Two-dimensional infrared spectroscopy for protein analysis

Elizabeth Mary Gardner
Department of Chemistry
Imperial College London

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London
November 2009
Declaration

The research presented in this thesis was carried out in accordance with the regulations of Imperial College London. The work is original, except where indicated by special reference in the text, and the contributions of colleagues are clearly acknowledged. Any views expressed in this thesis are those of the author and in no way represent those of Imperial College London. No part of this thesis has been submitted for any other degree and has not been presented to any other University for examination either in the United Kingdom or overseas.

SIGNED: ..............................................................................

DATE: ..................................................................................
Abstract

A number of forms of coherent multi-dimensional vibrational spectroscopy (CMDVS) have been identified as being useful for addressing a range of biological problems. Here a particular member of this family of spectroscopies, electron-vibration-vibration two-dimensional infrared (EVV 2DIR) spectroscopy (also known as DOubly-Vibrationally Enhanced InfraRed (DOVE-IR)), is explored for its possible utility for two particular bioanalytical applications; protein identification and the study of enzyme mechanisms. The main focus of this work is on the development of EVV 2DIR as a tool for high-throughput, label-free proteomics, in particular for protein identification and absolute quantification. The protein fingerprinting strategy is based on the identification of proteins through their spectroscopically determined amino acid compositions. To this end, spectral signatures of amino acid side chains (tyrosine, phenylalanine and tryptophan) have been identified, as well as those from CH$_2$ and CH$_3$ groups which have been found to be appropriate for use as internal references. The intensities of these cross peaks are measured to give proteins’ amino acid compositions in the form of amino acid / CH$_x$ ratios. Specialised databases comprising the amino acid / CH$_x$ ratios of proteins have been developed for achieving protein identifications using the EVV 2DIR data.

The second strand of this research considers the potential of triply resonant EVV 2DIR for studying protein structures and mechanisms. It is possible to use the electronic polarising properties of EVV 2DIR to good effect to achieve significant enhancement of the signal size when probing a chromophore. Here this effect is demonstrated for the case of bacteriorhodopsin (bR) membranes isolated from Halobacterium salinarium. The signal enhancement that is achievable from the retinal chromophore at the heart of bR makes it possible to study this whilst avoiding the surrounding protein.
Acknowledgments

I’d like to thank my supervisors, David Klug and Keith Willison – David for his endless supply of big ideas to add to the ever-growing ‘to do list’, and Keith for his wise and measured input. They have both been incredibly supportive, somehow managing to get me to this point, and I am so very grateful for the opportunities they have provided me with. I am also extremely thankful for Ian Gould’s assistance and ideas regarding the bacteriorhodopsin work.

Of course, I owe a huge debt of gratitude to the people with whom I worked on the 2DIR project – Paul Donaldson, Fred Fournier and Rui Guo. Everything that has been achieved to date can be attributed to an amazing team effort, and I am very grateful for the assistance and input provided by my 2DIR colleagues. A worthy acknowledgement must also go to Darek Kedra, without whom I would have become very stuck on the bioinformatics work. Importantly, many thanks go to the rest of the Klug Empire for providing such a great place to work over the past few years ….. Chris Barnett for being the most absurd person I have ever met; Je-Wen Liou for being my first mentor in the group; Laura Barter and Alison Stubbings for always being so kind and encouraging; Miriam Goyder for understanding the importance of warming the pot; Rafal Marszalek for his paradigm-shifting efforts; John Phillips for brightening my Christmases with The Mistletoes; Katy Gardner for always doing her washing up; and the many others who have also been supportive and so much fun – Gabriel Altschuler, Alex Cowan, Christian Loeffeld, Tita Miele, Tom Seaby, Junwang Tang and Bo Zhang.

Finally I must express my immense gratitude to my lovely family – my parents, my sister and Viscount Bosham (who will no doubt enjoy eating this) – for their support. And thank you to Ste, without whose help and encouragement I would never have got this thesis out.
Contents

Chapter 1: Introduction and thesis outline

1.1: Proteomics .................................................................17
  1.1.1: Proteomic goals and applications ..................................17
  1.1.2: Proteomic approaches and technologies .........................19
1.2: Protein identification techniques ..................................26
  1.2.1: Mass spectrometry-based protein identification ...............26
  1.2.2: Amino acid compositional analysis ..............................31
1.3: Optical spectroscopies in protein analysis .........................34
  1.3.1: Vibrational spectroscopies ....................................34
  1.3.2: Coherent multi-dimensional vibrational spectroscopies ........35
1.4: EVV 2DIR for protein analysis ..................................36
  1.4.1: EVV 2DIR protein fingerprinting strategy ........................37
  1.4.2: High-throughput protein identification with EVV 2DIR ......38
1.5: Thesis outline .....................................................39

Chapter 2: Electron-vibration-vibration two-dimensional infrared spectroscopy

2.1: Non-linear spectroscopy ............................................44
  2.1.1: Introduction to non-linear spectroscopy .....................44
  2.1.2: Basic theory of non-linear spectroscopy .....................45
2.2: Electron-vibration-vibration two-dimensional infrared spectroscopy ................47
  2.2.1: Introduction to EVV 2DIR ..................................47
  2.2.2: Principles of EVV 2DIR ....................................49
  2.2.3: EVV 2DIR properties and capabilities ......................52
  2.2.4: Electronically resonant EVV 2DIR .........................54
2.3: The EVV 2DIR spectrometer ......................................55
  2.3.1: The commercial laser system ..................................56
  2.3.2: The optical parametric amplifiers (OPAs) ..................57
  2.3.3: Delivery optics set-up ........................................59
  2.3.4: Signal detection and data collection .........................60
2.4: EVV 2DIR experimental methods ............................62
  2.4.1: Operational procedures .....................................62
  2.4.2: Imaging with EVV 2DIR ..................................63
Chapter 3: Protein identification and absolute quantification using EVV 2DIR

3.1: The EVV 2DIR protein fingerprinting strategy ................................................................. 64
3.2: Experimental methods and details ..................................................................................... 69
  3.2.1: Sample preparation ............................................................................................................. 69
  3.2.2: EVV 2DIR experimental methods ..................................................................................... 76
3.3: EVV 2DIR fingerprinting – proof of principle demonstration ............................................. 82
  3.3.1: Peptide EVV 2DIR spectra ................................................................................................. 82
  3.3.2: Spectral signatures of Tyr, Phe and CH$_2$ in peptides ....................................................... 84
  3.3.3: Quantification of amino acids levels in peptides .............................................................. 86
  3.3.4: Conclusions of peptide fingerprinting studies ................................................................. 91
3.4: Protein fingerprinting with EVV 2DIR ............................................................................... 92
  3.4.1: Protein EVV 2DIR spectra ................................................................................................. 93
  3.4.2: Spectral features of Tyr, Phe, Trp and CH$_3$ in proteins ................................................... 94
  3.4.3: Quantification of amino acid levels in proteins ................................................................. 96
  3.4.4: Conclusions of protein fingerprinting studies ................................................................. 98
3.5: Protein absolute quantification using EVV 2DIR .............................................................. 100
  3.5.1: Sample preparation and data acquisition .......................................................................... 101
  3.5.2: Absolute quantification procedure .................................................................................... 102

Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

4.1: Introduction ......................................................................................................................... 104
4.2: Protein databases and analysis tools ................................................................................... 106
  4.2.1: Protein databases ............................................................................................................... 108
  4.2.2: Search and analysis tools for protein databases ............................................................... 112
4.3: Databases for protein identification using EVV 2DIR ....................................................... 115
  4.3.1: Database construction ........................................................................................................ 115
  4.3.2: Database searching ............................................................................................................. 116
4.4: How many amino acid ratios to identify a protein? ............................................................ 118
  4.4.1: Narrowing down a search .................................................................................................. 118
  4.4.2: Combinations of amino acid ratios .................................................................................. 122
  4.4.3: Distribution of database hits ............................................................................................. 124
4.5: Database searches using EVV 2DIR data .......................................................................... 129
  4.5.1: Ranking of protein results ................................................................................................. 129
  4.5.2: Using the AAMplIdent tool with EVV 2DIR data ............................................................ 132
## Chapter 5: Electronically-resonant EVV 2DIR studies of bacteriorhodopsin

### 5.1: Electronically enhanced EVV 2DIR ................................................................. 135
   5.1.1: Initial studies of dye molecules ................................................................. 135
   5.1.2: Preliminary bacteriorhodopsin studies ...................................................... 137

### 5.2: Bacteriorhodopsin; structure, function and photocycles ................................. 139
   5.2.1: Structure and function .............................................................................. 139
   5.2.2: Light-adapted and dark-adapted bR ......................................................... 141
   5.2.3: The photocycles of bR<sub>568</sub> and bR<sub>548</sub> .............................................. 143
   5.2.4: Hydration level effects on the photocycles of bR ........................................ 150

### 5.3: ER EVV 2DIR spectroscopy of bR’s retinal chromophore ............................... 152
   5.3.1: Experimental methods and considerations .................................................. 152
   5.3.2: The ethylenic region ............................................................................... 154
   5.3.3: The fingerprint region .............................................................................. 158

### 5.4: Studying the intermediates of bR<sub>568</sub>’s photocycle using ER EVV 2DIR ........ 160
   5.4.1: Experimental methods and considerations .................................................. 161
   5.4.2: ER EVV 2DIR studies of bR<sub>568</sub>’s M and O intermediates ....................... 163

### 5.5: Theoretical studies of bacteriorhodopsin’s retinal ......................................... 168
   5.5.1: Theoretical methods and considerations ..................................................... 170
   5.5.2: Comparison of theoretical and experimental spectra ................................. 172

### 5.6: Sensitivity of electronically-resonant EVV 2DIR .............................................. 175
   5.6.1: Experimental methods and procedures ..................................................... 176
   5.6.2: Limit of detection determination ............................................................... 177
   5.6.3: Level of signal enhancement achievable .................................................... 178

## Chapter 6: Conclusions and future work

### 6.1: Protein identification and absolute quantification using EVV 2DIR .................... 181

### 6.2: Electronically-resonant EVV 2DIR studies of bacteriorhodopsin ....................... 183

### 6.3: Outlook for EVV 2DIR in bioanalysis ............................................................ 185
   6.3.1: Protein analysis ......................................................................................... 185
   6.3.2: EVV 2DIR as part of a high-throughput proteomic tool .............................. 189
   6.3.3: Future bacteriorhodopsin studies .............................................................. 190
   6.3.4: Measurement of post-translational modifications ........................................ 192
   6.3.5: Imaging with EVV 2DIR ......................................................................... 194
   6.3.6: Detection of protein-protein interactions ................................................... 195

## References ............................................................................................................. 195

## Publications .......................................................................................................... 211
List of figures

Chapter 1

Figure 1-1: A protein interaction map for *Drosophila melanogaster*, depicting ~30% of its proteins and ~10% of the interactions.

Figure 1-2: The typical distribution of protein abundances for *S. cerevisiae*.

Figure 1-3: Schematic explaining peptide mass fingerprinting (PMF).

Figure 1-4: Schematic explaining fragment ion analysis.

Figure 1-5: Schematic explaining amino acid composition analysis.

Figure 1-6: An FTIR spectrum of Concanavalin A, showing the broad features of the amide bond.

Figure 1-7: The procedure for high-throughput, on-line EVV 2DIR protein identification.

Chapter 2

Figure 2-1: EVV 2DIR spectroscopy is a four wave mixing (FWM) process, requiring the overlap of three pulsed laser beams – two infrared and one visible – on the sample.

Figure 2-2: An example EVV 2DIR spectrum of low density polyethylene.

Figure 2-3: The wave mixing energy level diagrams and coherence pathways depicting the four wave mixing EVV processes.

Figure 2-4: The pulse ordering used to select the EVV-IR 1 coherence pathway.

Figure 2-5: The WMEL diagram and coherence pathway for the EVV-IR 1 four wave mixing process, used for the work presented in this thesis.
List of figures

**Figure 2-6:** The commercial laser system used to provide the input for the optical parametric amplifiers that then generate the tuneable infrared beams.

**Figure 2-7:** The principles of optical parametric generation (OPG).

**Figure 2-8:** The optical set-up of the ‘EVV 2DIR spectrometer’ used to send the one visible and two infrared picosecond beams to the sample.

**Chapter 3**

**Figure 3-1:** Form of calibration curve to be produced using known protein samples; the ratios measured by EVV 2DIR ($\sqrt{I_{AA}} / \sqrt{I_{Ref}}$) are plotted as a function of the actual number ratios ($N_{AA} / N_{Ref}$) for each protein.

**Figure 3-2:** The side chains of the twenty amino acids. The label-free strategy presented in this chapter uses spectroscopic features of these to produce a novel type of protein fingerprint.

**Figure 3-3:** (a) Drop-coated deposits of LRRFFLG. (b) A microscope image of the edge of a drop-coated deposit of LRRYYLG.

**Figure 3-4:** Relative levels of FWM signal observed (at 1460 / 2980 cm$^{-1}$ and with delays of $T_{12} = 1.5$ ps, $T_{23} = 1.5$ ps) when a BSA drop-coated deposit was translated across the input laser beams.

**Figure 3-5:** Theoretical EVV 2DIR spectra of all the amino acid side chains; included are those for the three possible protonation states of histidine.

**Figure 3-6:** EVV 2DIR spectra of (a) glycine and (b) arginine crystals. Their infrared absorption spectra are also shown, with the positions of the EVV 2DIR cross peaks labelled.

**Figure 3-7:** EVV 2DIR spectra of the eight peptides used for the proof-of-principle study of the EVV 2DIR fingerprinting strategy.

**Figure 3-8:** Features identified in the peptide spectra as being suitable for performing the fingerprinting procedure.

**Figure 3-9:** A comparison of theoretical spectra of toluene and methyl-phenol, used as analogues for phenylalanine and tyrosine respectively, with the experimental spectra of LRRFSLGF and LRRYSLGY.
Figure 3-10: (a) Correlation between the known phenylalanine to CH$_2$ ratios ($N_{\text{Phe}} / N_{\text{CH}_2}$), and the ratios measured with EVV 2DIR ($\sqrt{I_{\text{Phe}}} / \sqrt{I_{\text{CH}_2}}$) of the eight peptide samples. (b) Correlation between the known tyrosine to CH$_2$ ratios ($N_{\text{Tyr}} / N_{\text{CH}_2}$), and the ratios measured with EVV 2DIR ($\sqrt{I_{\text{Tyr}}} / \sqrt{I_{\text{CH}_2}}$) of the eight peptide samples.

Figure 3-11: Example EVV 2DIR protein spectra of two proteins, covering the 1430 – 1550 cm$^{-1}$ / 2800 – 3200 cm$^{-1}$ region.

Figure 3-12: Measured amino acid to internal reference cross peak intensity ratios ($\sqrt{I_{\text{AA}}} / \sqrt{I_{\text{CH}_3}}$) plotted as a function of the known ratios ($N_{\text{AA}} / N_{\text{CH}_3}$) for the ten proteins analysed.

Figure 3-13: Differentiation maps of the ten proteins studied for (a) when the tyrosine peak only is used, and (b) when all four amino acid ratios are used.

Figure 3-14: EVV 2DIR image of five concanavalin A films mapped on the CH$_3$ cross peak at 1485 / 2930 cm$^{-1}$.

Figure 3-15: (a) Square root of the EVV 2DIR signal level as a function of the number of protein molecules that were probed. (b) The signal level per pixel of the concanavalin films plotted as a function of the number of molecules per pixel.

Chapter 4

Figure 4-1: The average percentage amino acid compositions across all sequences of the UniProtKB / SwissProt protein database.

Figure 4-2: (a) Distribution of hits that are returned when the ratios of three amino acids are input with a 10% accuracy interval for each entry in the human protein database of ~ 33,000 sequences. (b) Distribution of hits returned when, in addition to the three ratios, the molecular weight of each protein is also input at ± 10%.

Figure 4-3: (a) Distribution of hits that are returned when the ratios of five amino acids are input with a 10% accuracy interval for each entry in the human protein database of ~ 33,000 sequences. (b): Distribution of hits returned when, in addition to the five ratios, the molecular weight of each protein is also input at ± 10%.
Chapter 5

**Figure 5-1:** An example ER EVV 2DIR spectrum collected of a thin film of the red-absorbing dye, IR746.

**Figure 5-2:** The retinal chromophore located at the centre of the photosynthetic protein bacteriorhodopsin.

**Figure 5-3:** The seven trans-membrane helices of bacteriorhodopsin, with the retinal chromophore at their centre.

**Figure 5-4:** The conformations of the bR$_{568}$ (all-trans, 15-anti) and bR$_{548}$ (13-cis, 15-syn) retinal isomers.

**Figure 5-5:** (a) The absorption spectra of bR$_{568}$, bR$_{548}$ and a dark-adapted solution with a $\lambda_{max}$ at ~ 560 nm. (Reproduced from Sperling et al.) (b) The isomeric compositions of light-adapted and dark-adapted bacteriorhodopsin.

**Figure 5-6:** The photocycles of bR$_{568}$ and bR$_{548}$ – the details from a variety of static and time-resolved studies were combined to provide a full picture of both cycles.

**Figure 5-7:** The molecular mechanism of proton pumping in bacteriorhodopsin, reproduced from Kuhlbrandt.

**Figure 5-8:** The bR$_{568}$ photocycle providing details of the retinal’s intermediate conformations and the timescales over which they evolve.

**Figure 5-9:** The retinal intermediates of the bR$_{548}$ photocycle.

**Figure 5-10:** Photograph of dry purple membrane films deposited on a glass microscope slide.

**Figure 5-11:** Three example ethylenic region ER EVV 2DIR spectra measured of bacteriorhodopsin’s retinal chromophore.

**Figure 5-12:** Resonance Raman spectra of bacteriorhodopsin reproduced from Smith et al.

**Figure 5-13:** Two example fingerprint region ER EVV 2DIR spectra measured of bacteriorhodopsin’s retinal chromophore.
List of figures

**Figure 5-14:** An example sample cell used to maintain a PM film at a particular relative humidity whilst 2DIR spectra were measured.

**Figure 5-15:** Two example ethylenic region ER EVV 2DIR spectra of bacteriorhodopsin’s retinal measured of PM films at 65% RH.

**Figure 5-16:** Two example ethylenic region ER EVV 2DIR spectra of bacteriorhodopsin’s retinal measured of PM films at 100% RH.

**Figure 5-17:** Theoretical EVV 2DIR spectrum of the planar all-trans, 15-anti retinal protonated retinal Schiff base molecule, which serves as a model for the O$_{640}$ intermediate of the bR$_{568}$ photocycle.

**Figure 5-18:** EVV 2DIR images of (a) concanavalin A and (b) bacteriorhodopsin drop-coated deposits. The deposits are each labelled with the total number of molecules they contained.

**Figure 5-19:** (a) Square root of the EVV 2DIR signal level as a function of the number of protein molecules that were probed. (b) The signal level per pixel of the films, plotted as a function of the number of molecules per pixel for both concanavalin A and bR’s retinal.

**Chapter 6**

**Figure 6-1:** A comparison of the EVV 2DIR spectra of the peptide LRRYLYLG with its phosphorylated version, LRRpYpYLG.

**Figure 6-2:** EVV 2DIR and visible microscope images of a histological mouse kidney section.
List of tables

Chapter 3

Table 3-1: Mode assignments for the main cross peaks observed in the EVV 2DIR peptide spectra of Figure 3-7.

Table 3-2: The standard deviations for the dispersion of the data points in Figure 3-10 from linear behaviour.

Table 3-3: Mode assignments for the main cross peaks observed in the EVV 2DIR protein spectra of Figure 3-11.

Chapter 4

Table 4-1: A comparison between the numbers of entries in the UniProtKB-SwissProt, Ensembl and International Protein Index (IPI) databases.

Table 4-2: Average number of protein hits from the human protein database when it is queried using only a single amino acid / CH$_2$ ratio with accuracy levels of ± 5%, ± 10%, ± 15% and ± 30%.

Table 4-3: The average number of protein hits output when various pairs of amino acid ratios (with a ± 10% error range) are used to search the human protein database.

Table 4-4: The average number of protein hits output when combinations of two, three and four amino acid ratios (with ± 10% error ranges) are used to search the human protein database.

Table 4-5: Average numbers of hits output from the CH$_2$ and CH$_3$ ratio databases when searching with three and four amino acid ratios in combination.

Table 4-6: The similarity scores and rankings of the dataset of ten test proteins when searches where performed using EVV 2DIR measured ratios of (a) lysozyme and (b) BSA.

