Chen, H. L. & Chen, W. L. 2006. A simple and direct isolation of whey components from raw milk by gel filtration chromatography and structural characterization by fourier transform Raman spectroscopy. *Talanta,* **69,** 1269-1277.

Lindon, L. C. 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in nuclear magnetic resonance spectroscopy,* **39,** 1-40.

Liu, J. 2008. Toward high-throughput and reliable peptide identification via MS/MS spectra. *Methods Mol Biol,* **484,** 333-344.

Long, D. A. 2002. *The Raman effect: A unified treatment of the theory of Raman scattering by molecules,* Chichester, John Wiley and Sons.

Lord, R. C. & Yu, N. T. 1970. Laser-excited Raman spectroscopy of biomolecules .1. Native lysozyme and its constituent amino acids. *Journal of Molecular Biology,* **50,** 509-516.

Markert, Y., Koditz, J., Ulbrich-Hofmann, R. & Arnold, U. 2003. Proline versus charge concept for protein stabilization against proteolytic attack. *Protein Engineering,* **16,** 1041-1046.

Matsuo, K., Yonehara, R. & Gekko, K. 2005. Improved estimation of the secondary structures of proteins by vacuum-ultraviolet circular dichroism spectroscopy. *Journal of Biochemistry,* **138,** 79-88.

McCavit, T. L. 2012. Sickle cell disease. *Pediatrics in review / American Academy of Pediatrics,* **33,** 195-206.

McColl, I. H., Blanch, E. W., Gill, A. C., Rhie, A. G. O., Ritchie, M. A., Hecht, L., Nielsen, K. & Barron, L. D. 2003. A new perspective on beta-sheet structures using vibrational Raman optical activity: From poly-l-lysine to the prion protein. *Journal of the American Chemical Society,* **125,** 10019-10026.

McCreery, R. L. 2000. *Raman spectroscopy for chemical analysis,* Chichester, Wiley Interscience.

McGovern, A. C., Broadhurst, D., Taylor, J., Kaderbhai, N., Winson, M. K., Small, D. A., Rowland, J. J., Kell, D. B. & Goodacre, R. 2002. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: Application to gibberellic acid production. *Biotechnology and Bioengineering,* **78,** 527-538.

Meier, E. R. & Miller, J. L. 2012. Sickle cell disease in children. *Drugs,* **72,** 895-906.

Melnik, T. 2011. Studies of Irreversible Heat Denaturation of Green Fluorescent Protein by Differntial Scanning Microcalorimetry. *Thermochimica Acta,* **512**, 71-75.

Morris, H. R. 1980. Biomolecular structure determination by mass-spectrometry. *Nature,* **286,** 447-452.

Morris, H. R., Thompson, M. R., Osuga, D. T., Ahmed, A. I., Chan, S. M., Vandenheede, J. R. & Feeney, R. E. 1978. Antifreeze glycoproteins from blood of an antarctic fish- strusture of proline-containing glycopeptides. *Journal of Biological Chemistry,* **253,** 5155-5162.

Mouls, L., Aubagnac, J. L., Martinez, J. & Enjalbal, C. 2007. Low energy peptide fragmentations in an ESI-Q-TOF type mass spectrometer. *Journal of Proteome Research,* **6,** 1378-1391.

Mrozek, M. F., Zhang, D. & Ben-Amotz, D. 2004. Oligosaccharide identification and mixture quantification using Raman spectroscopy and chemometric analysis. *Carbohydrate Research,* **339,** 141-149.

Murayama, M. 1972. Thermal or endo thermic aggregation of sickle cell hemo globin during sickling. *Advances in Experimental Medicine and Biology,* **28,** 243-251.

Naes, T. & Isaksson, T. 1992. Locally weighted regression in diffuse near-infrared transmittance spectroscopy. *Applied Spectroscopy,* **46,** 34-43.

Naidu, K. T. & Prabhu, N. P. 2011. Protein-surfactant interaction: Sodium dodecyl sulfate-induced unfolding of ribonuclease a. *Journal of Physical Chemistry B,* **115,** 14760-14767.

Nalbandi, R. M., Henry, R. L., Lusher, J. M., Camp, F. R. & Conte, N. F. 1971. Sickledex test for hemoglobin-S - critique. *Journal of the American Medical Association,* **218,** 1679-1986.

Nie, S. & Emory, S. R. 1997. Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. *Science,* **275** 1102-1106.