Table 4-7: A comparison of the two different methods of database searching.
List of tables

Table 4-8: The known amino acid ratios and molar percentages of lysozyme compared with the ratios measured with EVV 2DIR and the simulated molar percentages.

Table 4-9: The list of twenty proteins from the G. Gallus UniProtKB / SwissProt database that were returned from an AACompIdent search using simulated data of the protein lysozyme.

Chapter 5

Table 5-1: The conformations and protonation states of the four different retinal photocycle intermediates studied to date: the bR_568 and bR_548 isomers, and the M_{412} and O_{640} intermediates of the bR_{568} photocycle.

Table 5-2: Results of DFT calculations of the Raman and IR spectra of four different bacteriorhodopsin photocycle intermediates; bR_{568}, bR_{548}, M_{412} and O_{640}.

Table 5-3: Summary of sensitivity achievable with triply resonant EVV 2DIR, compared with that of the non-electronically enhanced case.
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2DE</td>
<td>Two-dimensional gel electrophoresis</td>
</tr>
<tr>
<td>2D-NMR</td>
<td>Two-dimensional nuclear magnetic resonance</td>
</tr>
<tr>
<td>AA</td>
<td>Amino acid</td>
</tr>
<tr>
<td>AA / CH₂</td>
<td>Ratio of number of amino acids to number of CH₂ groups in a protein</td>
</tr>
<tr>
<td>AA / CH₃</td>
<td>Ratio of number of amino acids to number of CH₃ groups in a protein</td>
</tr>
<tr>
<td>BBO</td>
<td>β-Barium borate</td>
</tr>
<tr>
<td>C / CHₓ</td>
<td>Ratio of number of cysteine residues to CHₓ groups in a protein</td>
</tr>
<tr>
<td>CaF₂</td>
<td>Calcium fluoride</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding sequence</td>
</tr>
<tr>
<td>CID</td>
<td>Collision-induced dissociation</td>
</tr>
<tr>
<td>CHₓ</td>
<td>CH₂ and / or CH₃ group</td>
</tr>
<tr>
<td>CMDVS</td>
<td>Coherent multi-dimensional vibrational spectroscopy</td>
</tr>
<tr>
<td>Cys / CHₓ</td>
<td>Ratio of number of cysteine residues to CHₓ groups in a protein</td>
</tr>
<tr>
<td>DA-bR</td>
<td>Dark-adapted bacteriorhodopsin</td>
</tr>
<tr>
<td>DCDR</td>
<td>Drop coating deposition Raman</td>
</tr>
<tr>
<td>DDJB</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>DESI</td>
<td>Desorption electrospray ionisation</td>
</tr>
<tr>
<td>DFM</td>
<td>Difference frequency mixing</td>
</tr>
<tr>
<td>DFT</td>
<td>Density functional theory</td>
</tr>
<tr>
<td>DOVE-FWM</td>
<td>Doubly vibrationally enhanced four wave mixing</td>
</tr>
<tr>
<td>DOVE-IR</td>
<td>Doubly vibrationally enhanced infrared spectroscopy</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionisation</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag</td>
</tr>
<tr>
<td>EVV 2DIR</td>
<td>Electron-vibration-vibration two-dimensional infrared spectroscopy</td>
</tr>
<tr>
<td>ExPASy</td>
<td>Expert Protein Analysis System</td>
</tr>
<tr>
<td>F / CHₓ</td>
<td>Ratio of number of phenylalanine residues to CHₓ groups in a protein</td>
</tr>
<tr>
<td>FT-ICR</td>
<td>Fourier transform ion cyclotron resonance mass spectrometry</td>
</tr>
<tr>
<td>FTIR</td>
<td>Fourier transform infrared spectroscopy</td>
</tr>
<tr>
<td>FWHM</td>
<td>Full width half max</td>
</tr>
<tr>
<td>FWM</td>
<td>Four wave mixing</td>
</tr>
<tr>
<td>H / CHₓ</td>
<td>Ratio of number of histidine residues to CHₓ groups in a protein</td>
</tr>
<tr>
<td>His / CHₓ</td>
<td>Ratio of number of histidine residues to CHₓ groups in a protein</td>
</tr>
<tr>
<td>HPLC</td>
<td>High performance / pressure liquid chromatography</td>
</tr>
<tr>
<td>IEF</td>
<td>Isoelectric focusing</td>
</tr>
<tr>
<td>IPI</td>
<td>International Protein Index</td>
</tr>
<tr>
<td>IR</td>
<td>Infrared</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LA-bR</td>
<td>Light-adapted bacteriorhodopsin</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>MALDI</td>
<td>Matrix assisted laser desorption / ionisation</td>
</tr>
<tr>
<td>MS / MS</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>Mw</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass-to-charge ratio</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>OPA</td>
<td>Optical parametric amplifier</td>
</tr>
<tr>
<td>OPG</td>
<td>Optical parametric generation</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>Phe / CH₅</td>
<td>Ratio of number of phenylalanine residues to CH₅ groups in a protein</td>
</tr>
<tr>
<td>pI</td>
<td>Isoelectric point</td>
</tr>
<tr>
<td>PIR-PSD</td>
<td>Protein Information Resource – Protein Sequence Database</td>
</tr>
<tr>
<td>PKU</td>
<td>Phenylketonuria</td>
</tr>
<tr>
<td>PM</td>
<td>Photomultiplier</td>
</tr>
<tr>
<td>PM</td>
<td>Purple membrane</td>
</tr>
<tr>
<td>PMF</td>
<td>Peptide mass fingerprinting</td>
</tr>
<tr>
<td>PPP</td>
<td>Parallel, parallel, parallel (laser beam polarisations)</td>
</tr>
<tr>
<td>PPS</td>
<td>Parallel, parallel, perpendicular (laser beam polarisations)</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>RH</td>
<td>Relative humidity</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulphate-polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SIB</td>
<td>Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SRS</td>
<td>Sequence Retrieval System</td>
</tr>
<tr>
<td>TAIR</td>
<td>The Arabidopsis Information Resource</td>
</tr>
<tr>
<td>TAP</td>
<td>Tandem affinity purification</td>
</tr>
<tr>
<td>TOF</td>
<td>Time-of-flight</td>
</tr>
<tr>
<td>TQ</td>
<td>Triple quadrupole</td>
</tr>
<tr>
<td>TrEMBL</td>
<td>Translated EMBL</td>
</tr>
<tr>
<td>Trp / CH₅</td>
<td>Ratio of number of tryptophan residues to CH₅ groups in a protein</td>
</tr>
<tr>
<td>Tyr / CH₅</td>
<td>Ratio of number of tyrosine residues to CH₅ groups in a protein</td>
</tr>
<tr>
<td>UniProtKB</td>
<td>Universal Protein Resource Knowledgebase</td>
</tr>
<tr>
<td>W / CH₅</td>
<td>Ratio of number of tryptophan residues to CH₅ groups in a protein</td>
</tr>
<tr>
<td>WMEL</td>
<td>Wave mixing energy level</td>
</tr>
<tr>
<td>Y / CH₅</td>
<td>Ratio of number of tyrosine residues to CH₅ groups in a protein</td>
</tr>
</tbody>
</table>
Chapter 1:
Introduction and thesis outline

1.1: Proteomics

1.1.1: Proteomic goals and applications

Proteins are functionally the most relevant components of biological systems. They are ultimately responsible for all cellular processes, which they achieve through interactions with each other, as well as with a huge variety of other biomolecules. The majority of biochemical reactions are catalysed by enzymes, whilst other proteins are involved with regulation and communication, or play important structural and mechanical roles. Proteins are also the most significant molecules therapeutically; they are the molecular targets for the majority of drugs and are used as biomarkers for diagnosing or predicting disease.

Although the fields of genomics and transcriptomics have contributed greatly to molecular biology research over the past twenty years, and have helped to provide a much more detailed understanding of how biological systems work, by studying genes or mRNA it is only possible to learn about protein function indirectly. Genes only supply us with the codes used by cells to manufacture its proteins, and do not provide us with information about how proteins function in cells, nor of their
abundances and locations. Using sequence information alone it is not possible to accurately predict a protein’s structure and complex molecular interactions and localisations, all of which contribute to determining its function. Most protein diversity and activity is generated through post-translational modifications and these too cannot be predicted from the transcript level.

As the number of fully sequenced genomes has rapidly grown, the focus of biological research has therefore shifted to the study of the proteins that they encode. Increased effort is now being directed at the identification and characterisation of the complete set of proteins produced by a given cell or organism - its proteome. This not only comprises the proteins initially manufactured, but also those that are produced as the proteins interact and undergo post-translational modifications. The proteome varies over time as the cell responds to changes in internal and external conditions by regulating its protein levels and activities. Proteomic experiments thus aim at the large-scale, systematic determination of the identity, structures, functions and interactions of the entire complement of proteins found in a particular biological system. The grand ambition of proteomics is to be able to achieve a comprehensive map of an organism’s components at the protein level, and to use this to form a complete picture of how its biological processes occur and are maintained.

Proteomic information is used to aid the discovery and analysis of protein networks, which enable global views of cellular processes to be formed. Through the monitoring of proteins, it is possible to elucidate how cellular pathways change in response to different biological conditions, disease states or drug actions. The emerging discipline of ‘Systems Biology’ makes much use of proteomic data, combining it with other global biological information to model and ultimately predict biological functions and events. While in the past molecular biology research
had focused on the isolation and characterisation of individual genes and proteins, the information produced by large scale genomic and proteomic projects has enabled the move away from a reductionist viewpoint to the systematic analyses of biological systems as a whole. Proteomic research is also now increasingly being applied to the field of biomarker discovery, with the aim of being able to distinguish which proteins represent healthy or disease states. It is hoped that the use of prognostic and diagnostic biomarkers will help medicine to move from reactive to more preventative and predictive approaches. Although the number of biomarkers in routine clinical use is presently low, this area shows great promise with biomarkers already being exploited to aid drug development.

1.1.2: Proteomic approaches and technologies

Proteomic approaches

The field of proteomics is diverse with many overlapping branches. Whatever form a strand of research takes, a central task will be the cataloguing of the proteins present in a particular sample. Proteins can then be analysed in a variety of contexts – some studies concentrate on determining sequence and structures, whereas others aim to elucidate the proteins’ localisations, interactions and modifications. In protein-expression profiling, the aim is to identify the proteins differentially expressed by a specific cell or tissue. The field of interactomics, on the other hand, concentrates on the analysis of protein interactions, both between themselves, leading to the formation of protein complexes, as well as with nucleic acids and small molecules. A recent yeast-interactome study, for example, identified nearly 2,800 interactions among 1,124 proteins, and it is estimated that in yeast, each protein ‘sees’ on average five other proteins. [1] The study of interactions not only helps to elucidate how individual proteins function, but also how they fit in with and
contribute to cellular pathways and networks. Figure 1.1 shows an example ‘protein interaction map’ for *Drosophila melanogaster*. [2, 3] Another important aspect of proteomics is the study of proteins that have undergone post-translational modifications (PTMs), leading to changes in their function. The most common PTM is reversible phosphorylation, which plays a vital role in controlling many complex biological processes – this has led to intense interest in the field of phosphoproteomics.

**Figure 1-1:** A protein interaction map for *Drosophila melanogaster*, depicting ~ 30% of its proteins and ~ 10% of the interactions. The proteins are colour-coded according to their families and interactions, and are also grouped by predicted sub-cellular localisation. Reproduced from *Giot et al.* [2, 3]
Proteomic technology

Proteome determination is an intensive and challenging task; there are vast numbers of proteins involved, all with hugely diverse molecular properties. Techniques are required that can achieve protein identification, quantification and characterisation on a large-scale and at high-throughput, and it is of the utmost importance that the results provided by these are reproducible and unambiguous. A wide range of proteomic technologies have emerged over the past decade, and these continue to be developed as the challenges of automation and integration are met. All of the techniques have their own strengths and weaknesses, and it is clear that no particular one is suitable for every application.

Traditional proteomic approaches were based upon the separation of complex protein mixtures using two-dimensional gel electrophoresis, 2DE. Typically this would involve isoelectric focusing (IEF) in the first dimension gel and sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) in the second, with protein identification being achieved using Edman sequencing or amino acid composition analysis (see Section 1.2.2). Gel electrophoresis is however not suited to large-scale studies and more recent methodologies have increased sample throughput by moving to capillary electrophoresis (CE) or multi-dimensional high pressure / performance liquid chromatography (HPLC) for separation. Undoubtedly though, the major breakthrough in proteomic technology came in the 1990s when mass spectrometry was adapted for protein identification, and algorithms were designed for searching protein databases with the acquired data. Mass spectrometry is now the method of choice for high-throughput protein identification, whether it is coupled with gel or capillary electrophoresis, liquid chromatography or affinity purification.
A number of different forms of mass spectrometry apparatus have been developed, which are used for both ‘bottom-up’ and ‘top-down’ proteomic approaches. The majority of mass spectrometry experiments, namely peptide mass fingerprinting, peptide fragment fingerprinting and \textit{de novo} sequencing, take the bottom-up strategy. This, for which a number of commercial instruments are widely available and used, involves the digestion of the proteins by an enzymatic or chemical process followed by analysis of the resulting peptides or their fragments. In top-down approaches, intact protein samples are analysed, providing the opportunity to study native proteins. Such techniques are at the forefront of advances in novel mass spectrometry, for example having recently been used to study the surfaces of living cells and tissues [4, 5]. The bottom-up methods have however underpinned proteomic research over the past ten years. A more detailed explanation of protein identification techniques, in particular those that are mass spectrometry based, is provided in Section 1.2.

Although mass spectrometry may be the workhorse for protein identification, a number of other bioanalytical tools are used for their characterisation. Structural proteomics generally employs techniques such as x-ray crystallography, cryo-electron microscopy and multi-dimensional nuclear magnetic resonance spectroscopy (2D-NMR) to study protein three-dimensional structures. The aim of functional proteomics is to directly test the functions of proteins on a large-scale, for which high-throughput assay chips are employed. Novel antibody microarray technologies can now detect and identify proteins on the single molecule level via fluorescence labelling. Interactomic research tends to entail the use of large-scale yeast two-hybrid experiments; most recently, tandem affinity purification (TAP) approaches have proved particularly useful. [6] Finally, protein localisations can be determined using electron or the various light microscopies to image cells.
The enormous complexity of biological samples makes dynamic range and sensitivity key issues in proteomics. Although the human genome comprises ~30,000 genes, in a cell this may represent more than 100,000 distinct proteins when post-translational modifications and alternative splicing are taken into consideration. The proteins have vastly diverse molecular properties and are present at a wide range of concentrations, with the dynamic range of protein abundances within cells ranging from $10^5$ (for tissues) up to $10^9$ (for body fluids such as serum). Ideally, a proteomic technique therefore needs to be capable of operating over these large dynamic ranges. The isolation and identification of low-abundance proteins is also particularly difficult – for single cell studies, highly sensitive techniques are required that are able to pick out these from the high abundance proteins under which they are usually hidden.

One of the main factors contributing to the success of proteomic research has been the incredible advances in bioinformatics that have occurred concurrently with those in technology. The handling of the vast amounts of data produced is alone a huge feat, and much bioinformatics concentrates on the construction of databases and the standardisation of their data. Equally important and labour intensive is the creation of interfaces and gateways for the databases, so that they can be easily accessed and mined. The major challenge in proteomics however, for which the work of bioinformaticians is key, is the refinement and manipulation of data in order to extract biologically meaningful information and insight.

Whilst it is apparent that during the past decade proteomics has benefited from incredible technological advances, with a wealth of bioanalytical techniques having been made available to life scientists, even the most widely used of these have their limitations. 2D-NMR, for example, is currently unparalleled in its ability for determining the structure of proteins in solution, but it is a very low sensitivity
technique and so concentrated samples are required. Advances in mass spectrometry have led to some methods being able to detect and reliably identify proteins at the femtomole level (corresponding to \( \sim 10^{8} \) protein molecules). Although mass spectrometry may therefore be considered a high sensitivity method, it still falls short of the requirements for performing single cell analyses. Less than 5% of the proteins in a cell are present at copy numbers of more than \( 10^{5} \); in fact more than 50% of proteins are present at levels of only \( 10^{3} \) to \( 10^{4} \) copies. [7] Figure 1-2 shows the typical distribution of protein abundances for yeast cells. Mass spectrometry also suffers from poor a dynamic range; the most up-to-date techniques can only cover concentration ranges of \( 10^{6} \), with most operating over a range of \( 10^{4} \). Additionally, the fragmentation processes used are not always reproducible and it is a destructive technique, meaning that post-analysis sample recovery is not possible. Other technologies also suffer from high rates of false-positive and false-negative results. This is particularly a problem for the yeast two-hybrid system used for the detection of binary protein interactions.

Figure 1-2: The typical distribution of protein abundances for *S. cerevisiae*. About 25% of the proteins are present at levels of less than \( 10^{3} \) copies, and 50% are present at \( 10^{3} \) to \( 10^{4} \) molecules per cell. Data taken from Sem. [7]
Novel proteomic technologies

Along with the continued development of the established proteomic methodologies, there is clearly a need for new technical disciplines to emerge which can offer improved sensitivity, resolution and repeatability. The broad range of applications for proteomic information, from biomarker discovery and clinical diagnostics to the provision of data for systems biology and fundamental biological research, provides strong motivation for the increased effort now being directed at the development of novel proteomic tools with greater capability. Although some of these may challenge the traditional approaches, it is most likely that they will instead complement them. Different approaches will always have their own limitations, and by using them together it is possible to obtain distinct, but overlapping sets of data which provide more information than one method can in isolation.

The work presented in this thesis involves the development of a form of multidimensional optical spectroscopy as a novel tool for protein analysis, that it is hoped will be able to complement the established approaches. To date this technique has been used to investigate enzyme mechanisms (Chapter 5), and initial investigations into protein post-translational modifications and protein-protein interactions using it have proved promising (Chapter 6). The initial focus of the development of this technique, however, was on its capability for protein identification and quantification (Chapter 3). The following section therefore contains a brief explanation of the most widely used methods of protein identification using mass spectrometry, as well as of a largely superseded method, amino acid compositional analysis, on which the new optical fingerprinting strategy is loosely based.
1.2: Protein identification techniques

Although the overarching aim of many proteomic experiments may be to catalogue every single protein expressed in a particular cell, technologies are not yet far enough advanced to enable this. Typically smaller sets of proteins, isolated in specific functional contexts – sub-proteomes – are selected for systematic analysis. Before the advent of mass spectrometry for protein identification, this was normally achieved through separation by gel electrophoresis followed by Edman sequencing or amino acid composition analysis. A number of different mass spectrometry approaches are now routinely used, which all in some way involve the identification of proteins from their mass spectra using protein sequence databases. Prior to analysis, separation tends to be accomplished using high-throughput electrophoretic or chromatographic techniques.

1.2.1: Mass spectrometry-based protein identification

For mass spectrometric analysis, protein or peptide analytes are ionised, and measurements of the mass-to-charge ratios (m/z) of their ions are carried out in the gas phase. The two most commonly used techniques for the volatisation and then ionisation of the analytes are matrix-assisted laser desorption / ionisation (MALDI) and electrospray ionisation (ESI). [8, 9] For measuring the m/z ratios of the protein or peptide ions, a mass analyser is used, of which there are four basics types – the triple quadrupole (TQ), time-of-flight (TOF), ion trap and Fourier transform ion cyclotron resonance (FT-ICR) analysers.

In MALDI-MS, laser pulses are used to sublimate and then ionise intact peptide or protein samples from a dry, crystalline matrix. MALDI is typically coupled with a TOF analyser, and such MALDI-TOF systems are predominantly used for the
analysis of single peptides or relatively simple mixtures, such as those derived from a single spot of a 2DE gel. ESI techniques ionise the analytes out of solution and so are easily integrated with an online liquid-phase separation method, usually liquid chromatography (LC) or sometimes capillary electrophoresis (CE). ESI instruments are mostly used with triple quadrupole and ion trap mass analysers to generate spectra of peptide fragment ions – so-called collision-induced (CID) spectra. The most commonly employed system, tandem MS coupled with LC (LC-MS/MS), is used for the analysis of complex peptide mixtures. Other configurations of ion sources and mass analysers are now starting to emerge for protein analysis, and may in the future provide higher sensitivity, resolution and accuracy than the two established methods explained here. [8-10]

**Bottom-up mass spectrometry approaches**

There are two main approaches for protein identification with mass spectrometry:

(i) *Peptide mass fingerprinting (PMF)*, also referred to as peptide mapping, typically uses MALDI-TOF to analyse the intact peptide ions of single proteins or simple mixtures.