Nienhuis, A. W. & Bunn, H. F. 1974. Hemoglobin switching in sheep and goats - occurrence of hemoglobins a and c in same red-cell. *Science,* **185,** 946-948.

Noda, I. 1989. Two-dimensional infrared-spectroscopy. *Journal of the American Chemical Society,* **111,** 8116-8118.

Noda, I. 1993. Generalized 2-dimensional correlation method applicable to infrared, Raman, and other types of spectroscopy. *Applied Spectroscopy,* **47,** 1329-1336.

Noda, I. & Ozaki, Y. 2004. *Two-dimensional correlation spectroscopy: Applications in vibrational and optical spectroscopy,* Chichester, John Wiley and Sons Ltd.

Novotny, J., Tonegawa, S., Saito, H., Kranz, D. M. & Eisen, H. N. 1986. Secondary, tertiary, and quaternary structure of T-cell-specific immunoglobulin-like polypeptide-chains. *Proceedings of the National Academy of Sciences of the United States of America,* **83,** 742-746.

Ohtake, S. & Wang, Y. J. 2011. Trehalose: Current use and future applications. *Journal of Pharmaceutical Sciences,* **100,** 2020-2053.

Oladepo, S. A., Xiong, K., Hong, Z., Asher, S. A., Handen, J. & Lednev, I. K. 2012. UV resonance Raman investigations of peptide and protein structure and dynamics. *Chemical Reviews,* 112**,** 2604-2628.

Oleinikov, V., Kryukov, E., Kovner, M., Ermishov, M., Tuzikov, A., Shiyan, S., Bovin, N. & Nabiev, I. Sialylation sensitive bands in the raman spectra of oligosaccharides and glycoproteins. 24th European Congress on Molecular Spectroscopy, Aug 23-28 1998 Prague, Czech Republic. Elsevier Science Bv, 475-480.

Otto, M. 1999. *Chemometrics: Statistics and computer application in analytical chemistry,* Weinheim, Wiley-VCH.

Padliya, N. D. & Wood, T. D. 2008. Improved peptide mass fingerprinting matches via optimized sample preparation in maldi mass spectrometry. *Analytica Chimica Acta,* **627,** 162-168.

Peng, J., Peng, S., Jiang, A., Wei, J., Li, C. & Tan, J. 2010. Asymmetric least squares for multiple spectra baseline correction. *Analytica Chimica Acta,* **683,** 63-68.

Pierce-Biotechnology. 2010. *In-solution tryptic digestion and guanidination kit.* Online. 2010.

Prevelige, P. E., Thomas, D., Aubrey, K. L., Towse, S. A. & Thomas, G. J. 1993. Studies of virus structure by Raman-spectroscopy .37. Subunit conformational-changes accompanying bacteriophage-p22 capsid maturation. *Biochemistry,* **32,** 537-543.

Pumphrey, J. G. & Steinhar.J 1973. Light-scattering studies of differences in aggregation behavior of normal and sickle-cell hemoglobins. *Abstracts of Papers of the American Chemical Society,* **26,** 160-160.

Raman, C. V. & Krishnan, K. S. 1928. A new type of secondary radiation. *Nature,* **121,** 501-502.

Rapp, E., Charvat, A., Beinsen, A., Plessmann, U., Reichl, U., Seidel-Morgenstern, A., Urlaub, H. & Abel, B. 2009. Atmospheric pressure free liquid infrared MALDI mass spectrometry: Toward a combined ESI/MALDI-liquid chromatography interface. *Analytical Chemistry,* **81,** 443-452.

Redwan, E. M. 2007. Cumulative updating of approved biopharmaceuticals. *Human Antibodies,* **16,** 137-158.

Rhiel, M., Ducommun, P., Bolzonella, I., Marison, I. & von Stockar, U. 2002. Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. *Biotechnology and Bioengineering,* **77,** 174-185.

Roberts, C. J., Das, T. K. & Sahin, E. 2011. Predicting solution aggregation rates for therapeutic proteins: Approaches and challenges. *International Journal of Pharmaceutics,* **418,** 318-333.

Royer, C. A. 2006. Probing protein folding and conformational transitions with fluorescence. *Chemical Reviews,* **106,** 1769-1784.

Sahin, E. & Roberts, C. J. 2012. Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms. *Methods in molecular biology,* **899,** 403-423.

Samanta, A., Paul, B. K. & Guchhait, N. 2011. Spectroscopic probe analysis for exploring probe-protein interaction: A mapping of native, unfolding and refolding of protein bovine serum albumin by extrinsic fluorescence probe. *Biophysical Chemistry,* **156,** 128-139.

Sane, S. U., Cramer, S. M. & Przybycien, T. M. 1999. A holistic approach to protein secondary structure characterization using amide i band raman spectroscopy. *Analytical Biochemistry,* **269,** 255-272.

Sane, S. U., Wong, R. & Hsu, C. C. 2004. Raman spectroscopic characterization of drying-induced structural changes in a therapeutic antibody: Correlating structural changes with long-term stability. *Journal of Pharmaceutical Sciences,* **93,** 1005-1018.

Savitzky, A. & Golay, M. J. E. 1964. Smoothing + differentiation of data by simplified least squares procedures. *Analytical Chemistry,* **36,** 1627-1635.

Schenk, J., Marison, I. W. & von Stockar, U. 2007. Simplified fourier-transform mid-infrared spectroscopy calibration based on a spectra library for the on-line monitoring of bioprocesses. *Analytica Chimica Acta,* **591,** 132-140.

Scheraga, H. A. 2011. Ribonucleases as models for understanding protein folding. *In:* Nicholson, A. W. (ed.) *Ribonucleases.*

Sellick, C. A., Hansen, R., Jarvis, R. M., Maqsood, A. R., Stephens, G. M., Dickson, A. J. & Goodacre, R. 2010. Rapid monitoring of recombinant antibody production by mammalian cell cultures using fourier transform infrared spectroscopy and chemometrics. *Biotechnology and Bioengineering,* **106,** 432-442.

Serrano, A. L., Waegele, M. M. & Gai, F. 2012. Spectroscopic studies of protein folding: Linear and nonlinear methods. *Protein Science,* **21,** 157-170.

Shao, J., Lin, M., Li, Y., Li, X., Liu, J., Liang, J. & Yao, H. 2012. In vivo blood glucose quantification using raman spectroscopy. *PLOS One,* **7,** 1-6.

Sharma, A. S., S. G. 1999. *Introduction to fluorescence spectroscopy,* Chichester, J Wiley and Sons

Shashilov, V. A. & Lednev, I. K. 2010. Advanced statistical and numerical methods for spectroscopic characterization of protein structural evolution. *Chemical Reviews,* **110,** 5692-5713.

Shaw, A. D., Kaderbhai, N., Jones, A., Woodward, A. M., Goodacre, R., Rowland, J. J. & Kell, D. B. 1999. Noninvasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics. *Applied Spectroscopy,* **53,** 1419-1428.

Shvartsburg, A. A., Tang, K. & Smith, R. D. 2009. Two-dimensional ion mobility analyses of proteins and peptides. *Methods in molecular biology.* Humana Press Inc.

Siamwiza, M. N., Lord, R. C., Chen, M. C., Takamatsu, T., Harada, I., Matsuura, H. & Shimanouchi, T. 1975. Interpretation of doublet at 850 and 830 cm^{-1} in Raman-

spectra of tyrosyl residues in proteins and certain model compounds. *Biochemistry,* **14,** 4870-4876.

Silverstein, A. & Nunn, L. 1997. *Sickle cell anaemia*, Enslow Publishers Inc.
Smekal, A. 1923. The quantum theory of dispersion. *Naturwissenschaften,* **43,** 873.

Smith, E. & Dent, G. 2005. *Modern Raman spectroscopy-a practical approach* Chichester, J Wiley and Sons.

Socrates, G. 2001. *Infrared and Raman characteristic group frequencies,* Chichester, John Wiley & Sons Ltd.

Sola, R. J. & Griebenow, K. 2009. Effects of glycosylation on the stability of protein pharmaceuticals. *Journal of Pharmaceutical Sciences,* **98,** 1223-1245.

Soldatova, A. V., Ibrahim, M., Olson, J. S., Czernuszewicz, R. S. & Spiro, T. G. 2010. New light on no bonding in fe(III) heme proteins from resonance Raman spectroscopy and DFT modeling. *Journal of the American Chemical Society,* **132,** 4614-4625.

Spink, C. H. 2008. Differential scanning calorimetry. *In:* Correia, J. J. & Detrich, H. W. (eds.) *Biophysical tools for biologists: Vol 1 in vitro techniques.* San Diego: Elsevier Academic Press Inc.

Stryer, L., Berg, J. & Tymoczke, J. 2002. *Biochemistry*, W.H. Freeman & Co ltd.

Sun, X., Chiu, J. F. & He, Q. Y. 2005. Application of immobilised metal affinity chromatography in proteomics. *Expert Reviews in Proteomics,* **2,** 649-657.