(ii) *Peptide fragment ion analysis* typically uses LC-MS/MS to analyse the fragmented peptide ions of complex protein mixtures.

Both of these are referred to as bottom-up proteomic schemes, and achieve protein identification by matching measured mass spectra with the theoretical mass spectra of database proteins. The experimental processes involved are summarised in the schematics of Figures 1-3 and 1-4 respectively.

Even though both PMF and fragment ion analysis may be considered reliable protein identification methods, there are a number of factors that can result an
incorrect protein match. Firstly, mis-cleavages can occur during the digestion step; some cleavages may not reach completion, whereas others may occur spuriously at other sites than expected. Mass changes to the proteins following post-translational modifications should be predicted by the search algorithms, though those that can occur during the extraction and separation processes may interfere with an identification. The inference of protein identify from a peptide hit is not always straightforward either; a peptide can be derived from many different proteins, including variants due to single nucleotide polymorphisms (SNPs). These methods also rely on having a high quality database with good coverage to search; any errors in the sequences will lead to incorrect theoretical spectra being generated. One way that PMF data can be improved is by using orthogonal data sets – by carrying out separate sample digestions using two or more different proteases, datasets for the same protein can be compared. Identifications by fragment ion analysis can be improved by the use of sequence tags.

**Top-down mass spectrometry approaches**

Although the two bottom-up approaches explained above have become established as the primary workhorses in proteomics, top-down mass spectrometric methods are beginning to emerge. For example, MALDI and desorption ESI (DESI) techniques have recently been developed for directly analysing the surfaces of living cells and tissues. Using these approaches, it is possible to measure the distributions across a sample of components with specific m/z values. The mass spectrometry image can then be superimposed onto the corresponding optical image for comparison. [4, 5, 11-14] Comprehensive assessments of mass spectrometry based proteomics can be found, for instance, in the reviews of Aebersold. [8-10]
Figure 1-3: Peptide mass fingerprinting (PMF), also referred to as peptide mapping, identifies proteins by virtue of the masses of their intact constituent peptides. The required sample is a single purified protein, or a simple mixture, most often an individual spot excised from a 2DE gel. The sample is digested with a cleavage reagent, for example trypsin, that cleaves the protein at specific sites, and the peptides produced are then submitted for MALDI-TOF analysis to determine their masses. The protein that the peptides are derived from is identified by correlative searching of a protein database. A search algorithm first carries out an *in silico* digestion of all the database entries, at the same places as the cleavage reagent of choice. The database, now comprised of all the theoretical peptide masses of the proteins, is compared with the experimentally-determined peptide masses of the sample. A list of protein hits is returned, in which each database protein is ranked in order of how well it correlates with the measured data. [8, 15-17]
**Figure 1-4:** Fragment ion analysis is typically used for the identification of complex protein mixtures which have been isolated from cell lysate or tissue by biochemical fractionation or affinity selection. Following digestion of the proteins using a protease such as trypsin (as for PMF analysis), the resulting peptide mixtures are separated using one or more steps of liquid chromatography and are then eluted into an ESI source. Following ionisation, two mass analysers are used in series (so-called tandem mass spectrometry, MS/MS) to create fragment ions from the individual peptides. The CID spectra produced by LC-MS/MS can be used in two different ways to achieve protein identification – by carrying out database mass searches, or by partial *de novo* sequencing. Using search engines such as *Sequest* [18] or *Mascot* [16, 19], the observed peptide fragmentation spectra are compared with theoretical fragmentation patterns derived from the peptides of each of the proteins in a database. Scores are calculated which represent the statistical significance of a match between the experimental spectrum and a predicted spectrum, and a list of potential peptide matches is produced. These are then finally associated with the proteins from which they are derived. As well as mass information, CID spectra can provide peptide sequence details. Although it can’t be used to derive a full peptide sequence *de novo*, the pattern of a CID spectrum can be used to create a short unambiguous amino acid sequence for a peptide. Such sequence tags can be used in combination with the mass spectrum to carry out database searching. By combining sequence and mass information, CID spectra provide more clear-cut identifications than those from PMF. [8, 15-17]
1.2.2: Amino acid compositional analysis

Before being superseded by high-throughput mass spectrometric protein fingerprinting, amino acid compositional analysis was an essential method for identifying proteins. [17, 20-25] It is still however routinely used for the identification of low mass proteins, for which mass spectrometry is not always suitable, or of those separated by 2D gel electrophoresis. The amino acid composition is defined as the molar percentages of each of the different residues, and is characteristic for each peptide and protein. As well as being capable of protein identification, amino acid compositions can be used for predicting protein properties such as secondary structure content [26, 27] or relatedness to other proteins. [28] Although no longer used for proteomic studies, amino acid analysers are still widely used in hospitals to diagnose inherited metabolic disorders; plasma and urine amino acid profiles give the best indication of the presence of these. The most common is phenylketonuria (PKU) which leads to an accumulation of phenylalanine in the blood, and untreated leads to mental retardation. All babies are now screened at birth for PKU by a simple test of the percentage composition of phenylalanine in their blood.

The experimental scheme for composition determination involves hydrolysis of the protein substrate into its constituent amino acids, followed by separation and derivatisation of these before quantification can occur. Modern amino acid analysers can be programmed to carry out both the separation and quantification steps. Amino acid composition based identification is performed using the AACompIdent algorithm provided by ExPASy. [29, 30] This compares the experimentally determined compositions of the target protein to the entries of a database – a list of hits is produced of the proteins whose compositions are most closely matched to that
measured of the target. The experimental methods of compositional analysis are explained more fully in Figure 1-5.

Amino acid compositional analysis is capable of detection down to the picomole (~10^{11} molecules), making it considerably less sensitive than mass spectrometry methods. It is also a destructive technique, involving complex and time-consuming processes – the hydrolysis step alone takes 1 to 24 hours. Moreover, the hydrolysis procedure, for which no single method for all residues exists, is the least controllable part of the process and is the major cause of incorrect identifications. There are a number of ‘non-reliable’ residues, particularly those of low-abundance, whose quantification is imprecise. Cleavage of bonds between aliphatic residues is difficult to achieve, whilst a number of amino acid side chains are themselves hydrolysed when acid-hydrolysis is used. When performing database searches using AACompIdent, identification can be aided by including other properties in a search. These include mass and isoelectric point, or a short sequence tag, which can be determined prior to compositional analysis using 2DE and Edman sequencing respectively. [21, 25]
Figure 1-5: Schematic explaining the procedure for amino acid composition analysis. Individual proteins are taken from 2DE gels and then hydrolysed into their constituent amino acids, typically by boiling in highly concentrated HCl for about 24 hours. The free amino acids are labelled with a chromophore or fluorophore (for example ninhydrin or fluorescamine) and then separated by HPLC. Using a reference panel, the amino acids are detected and quantified as they are eluted from the column. The height of a particular residue’s peak is proportional to its molar percentage in the protein. Using the AACompIdent tool [29, 30], a protein’s amino acid composition is compared to those of the proteins in the UniProt / SwissProt database [31-33]. A list of hits, representing the proteins that are the best-match to the measured data, is returned. [17, 21, 25]
Chapter 1: Introduction and thesis outline

1.3: Optical spectroscopies in protein analysis

1.3.1: Vibrational spectroscopies

Although optical spectroscopies appear to have significant potential for protein analysis, the conventional approaches such as infrared absorption and Raman scattering suffer greatly from spectral congestion. The extraction of structural information from the linear vibrational spectra of small molecules can be relatively straightforward, but this is not the case for the complex molecules and mixtures that are found in biological systems. Large molecules like proteins have thousands of bonds and possible vibrational modes, and their spectra therefore consist of broad bands comprised of many overlapping peaks. Additionally, much of the desired data is hidden by the most dominant spectral features, namely the strong amide modes of the protein backbone (Figure 1-6). The extraction of useful information, in particular the assignment of specific modes, from such over-congested spectra can be extremely difficult and imprecise.

![Figure 1-6: An FTIR spectrum of concanavalin A, showing the broad features of the amide bond. The mode assignments are as follows: Amide I: CO stretching, Amide II: NH bending and CN stretching, Amide III: CN stretching and NH bending, Amides A & B: NH stretching in resonance with first amide II overtone.](image)
Despite the limitations, there are ways in which vibrational spectroscopies can be employed in bioanalysis. To a certain extent spectral congestion can be alleviated though either post hoc analyses of spectra, or through the controlled variation of measurement parameters. Protein secondary structure determination from the Amide I band using multivariate analysis methods, for example, is a long-established technique. [34-38] It is also possible to extract information about specific amino acids or individual bonds through the use of reaction-induced difference spectroscopy and/or isotopic labelling [39-43]. The problems caused by the strong water absorptions that obscure mid-infrared protein spectra can now be overcome by using the widely-employed attenuated total reflectance (ATR) technique. More recently, a particular form of Raman spectroscopy has successfully been applied to the measurement of protein phosphorylation levels. [44, 45]

1.3.2: Coherent multi-dimensional vibrational spectroscopies

It is clear that the one-dimensional vibrational spectroscopies do have something to offer with regards to protein analysis. In general though the methods for overcoming spectral congestion involve time-consuming and complicated data collection and analysis, and even then only achieve this to a limited degree. In the past ten years, various multi-dimensional infrared spectroscopic techniques have therefore emerged as alternatives to the conventional approaches. [46-49] Coherent multi-dimensional vibrational spectroscopies (CMDVS) are sometimes referred to as optical analogues to multi-dimensional NMR, since their underlying principles are broadly similar – whilst two-dimensional NMR measures coupled nuclear spins, CMDVS probe coupled vibrations. These spectroscopies have become practical and realisable due to advances in ultrafast laser technology. By using sequences of short optical pulses which select only coupled vibrational modes, these techniques spread out the spectral information over two or more dimensions. A central motivation behind the
development of these multi-dimensional methods has therefore been the decongestion and simplification of spectra that they can afford. This is a particularly attractive and powerful attribute for protein analysis.

Several forms of coherent multi-dimensional spectroscopy techniques using different optical pulse sequences and combinations have emerged as useful tools. Whilst each of these methods have particular strengths and capabilities, they all benefit from spectral decongestion, high sensitivity and fast timescales. The most widely used two-dimensional infrared (2DIR) spectroscopies are the pump-probe [50] and three pulse photon echo [51-56] techniques. These have successfully been applied to the investigation of structure and dynamics in a range of systems containing amide, carbonyl and hydroxyl groups. There have been numerous examples of the application of the photon echo approach to the study of proteins, peptides and DNA, including the structure determination of peptides through the backbone amide I mode [57-64], and the observation of hydrogen-bond dynamics [62, 63, 65-70]. For the work presented in this thesis, however, an alternative 2DIR method was adopted and its utility for a range of biological applications explored. The approach, originally demonstrated by J. C. Wright and colleagues [49, 71-84], is known as Electronic-Vibration-Vibration Two-Dimensional InfraRed Spectroscopy, EVV 2DIR. The theory, principles and practicalities of EVV 2DIR are covered in more detail in Chapter 2.

1.4: EVV 2DIR for protein analysis

Any new proteomic technology needs to be capable of delivering high-throughput and reliability and, ideally, be able to combine the sensitivity and selectivity of mass spectrometry with NMR’s ability to provide detailed structural information. To this
end, D. R. Klug’s research group at Imperial College London is currently pursuing EVV 2DIR as a spectroscopic technique for bioanalysis. EVV 2DIR is a third order non-linear spectroscopy that probes only the vibrational modes of a sample that are coupled to one another in some way, thus spreading out the spectral information into two dimensions. It is this ability for spectral decongestion, and consequently easier interpretation, that makes EVV 2DIR a viable technique for protein analysis. Using this approach, the congestion of protein infrared spectra is relieved sufficiently to allow assignment of vibrational features to specific amino acid side chains. The work covered in this thesis demonstrates two particular uses for the technique: as a protein identification and quantification tool and as a method for studying enzyme mechanisms.

1.4.1: The EVV 2DIR protein fingerprinting strategy

The proposed protein fingerprinting strategy presented here is based on using EVV 2DIR to determine the amino acid compositions of proteins. This novel form of compositional analysis establishes amino acid content by only measuring spectroscopic features of the proteins. A particular EVV 2DIR cross peak is used as an internal reference, and its intensity is ratioed with the intensities of the cross peaks of the different residues. Identifications are achieved by comparing the spectroscopically-determined compositions, the amino acid / reference ratios, with those of the proteins in a database. EVV 2DIR is also able to perform absolute quantification of protein levels, something of major importance in the field of proteomics but rather difficult and time consuming to achieve with mass spectrometry. Whilst the initial goal for this technique is to be able to perform protein identification and quantification, in the future EVV 2DIR will be used to analyse the identified proteins in greater detail. Research into the probing of enzyme mechanisms has proved fruitful, and it is also already being used as an imaging
technique. Initial studies into the extent and type of protein post-translational modifications using the technique augur well for advancements in this important area of proteomics. Additionally, EVV 2DIR has the potential for studying protein-protein interactions, for which preliminary investigations are underway.

Protein analysis with EVV 2DIR can be qualified as a top-down label-free method. It does not require intensive sample preparation – no chemical or biochemical modification of the proteins is needed. There are no mass restrictions on the proteins, and they remain intact when analysed. Moreover, EVV 2DIR is a non-destructive technique, so the samples can be kept for reanalysis in the light of further information. This spectroscopy also benefits from potentially high sensitivity; it employs high intensity lasers and the measured output is visible light, the detection of which is far more efficient than of that in the infrared. Since it is an ultrafast method, EVV 2DIR may be useful for exploring, for example, protein folding dynamics. While NMR has only limited capabilities for this due to its millisecond time resolution, EVV 2DIR operates on picosecond timescales.

1.4.2: High-throughput protein identification with EVV 2DIR

Ultimately, the goal is to make the EVV 2DIR apparatus part of a suite of technologies for label-free high-throughput proteomics. Other sections of the Klug research group are concentrating on the development of separation techniques to be coupled with the 2DIR spectrometer. The aim is to perform two-dimensional separations, most probably HPLC followed by CE, of complex protein mixtures prior to 2DIR analysis. Using electrodeposition, the separated proteins eluted from the CE can be written onto stainless steel plates, ready to be submitted for protein identification. [85] These protein samples are not destroyed by the 2DIR procedure and can be reanalysed over and over again. The EVV 2DIR protein fingerprinting
procedure is outlined in the schematic of Figure 1-7. As for any bioanalytical technique, EVV 2DIR may be expected to excel for use in some applications, whilst be of limited use for others. Any new technology will not be able to compete on every front with those already established; rather huge benefits are gained from having the approaches complement each other. By using various techniques to consider a biological problem from different viewpoints, added insight can be gained. Complementary datasets provide much more information than one in isolation.

1.5: Thesis outline

The research presented in this thesis is part of a much larger EVV 2DIR project that encompasses many different and overlapping aspects of protein analysis. As such, much of the work was completed in conjunction with other members of the EVV 2DIR team at Imperial College London. In particular, the 2DIR experiment had already been built, and the first meaningful spectra recorded, by P. M. Donaldson before the work presented here commenced. Subsequent developments and alterations to the apparatus were largely performed by P. M. Donaldson, F. Fournier and C. J. Barnett. The proof of principle protein fingerprinting results covered in Chapter 3 were achieved jointly with F. Fournier, whilst all theoretical spectra and mode assignments were provided by R. Guo and I. R. Gould. The Bioinformatics Support Service at Imperial College London [86], specifically D. Kedra, enabled the bioinformatics studies outlined in Chapter 4.

Chapter 2 gives a brief introduction to the basis and principles of EVV 2DIR spectroscopy. The design of the laser system that comprises the `2DIR
Figure 1-7: The procedure for high-throughput, on-line EVV 2DIR fingerprinting. A complex protein mixture (requiring no chemical or biochemical preparation steps) undergoes a two-dimensional separation process involving HPLC followed by CE. The separated proteins are eluted from the CE instrument and are written onto stainless steel plates by an electro-deposition procedure. The individual protein spots are then submitted for EVV 2DIR analysis for their amino acid compositions to be measured. These are expressed in terms of the ratios of the numbers of each of the residues to the total number of \( \text{CH}_3 \) groups in the proteins, AA / CH\(_3\). These ratios are used to search a database comprised of proteins’ AA / CH\(_3\) ratios, and a number of hits are returned with scores indicating the degree of similarity between them and the target protein.
Chapter 1: Introduction and thesis outline

spectrometer’ is explained, and details of the experimental procedures are given. More in depth explanations of the theoretical background to EVV 2DIR are beyond the scope of this thesis and can be found in the publications of J. C. Wright [49, 71-81, 87, 88] and D. R. Klug [89-94]. Additionally, the PhD thesis of P. M. Donaldson provides comprehensive descriptions of the apparatus used. [95-98]

Chapter 3 presents the proof of principle of the EVV 2DIR protein fingerprinting strategy, based on determining amino acid compositions by measuring spectroscopic features of the samples. The identification method, in particular the practicality of using an internal reference to achieve relative quantification of amino acid levels, was initially demonstrated in peptides. Subsequently the approach was applied to a set of ten proteins and used to differentiate them. Also explained in this chapter is how EVV 2DIR can be used to achieve absolute quantification of protein levels.

Chapter 4 covers the bioinformatics work carried out concurrently with the experimental validation of the protein identification method. This included the creation of protein databases (and their interfaces) suitable for searching with EVV 2DIR compositional data. The results of tests are presented which demonstrate that protein identification can be achieved using spectroscopically measured amino acid ratios. Finally, initial work into the development of a system for scoring protein database results is explained.

Chapter 5 is concerned with the use of triply resonant EVV 2DIR spectroscopy, a method that can achieve an increase in sensitivity of several orders of magnitude over the usual doubly-resonant technique. This was demonstrated using the retinal chromophore at the centre of the photosynthetic protein bacteriorhodopsin. The first half of this chapter therefore provides an introduction to the structure and functions of this protein.
Finally, in Chapter 6, the achievements to date regarding the implementation of the protein fingerprinting strategy are summarised, along with the findings of the bacteriorhodopsin study. How this area of research is to be carried forward to achieve the first protein identification with EVV 2DIR spectroscopy is outlined. Other forms of protein analysis that are also being developed are explained – these include imaging, the measurement of protein phosphorylation levels and the study of protein-protein interactions.
Chapter 2: Electron-vibration-vibration two-dimensional infrared spectroscopy

Over the past decade several forms of two-dimensional infrared (2DIR) spectroscopy have been developed, including the photon-echo technique and infrared pump-probe spectroscopy. The method adopted here, however, for protein analysis is electron-vibration-vibration two-dimensional infrared spectroscopy (EVV 2DIR), also known as doubly vibrationally enhanced four-wave mixing (DOVE FWM). This chapter first briefly covers the basic theory and capabilities of this non-linear spectroscopy, before explaining the experimental set-up of the apparatus used for the protein fingerprinting and bacteriorhodopsin results presented in Chapters 3 and 5 respectively.

The overarching aim of this project was the application of EVV 2DIR spectroscopy to proteomic research. Whilst this did require an in-depth understanding of the technique, other members of the research team were responsible for concentrating on its technical development. Therefore only an introduction to this spectroscopy is presented here, and far more detailed explanations of its underlying principles can be found in the numerous publications of J. C. Wright [49, 71-81, 87] and D. R.
Klug [89-94]. Additionally, the theses of P. M. Donaldson [95-98] provide comprehensive descriptions of the construction of the instrument used, and the steps taken to achieve the first meaningful EVV 2DIR spectra.

2.1: Non-linear spectroscopy

2.1.1: Introduction to non-linear spectroscopy

Spectroscopy is based on the molecular polarisations that are induced in a sample by the electric field component of electromagnetic radiation fields. At the low field intensities used in the most commonly encountered optical spectroscopies, the dependence of the induced polarisation on the applied field is linear. With the high field intensities produced by pulsed lasers, however, the polarisation acquires distortions, creating new fields that radiate from it. In these cases, the relationship between the polarisation and the applied field becomes non-linear. Non-linear spectroscopy is therefore performed by using multiple laser beams to induce a non-linear polarisation in a sample, then observing the new fields that emerge. The generated field can be measured as a function of a variety of properties of the input fields, for example the amplitude, phase, time or frequency. In $n$th order non-linear spectroscopy, $n$ input beams are focused into a sample. The new field that is launched by the non-linear polarisation is the detected $(n + 1)^{th}$ beam, and so these techniques are referred to as $(n + 1)$-wave mixing.