Sundararajan, N., Mao, D. Q., Chan, S., Koo, T. W., Su, X., Sun, L., Zhang, J. W., Sung, K. B., Yamakawa, M., Gafken, P. R., Randolph, T., McLerran, D., Feng, Z. D., Berlin, A. A. & Roth, M. B. 2006. Ultrasensitive detection and characterization of posttranslational modifications using surface-enhanced Raman spectroscopy. *Analytical Chemistry,* **78,** 3543-3550.

Takekiyo, T., Takeda, N., Isogai, Y., Kato, M. & Taniguchi, Y. 2006. Pressure stability of the α-helix structure in a de novo designed protein (α-l-α)2 studied by FTIR spectroscopy. *Biopolymers,* **85,** 185-188.

Tarentino, A. L., Gomez, C. M. & Plummer, T. H. 1985. Deglycosylation of asparagine-linked glycans by peptide-n-glycosidase-f. *Biochemistry,* **24,** 4665-4671.

Taylor, M. E. 2006. *Introduction to glycobiology,* Oxford, Oxford University Press.

Teng-umnuay, P., Morris, H. R., Dell, A., Panico, M., Paxton, T. & West, C. M. 1998. The cytoplasmic f-box binding protein skp1 contains a novel pentasaccharide linked to hydroxyproline in dictyostelium. *Journal of Biological Chemistry,* **273,** 18242-18249.

Tienta Sciences, I. 2004a. Application guide AG-001 raman detection of proteomic analytes.

Tienta Sciences, I. 2004b. Application guide AG-002 reduction of fluorescent impurities in protein samples for raman spectroscopy.

Torreggiani, A., Barata-Vallejo, S. & Chatgilialoglu, C. 2011. Combined Raman and IR spectroscopic study on the radical-based modifications of methionine. *Analytical and Bioanalytical Chemistry,* **401,** 1231-1239.

Truong, K. & Ikura, M. 2001. The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. *Current Opinion in Structural Biology,* **11,** 573-578.

Tsien, R. Y. 1998. The green fluorescent protein. *Annual Review of Biochemistry,* **67,** 509-544.

Tuma, R. 2005. Raman spectroscopy of proteins: From peptides to large assemblies. *Journal of Raman Spectroscopy,* **36,** 307-319.

Tuma, R., Russell, M., Rosendahl, M. & Thomas, G. J. 1995. Solution conformation of the extracellular domain of the human tumor-necrosis-factor receptor probed by Raman and UV resonance Raman spectroscopy- structural effects of an engineered peg linker. *Biochemistry,* **34,** 15150-15156.

Verma, S. P. & Wallach, D. F. H. 1977. Changes of raman-scattering in CH-stretching region during thermally induced unfolding of ribonuclease. *Biochemical and Biophysical Research Communications,* **74,** 473-479.

Villar, S. E. J., Edwards, H. G. M. & Benning, L. G. 2006. Raman spectroscopic and scanning electron microscopic analysis of a novel biological colonisation of volcanic rocks. *Icarus,* **184,** 158-169.

Walters, J., Milam, S. L. & Clark, A. C. 2009. Practical approaches to protein folding and assembly: Spectroscopic strategies in thermodynamics and kinetics. *Methods in enzymology: Biothermodynamics.* San Diego: Elsevier Academic Press Inc.

Wang, X., Li, Y., Quan, D., Wang, J., Zhang, Y., Du, J., Peng, J., Fu, Q., Zhou, Y., Jia, S., Wang, Y. & Zhan, L. 2012. Detection of hepatitis B surface antigen by target-induced aggregation monitored by dynamic light scattering. *Analytical Biochemistry,* **428,** 119-125.

Webster, G. T., Tilley, L., Deed, S., McNaughton, D. & Wood, B. R. 2008. Resonance Raman spectroscopy can detect structural changes in haemozoin (malaria pigment) following incubation with chloroquine in infected erythrocytes. *FEBs Letters,* **582,** 1087-1092.

Webster, S. 2010. Screening proteins for manufacturability *Genetic Engineering and Biotechnology News.*

Weiss, W. F., Young, T. M. & Roberts, C. J. 2009. Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *Journal of Pharmaceutical Sciences,* **98,** 1246-1277.

Welch, W. J. 1990. Construction of permutation tests. *Journal of the American Statistical Association,* **85,** 693-698.

Wen, J., Arakawa, T. & Philo, J. S. 1996. Size-exclusion chromatography with on-line light-scattering, absorbance, and refractive index detectors for studying proteins and their interactions. *Analytical Biochemistry,* **240,** 155-166.

Wen, Z. Q. 2007. Raman spectroscopy of protein pharmaceuticals. *Journal of Pharmaceutical Sciences,* **96,** 2861-2878.

Wen, Z. Q., Cao, X. C. & Vance, A. 2008. Conformation and side chains environments of recombinant human interleukin-1 receptor antagonist (rh-iL-1 ra) probed by Raman, Raman optical activity, and UV-resonance raman spectroscopy. *Journal of Pharmaceutical Sciences,* **97,** 2228-2241.

Westerhuis, J. A., Kourti, T. & MacGregor, J. F. 1998. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics,* **12,** 301-321.

Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U. & Bjorck, L. 1994. 3-dimensional solution structure of an immunoglobulin light chain-binding domain of protein-l - comparison with the IgG-binding domains of protein-g. *Biochemistry,* **33,** 14011-14017.

Wold, S., Sjostrom, M. & Eriksson, L. 2001. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **58,** 109-130.

Wood, B. R., Hammer, L., Davis, L. & McNaughton, D. 2005. Raman microspectroscopy and imaging provides insights into heme aggregation and denaturation within human erythrocytes. *Journal of Biomedical Optics,* **10,** 23-29.

Wood, B. R., Stoddart, P. R. & McNaughton, D. 2011. Molecular imaging of red blood cells by raman spectroscopy. *Australian Journal of Chemistry,* **64,** 593-599.

Wyatt, P. J. 1993. Light-scattering and the absolute characterization of macromolecules. *Analytica Chimica Acta,* **272,** 1-40.

Yamamoto, S. 2012. Conformational analyses of peptides and proteins by vibrational Raman optical activity. *Analytical and Bioanalytical Chemistry,* **403,** 2203-2212.

Yu, P. Q. 2005. Multicomponent peak modeling of protein secondary structures: Comparison of gaussian with lorentzian analytical methods for plant feed and seed molecular biology and chemistry research. *Applied Spectroscopy,* **59,** 1372-1380.

Zheng, R., Zheng, X. J., Dong, J. & Carey, P. R. 2004. Proteins can convert to beta-sheet in single crystals. *Protein Science,* **13,** 1288-1294.

Zheng, Y. F., Guo, Z. H. & Cai, Z. W. 2009. Combination of beta-elimination and liquid chromatography/quadrupole time-of-flight mass spectrometry for the determination of O-glycosylation sites. *Talanta,* **78,** 358-363.

Zhu, F., Isaacs, N. W., Hecht, L. & Barron, L. D. 2005a. Polypeptide and carbohydrate structure of an intact glycoprotein from Raman optical activity. *Journal of the American Chemical Society,* **127,** 6142-6144.

Zhu, F., Isaacs, N. W., Hecht, L., Tranter, G. E. & Barron, L. D. 2006. Raman optical activity of proteins, carbohydrates and glycoproteins. *Chirality,* **18,** 103-114.

Zhu, F. J., Isaacs, N. W., Hecht, L. & Barron, L. D. 2005b. Polypeptide and carbohydrate structure of an intact glycoprotein from raman optical activity. *Journal of the American Chemical Society,* **127,** 6142-6143.

Zikan, J., Novotny, J., Trapane, T. L., Koshland, M. E., Urry, D. W., Bennett, J. C. & Mestecky, J. 1985. Secondary structure of the immunoglobulin j-chain. *Proceedings of the National Academy of Sciences of the United States of America,* **82,** 5905-5909.

Appendix

1. Results from Preliminary Investigations into Glycan Characterisation

2. Work published during the course of this Ph.D:

   A. Monitoring the glycosylation status of Proteins Using Raman Spectroscopy.

      V. L. Brewster, L. Ashton & R. Goodacre. *Analytical Chemistry, 83: 6074-6081,* 2011.

   B. Using natural green fluorescence to monitor protein unfolding in GFP mutants using Optim 1000.

      V.L. Brewster, R. Goodacre, C. Dodd and S. Webster. Application note for Avacta Analytical.

   C. Application of 2D Correlation Analysis to Optim 1000 Data.

      V.L. Brewster, L. Ashton, R. Goodacre, C. Dodd and S. Webster. Application note for Avacta Analytical.

   D. Application of Multivariate Data Analysis Strategies to Optim 1000 Data.

      V.L. Brewster, E.S.Correa, R. Goodacre, C. Dodd and S. Webster. Application note for Avacta Analytical.