When the input beams excite electronic or vibrational resonances in a sample, the output field is enhanced. This is the basis of frequency domain non-linear spectroscopy, of which EVV 2DIR is an example. In general, much larger enhancements occur when the applied fields are in resonance with electronic, rather
than vibrational, states. However, if there are vibrational modes that are resonant with the beams and the electronic states are non-resonant, then the vibrational enhancements will dominate. When more than one state is resonant the enhancements gained from these resonances are additive and, more importantly, when these states are coupled the enhancements are multiplicative.

### 2.1.2: Basic theory of non-linear spectroscopy

Classically, the polarisation, $P$, induced in a sample may be expressed by a series expansion in powers of the net input electric field, $E$ (the sum of all the electric fields of the light beams exciting the sample):

$$ P = \varepsilon_0 \left( \chi^{(1)} E + \chi^{(2)} EE + \chi^{(3)} EEE + \ldots \right) $$

Equation 2-1

The electric susceptibility, $\chi$, is a constant of proportionality that represents how easily the sample polarises in response to the applied field. $\chi^{(n)}$ are the $n^{th}$-order components of this characteristic susceptibility – molecular polarisation is dependent upon the orientation of all of the incident fields and so the $\chi^{(n)}$ terms have tensor character. At low field intensities the higher order terms in this expansion are negligible and the first term is dominant. In such cases there is a linear relationship between the induced bulk polarisation and the strength of the incident electric field:

$$ P = \varepsilon_0 \chi^{(1)} E $$

Equation 2-2

At the very high field intensities produced by pulsed laser systems, however, the higher order terms in the expansion become significant, leading to non-linear
responses in the sample. The induced polarisation is thus non-linear with respect to the input electric field. In \( n \)th order spectroscopy, the \((n + 1)\)-wave mixing process is used to probe \( \chi^{(n)} \). In third order non-linear spectroscopies, for example, \( \chi^{(3)} \) is probed, with the first two fields modifying the optical properties of the sample that the third field then experiences.

Quantum mechanically, a molecule’s polarisation is created by the transition dipole moment of its coherence. A coherence may be considered a quantum mechanical entanglement of states, created when they are perturbed by an applied electromagnetic field. [49] For example, if \( \psi_a \) and \( \psi_b \) are two quantum states of a molecule, when an electromagnetic field perturbation mixes the two states the superposition of these generates a coherence (where \( c_a(t) \) and \( c_b(t) \) are the time-dependent amplitudes for the molecules being in states \( a \) and \( b \)):

\[
\psi(t) = c_a(t)\psi_a e^{i\omega_a t} + c_b(t)\psi_b e^{i\omega_b t}
\]

Equation 2-3

The probability of finding the molecule in state \( a \) or state \( b \) is given by \( \langle \psi | \psi^* \rangle \), with the populations and coherences often being manipulated as elements of a density matrix. The diagonals, here \( c_a(t)c_a^*(t) \) and \( c_b(t)c_b^*(t) \), give the populations of the states, and the off-diagonals, \( c_a(t)c_b^*(t) \) and \( c_b(t)c_a^*(t) \), represent the coherences – the quantum mechanical superpositions of the states. The coherences are time-dependent states and have dipole moments that oscillate at a frequency of \( (\omega_a - \omega_b) \).

An oscillating coherence couples with a field, absorbing or emitting light depending upon its phase relative to the field. In linear spectroscopy, refraction and absorption are thus described by the real (the part of the polarisation that is in-phase with the field) and imaginary (the out-of-phase part) terms of a coherence. In non-linear spectroscopies, interaction with other fields leads to further evolution of the
coherence through further transitions of $\psi_a$ or $\psi_b$ to other states. If the final coherence has a dipole moment, the non-linear polarisation that is associated with the oscillating charge launches an output field in a process analogous to refraction. It is this signal field which is detected in non-linear spectroscopies.

2.2: Electron-vibration-vibration two-dimensional infrared spectroscopy

2.2.1: Introduction to EVV 2DIR

EVV 2DIR signal is measured as a function of two infrared frequencies, with the spectral information spread out into a two-dimensional landscape. The cross peaks that appear in EVV 2DIR spectra represent only coupled vibrational modes of the sample, resulting in the spectra being significantly simplified compared to their linear vibrational counterparts. Initial EVV 2DIR studies used small model systems, such as acetonitrile and CS$_2$, to test the technique, and coupling between two vibrations was first demonstrated between the CC and CN stretching modes of acetonitrile. [78, 80] The work presented here demonstrates that, by spreading the spectral information into two dimensions and only probing coupled vibrations, EVV 2DIR can be used to massively decongest the spectra of large molecules, in particular those of proteins.

The EVV 2DIR technique is a third-order non-linear spectroscopy, involving four (three input and one output) electric fields. When the three pulsed laser beams are focused into the sample, their high intensity fields induce non-linear polarisations which mix and radiate the detected output field (Figure 2-1). Two of the input beams
are in the infrared (IR) and are used to excite coupled molecular vibrations. The third is a visible beam and this is essentially used to probe the polarisation generated by the two IR beams. Thus the coupling information of the two vibrations is ‘read-out’ by the visible beam when it is scattered from the induced polarisation. EVV 2DIR can therefore be considered a hybrid IR-Raman method where the final signal strength depends on both IR and Raman processes. The non-linear signal generated is in the visible part of the optical spectrum; this is a particular advantage of this variant of 2DIR, as the detection of visible photons is much more efficient than of those in the infrared.

**Figure 2-1:** EVV 2DIR spectroscopy is a four wave mixing (FWM) process, requiring the overlap of three pulsed laser beams – two infrared and one visible – on the sample. The oscillating non-linear polarisation that is induced launches a visible output field. This third order non-linear signal is detected as a function of the scanned infrared frequencies.

To measure EVV 2DIR spectra, two tuneable infrared beams are scanned across the vibrational spectrum and a particular output frequency is monitored as a function of the two infrared frequencies. When the frequencies are independently in resonance with coupled vibrational transitions of the sample, the output non-linear signal is enhanced. The signal is measured as a function of both IR frequencies, and the resulting spectra are presented as two dimensional intensity maps, an example of which is shown in Figure 2-2. The cross peaks observed in the 2D spectra therefore represent multiplicatively enhanced four wave mixing (FWM) signal from two
coupled vibrational modes of the sample. Coupling between two modes can occur when excitation of one mode causes changes to the other. Thus, the two coupled modes may be linked by inter- or intra-molecular interactions, or be a fundamental mode and its overtone, or a fundamental mode and one of its combination bands. As this particular version of 2DIR is a homodyne spectroscopy, the total signal is proportional to the square of the number of molecules being probed by the beams.

This four wave mixing process was originally termed doubly vibrationally enhanced (DOVE-FWM), since enhanced signal is detected when the two IR beams are in resonance with two coupled vibrational modes. Infrared double resonances are however now a common feature of 2DIR spectroscopy. Therefore to emphasise the distinction between this version of 2DIR and other formats, such as the photon echo technique, the measurements are now referred to as being of electron-vibration-vibration (EVV) couplings.

Figure 2-2: An example EVV 2DIR spectrum of low density polyethylene in which the cross peaks are labelled with their assignments. Both cross peaks arise from the coupling of a CH₂ deformation, \( \delta(\text{CH}_2) \), with a combination band of this deformation and the CH₂ stretch, \( \delta(\text{CH}_2) + \nu(\text{CH}_2) \). The cross peak at 1470 / 4350 cm\(^{-1}\) is attributed to the symmetric CH₂ stretch, \( \nu_s(\text{CH}_2) \), and that at 1470 / 4320 cm\(^{-1}\) to the asymmetric CH₂ stretch, \( \nu_{as}(\text{CH}_2) \).
2.2.2: Principles of EVV 2DIR

The frequencies of the three fields that are focused in a sample to achieve EVV four wave mixing are not equal, making it a form of non-degenerate mixing. There are twelve possible schemes for mixing these fields that can access two vibrational resonances and bring about the FWM EVV process; thus far it has only been observed using two of these. These two schemes are termed EVV-IR and EVV-Raman, and the resonances involved in these processes are best visualised using the wave mixing energy level (WMEL) diagrams and coherence pathways shown in detail in Figure 2-3. In brief, four wave mixing signal is generated by using the two infrared excitation pulses (E$_\alpha$ and E$_\beta$) to first induce two resonant vibrational coherences in a sample. These are followed by the visible pulse (E$_\gamma$) which induces a non-resonant coherence and drives the output field (E$_\delta$), detected as the FWM signal.

Figure 2-3: The wave mixing energy level diagrams and coherence pathways depicting the four wave mixing EVV processes. The levels labelled $a$ and $b$ represent resonant vibrational states of the sample, and the upper levels, labelled $e$, represent non-resonant (virtual) electronic states. Full arrows indicate ket transitions and dotted arrows bra transitions. The four electric fields involved in the processes are labelled in Greek letters in ascending order of frequency; the three input fields E$_\alpha$, E$_\beta$, and E$_\gamma$, have frequencies of $\omega_\alpha$, $\omega_\beta$ and $\omega_\gamma$, and the output field, E$_\delta$, has a frequency of $\omega_\delta$. E$_\alpha$ and E$_\beta$ correspond to the separately tuneable infrared beams, and E$_\gamma$ to the fixed visible beam. The output field, E$_\delta$, is only generated if E$_\alpha$ and E$_\beta$ are resonant with coupled vibrational modes. The fields are also labelled numerically to indicate the time ordering of the processes: beam 1 is the first to arrive at the sample.
The WMEL diagrams of Figure 2-3 show three possible coherence pathways: EVV-IR 1, EVV-IR 2 and EVV-Raman. The arrows between the levels of these diagrams do not represent absorptions and emissions, rather the creation and destruction of coherences in the system. The evolutions of the coherences, from the initial state population, $gg$, to the final state coherence, are described by the coherence pathways. When the input pulses are overlapped in time, the three pathways coexist and interfere. Selection of a certain process is made possible with time ordering of the pulsed input fields and the introduction of time delays between each of them.

Using delay stages, it is possible to arrange for the $E_\alpha$ pulse to arrive first so that the $gg \rightarrow ga \rightarrow ba \rightarrow ea \rightarrow aa$ (EVV-IR 1) pathway can be selected. The $gg \rightarrow bg \rightarrow ba \rightarrow ea \rightarrow aa$ (EVV-IR 2) and $gg \rightarrow bg \rightarrow ag \rightarrow eg \rightarrow gg$ (EVV-Raman) pathways are only accessible when the $E_\beta$ pulse is timed to arrive first. It is thus possible to switch pathways on and off by changing the pulse orderings and timings.

![Figure 2-4: The pulse ordering used to select the EVV-IR 1 coherence pathway. The delay between the first and second (IR) pulses is denoted $T_{12}$. The third (visible) pulse arrives with a delay $T_{23}$ after the second pulse.](image)

For the work presented in this thesis, the EVV-IR 1 process was used. In order to select this pathway, the pulses are set to arrive at the sample in the following order: 1: $E_\alpha$, 2: $E_\beta$ and 3: $E_\gamma$. Time delays of $T_{12}$ between the two infrared pulses, and $T_{23}$ between the second infrared pulse and the visible pulse are implemented (shown in Figure 2-4). A coherence at $E_\alpha$ is thus created which, after a time delay of $T_{12}$,
interacts with the coherence at $E_\beta$. After a further delay of $T_{23}$, the visible pulse, $E_\gamma$, then induces a Raman-like transition. Finally, the output field radiates from the induced non-linear polarisation and this signal is detected at the frequency: $\omega_\delta = \omega_\gamma + \omega_\beta - \omega_\alpha$. The WMEL diagram and coherence pathway for the EVV-IR 1 process are shown in more detail in Figure 2-5.

![WMEL diagram and coherence pathway for the EVV-IR 1 process](Figure 2-5)

**Figure 2-5:** The wave mixing energy level diagram and coherence pathway for the EVV-IR 1 four wave mixing process, used for the work presented in this thesis. The signal is detected at a frequency of $\omega_\delta = \omega_\gamma + \omega_\beta - \omega_\alpha$.

### 2.2.3: EVV 2DIR properties and capabilities

#### Phase matching

In order to ensure that the signal generated by the four wave mixing process is intense and directional, it is necessary to phase match the non-linear polarisation with the output field it launches. Due to the variation of refractive index with frequency, the different frequency fields experience different phase shifts in the interaction length of the sample. The phase of the oscillating non-linear polarisation – which depends on the phases of the input beams – will therefore generally be
different from that of the output field. If the two are out of phase, then interference occurs along the interaction region and the output field is reduced. The wavevectors of the three input beams depend on their direction and refractive indices, and the vector sum of these gives the wavevector of the non-linear polarisation \( \vec{k}_4 \). To correct for dephasing, the angles of the input beams can therefore be arranged so that the non-linear polarisation and the field it creates oscillate in phase. This also has the added advantage that the output signal emerges from the sample at a different angle from the inputs, making detection simpler.

\[
\vec{k}_4 = \vec{k}_1 - \vec{k}_2 + \vec{k}_3
\]

Equation 2-3

**Temporal suppression of background signal**

Non-resonant background signal from the sample itself, or from sample substrates, can be suppressed through the use of appropriately long delays between the input pulses. The dephasing times for the vibrational coherences involved in EVV four wave mixing can be several picoseconds (normally \( \sim 2 \) to \( 5 \) ps). Fortunately, the non-resonant electronic polarisations which give rise to the background signal have much shorter lifetimes and relax over femtosecond timescales. Therefore ultrafast lasers are used which deliver pulses that are shorter in duration (\( \sim 1.5 \) ps) than the lifetimes of the vibrational coherences. The time delay, \( T_{23} \), introduced between the two IR pulses and the visible pulse thus allows for the non-resonant electronic responses driven by \( E_\alpha \) and \( E_\beta \) to have died away before \( E_\gamma \) arrives. Because of their longer lifetimes, the resonant vibrational coherences do not diminish during this time delay and \( E_\gamma \) can therefore launch their output, \( E_\delta \). FWM signal, with limited background signal, is then detected.
Spectral decongestion

Like other 2DIR techniques, EVV 2DIR offers significant spectral decongestion over linear vibrational spectroscopies by probing only coupled vibrations and spreading out the spectral information into two dimensions. This is a particular benefit when dealing with large molecules such as proteins, and will be demonstrated in Chapters 3 and 5. The independent timings of the infrared and visible pulses in EVV 2DIR however afford an additional benefit over the pump-probe and echo variants of 2DIR, namely that it is possible to break the fixed transform relationship between wavelength and time used in these techniques. This means that, by manipulation of the pulse timings, pathways which are always present in other forms of 2DIR can be switched off. Through appropriate ordering of the pulses only the coherence pathway of interest can be selected (Figure 2-3) and interference from other non-linear processes (as well as from the electronic non-resonant background) can be minimised. Additional control over spectral decongestion can be achieved through variation of the orientation of the visible electric field. The polarisations of the three input beams select which components of the susceptibility tensor will be probed and therefore can be used to extinguish certain vibrational modes, thus further decreasing the spectral congestion.

2.2.4: Electronically resonant EVV 2DIR

It is possible to achieve significant enhancements of EVV 2DIR signal if the sample has a suitable electronic transition that may also be probed. The final ‘read-out’ step of the EVV process is essentially half of a conventional Raman scattering event but coherent (Figure 2-5). One of the consequences of this is that the electronic properties of the molecular system being probed also affect the signal, and the
experiment can therefore be made triply resonant if the visible beam, $E_\gamma$, is tuned towards an electronic resonance of the system.

EVV 2DIR signal is generated when the infrared fields, $E_\alpha$ and $E_\beta$, are resonant with coupled vibrational transitions of the sample. When the visible field, $E_\gamma$, is additionally resonant with an electronic state of the sample that is coupled to the vibrational transitions, then the FWM signal is further enhanced. In such cases the technique becomes triply resonant and the resulting spectra represent electron-vibration-vibration couplings. This can lead to an increase in the sensitivity of the method by several orders of magnitude.

The increased sensitivity afforded by electronically resonant EVV 2DIR is of great importance for biological applications where the number of proteins, metabolites, or drugs molecules to be detected is low (typically picomoles to femtomoles). The possibility of enhancing the signal in such cases is covered briefly in Chapter 6. Moreover, the technique has particular potential for homing-in on the active sites of proteins that contain groups with appropriate electronic transitions, for example metal ions or chromophores. Due to the triple enhancement of the FWM signal that could be achieved from such groups, it should be possible to selectively study these whilst avoiding the surrounding protein. In fact, Chapter 5 of this thesis presents the first example of such a triply resonant experiment on a protein active site, namely the retinal chromophore that lies at the heart of the protein bacteriorhodopsin (bR).

### 2.3: The EVV 2DIR spectrometer

The data presented in this thesis was collected using an existing EVV 2DIR experiment. The prototype, constructed by D. J Palmer [99], was based on that first
built by the J. C. Wright group. Extensive alterations were made to this by P. M. Donaldson, who successfully developed it into a working spectrometer. [95-98] Further modifications were made to the experiment with F. Fournier, and LabView programming was provided by C. J. Barnett. This section provides a description of the set-up of the spectrometer and a general explanation of its operation. Further details of the operational protocols and daily maintenance procedures required to collect spectra are covered in Section 2.4. Brief details of a second, improved EVV 2DIR spectrometer that has since been built are supplied in Chapter 6.

For explanation purposes, the ‘EVV 2DIR Spectrometer’ may be roughly divided into four sections:

1. The commercial laser system that provides the visible beam and generates the pump for the OPAs.
2. The Optical Parametric Amplifiers (OPAs) that produce the two tuneable infrared beams.
3. The optical set-up used to send the one visible and two infrared beams to the sample.
4. The sample area, system for the detection of the FWM signal and Labview-controlled collection of the data.

2.3.1: The commercial laser system

Ultrafast laser systems are used to provide the high-energy short pulses required in non-linear spectroscopies, and the commercial development of these systems was crucial to the realisation of such techniques. EVV 2DIR employs picosecond laser pulses – one fixed frequency visible pulse and two frequency tuneable infrared pulses. The Spectra-Physics [100] laser system used here to generate the visible
beam and the input for the OPAs (that provide the tuneable IR beams) is outlined in Figure 2-6.

Figure 2-6: The commercial laser system used to provide the input for the optical parametric amplifiers that then generate the tuneable infrared beams. In brief, the initial optical power is provided by the pump laser, a Spectra Physics Millennia (a Nd:YVO\textsubscript{4}, continuous wave (CW) laser that generates a 530 nm output of about 4 W). This provides the energy for the gain medium of the seed laser, a Spectra Physics Tsunami. The output of this mode-locked Ti:Sapphire is a pulse train of ~ 90 fs pulses at ~ 80 MHz, with a wavelength of 790 nm. The energy per pulse of ~ 6 nJ is not sufficient for performing optical parametric amplification, so an amplifier is then used to boost the energies of the pulses. The chirped pulse regenerative amplifier (a Spectra Physics Spitfire) first stretches the duration of the pulses to more than 200 ps, then amplifies the broadened pulses before compressing them back to ~ 1.5 ps. The Spitfire derives its gain from a separate pump laser – a Spectra Physics Evolution Nd:YLF. The output of the regenerative amplifier is a beam of ~ 1.5 ps pulses with a 1 kHz repetition rate and an average energy per pulse of ~ 800 µJ.

2.3.2: The optical parametric amplifiers (OPAs)

The output of the regenerative amplifier is a beam of ~ 1.5 ps pulses with a 1 kHz repetition rate and an average energy per pulse of ~ 800 µJ. This is used to pump two Spectra Physics 800C OPAs; the first of these contains a 50% beamsplitter to allow for half of the input pulse energy to be delivered to the second. The two OPAs convert the fixed frequency visible beam into the independently tuneable infrared
beams, $\omega_\alpha$ and $\omega_\beta$, required for EVV 2DIR. The conversion process of the visible pulses into those in the infrared is known as optical parametric generation (OPG), and is explained in Figure 2-7.

![Figure 2-7: The principles of optical parametric generation (OPG). A non-linear optical crystal is pumped with the input visible pulse. Each photon of the incident pulse is divided into two photons, with conservation of energy requiring that the sum of the energy of these two photons equals that of the pump photon. The two new generated fields are termed the signal and the idler, and their wavelengths depend on the phase matching in the non-linear crystal. The wavelengths of the signal and idler may therefore be tuned by adjustment of the angle between the axes of the crystal and the incident pump photon.](image)

The input pump pulses for the OPAs are at 790 nm (12,650 cm$^{-1}$) and the optical parametric generation process is performed using a $\beta$-Barium Borate ($\beta$-BaB$_2$O$_4$ or BBO) non-linear crystal. Through adjustment of the BBO orientation, the generated signal and idler fields cover the near-infrared frequency ranges of $\sim 6300 – 9100$ cm$^{-1}$ and $\sim 3500 – 6300$ cm$^{-1}$, respectively. Depending on the frequencies required, the idler beam is either used directly or the difference frequency mixing (DFM) process is used to convert the signal and idler to lower (mid-infrared) frequencies. In such cases, the signal and idler are crossed in a DFM crystal, here silver thiogallate (AgGaS$_2$), to generate a new field whose frequency is the difference between the two inputs. This provides a mid-infrared beam, tuneable over 1100 – 3000 cm$^{-1}$.