   E. Monitoring Guanidinium-Induced Structural Changes in Ribonuclease Proteins Using Raman Spectroscopy and 2D Correlation Analysis

      V.L.Brewster, L. Ashton and R. Goodacre. *Analytical Chemistry,*85:3570-3575, 2013.

   F. Raman Spectroscopic techniques for biotechnology and bioprocessing.

      V.L.Brewster, R. Jarvis and R. Goodacre, *European Pharmaceutical Review, 1: 48-52*, 2009.

   G. Fingerprinting Food: current technologies for the detection of food adulteration and contamination.

      D.I.Ellis, **V.L.Brewster,** W.B.Dunn, J.W.Allwood, A.P. Golovanov, an R Goodacre. Chemical Society Reviews, DOI:10.1039/C2CS35138B, 2012.

Towards Glycan Characterisation by Raman Spectroscopy.

Preliminary Investigations have shown the ability to distinguish between various monosaccharaides and glycan fragments based on their Raman spectra (Figure 1 and 2). We then investigated the Raman spectra of whole glycans, which were supplied by Chris Jones from NIBSC, where PCA was able to differentiate between glycans easily, even glycans with the same sugars in different spatial arrangements and ratios (Figure 3).

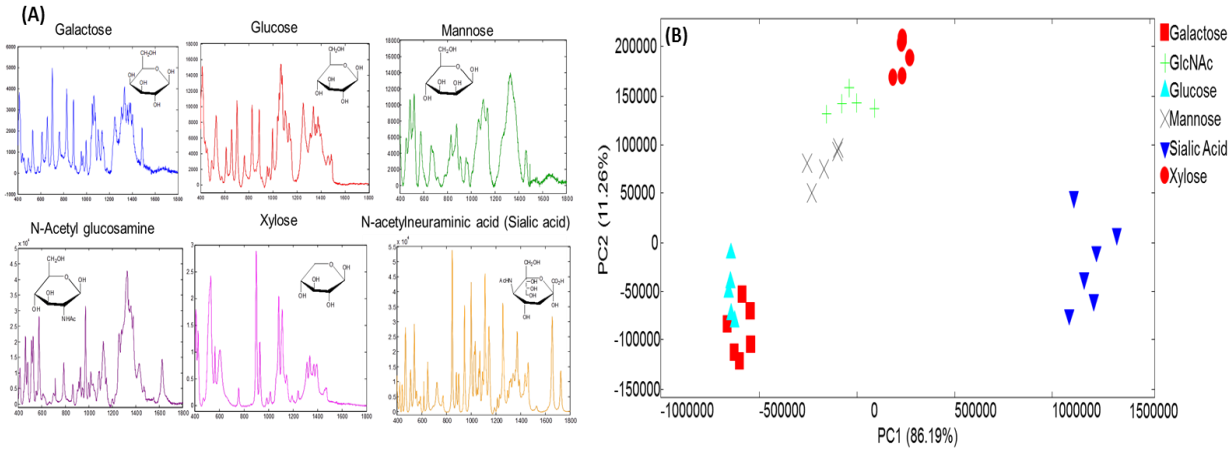

Figure 1: (A) Average Raman spectra and structures of 6 different monosaccharaides, and **(B)** PCA Scores plot of Raman data from monosaccharaides.

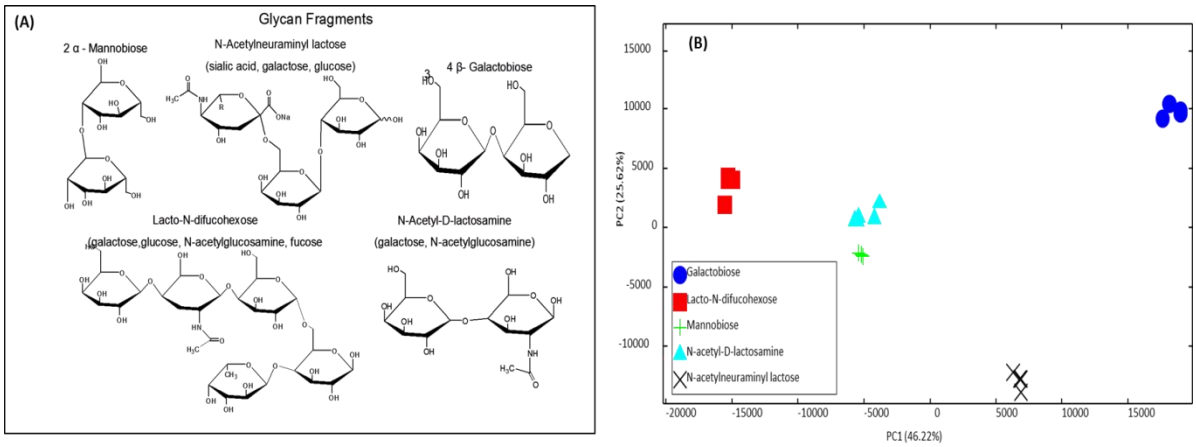

Figure 2: (A) Structures of Glycan fragments analysed, and **(B)** PCA Scores plot of Raman data from glycan fragments .

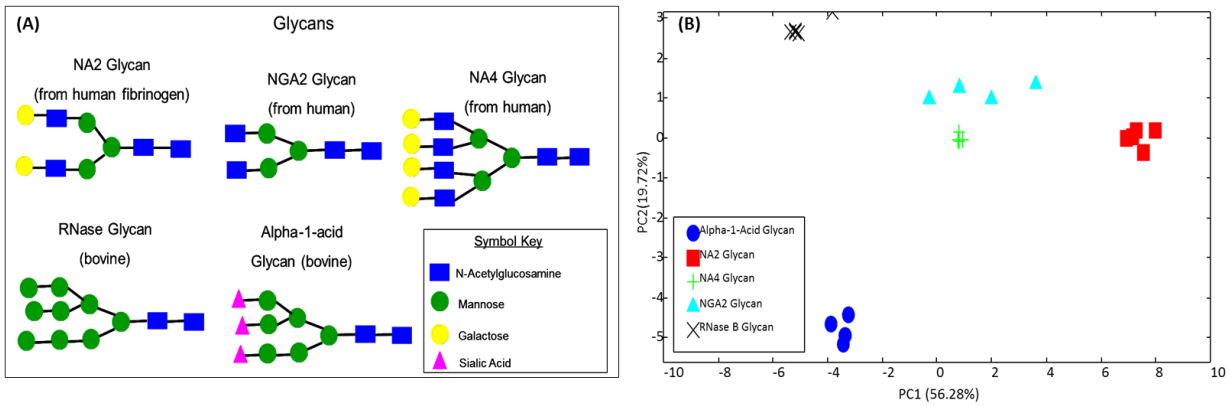

Figure 3: (A) Structures of whole glycans analysed, and **(B)** PCA Scores plot of Raman data from whole glycans.

When comparing all data from monosaccharaides, glycan fragments and whole glycans (Figure 4), the monosaccharaides appear to all fall at the left hand side of the PCA scores plot, indicating that separation a cross PC1 may be partly due to glycosidic bond vibrations. This was investigated further by comparing the Raman spectra of mannose and mannobiose (Figure 5), where bands which can be specifically assigned to the C-O stretching of the C-O-C glycosidic linkages are observed at ~708, ~900 and ~1009 cm^{-1}.

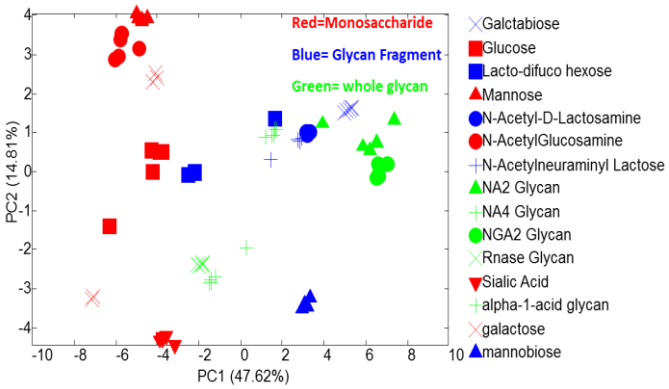

Figure 4: PCA plot of Raman data from all sugars.

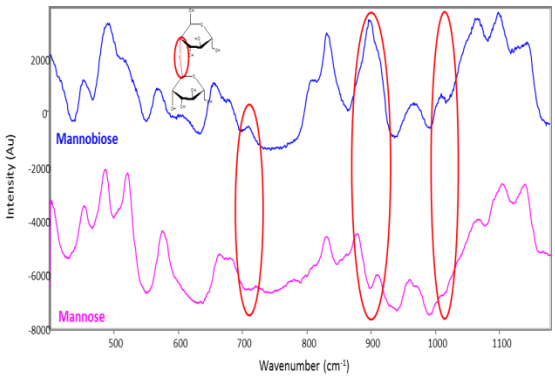

Figure 5: Raman spectra of Mannose and Mannosbiose, with bands due to glycosidic link highlighted.