In summary, the OPAs are used to provide tuneable infrared beams covering a frequency range of 5000 – 1000 cm$^{-1}$ ($\sim 2 – 10$ $\mu$m). The output pulses are of $\sim 1.5$ ps duration with a 1 kHz repetition rate, and have an average energy per pulse of $\sim 4$
Chapter 2: Electron-vibration-vibration two-dimensional infrared spectroscopy

– 12 µJ. To scan over the required frequencies, the OPAs are fitted with actuators for rotating the orientations of the crystals and gratings. These are computer-controlled, and frequency scanning and calibrations are performed using LabView.

2.3.3: Delivery optics set-up

The remainder of the output from the regenerative amplifier, not used for the OPA process, provides the visible beam. To make it comparable to the infrared beams, and also to prevent sample damage, the average energy per pulse is reduced (using a variable attenuator) from ~ 300 µJ to ~ 0.5 – 5 µJ. The final output from the laser system and OPAs is therefore three picosecond pulses – one fixed frequency visible beam (ωγ) and two tuneable IR beams (ωα and ωβ). A schematic illustrating the layout of the delivery optics used to bring these together at the sample is provided in Figure 2-8.

To clean up their profiles, all three of the beams are spatially filtered using pinholes. The IR beams are expanded using telescopes before being focused, in order to achieve narrower beam waists in the sample. CaF₂ lenses are used to focus all three of the beams, giving waists at the sample of ~ 100 µm for the IR and ~ 70 µm for the visible. The beams are spatially overlapped at the focus of the sample using a 100 µm pinhole. Temporal overlap of the ωγ and ωβ beams is carried out by sum frequency generation in a BBO crystal, and then their overlap with ωα is found by detecting non-resonant four wave mixing signal from a CaF₂ window. The optical path of the ωα beam is purged with nitrogen to minimise the atmospheric absorption of the beam (water vapour absorbs IR radiation between 3500 – 4000 cm⁻¹ and 1250 – 2000 cm⁻¹) which can cause pulse distortions and artefacts in the 2D spectra. The humidity measured at the sample position is < 1 %, and the temperature of the lab is also controlled at 19 °C ± 0.5 °C.
The two IR beams are horizontally polarised with respect to the bench, whilst the visible beam polarisation can be rotated depending on the experiment. The detected FWM signal is unpolarised. The polarisation states are denoted using the usual S and P notation; this describes the orientation of the electric fields relative to the plane of propagation. When all three input beams are polarised in the plane of propagation, the polarisations are denoted PPP. When the visible beam is polarised orthogonally with respect to the IR beams, the denotation is PPS.

To implement time delays between the three pulses, there are adjustable delay stages on the $\omega_\beta$ and $\omega_\gamma$ beam paths. These are controlled through a motion control box using LabView. For the experiments in this thesis, the pulse sequence was 1: $\omega_\alpha$, 2: $\omega_\beta$, 3: $\omega_\gamma$, with delays of $T_{12}$ between $\omega_\alpha$ and $\omega_\beta$, and $T_{23}$ between $\omega_\beta$ and $\omega_\gamma$. The scanning of the infrared beams across a spectral region causes beam pointing changes and this results in changes in the time delays between the pulses. A computer-controlled procedure is therefore used to scan the delay stages during the spectral scanning; any delay changes are then corrected for and $T_{12}$ and $T_{23}$ kept constant at their set values.

### 2.3.4: Signal detection and data collection

The three beams are focused into the sample at angles that have been calculated to give the optimum phase matching for the FWM process. The $\omega_\beta$ beam is approximately at normal incidence to the sample, and the $\omega_\gamma$ and $\omega_\alpha$ beams have angles of incidence relative to the $\omega_\beta$ beam of about -2° and 8° respectively. The FWM output signal ($\omega_\delta = \omega_\gamma + \omega_\beta - \omega_\alpha$) emerges from the sample in a different direction to the input beam, and so can be isolated using an aperture and a series of filters to remove any residual probe beam. It is then detected in the 690 – 725 nm spectral range using a photomultiplier (PM) tube, which can be used in photon-
Chapter 2: Electron-vibration-vibration two-dimensional infrared spectroscopy

The optical set-up of the 'EVV 2DIR spectrometer' used to send the one visible and two infrared picosecond beams to the sample.

**Figure 2-8:** The optical set-up of the 'EVV 2DIR spectrometer' used to send the one visible and two infrared picosecond beams to the sample.
counting mode. For reduction of noise, the PM is placed in a light-proof box and all experiments are performed in low ambient light conditions. Experiments may be performed in either transmission or reflection, for which a polished stainless substrate is generally used (see Section 3.2). For collection of reflected signal, a set of extra mirrors is simply inserted between the sample and the detector.

Once FWM signal has been found and optimised from a sample, the collection of data is mainly automated through LabView. For a particular spectral region, $\omega_\beta$ is scanned for each $\omega_\alpha$ frequency with an increment of $5 \text{ cm}^{-1}$ for both beams. The signal is recorded in the form of two-dimensional spectra, in which the FWM signal is a function of $\omega_\alpha$ and $\omega_\beta$.

### 2.4: EVV 2DIR experimental methods

#### 2.4.1: Operational procedures

Although the lasers and OPAs are commercial systems, the technology is not yet sufficiently reliable for them to be left to run without adjustments being made on a regular basis. The performance of these can be quite variable and, although the laser systems can sometimes be relatively trouble-free for weeks at a time, faults occur frequently. Such disruptions are however becoming less commonplace, due to the improved reliability achieved through a number of alterations mainly implemented by F. Fournier and P. M. Donaldson. For day-to-day operation of the EVV 2DIR experiment, monitoring of the beam energy, quality and alignment is required, along with general fault diagnosis and trouble-shooting.
The spectral space that is potentially accessible using the OPAs is very large and could cover a range of ~ 1100 cm\(^{-1}\) to ~ 9000 cm\(^{-1}\). However, only small regions, roughly up to 500 cm\(^{-1}\) for each of the IR beams, can be scanned at a time. Changing spectral region is not a trivial task, with significant realignment and recalibration of the OPAs and delivery optics required. Using the current DFM crystal (AgGaS\(_2\)), the lowest accessible frequency is 1100 cm\(^{-1}\). To be able to cover frequencies lower than this, different non-linear crystals are required for the OPAs – this is currently being implemented on a second version of the EVV 2DIR experiment (detailed in Chapter 6).

For the work presented in this thesis, three main spectral regions were surveyed:

\[ \omega_a: 1300 – 1600 \text{ cm}^{-1}; \quad \omega_b: 2700 – 3200 \text{ cm}^{-1} \]

\[ \omega_a: 1200 – 1400 \text{ cm}^{-1}; \quad \omega_b: 2600 – 3000 \text{ cm}^{-1} \]

\[ \omega_a: 1300 – 1600 \text{ cm}^{-1}; \quad \omega_b: 4200 – 4600 \text{ cm}^{-1} \]

Cross peaks observed in the two-dimensional spectra are given in terms of their two IR frequencies as follows: \( \omega_a / \omega_b \text{ cm}^{-1} \).

### 2.4.2: Imaging with EVV 2DIR

To perform EVV 2DIR imaging, the two IR frequencies are set on a cross peak of interest (\( \omega_a / \omega_b \text{ cm}^{-1} \)), specific to a particular chemical group, and the sample is spatially scanned (using two automated translation stages, controlled by LabView) across the input beams. The detected signal level is mapped as a function of the beam position on the sample. For example, the cross peak attributed to the CH\(_3\) group of proteins (Chapter 3) may be used to monitor protein levels across a sample. To date this technique has been used to image thin protein films (Chapters 3 and 5), and histological sections of mouse kidney (Chapter 6).
Chapter 3:
Protein identification and absolute quantification with EVV 2DIR

3.1: The EVV 2DIR protein fingerprinting strategy

While other aspects of the research presented here concern the development of EVV 2DIR for imaging and other proteomic applications, such as the study of enzyme mechanisms, protein-protein interactions and post-translational modifications [93], the initial objective for the technique is its use as a protein identification tool. For reasons of sensitivity and spectral decongestion, EVV 2DIR is an attractive method for protein fingerprinting. This chapter covers the proof of principle of a label-free, high-throughput protein identification strategy using this spectroscopy. The key proposition of the approach is that identification can be performed using spectroscopically-determined amino acid content, relative to an internal reference. The fingerprint of a protein in this case is the distribution of its relative amino acid quantities. Although it is known that amino acid composition analysis can be used to identify proteins (without the need of sequence data), this technique is no longer widely employed. The experimental methods used for composition determination require hydrolysis of the protein substrate, followed by separation and derivatisation
of its amino acids before quantification can occur (see Section 1.2). [20-25, 101, 102] In contrast, the identification scheme outlined here requires no chemical or biochemical preparation steps, and achieves quantification by measuring only spectroscopic features of the proteins.

The core aim of the strategy is to establish the relative quantities of amino acids in a protein by using the intensities of the cross peaks in its EVV 2DIR spectrum. To achieve this, the intensity of an internal reference peak will also need to be measured. This internal standard could be another amino acid, but in this case is to be a protein aliphatic group. The relative quantity of an amino acid in a sample will be established by ratioing the intensity of its cross peak ($I_{AA}$) with that of the internal reference ($I_{Ref}$). As EVV 2DIR is a homodyne technique (the measured signal is proportional to the square of the number of molecules being probed, $I_{FWM} \propto N^2$), the square roots of the intensities will be used. The fingerprint of a protein will therefore comprise the cross peak intensity ratios ($\sqrt{I_{AA}} / \sqrt{I_{Ref}}$) for a number of its different amino acids. For a sample to be identified, these characteristic ratios will be used to perform a search of a specialised protein database. Instead of this database containing protein sequences, it will contain ratios of the actual numbers of amino acids and reference groups ($N_{AA} / N_{Ref}$) in each protein.

For this strategy to work, a linear correlation between the experimentally-determined ratios and number ratios of known protein samples first needs to be demonstrated:

$$\sqrt{\frac{I_{AA}}{I_{Ref}}} \propto \frac{N_{AA}}{N_{Ref}}$$

Measured peak ratio       Known number ratio

Equation 3-1
Once this has been achieved, calibration graphs will be constructed using data of known proteins. These will be used to convert relative peak intensities of unknown protein samples into the number ratios required to perform a database search. The type of calibration curve is shown in Figure 3-1, in which 2DIR measured ratios are plotted as a function of the number ratios.

![Figure 3-1](image)

**Figure 3-1**: Form of calibration curve to be produced using known protein samples; the ratios measured by EVV 2DIR ($\frac{I_{AA}}{I_{Ref}}$) are plotted as a function of the actual number ratios ($\frac{N_{AA}}{N_{Ref}}$) for each protein. These will be used to convert relative peak intensities of unknown protein samples into the number ratios required for database searching.

The first step towards realising the protein fingerprinting strategy was to identify 2DIR spectral features which are amino acid specific. For a cross peak to be suitable, it must correspond to a unique residue and also be free of any interferences that could affect its amplitude. Spectral congestion and interference were therefore minimised when finding, and subsequently using, cross peaks through the use of different delays and polarisations. Delays enable the selection of the coherence pathways corresponding to the desired EVV 2DIR process (here EVV-IR1), minimising other non-linear processes and reducing the electronic non-resonant background, whilst different beam polarisations can be used to extinguish certain vibrational modes.
Amino acids, peptides and proteins were used to find, and then attribute, cross peaks to amino acid side chains. To date, 2DIR spectral signatures of three amino acids have been discovered: tyrosine (Tyr, Y), phenylalanine (Phe, F) and tryptophan (Trp, W). Assignment of the modes associated with these observed features was achieved through DFT calculations. The spectral signatures of suitable internal references were also identified. These are the CH$_2$ (methylene) and CH$_3$ (methyl) groups (CH$_x$) that occur in most amino acid side chains, shown in Figure 3-2.

**Figure 3-2:** The side chains of the twenty amino acids, reproduced from Mathews. [103] The label-free strategy presented in this chapter uses spectroscopic features of these to produce a novel type of protein fingerprint. 2DIR spectral features associated with amino acid CH$_2$ or CH$_3$ groups (labelled) serve as the internal standards, as all residues have at least one of these groups.
The linear relationship between the intensity ratios of the cross peaks \( \sqrt{I_{AA}} / \sqrt{I_{Ref}} \), and the ratios of the actual numbers of amino acids and CH\(_x\) present \( N_{AA} / N_{Ref} \) in known peptide and protein samples was demonstrated. Initial work, covered in Section 3.3, was carried out with peptides and poly-amino acids. [95, 96, 99] This resulted in the identification of exploitable cross peaks for phenylalanine, tyrosine and the CH\(_2\) group. [91] Although this work showed that it is possible to quantify relative amino acid levels for short peptides, there was always the possibility that primary, secondary or tertiary structural effects would prevent such measurements on proteins. Subsequent studies using proteins, covered in Section 3.4, however demonstrated that these structural sensitivities are not limiting factors, and the quantification of relative amino acid levels in proteins was achieved. The EVV 2DIR spectra of the peptides obtained in the previous work were an aid in the identification of the features observed in the protein spectra, which were found to be very similar. Spectral features of three amino acids – phenylalanine, tyrosine and tryptophan - and of the CH\(_3\) group, suitable for use as an internal reference, were identified and validated in proteins. The use of different polarisation schemes to provide additional information was also demonstrated for phenylalanine. [92]

Protein identification is to be accomplished by comparing the spectroscopically-determined amino acid ratios of a protein to the contents of a custom-made database, consisting of proteins’ \( N_{AA} / N_{Ref} \) ratios. Bioinformatics studies have revealed that a protein fingerprint need only comprise the relative levels of five to ten different residues to achieve an identification. The details of the construction of the database and the bioinformatics studies are covered in Chapter 4.

The protein CH\(_3\) spectral feature, identified for use as an internal reference for amino acid content determination, can also be employed for performing simple absolute quantification of protein levels. This is an important property of EVV 2DIR
as it is increasingly an important issue in many proteomic applications, but one that is relatively difficult to achieve with mass spectrometry. The quantification procedure proposed here is however relatively straightforward to achieve, the details of which are covered in Section 3.5.

3.2: Experimental methods and details

3.2.1: Sample preparation

One of the first issues to be tackled was how the protein samples are to be presented to the EVV 2DIR experiment. A reliable protocol that can produce repeatable spectra is required. Previous studies by the group on amino acids had used saturated solutions of these molecules. [95, 96, 99] This was also attempted for proteins but, even when using the most concentrated of solutions, any protein signals were completely drowned out by those from the water. A methodology for preparing the proteins as dry films, to avoid the complications caused by water and to concentrate the analytes, was therefore identified. This approach is also employed in drop coating deposition Raman spectroscopy (DCDR) to provide increased sensitivity, and has been used to demonstrate a number of biologically important measurements, including that of protein phosphorylation levels. [44, 45]

The drop-coating deposition technique

The drop coating deposition technique was first developed for obtaining Raman spectra of dilute protein solutions. [104-107] Based on the characteristic deposition pattern that a drop of solution makes when left to dry (the so-called coffee-ring effect), this method is a simple way of pre-concentrating a protein analyte to achieve
increased signal. The coffee-ring effect occurs when the pinning of the contact line 
of a drying drop causes a form of capillary-flow to be set up. When the liquid from 
the outer edge of a drop evaporates, it is replenished by liquid that flows from the 
interior of the drop. This outward flow of the solvent carries the solute material to 
the edge of the drop, resulting in the formation of a ring that contains nearly all of 
the deposited material. [108, 109]

Droplets of low concentration protein solutions form well-defined rings of the 
protein material when the buffer evaporates. The reproducibility of Raman spectra 
recorded of deposits made using a variety of substrates (including quartz slides, 
CaF$_2$ slides and polished stainless steel plates), and using solutions of a range of 
concentrations (mM to µM) and volumes (mL to nL) has been demonstrated. No 
significant spectral variations are observed between different types of deposits of a 
particular protein analyte. The protein samples also remain stable and, after a 
number of weeks, the Raman spectra recorded are virtually identical to those taken 
from a freshly-made deposit. [104, 105]

Proteins in drop-coated films are preserved in a gel-like condition, with a glassy 
skin forming at the droplet surface. The proteins remain partially hydrated and their 
solution structure is preserved; the Raman spectra of dried deposits and concentrated 
solutions have no significant differences. Despite the proteins retaining a partial 
level of hydration, the spectral interference from water is insignificant, making 
solvent subtraction unnecessary. The type of buffer used also has no effect on the 
Raman spectra obtained, since during the drying the different solutes become 
segregated and the proteins separate from the buffer compounds. The buffer salts 
tend to crystallise towards the centre of the dried drop, and the coffee ring contains 
almost pure protein material. [106, 107]
Preparation of drop-coated protein films for EVV 2DIR analysis

Drop-coated deposits were made either from single droplets of a solution, or by co-depositing multiple drops on the same substrate spot in order to increase the sample concentration. The protein and peptide solutions used were either made with only ultrapure water, or using a variety of buffers. Typically the drops were left to dry at room temperature for 15 – 30 minutes. The samples were further dried, however, when in the sample compartment of the EVV 2DIR experiment which was purged with nitrogen to give a humidity level of < 1 %. Figure 3-3 shows an example of the drop-coated deposits of peptides and proteins used for the results presented in Sections 3.3.

![Figure 3-3](image)

**Figure 3-3:** (a) Drop-coated deposits of LRRFFLG, produced by co-depositing two 1.5 µL drops of a 50 mM solution of the peptide on a glass microscope slide. (b) A microscope image of the edge of a drop-coated deposit of LRRYYLG. The overall diameter of the deposit is ~ 5 mm, the outer ‘coffee-ring’ being ~ 0.5 mm wide.

As observed in drop-coating deposition Raman (DCDR) experiments, no differences were seen in the 2DIR spectra taken of deposits made from solutions of the same proteins with or without a buffer. Once dry the samples also remain stable for a long period of time; deposits prepared for the work presented here have shown no spectral changes after more than two years. Figure 3-4 shows the relative levels of FWM signal observed from the coffee-ring and inner section of a drop-coated protein film. The amount of signal obtained from the centre of the deposit is minimal and not significantly higher than that from the substrate, indicating that the majority of the protein sample is deposited in the outer ring.
Figure 3-4: Relative levels of FWM signal observed (at 1460 / 2980 cm⁻¹ and with delays of $T_{12} = 1.5$ ps, $T_{23} = 1.5$ ps) when a BSA drop-coated deposit (two 2 µl co-deposited drops of a 0.75 mM solution) was translated across the input laser beams. The same level of signal is obtained from the centre of the drop as from the substrate, indicating that all protein material is deposited in the coffee ring.

Although most protein and peptide solutions were made using unbuffered ultrapure water (pH 5), some experiments were carried out with buffered solutions. For example, 10 mM maleic acid at pH 2.5 was used for attempting to acid denature the proteins and disrupt their higher order structures. This is desirable for the purposes of fingerprinting, so that the amino acids being measured are as unaffected as possible by their surrounding structures. However no differences were observed between the amino acid cross peaks of protein deposits made using water or maleic acid.

Initially the 2DIR work was carried out in transmission. For this approach, standard quartz microscope slides (1 – 1.2 mm thick) and coverslips (0.2 – 0.5 mm thick) were used as substrates. Through the implementation of delays the non-resonant signal from the glass was mostly eliminated, though in some spectral regions, and
with less concentrated samples, the glass signal did become more problematic. Non-resonant background was therefore minimised by using the thinner substrate, though the microscope slides were more commonly used as they are less fragile and easier to use in the experiment.

As it is intended that the EVV 2DIR experiment will eventually be coupled with protein separation systems, sample wells etched into glass slides were also explored as a way of collecting eluted protein samples. The approach currently being developed in the group however involves the electrodeposition of proteins following separation by capillary electrophoresis. As a conductive surface is required for this track writing procedure of CE-separated protein samples (Sections 1.4 & 6.3), later 2DIR experiments were carried out in reflection so that a polished stainless steel substrate could be used. In principle, working in reflection also reduces contributions from non-resonant background since the input beams do not propagate through the substrate. The drop-coated deposits formed on the steel substrates had the same properties as those formed on the glass slides, with minimal spectral variations between the two.

Before the drop-coating technique was established as the most suitable for measuring EVV 2DIR spectra of proteins, other methods were also trialled. These included polyacrylamide gels, with the aim of measuring protein signals directly from these following protein separation. Additionally, P. M. Donaldson tried using nanoporous TiO$_2$ films, which act as a matrix to support the protein samples. [110-112] EVV 2DIR signal was observed in reflection from proteins adsorbed in these films. [90]
Peptide and protein samples

In initial work to locate side chain cross peaks, amino acids (in the form of single crystals) and poly-amino acids were analysed – all purchased from Sigma Aldrich. Arginine and glycine crystals were grown by simply leaving saturated solutions of these amino acids in Petri dishes. Solid protein samples, including nylon and collagen, were also used – these gave strong reproducible signals and were easy to handle.

Studies then continued with commercially available peptides of varying sizes (from 3-mers to 20-mers), chosen according to their amino acid levels. These were mainly purchased from Anaspec. [113] During early work, solutions for drop-coating were either made as concentrated as possible, or to the concentration that was found to produce the most satisfactory films for a particular sample. As the EVV 2DIR technique progressed, however, it became possible to use increasingly lower sample concentrations. Not all of the peptides tested were sufficiently soluble in water to form drop-coated deposits with a ‘coffee-ring’ effect. These instead tended to form uniform, cloudy films from which it was more difficult, or not possible, to obtain FWM signal.

Three off-the-shelf peptides in particular – a kemptide (LRRASLG), endomorphin 2 (YPFF) and [Leu-5]-enkephalin (YGGFL) – formed good drop-coated deposits from which it was relatively straightforward to attain reasonable signal levels. Using these samples, the characteristic cross peaks of tyrosine, phenylalanine and the CH$_2$ group were first identified. Derivatives of these peptides were therefore then designed, in which extra Y and F residues were either added to the termini or used to replace other residues, in order to produce peptides with differing levels of these amino acids. These peptides were made by an internal synthesis service at Imperial
Chapter 3: Protein identification and absolute quantification with EVV 2DIR

College London. [114] The syntheses followed a standard Fmoc / tBu strategy, and the peptides were purified to more than 98% homogeneity using reverse-phase HPLC, then characterised by MALDI-TOF MS.

The sequences of the eight peptides used for the first proof of principle study of the EVV 2DIR fingerprinting technique presented in Section 3.3, are as follows: LRRFSLGF, LRRFFLG, LRRYSLGY , LRRYYLG , YPFF, YGGFL, YGGFF, YGGFY. Solutions of these peptides were made in ultrapure water at a concentration of 50 mM. Each peptide film was formed by co-depositing two 2 µl drops of a solution on a 1 mm thick microscope slide, with the first drop being left to dry at room temperature for 30 minutes before the second drop was added. The samples were further dried when placed in the < 1 % humidity environment of the EVV 2DIR sample compartment.

When studies progressed onto locating amino acid cross peaks in proteins, many commercially available proteins ranging in size from 10 kDa to 90 kDa were used (all from Sigma Aldrich). As with the peptides, the protein films were initially made from solutions of various concentrations and volumes for carrying out preliminary work to locate and validate side chain cross peaks. This resulted in the discovery of new peaks for tryptophan and the CH$_3$ group.

For the demonstration of EVV 2DIR protein fingerprinting covered in Section 3.4, a set of ten proteins was used. Protein solutions were made using 10 mM maleic acid (pH 2.5) at concentrations of 10 – 50 mg/ml. The protein films were produced in the same manner as the peptide samples, the only difference being that 1.5 µl drops were deposited. The ten proteins used and their solution concentrations were as follows (all were purchased from Sigma Aldrich, their product numbers being in brackets): 0.8 mM albumin from bovine serum (A4503), 1.2 mM albumin from
Chapter 3: Protein identification and absolute quantification with EVV 2DIR

chicken egg white (05461), 0.6 mM aldolase from rabbit muscle (05518), 0.2 mM alkaline phosphatase from bovine intestinal mucosa (P7640), 2.0 mM α-chymotrypsin from bovine pancreas (27270), 2.0 mM concanavalin A from Jack Bean (C2010), 1.8 mM α-lactalbumin from bovine milk (L5385), 1.4 mM β-lactoglobulin B from bovine milk (L3908), 3.5 mM lysozyme from chicken egg white (62970) and 1.5 mM pepsin from porcine gastric mucosa (P6887).

3.2.2: EVV 2DIR experimental methods

Identifying spectral regions of interest

Before protein fingerprinting research commenced, the EVV 2DIR technique was tested using various model compounds. These included small organic molecules (benzene and hexane) and plastics (polyethylene, polypropylene and polystyrene). Other team members carried out experiments in parallel with the protein studies – for example using EVV 2DIR to measure the Fermi resonance couplings in benzene – which continued to provide further insight into the technique. Previous work with proteins carried out by P. M. Donaldson was also consulted and helped to guide the way for locating amino acid cross peaks. Theoretical studies complemented the experimental work; a rigorous comparison between quantum mechanical simulations of benzene EVV 2DIR spectra and those measured experimentally was performed by Donaldson and Guo. [89, 90]

To locate and assign amino acid cross peaks, a number of approaches were used. At the time it was only possible to scan small regions of about 400 cm\(^{-1}\) in the \(\omega_\alpha\) direction, and about 600 cm\(^{-1}\) in the \(\omega_\beta\) direction. Moving between different regions was not trivial, as it required adjustments to the OPAs, beam realignments and recalibrations of the wavelengths. Although changes are continuing to be
implemented to the EVV 2DIR experiment to enable the more widespread and rapid scanning of regions, the practical difficulties at the time meant that judicious choices had to be made regarding the spectral areas to survey for amino acid features. The ongoing studies of model compounds were extremely useful for this purpose, and published infrared and Raman spectra of proteins were also consulted. [39] Quantum mechanical calculations of the EVV 2DIR spectra of amino acid side chains, performed by R. Guo, were also extremely useful for predicting the locations of prospective amino acid signatures. A composite of the complete set of more than twenty theoretical spectra is shown in Figure 3-5. The spectra were, and continue to be, helpful for identifying suitable frequencies to scan. In particular, as a set, they provide guidance on which cross peaks might be more isolated and free of interference from others.

![Figure 3-5: Theoretical EVV 2DIR spectra of all the amino acid side chains; included are those for the three possible protonation states of histidine. The DFT calculations were performed by R. Guo using model di-peptides, consisting of only one side chain residue and with CH₃ termini.](image-url)
Spectral regions studied

The first two-dimensional region to be surveyed for amino acid cross peaks covered a mid-infrared range of frequencies in one direction (ωα: 1300 – 1600 cm\(^{-1}\)), and a near-infrared range in the other (ωβ: 4200 – 4900 cm\(^{-1}\)). Fundamental vibrational modes occur in the mid-infrared, and the overtones and combination bands of these fundamentals occur in the near-infrared. Cross peaks found in the 1300 – 1600 cm\(^{-1}\) / 4200 – 4900 cm\(^{-1}\) frequency ranges therefore tend to be due to the coupling between a fundamental mode and its overtone, or between a fundamental and one of its combination bands.

A cross peak at 1485 / 4380 cm\(^{-1}\) was observed in the 2DIR spectra of some of the kemptide-derived peptides. This was attributed to the CH\(_2\) group found in the majority of residues. The two coupled modes leading to this feature were assigned as a fundamental CH\(_2\) deformation, δ(CH\(_2\)), at 1485 cm\(^{-1}\), and a combination band of this deformation with a second fundamental mode, the CH\(_2\) stretch, ν(CH\(_2\)) + δ(CH\(_2\)), at 4380 cm\(^{-1}\). Only very weak signal was achieved from this peak, and was only observed from two of the kemptides. No other cross peaks from any of the peptide samples were discovered in this range of frequencies.

Two amino acid crystals – glycine and arginine – were also studied in this region. The CH\(_2\) cross peak observed for the kemptides at 1485 / 4380 cm\(^{-1}\) was easily located from both crystals with unsurprisingly higher levels of signal. For arginine a feature, associated with the guanidyl group of its side chain, was also found at 1560 / 4460 cm\(^{-1}\). This cross peak was not observed in the peptide samples and, as finding any signal from these was proving difficult, a new spectral region was considered. The 2DIR spectra of the two amino acids are shown below in Figure 3-6, along with their infrared absorption spectra for comparison.
**Figure 3-6:** EVV 2DIR spectra of (a) glycine and (b) arginine crystals covering the frequency ranges 1400 – 1600 cm\(^{-1}\) / 4300 – 4550 cm\(^{-1}\) (delays of \(T_{12} = 2\) ps and \(T_{23} = 2\) ps). The cross peaks are labelled with the assignments for their coupled vibrational modes. The infrared absorption spectra (measured in a KBr pellet) are also shown, with the positions of the EVV 2DIR cross peaks labelled.
For glycine there are two cross peaks attributed to the coupling of the CH$_2$ deformation, $\delta$(CH$_2$), with a combination band of the deformation and the CH$_2$ stretch, $\nu$(CH$_2$) + $\delta$(CH$_2$). One of these cross peaks, found at 1450 / 4420 cm$^{-1}$, arises from the symmetric CH$_2$ stretch, $\nu_s$(CH$_2$), while the second one at 1450 / 4445 cm$^{-1}$ is due to the asymmetric stretch, $\nu_{as}$(CH$_2$). The frequency shift of the glycine methylene features compared to those of the kemptides and arginine is not surprising, since its CH$_2$ group is of a different nature – all other amino acids have a CH$_2$ as part of their side chains, whereas glycine’s forms part of its backbone.

The arginine cross peak at 1560 / 4460 cm$^{-1}$ is associated with the coupling of a complex mode of its guanidyl group with a combination band. The 1560 cm$^{-1}$ mode is attributed to the guanidyl group as a whole (see Figure 3-2), and is due to various stretches and deformations of the NH$_2$ and CN groups. This complex mode couples with a combination band, at 4460 cm$^{-1}$, involving itself and the CH$_2$ stretch. For the pulse delays used here ($T_{12} = 2$ ps and $T_{23} = 2$ ps), arginine’s $\delta$(CH$_2$) / $\nu$(CH$_2$) + $\delta$(CH$_2$) cross peak is observed at 1470 / 4390 cm$^{-1}$. However, the structure of this peak was subsequently shown by P. M. Donaldson to undergo complex changes upon increasing $T_{12}$ delay. The three CH$_2$ groups of arginine’s side chain are non-identical, thus giving rise to three different modes. The multiple cross peaks fall under the laser excitation bandwidth and, with changes in the $T_{12}$ pulse delay, quantum beating of the coherences can be observed in the EVV 2DIR spectra. [90]

The hunt for new amino acid vibrational signatures continued in the frequency ranges of 1350 – 1600 cm$^{-1}$ / 2700 – 3200 cm$^{-1}$. Strong signal was observed from both peptide and protein samples at 1450 / 2850 cm$^{-1}$, and this cross peak was assigned to two CH$_2$ modes – a deformation, $\delta$(CH$_2$), and its first overtone, 2$\delta$(CH$_2$). Calculations of the side chain spectra indicated that this region could be potentially very fruitful, with strong aromatic side chain cross peaks predicted (Figure 3-5).
Indeed, it was in this region that spectral features attributed to the aromatic amino acids, phenylalanine, tyrosine and tryptophan were observed, along with more CH$_2$ and also CH$_3$ cross peaks. As more data was collected from peptide and then protein samples, it became apparent that these cross peaks would be suitable for the amino acid quantification procedures that form part of the protein fingerprinting strategy. The validation of the use of these aromatic features, and of methyl and methylene groups as internal standards, is covered in Sections 3.3 and 3.4.

**EVV 2DIR experimental procedures**

For the collection of all EVV 2DIR spectra presented in this chapter, the following pulse sequence was used: 1: $\omega_\alpha$ 2: $\omega_\beta$ 3: $\omega_\gamma$. This order was chosen to select the $gg \rightarrow ga \rightarrow ba \rightarrow ea \rightarrow aa$ coherence pathway for the EVV-IR 1 process. The implementation of delays between the pulses limited any interference from the other EVV-IR pathway.

For the peptide spectra shown in Section 3.3, delays of $T_{12} = 1.5$ ps and $T_{23} = 1$ ps were used. These gave a high enough level of FWM signal from the peptides whilst reducing the non-resonant background. All three beams were polarised in the plane of propagation. The exact regions scanned by $\omega_\alpha$ and $\omega_\beta$ were 1350 – 1570 cm$^{-1}$ and 2700 – 3200 cm$^{-1}$ respectively, and both were scanned with 5 cm$^{-1}$ increments.

For the protein spectra presented in Section 3.4, delays of $T_{12} = 2$ ps and $T_{23} = 1$ ps were used. The exact regions scanned by $\omega_\alpha$ and $\omega_\beta$ were 1425 – 1550 cm$^{-1}$ and 2800 – 3200 cm$^{-1}$ respectively, both with 5 cm$^{-1}$ increments. Following the peptide studies, the sensitivity of the EVV 2DIR experiment had been improved through the implementation of photon counting. This allowed the protein data to be collected using longer time delays between the laser pulses, thus producing less congested
Chapter 3: Protein identification and absolute quantification with EVV 2DIR

spectra. The congestion issue was also addressed by polarising the visible input beam orthogonally with respect to the infrared beams; this helped to decongest the spectra even more, rendering more of the peaks exploitable. Thus two different sets of polarisations – PPP (when all beams are polarised in the plane of propagation) and PPS (when the electric field of the visible is perpendicular to the plane of propagation) – were used.

3.3: EVV 2DIR fingerprinting – a proof of principle demonstration in peptides

A set of eight short peptides were used to locate and then assign vibrational signatures of amino acid side chains in EVV 2DIR spectra. These cross peaks were used to perform the first demonstration of the quantification of amino acid levels using two-dimensional infrared spectroscopy. [91]

3.3.1: Peptide EVV 2DIR spectra

The two-dimensional infrared spectra of the set of eight peptides used for this study (LRRFSLGF, LRRFFLG, LRRYSLGY, LRRYYLG, YPFF, YGGFF, YGGFL and YGGFY) are shown in Figure 3-7. These spectra all have the strong CH\textsubscript{2} peak at 1450 / 2850 cm\textsuperscript{-1} in common. Also present are some weaker features centred around 1475 / 2920 cm\textsuperscript{-1} and, more importantly, a set of complex features in the 1440 – 1550 cm\textsuperscript{-1} / 3100 – 3150 cm\textsuperscript{-1} frequency range. These peaks in the ‘aromatic region’ were not observed in the spectra of other peptide samples tested that contained no Y or F residues, for example LRRASLG.
Figure 3-7: EVV 2DIR spectra of the eight peptides used for the proof-of-principle study of the EVV 2DIR fingerprinting strategy. For all spectra, delays of $T_{12} = 1.5$ ps and $T_{23} = 1$ ps were used, and the two infrared beams, $\omega_\alpha$ and $\omega_\beta$, were scanned with 5 cm$^{-1}$ increments ($45 \times 101$ points per spectrum). The spectra are presented linearly scaled and, though the colour scales are linear, they are not identical from one spectrum to another to help highlight the observed features. The peak at 1450 / 2850 cm$^{-1}$, common to all the spectra, is attributed to aliphatic CH$_2$ vibrational modes, and is taken as an internal reference. The complex features in the 1440 – 1550 cm$^{-1}$ / 3100 – 3150 cm$^{-1}$ frequency range are due to aromatic vibrational modes of the side chains of tyrosine and phenylalanine.
For peptides in which phenylalanine is the only aromatic amino acid present (LRRFSLG and LRRFFLG), five peaks are observed in the aromatic range. Three of these form a triplet along $\omega_\alpha \approx 1515 \text{ cm}^{-1}$ at $\omega_\beta = 3110$, 3080 and 3040 cm$^{-1}$; the two other peaks are at 1470 / 3050 cm$^{-1}$ and 1480 / 3070 cm$^{-1}$. The spectra of the peptides that contain tyrosine, but no phenylalanine residues (LRRYSLGY and LRRYYLG), have two cross peaks in the aromatic region; a very weak one at 1470 / 3040 cm$^{-1}$ and a much stronger one at 1530 / 3130 cm$^{-1}$. For the amino acid quantification procedure, three of the phenylalanine peaks (the two lower frequency peaks, and the most intense of the triplet) and the stronger of the two tyrosine peaks were used.

### 3.3.2: Spectral signatures of tyrosine, phenylalanine and CH$_2$ in peptides

The spectral signatures identified in the peptide spectra are shown in Figure 3-8. As well as the tyrosine, phenylalanine and methylene peaks, the two features that were observed at $\sim 1475 / 2920 \text{ cm}^{-1}$ are shown. The peak at 1475 / 2930 cm$^{-1}$ has tentatively been assigned to leucine (L) CH$_3$ groups, whilst the one at 1460 / 2900 cm$^{-1}$ remains unassigned.

The tyrosine and phenylalanine cross peaks were assigned their specific vibrational modes using DFT calculations (performed by R. Guo) of the EVV 2DIR spectra of benzene derivatives and of aromatic amino acid side chain models. Figure 3-9 shows the theoretical spectra of toluene and methyl-phenol, used as analogues for phenylalanine and tyrosine respectively, compared with the aromatic regions of the experimental peptide spectra of LRRFSLG and LRRYSLGY.
Figure 3-8: Features identified in the peptide spectra as being suitable for performing the fingerprinting procedure. These are a CH$_2$ cross peak (to be used as an internal reference), three phenylalanine cross peaks and one tyrosine cross peak. Also shown is a lower frequency feature attributed to leucine. The unassigned peak is possibly due to a weak contribution from glycine. The complete mode assignments are detailed in Table 3-1.

Figure 3-9: A comparison of theoretical spectra of toluene and methyl-phenol, used as analogues for phenylalanine and tyrosine respectively, with the experimental spectra of LRRFSLGF and LRRYSGLY. Figure courtesy of Guo and Fournier.
The observed aromatic peaks in the peptide spectra have contributions from two ring stretching modes: CC + HCC in plane ($\nu_{13}$) and CC + HCC in-plane + CCC in-plane ($\nu_{16}$). The full list of cross peak frequencies and their mode assignments is shown in Table 3-1. Along $\omega_{\alpha}$ the vibrational modes are fundamentals, whereas those in the $\omega_{\beta}$ direction are either overtones or combination bands involving the fundamentals.

<table>
<thead>
<tr>
<th>Amino acid or group</th>
<th>Cross peak frequencies (cm$^{-1}$)</th>
<th>Mode assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH$_2$</td>
<td>1450 / 2850</td>
<td>$\delta$(CH$_2$) / 2$\delta$(CH$_2$)</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>1530 / 3130</td>
<td>$\nu_{13} / \nu_{13} + \nu_{16}$</td>
</tr>
<tr>
<td>Phenylalanine 1</td>
<td>1515 / 3110 (Not used)</td>
<td>$\nu_{13} / \nu_{13} + \nu_{16}$</td>
</tr>
<tr>
<td></td>
<td>1515 / 3080 (Not used)</td>
<td>$\nu_{13} / \nu_{13} + (\delta$(CH$<em>2$) &amp; $\nu</em>{16}$)</td>
</tr>
<tr>
<td></td>
<td>1515 / 3040 (Not used)</td>
<td>Not assigned</td>
</tr>
<tr>
<td>Phenylalanine 2</td>
<td>1480 / 3070</td>
<td>($\delta$(CH$<em>2$) &amp; $\nu</em>{13}$) / ($\delta$(CH$<em>2$) &amp; $\nu</em>{13}$) + $\nu_{16}$</td>
</tr>
<tr>
<td>Phenylalanine 3</td>
<td>1470 / 3050</td>
<td>($\delta$(CH$<em>2$) &amp; $\nu</em>{13}$) / ($\delta$(CH$<em>2$) &amp; $\nu</em>{13}$) + ($\delta$(CH$<em>2$) &amp; $\nu</em>{16}$)</td>
</tr>
</tbody>
</table>

Table 3-1: Mode assignments, based on those of benzene, for the main cross peaks observed in the peptide spectra. $\nu_{13}$ and $\nu_{16}$ are two ring stretching modes, CC + HCC in-plane and CC + HCC in-plane + CCC in-plane respectively. $\delta$ denotes a deformation, and $\nu$ denotes a stretch. Bracketed modes indicate complex modes that have two contributions. Combination bands are indicated with an addition sign.