Subsequently, we also compared the Raman spectra of the RNase B glycan and the α-1-acid glycoprotein glycan to the spectra of the component monosaccharaides and a mixture of these sugars in the relevant ratios (Figure 6 and 7). Spectra of monosaccharide mixtures and sugars differ vastly, with sharp peaks which can be assigned to glycosidic bond vibrations.

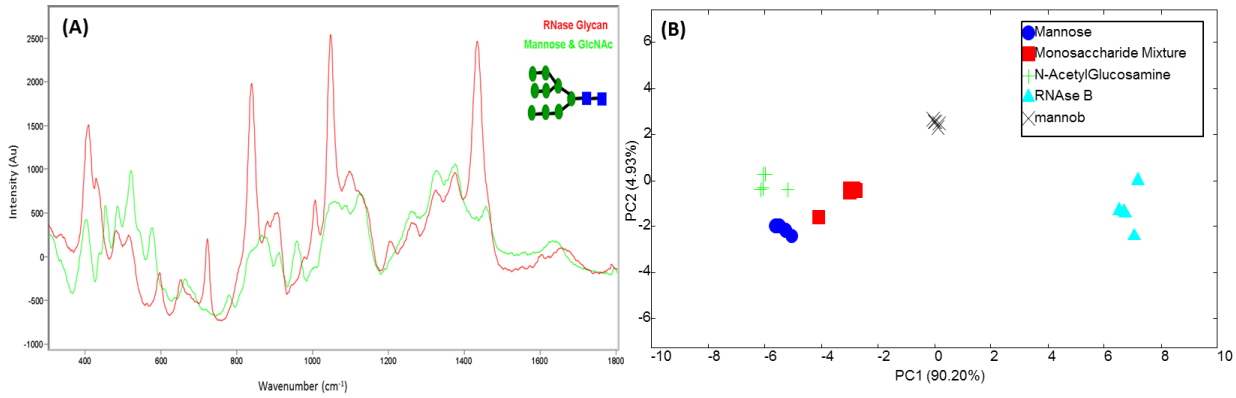

Figure 6: (A) Raman Spectra of the RNase Glycan and a mixture of GlcNAc and Mannose, and **(B)** PCA scores plot for Raman data of RNase glycan, Mannose, GlcNAc and the monosaccharaide mixture.

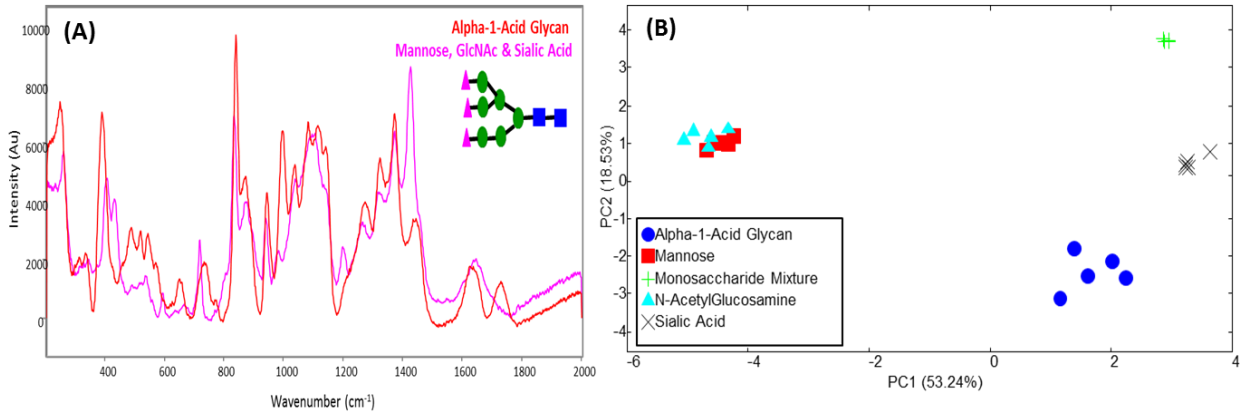

Figure 7: (A) Raman Spectra of the α-1-acid glycan and a mixture of GlcNAc, Sialic acid and Mannose, and **(B)** PCA scores plot for Raman data of α-1-acid glycan, Mannose, GlcNAc, Sialic acid and the monosaccharaide mixture.