3.3.3: Quantification of amino acids levels in peptides

The key objective of the fingerprinting strategy is to ascertain the identities of unknown peptides and proteins through their spectroscopically-determined amino acid compositions. The spectra of the eight peptides were used to validate the strategy of ratioing the intensity of an amino acid side chain cross peak with that of an internal reference peak (in this case CH$_2$). Using the measurements from these
samples, the linear correlation between experimentally-determined and actual amino acid / CH$_2$ ratios of peptides was demonstrated for phenylalanine and tyrosine.

**Peak intensity measurements**

Values for the cross peak intensities were obtained in two different ways. The first of these was a simple peak amplitude measurement, which demonstrates the high through-put potential of the EVV 2DIR technique. This was simply achieved by measuring the signal level from the peak of interest ($S_{Peak}$), and then subtracting the level of signal measured from an off-resonance point ($S_{Off-res}$) on the spectrum ($I_{Peak} = S_{Peak} - S_{Off-res}$). The second type of measurement, performed by Fournier, involved calculating the integrated intensity of each cross peak of interest. In order to do this, the cross peaks were first fitted with two-dimensional Gaussian functions. Peak amplitudes and FWHM were then extracted from these fits to use for calculating the integral of each peak independently from the others. [91]

Although the intensities of the infrared beams are wavelength dependent, these variations are the same for each scan and so the peaks’ intensities are not affected and are directly comparable from one 2DIR scan to another. To test for variations in the measured intensity ratios from sample to sample, reproducibility tests were performed using eight different YPFF samples. For the experimentally-determined Phe / CH$_2$ ratios, a standard deviation of ± 16 % was estimated.

**Quantification procedure**

The EVV 2DIR intensity ratios for the three phenylalanine peaks and the one tyrosine peak were each plotted as a function of the known number ratios of these amino acids in the eight peptide samples. This was performed for the two types of
peak measurement; the peak amplitudes and the peak integrated intensities. In both cases the intensities were square-rooted before ratioing. The phenylalanine and tyrosine correlation curves are shown in Figures 3-10 (a) and 3-10 (b) respectively. The curves are fitted with linear functions, constrained to go through the origin; for example, the spectra of samples without any Phe residues (LRRYLYLG and LRRYSLGYY) do not show any features at the Phe frequencies, and so these peak intensities are zero.

It should be noted that the gradients of the linear fits are not equal to one; the peak intensity ratios measured for the four different aromatic cross peaks (√I_{AA} / √I_{CH2}) are greater than the peptides’ known number ratios (N_{AA} / N_{CH2}). For example, in the case of the Phe 3 cross peak of the YGGFF peptide, the peak intensity amplitude measurement gives a ratio of 1.0, whilst its number ratio equals 0.4. Although a one-to-one relationship between the actual and measured AA / CH\textsubscript{2} ratios might be expected, it is not surprising that this is not observed. This can be attributed to a couple of factors. First, the oscillator strengths of the aromatic stretching modes could be stronger than that of the CH\textsubscript{2} deformation mode. Therefore, one aromatic side chain could result in a higher level of FWM signal than a number of CH\textsubscript{2} groups. Additionally for this peptide, two of the CH\textsubscript{2} groups counted in the number ratio are from glycine residues. Due to the different nature of the glycine methylene group (rather than being on its side chain, it forms part of the backbone), the level of its contribution to the 1450 / 2850 cm\textsuperscript{-1} feature could differ greatly from that of the side chain methylene groups of the tyrosine and phenylalanine residues.

The gradients of the linear fits also differ for the three different phenylalanine cross peak measurements. Again, this might be expected due to the differing oscillator strengths of the various ring modes that are associated with these cross peaks. Although the gradients for the calibration graphs also differ between the amplitude
and integral measurements, there doesn’t appear to be a significant pattern indicating that one or other of these measurements produces a peak intensity ratio closer to the number ratio of the sample. (For the Phe 1 and Tyr calibration graphs, the integral measurement produces the smaller gradient. For the Phe 2 and Phe 3 cross peaks, however, the amplitude measurements result in the smaller gradients.) Nevertheless, it is not important that there is not a one-to-one relationship between the experimentally determined and actual AA/CH$_2$ ratios – as long as the correlation is linear, the EVV 2DIR fingerprinting strategy is practicable.

The standard deviations for the dispersion of the data points in the calibration curves of Figure 3-10 are given in Table 3-2. These deviations from the linear behaviour can have two different origins: structural effects (protein primary and secondary structure) and sample to sample variations. Such variations not only include differences in samples, but also instrumental changes. These can be fluctuations over time of the intensity profiles of the lasers, and also pointing changes which induce differences in the phase matching conditions. It is not clear yet if the precision is limited by structural effects or sample and instrumental variations. A comparison of the standard deviations resulting from the peak amplitude (average value ± 11.5%) and peak integral (average value ± 13.4%) measurements, show that the fitting procedure to calculate integrated intensities of peaks is not required. A simple cross peak amplitude measurement is sufficient, which also means that there is no need to collect full 2DIR spectra of samples. To carry out the quantification procedure, only measurement of signal at the frequencies of interest is required.
Figure 3-10: (a) Correlation between the known phenylalanine to CH$_2$ ratios, and the ratios measured with EVV 2DIR of the eight peptide samples. Data was collected from three different cross peaks: Phe 1 (1515 / 3110 cm$^{-1}$), Phe 2 (1480 / 3070 cm$^{-1}$) and Phe 3 (1470 / 3050 cm$^{-1}$). (b) Correlation between the known tyrosine to CH$_2$ ratios, and the ratios measured with EVV 2DIR of the eight peptide samples. Data was collected from one cross peak at 1530 / 3130 cm$^{-1}$.

The square root of the intensity of each amino acid feature relative to the intensity of the CH$_2$ feature is plotted as a function of the known amino acid / CH$_2$ number ratios of the samples. Two types of intensity measurement are shown: peak amplitudes and peak integrals. The data points for each of the peptide samples are shown by open circles and are labelled with their corresponding sequences. The error bars are the standard deviations from four repeat measurements, and the blue line is the linear fit constrained to go through the origin.
<table>
<thead>
<tr>
<th>Cross peak</th>
<th>Integral data (%)</th>
<th>Amplitude data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine (1530 / 3130 cm⁻¹)</td>
<td>± 11.5</td>
<td>± 12</td>
</tr>
<tr>
<td>Phenylalanine 1 (1515 / 3110 cm⁻¹)</td>
<td>± 13</td>
<td>± 10.5</td>
</tr>
<tr>
<td>Phenylalanine 2 (1480 / 3070 cm⁻¹)</td>
<td>± 15</td>
<td>± 10.5</td>
</tr>
<tr>
<td>Phenylalanine 3 (1470 / 3050 cm⁻¹)</td>
<td>± 14</td>
<td>± 13.5</td>
</tr>
<tr>
<td>Average value</td>
<td>± 13.4 (± 1.5)</td>
<td>± 11.5 (± 1.6)</td>
</tr>
</tbody>
</table>

Table 3-2: The standard deviations for the dispersion of the data points in Figure 3-10 from linear behaviour. Data is shown for both the peak integrated intensities and the simple peak amplitude measurements.

### 3.3.4 Conclusions of peptide fingerprinting studies

EVV 2DIR spectroscopic features corresponding to the aromatic vibrational modes of tyrosine and phenylalanine were observed from peptides. Importantly, also identified was a cross peak from the amino acid methylene (CH₂) group, suitable for use as an internal reference. Using the EVV 2DIR spectra collected from peptides containing differing amounts of phenylalanine and tyrosine, it was demonstrated that the ratios of the signal strengths of Phe and Tyr cross peaks, relative to that of the CH₂ peak, are proportional to the actual Phe / CH₂ and Tyr / CH₂ number ratios of peptide samples. This linear correlation, between the relative intensity of an amino acid vibrational feature and the relative amount of that residue present in a sample, demonstrates that relative quantification of amino acid content by EVV 2DIR is possible.

The peptide two-dimensional spectra in Figure 3-7 were collected at 0.5 seconds per data point, with each spectrum comprising 4545 points. The precisions achieved for the peak amplitude measurements (Table 3-2), however demonstrate that there is no need to acquire entire 2DIR spectra to perform a relative quantification procedure.
Once spectral features have been located, assigned and identified, the signal strength of a peak can be retrieved from one well-chosen data point. Providing the points have relatively little interference from surrounding features (in other words that the spectral features are resolved), the measurements of peak amplitudes are sufficiently accurate. Preliminary bioinformatics studies (Chapter 4) suggest that between five and ten amino acid / CH$_x$ ratios are required to achieve a protein identification. Since this would only require the measurement of signal strengths from six to ten points (each amino acid peak, plus that of the CH$_3$), data collection could be very rapid indeed. If the acquisition time for each peak is under 5 seconds, then the data required for a protein identification could be acquired in less than 50 seconds.

### 3.4: Protein fingerprinting with EVV 2DIR

Following the proof of principle of the EVV 2DIR fingerprinting strategy in peptides, the next step was to attempt the same types of measurements in proteins. Although the quantification of relative amino acid levels had been demonstrated in short peptides, it was possible that structural effects would prevent such measurements on proteins. The results presented in this section show that these structural sensitivities are not limiting factors, and that the quantification of relative amino acid levels in proteins can be achieved. [92]

EVV 2DIR spectral signatures for three aromatic amino acids, phenylalanine, tyrosine and tryptophan, were observed from proteins, along with two possible internal reference cross peaks of the methylene (CH$_2$) and methyl (CH$_3$) groups. The tryptophan peak had not previously been observed in peptide samples. Measurements of the intensities of the peaks were used to demonstrate quantification of the levels of the three aromatic amino acids in protein samples.
3.4.1: Protein EVV 2DIR spectra

EVV 2DIR spectra covering the 1430 – 1550 / 2800 – 3200 cm\(^{-1}\) region were collected for a set of ten protein samples, ranging in size from 14 kDa to 66 kDa (details of these can be found in Section 3.2). Each protein’s spectrum was measured using two different polarisation schemes; PPP (when all three input beams are polarised in the plane of propagation) and PPS (when the two infrared beams are polarised in the plane of propagation and the visible is orthogonal to these). The change in polarisation of the visible beam resulted in less congested spectra, thus making previously obscured peaks exploitable. Figure 3-11 shows three example protein spectra; an \(\alpha\)-chymotrypsin spectrum measured with the PPS polarisation combination, and two spectra obtained from pepsin using the PPS and PPP schemes.

All the protein spectra have an intense, structured peak at \(\sim 1475 / 2920\) cm\(^{-1}\) in common which has been attributed to a collection of various aliphatic modes. This was the first peak to be observed from protein samples, as the established peptide \(\text{CH}_2\) cross peak at 1450 / 2850 cm\(^{-1}\) was a useful guide. The PPP protein spectra have three other features at 1485 / 3070 cm\(^{-1}\), 1525 / 3120 cm\(^{-1}\) and 1545 / 3150 cm\(^{-1}\). The first two of these are due to vibrational modes of the phenylalanine side chain, and the latter has been attributed to tyrosine. The 1485 / 3070 cm\(^{-1}\) phenylalanine cross peak is also present in the PPS spectra. The previous observation of the vibrational signatures of these aromatic side chains in peptide spectra aided in their discovery from protein samples. Changing the polarisation of the visible beam resulted in the detection of a spectroscopic feature for a third aromatic amino acid, tryptophan. This cross peak at 1490 / 3020 cm\(^{-1}\) had not previously been observed in peptide spectra.
Figure 3-11: Example EVV 2DIR protein spectra of two proteins, covering the 1430 – 1550 cm\(^{-1}\) / 2800 – 3200 cm\(^{-1}\) region. Two different beam polarisation configurations were used when measuring the spectra; PPP (all beams have their fields in the plane of propagation) and PPS (the electric field of the visible beam is perpendicular to the plane of propagation). Pulse delays of \(T_{12} = 2\) ps, \(T_{23} = 1\) ps (established to be a compromise between signal strength and decongestion) were used for all spectra. Shown is a comparison of the pepsin spectra measured using the two different polarisation schemes. The spectrum of \(\alpha\)-chymotrypsin, taken using the PPS combination, shows the appearance of the tryptophan cross peak at ~ 1490 / 3020 cm\(^{-1}\). The spectra are presented linearly scaled and, though the colour scales are linear, they are not identical from one spectrum to another to help highlight the observed features.

3.4.2: Spectral features of tyrosine, phenylalanine, tryptophan and CH\(_3\) in proteins

The structured feature at around 1475 / 2920 cm\(^{-1}\) contains contributions from both methylene and methyl groups. As for the peptides, the coupled vibrational modes involved are the CH deformation, \(\delta(CH)\), and its first overtone, \(2\delta(CH)\). The peak at 1480 / 2960 cm\(^{-1}\) was attributed to the CH\(_3\) group in particular, and was found to
work well as an internal standard. A complementary experiment, not shown here, showed that a peak at 1460 / 2920 cm\(^{-1}\) can also be used as the internal reference.

The frequencies and mode assignments of the aromatic cross peaks are detailed in Table 3-3. Although the Tyr and Phe 1 features are only 20 – 30 cm\(^{-1}\) apart and appear not fully resolved, the cross contamination is sufficiently small for the cross peak intensity to be measured on resonance. Changing the polarisation of the visible beam from PPP to PPS renders these aromatic modes very weak – the Tyr and Phe 1 peaks can be seen in the PPP, but not the PPS, pepsin spectrum of Figure 3-11. The phenylalanine peak at 1485 / 3070 cm\(^{-1}\) (Phe 2) is observed in protein spectra measured using both beam polarisation combinations. Importantly, when the PPS scheme is used, as well as the Tyr and Phe 1 being extinguished, the congestion in the spectral region around Phe 2 is relieved. This reveals the clearly resolved tryptophan peak at 1490 / 3020 cm\(^{-1}\). The Trp feature can be seen in Figure 3-11 in the spectrum of the tryptophan-rich protein, \(\alpha\)-chymotrypsin. This cross peak has since been observed from peptides containing tryptophan, for example LRRWSLG.

<table>
<thead>
<tr>
<th>Amino acid cross peak</th>
<th>Cross peak frequencies (cm(^{-1}))</th>
<th>Mode assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyr (P)</td>
<td>1545 / 3150</td>
<td>(v_{13} / v_{13} + v_{16})</td>
</tr>
<tr>
<td>Phe 1(P)</td>
<td>1525 / 3120</td>
<td>(v_{13} / v_{13} + v_{16})</td>
</tr>
<tr>
<td>Phe 2(P) and Phe 2(S)</td>
<td>1485 / 3070</td>
<td>((\delta(CH_2) &amp; v_{13}) / (\delta(CH_2) &amp; v_{13}) + v_{16})</td>
</tr>
<tr>
<td>Trp (S)</td>
<td>1490 / 3020</td>
<td>Not yet assigned.</td>
</tr>
</tbody>
</table>

Table 3-3: Mode assignments, based on those of benzene, for the main cross peaks observed in the protein spectra. \(v_{13}\) and \(v_{16}\) are two ring stretching modes, CC + HCC in-plane and CC + HCC in-plane + CCC in-plane respectively. \(\delta\) denotes a deformation, and \(v\) denotes a stretch. Bracketed modes indicate complex modes that have two contributions. Combination bands are indicated with an addition sign.
The use of different visible beam polarisations provides additional information from the protein spectra. Both polarisation schemes produce cross peak data that can be used for the amino acid quantification procedure. In summary, the intensities of four spectral features corresponding to three amino acid residues were used: the Tyr (P) and Phe 1(P) cross peaks using the PPP polarisation scheme, and the Phe 2(S) and Trp (S) cross peaks using the PPS polarisations. This means that two independent measurements were made of phenylalanine which both contribute to the overall dataset for the protein. The internal reference CH₃ peak can be measured for both polarisation schemes.

### 3.4.3: Quantification of amino acid levels in proteins

#### Peak intensity measurements

The tests of two different methods for measuring peak intensities covered in Section 3.3.3 demonstrated that once spectral features have been identified and associated with amino acids, there is no need for entire two-dimensional spectra to be recorded for each protein. The signal intensity can simply be measured for the pairs of infrared frequencies corresponding to the peaks of interest.

For the measurements presented here, the intensity of each peak was recorded over a period of 120 seconds, with 5 seconds per point. An average signal was taken from these measurements, and the deduced standard deviations ensured that the conditions corresponded to the photon-counting regime. To minimize the photon-counting nonlinearity, the absolute signal was constrained at < 100 photons per second by using calibrated neutral density filters. The incident number of photons was then calculated from the detected number of photons and corrected for non-linearity.
Before ratioing, the signal levels were corrected from the non-resonant background. The measured ratios were calculated as follows, where $S_{AA}$ is the average signal measured for an amino acid peak, $S_{Off-res}$ is the average background signal (off resonance), and $S_{Ref}$ is the average signal of the reference peak:

$$\text{Measured ratio} = \frac{S_{AA} - S_{Off-res}}{\sqrt{S_{Ref} - S_{Off-res}}}$$

Equation 3-2

**Amino acid quantification**

The measured ratios for each of the four identified aromatic cross peaks, Tyr ($P$), Phe 1($P$), Trp ($S$) and Phe 2($S$), were calculated by dividing the square root of the intensity of each peak by the square root of the intensity of the CH$_3$ reference peak. Figure 3-12 shows the measured ratios plotted as a function of the actual amino acid to CH$_3$ ratios for each of the ten proteins analysed.

In order to assess the reproducibility, the measurements were repeated four times for each of the ten proteins – the error bars on the graphs in Figure 3-12 are the standard deviations from these measurements. The calculated average precisions on the amino acid / CH$_3$ ratios were deduced from the average experimental error bars and are denoted ‘Precision’. The horizontal dispersions of the data points compared to the linear fits are the average absolute differences and are denoted ‘Dispersion’. This is essentially the standard deviation due to variation of the individual protein points from the linear fit.
Figure 3-12: Measured amino acid to internal reference (CH$_3$) cross peak intensity ratios plotted as a function of the known ratios for the ten proteins analysed. Data is shown for four cross peaks (Phe 1(P), Phe 2(S), Tyr (P) & Trp (S)), corresponding to three different amino acids. Each data point comes from one of the ten proteins, and the solid lines are the linear fits constrained through the origin. The error bars are standard deviations from four repeat measurements. The calculated average precisions of the measured amino acid / CH$_3$ ratios and the horizontal dispersions of the data points from the linear fits are shown on each graph.

3.4.4: Conclusions of protein fingerprinting studies

Two-dimensional vibrational signatures of three aromatic amino acids – phenylalanine, tyrosine and tryptophan – were identified from proteins. The signal strengths of these cross peaks, relative to that of a methyl internal reference peak, were used to quantify the relative levels of the amino acids in a set of ten proteins. The direct correlation between the EVV 2DIR measured level of each amino acid and the known amount of that amino acid in the test proteins suggests that these cross peaks are not structurally sensitive, and that this approach can be applied efficiently to protein identification.
In some cases it appears there is a limitation on the precision with which the content of particular amino acids can be determined. This can be seen by the spread of the data points around the straight line fit in Figure 3-12 for tryptophan. This spread, or dispersion, is greater than that predicted from experimental precision alone. The implication is that there is some small residual structural sensitivity for this particular cross peak.

For amino acids where this limitation is reached (higher dispersion than precision, meaning that there is some residual structural effect influencing the cross peak intensity), improved identification capability will only be achieved by finding spectral features for more amino acids, rather than by further signal averaging for that particular cross peak. On the other hand, the use of different polarisations for phenylalanine shows that higher precisions can be achieved by taking more than one peak per amino acid if desired.

The amino acid quantification curves are the first steps towards using the EVV 2DIR technique for performing protein identification. For a protein to be identified, intensity ratios will need to be measured for five to ten different amino acid residues to form a characteristic fingerprint (see Chapter 4). The capability of the technique for differentiating proteins has already been demonstrated by Fournier [92] using the data set of amino acid / CH$_3$ ratios presented in this section. The approach defines the extent to which two proteins can be distinguished using the level of overlap of their amino acid / CH$_3$ ratios. The probability of two proteins being identical is calculated using overlap integrals of the ratios, and then represented in differentiation maps. The degree of identicality of two proteins is different depending on which amino acid ratio is being considered. Combining the overlap integrals for the different ratios demonstrates how the proteins become increasingly distinguishable from one another (less identical) as data for more amino acid ratios
are used. Figure 3-13 shows two differentiation maps for the set of ten proteins; the first shows the case for when only the tyrosine ratio is used to distinguish the proteins, and the second for when all four ratios are used.

![Figure 3-13: Differentiation maps of the ten proteins studied for (a) when the tyrosine peak only is used, and (b) when all four amino acid ratios are used. White squares correspond to normalized overlap integral values of 1 (completely indistinguishable) and black to 0 (completely distinguishable). The grey level of each square reflects the probability that the two proteins being compared are the same protein. The diagonal compares a protein with itself and so is white. The second map demonstrates how most of the ten proteins become completely distinguishable from one another when all four of their amino acid ratios are known. Diagram courtesy of Fournier, [92]](image)

### 3.5: Protein absolute quantification using EVV 2DIR

A significant property of EVV 2DIR for proteomic applications is its potential for simple absolute quantification of protein levels. Absolute quantification is becoming an increasingly important issue in proteomics, but has proved difficult to achieve with other established technologies such as mass spectrometry. Quantification with EVV 2DIR is however a relatively uncomplicated procedure which simply involves measuring the intensity of the CH$_3$ feature at 1485 / 2930 cm$^{-1}$. 
Chapter 3: Protein identification and absolute quantification with EVV 2DIR

By mapping the CH$_3$ peak intensity across a series of protein drop-coated films of varying concentrations, it is possible to demonstrate that the integrated intensity of the square rooted image of each protein deposit is proportional to the total number of protein molecules it contains. The intensity measurements can then be converted into signal levels per pixel (the area probed by the laser beams) and plotted as a function of the number of protein molecules in each pixel. This provides an effective calibration graph for determining the number of protein molecules being probed by the lasers from the level of EVV 2DIR signal being achieved.

This approach is possible because the average oscillator strengths for the CH$_3$ cross peaks appear to be the same for all proteins so far studied. It is presumed that this reflects the fact that the oscillator strength is an average over many CH$_3$ groups in each protein, and that these groups are relatively insensitive to structural effects.

3.5.1: Sample preparation and data acquisition

A series of five drop-coated films of concanavalin A were produced on a polished stainless steel substrate. Solutions with concentrations of 0.1, 0.2, 0.3, 0.4 and 0.5 mM were used, and the volume of each deposited droplet was 0.3 µl. This corresponded to approximately $2 \times 10^{13}$, $3.5 \times 10^{13}$, $5.5 \times 10^{13}$, $7 \times 10^{13}$ and $9 \times 10^{13}$ total protein molecules respectively in each of the five deposits. Due to the heterogeneity of the analyte density across the dried films, the EVV 2DIR signal on the CH$_3$ peak ($1485 / 2930$ cm$^{-1}$) was mapped across the whole of each of the five protein deposits. Images were acquired for four different sample sets, an example of which is shown in Figure 3-14.
Figure 3-14: EVV 2DIR image of five concanavalin A films mapped on the CH$_3$ cross peak at 1485 / 2930 cm$^{-1}$. Each of the films is labelled with the number of protein molecules it contained. The image was measured in reflection mode using delays set at $T_{12} = 1$ ps and $T_{23} = 1$ ps, a 100 µm step size and a 1 second acquisition time per point in photon counting mode. Due to the ‘coffee-ring effect’, the majority of the protein solute dries around the periphery of each deposit - this is reflected in the high levels of CH$_3$ signal observed from the outer rings compared to the centre of the films.

3.5.2: Absolute quantification procedure

The acquired images were first corrected for background signal and any photon counting non-linearity. For each film, the integrated intensity of its square-rooted image was then calculated and plotted as a function of its known protein content. The square-rooted integrated signal intensity of a deposit was shown to be proportional to the total number of concanavalin A molecules it contained – shown in Figure 3-15 (a). This proportionality has also been demonstrated using bovine serum albumin. [92, 93]

The integrated intensities of the films were then converted to signal levels per pixel by simply dividing them by the number of pixels (the interaction area of the beams and thus the selected step size of 100 µm by 100 µm) covered by each film.
Similarly, the total number of proteins in each film was divided by the total number of pixels it covered. A simple plot of signal / pixel against the number of proteins / pixel – Figure 3-15 (b) – provides a calibration curve which can be used to determine the number of protein molecules in the laser beams corresponding to any level of signal produced.

Figure 3-15: (a) Square root of the EVV 2DIR signal level as a function of the number of protein molecules that were probed. The integrated EVV 2DIR intensity (I) of the square-rooted image for each protein film is plotted against the total number of protein molecules (N_{ConA}) present in the deposited droplet. The error bars are standard deviations from repeats performed on four different sets of five concanavalin A samples. The solid line represents the linear fit. (b) The signal level per pixel of the concanavalin films plotted as a function of the number of molecules per pixel. This provides an effective calibration curve for determining the number of protein molecules corresponding to a particular level of signal.
Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

4.1: Introduction

The proposed EVV 2DIR protein identification strategy comprises two steps – the determination of amino acid relative levels by measuring the intensities of their spectroscopic signatures, followed by database searching using these measurements. The proof of principle of the spectroscopy aspect was demonstrated in Chapter 3, where the linear correlation between the known amino acid / CH₃ ratios, and those measured with EVV 2DIR, was confirmed for a set of ten proteins. Further to this, it was shown that these proteins could be distinguished from one another using the measured ratios. The bioinformatics challenge, which was addressed in parallel with the experimental work, however centred on how to identify unknown proteins using their spectroscopically-determined amino acid / CH₃ ratios.

Although now rarely used as it has been largely superseded by mass spectrometry techniques, there exists an established approach for performing the identification of proteins using their amino acid compositions (Section 1.2). In fact the ExPASy proteomics server provides a specific search tool for this purpose. [29] The work covered in this chapter, however, concentrated on determining whether amino acid
levels that use the CH$_3$ groups as an internal standard can also be used to identify proteins. To this end, a specialised database was constructed comprising proteins’ amino acid / CH$_3$ ratios. This was used to perform tests of the viability of the CH$_3$ group internal reference method, and to establish how many different EVV 2DIR-determined ratios are required to uniquely identify proteins. The first version of the database covered only the human proteome, but this was later expanded to include protein data sets for more than ten model organisms. A web-based interface for searching the database was also created. Finally, the EVV 2DIR data collected of the ten proteins’ phenylalanine, tyrosine and tryptophan levels shown in Chapter 3 was used to carry out simulated database searches. These provide an indication of the capability of the technique for protein identification even when as few as three amino acid ratios are known.

In order to construct the specialised ratios database, a source for the protein sequences was required. Before commencing work, it was therefore necessary to assess the numerous established protein databases currently available to biologists. The gateways to these, and tools for performing searches, were also appraised to provide guidance for the construction of a search facility. A brief summary of the main protein archives and their means of access are provided in the next section.

The work covered in Sections 4.3 and 4.4, regarding database construction and testing, was completed in conjunction with D. Kedra of the Imperial College Bioinformatics Support Service. [86]
4.2: Protein databases and analysis tools

What is considered to be the first protein sequence database, the Atlas of Protein Sequence and Structure [115], was created in the 1960s by Margaret Dayhoff, as a collection of sequences for investigating the evolutionary relationships between proteins. By the 1980s, sequence information had become increasingly abundant in literature, with vast amounts of data being produced across the research community. Laboratories realised that it would be advantageous to start pooling their results in a central databank, so that they could be accessed and used by other researchers. Database projects started to grow and these evolved into the many central repositories of biological data available today.

Database construction and maintenance involves two main processes; archiving and annotating. Archiving involves the gathering together and curation of the data, to make it accessible and in a fit state for use. Annotation procedures entail the interpretation and compilation of any available information about the data. There now exists a plethora of different databases containing nucleic acid sequences, protein sequences, protein structures and much more. The many different projects archive and annotate the available data in various ways. In recent years, however, some of the larger databases have joined to form consortiums or have brought in smaller projects. This has helped to achieve greater sequence coverage and to rationalise the wealth of data by providing standardised annotations and identifiers.

The three principle nucleic acid databases are EMBL-Bank (European Molecular Biology Laboratory) [116, 117], GenBank (from the National Center for Biotechnology Information, NCBI) [118, 119] and DDBJ (DNA Data Bank of Japan) [120]. These now form a triple partnership, EMBL / GenBank / DDJB, which provides a comprehensive public nucleic acid sequence archive. These each collect a
portion of the total sequence data reported worldwide each day, and then exchange, archive and annotate it. The sources of the DNA and RNA sequences are assembled from genome projects, scientific literature, direct author submissions and patent applications. [121] Of particular interest here, however, are the protein sequence databases, the most comprehensive and widely used of which are UniProtKB (renowned for its high level of annotation), Ensembl (supplies predicted gene transcripts using genome assemblies) and the International Protein Index (brings together results from gene prediction algorithms with experimentally-determined mRNA and protein sequences). These are covered in more detail in Section 4.2.1.

Access to the numerous different biological archives is provided through various information retrieval systems. Sequences can be retrieved from databases on the basis of a wide variety of search parameters, including standard identifiers, sequence patterns and annotation features. Analysis tools can be used to perform a vast number of tasks such as sequence comparisons, pattern recognition and structure analysis and prediction. The two most widely used search retrieval systems for molecular biology databases are Entrez [122] and the Sequence Retrieval System, (SRS) developed at the European Bioinformatics Institute (EBI) [123]. These are integrated with, and provide access, to over two hundred different databases. Of particular relevance to the EVV 2DIR fingerprinting strategy, however is the ExPASy (Expert Protein Analysis System) proteomics server, which is dedicated to the analysis of protein sequences and structures. Of the many identification tools it provides, the AACompIdent program is of significant interest as it identifies proteins by virtue of their amino acid compositions. ExPASy is covered in more detail in Section 4.2.2.
4.2.1: Protein databases

**UniProtKB (Universal Protein Resource Knowledgebase)**

The Universal Protein Resource Knowledgebase (UniProtKB) is part of UniProt [31-33], a central repository and access point of all publicly available information of protein sequence and function. The UniProt consortium was created in 2002 when three protein databases (Swiss-Prot, TrEMBL and PIR-PSD [124]) which had formerly coexisted with different protein sequence coverage, came together to join their information and standardise their annotations. UniProtKB not only collects together core data about its protein entries (amino acid sequence, name or description, taxonomic data and citation information), but also provides comprehensive, manual annotation (including details of function, classifications and cross-references). [32] The database is updated every three weeks, and data sets are freely available for download from the website. Release 13.3 (April 2008) of UniProtKB contains 6 million entries.

The UniProtKB consists of two sections: UniProtKB / TrEMBL and UniProtKB / Swiss-Prot. UniProtKB / TrEMBL contains computationally-generated protein sequences, the entries of which are automatically annotated and classified. The TrEMBL protein sequences are mainly derived from the translation of the coding sequences (CDS) present in the EMBL / GenBank / DDBJ nucleotide databases, along with those from other more specialised resources, including The Arabidopsis Information Resource (TAIR) [125] and the Saccharomyces Genome Database (SGD) [126]. The entries in TrEMBL are not reviewed and are considered preliminary. Overtime, the entries undergo manual annotation and curation, and are then converted to full UniProtKB / Swiss-Prot status. [33]
The UniProtKB / Swiss-Prot database brings together, and manually annotates, experimental results scanned from literature, direct submissions from authors, the sequences of PDB structures and the computer-predicted sequences from the TrEMBL database (from where they are then removed once they have been manually annotated and reviewed). The manual annotation process is performed by curators who analyse and compare all the available sequences for a given protein. The computer-predicted and experimentally-determined data are compared and reviewed, and, to achieve minimal redundancy, all protein sequences encoded by the same gene are merged into a single Swiss-Prot entry. Biological information is also extracted from literature in order to provide as much information as possible about a particular protein, including descriptions of its functions, structures and post-translational modifications. [32]

**Ensembl**

Whereas general nucleotide databases, such as the EMBL / GenBank / DDBJ triple partnership, focus on collecting individual sequences, the Ensembl project [127-129] brings together all the information available about particular eukaryotic species to produce genome databases. Ensembl is not involved with sequencing genomes, nor does is assemble them; for each species, the goal is to collect and annotate all the available information about its DNA sequences and then to link it to the master genome sequence. At its inception in 1999, Ensembl’s main focus was on the human, mouse, and rat genomes, however Release 54 (May 2009) provides annotated gene predictions for over 70 different species.

All Ensembl gene transcripts are based on proteins and mRNA from the UniProt / Swiss-Prot, UniProt / TrEMBL and Reference Sequence (RefSeq; a collection of DNA, RNA and protein sequences curated by the NCBI [130]) databases. These
proteins and mRNAs are aligned against a genomic sequence assembly, which is imported from the relevant sequencing centre or consortium. Transcripts that have overlapping coding sequence are clustered into the same gene, and for each transcript a list of the mRNAs and proteins it is based upon is provided. Annotations include information such as regulatory regions, conserved base pairs across species, and mRNA protein mappings to the genome. [128] The Ensembl databases are publicly available online, and can be downloaded as DNA, cDNA or protein sequences.

**International Protein Index (IPI)**

The International Protein Index (IPI) provides complete non-redundant proteome sets for some of the more studied eukaryotic organisms. [131, 132] The data sets, which are freely available to download from the IPI website, are built from the UniProtKB / Swiss-Prot, UniProtKB / TrEMBL, Ensembl and RefSeq databases.

The purpose of IPI is to construct complete protein sets for higher eukaryotic species whose full genomic sequences have been completely determined. Although UniProtKB aims to hold all available protein data for a species, it only accepts protein sequences that have been determined by direct experiment or have been translated from the results of the sequencing of individual DNA clones or RNA molecules. For organisms where the complete genome sequence has been determined, there are often large numbers of predicted protein sequences from projects such as Ensembl that are not accepted into the UniProt databases. In addition, a preliminary assembly of an organism’s genome may become available before its sequence has been completed – these can be used to make provisional protein predictions which are often not submitted to the EMBL / Genbank / DDBJ nucleotide sequence databases and thus also do not appear in UniProtKB. [132]
IPI therefore aims to bring together the data of the UniProtKB database with protein prediction from other sources, particularly Ensembl, in order to provide maximum coverage for each proteome. For minimal redundancy of the datasets, a clustering process is carried out in which transcripts are grouped together according to their sequence similarity and any other information gained from their annotations and cross-references. Each IPI entry therefore comprises a cluster of source entries, from which a master cluster is chosen to supply the sequence for that entry. An IPI dataset for a particular species is much larger than its counterpart from Ensembl and UniProtKB. This not only reflects the extra sources of sequences that IPI accepts compared to these databases, but also the diversity in sequences and annotation reported for a particular protein between all the data sources used. Often these will not fall within the threshold used by IPI to define sequence similarity and therefore will be represented by multiple IPI entries. [132] Table 4-1 compares the number of entries in UniProtKB / SwissProt, Ensembl and IPI for four widely-studied eukaryotes; human, cow, mouse, rat.

<table>
<thead>
<tr>
<th></th>
<th>UniProtKB-Swiss-Prot Release 57.4</th>
<th>Ensembl 49 Gene transcripts (Known protein-coding genes)</th>
<th>International Protein Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (H. sapiens)</td>
<td>20,330</td>
<td>48,400 (21,541)</td>
<td>71,884</td>
</tr>
<tr>
<td>Cow (B. taurus)</td>
<td>5,672</td>
<td>28,958 (16,938)</td>
<td>32,934</td>
</tr>
<tr>
<td>Mouse (M. musculus)</td>
<td>16,140</td>
<td>40,466 (22,010)</td>
<td>55,078</td>
</tr>
<tr>
<td>Rat (R. norvegicus)</td>
<td>7,384</td>
<td>34,704 (17,738)</td>
<td>40,087</td>
</tr>
</tbody>
</table>

Table 4-1: A comparison between the numbers of entries in the UniProtKB / SwissProt [31], Ensembl [129] and International Protein Index (IPI) [131] databases.
4.2.2: Search and analysis tools for protein databases

The Expert Protein Analysis System (ExPASy)

The Expert Protein Analysis System servers [29, 30], maintained by the Swiss Institute of Bioinformatics (SIB), provide a central resource for tools dedicated to the analysis, characterization and identification of proteins. The analytical packages can be employed to predict certain properties about a protein, which can then be used in further investigations. Protein identification software is essential to proteomic techniques such as 2D gels and mass spectrometry; ExPASy offers a large variety of such tools that enable a protein to be identified by comparing certain information about it with those proteins of the UniProtKB database.

The extensive collection of sequence analysis tools available through the ExPASy server can be used to perform such tasks as similarity searches of a protein sequence against those of a database (BLAST and ScanProsite), calculation of various protein physicochemical parameters (Compute pI/Mw and ProtParam), theoretical protease cleavage of proteins (PeptideMass and PeptideCutter), and prediction of protein structures and post-translational modifications. Examples of the protein identification tools include those that are designed for data obtained using mass spectrometry techniques (Aldente and Mascot), as well as those that use protein amino acid composition data (AACompIdent and MultiIdent).

The AACompIdent protein identification tool

The AACompIdent resource provided by ExPASy, which is used to identify proteins by virtue of their amino acid (AA) compositions (see Section 1.2), is of particular relevance for the EVV 2DIR protein fingerprinting strategy. This tool is very similar
to that required for performing identifications using EVV 2DIR data, and was used as a model when developing the amino acid / CH₃ ratios database and search facility.

The AACompIdent program compares the experimentally-determined percentage amino acid compositions of an unknown protein with the theoretical percentage compositions of all the proteins contained in the UniProtKB database. A score is calculated for each database protein as the sum of the squared difference in the amino acid (AA) molar percentage between it and the unknown protein for all the amino acids. This represents the degree of difference between the AA composition of the unknown protein and the database entry. [29, 30] Equation 4-1 shows how a score is calculated, where for example, %Ala_{protein} is the percentage composition of alanine of the database entry being scored, and %Ala_{unknown} is the empirically measured percentage alanine composition of the unknown protein. Once every protein entry in UniProtKB has been given a score, they are then ranked from the lowest score (the closest match to unknown protein) down to the highest (the worst match).

\[
\text{Protein Score} = [(\%\text{Ala}_\text{protein} - \%\text{Ala}_\text{unknown})^2 + (\%\text{Cys}_\text{protein} - \%\text{Cys}_\text{unknown})^2 + (\%\text{Gly}_\text{protein} - \%\text{Gly}_\text{unknown})^2 + \ldots \ldots + (\%\text{AA}_\text{protein} - \%\text{AA}_\text{unknown})^2]
\]

Equation 4-1

To submit an AACompIdent query to ExPASy, a search constellation must first be chosen. The different constellations allow for searches to be made using different combinations of amino acid molar percentages. The most common constellation for use when standard compositional analyses have been employed (Section 1.2) uses sixteen amino acids - it does not include Cys or Trp, and calculates Asn & Asp, and Glu & Gln together. A ‘free constellation’ is available which allows searching with any amino acid combination of choice. When submitting an enquiry, in addition to
entering amino acid molar percentages, other parameters can also be used to further aid the search. These include species of interest, molecular weight (Mw) and isoelectric point (pI). It is also possible to submit a sequence tag for the protein, which can be obtained by Edman sequencing prior to the compositional analysis. The high specificity of such data means greater confidence can be placed in an identification result. If the data for a calibration or control protein is additionally submitted, its composition can be used to compensate for any errors in the AA analysis procedure of the unknown protein, and again the confidence of the identification can be significantly increased. [30]

The results from an AACompIdent query are emailed to the researcher and comprise the top-twenty ranked proteins from the searched database. For high confidence to be placed in the number one ranked protein representing the correct identification, a couple of criteria must be considered. A protein result should have a score lower than thirty for it to be deemed a good fit to the unknown protein, with a score of zero signifying a perfect match and thus an unambiguous identification. [30] There must also be a large difference between the scores of the proteins ranked first and second. A study using a set of *E. coli* proteins showed that a factor of two in the score difference of the first two results indicated that the top-ranked protein gave the correct identification. The same study demonstrated that proteins are far more likely to be correctly identified if their Mw and pI data are also used to perform the database search. [21, 101]
4.3: Databases for protein identification using EVV 2DIR

4.3.1: Database construction

Source databases

The initial databases to be constructed comprised only human protein sequences, obtained as FASTA files from Ensembl release 44 (this release is based on the NCBI 36 assembly of the human genome, November 2005). Many of these sequences contain stop codons, selenomethionines and unknown amino acids – for the initial test databases any sequences containing these were therefore removed. (For later versions of the databases, any ambiguous residues were replaced with an average of the amino acid composition of the sequence). This reduced the number of proteins from ~ 48,000 to ~ 33,000. Two databases were constructed using this dataset; one populated with $AA/CH_2$ ratios and the other with $AA/CH_3$ ratios. These two databases were used for the tests presented in Section 4.4.

Later, the source archive for the databases was changed from Ensembl to the International Protein Index (IPI). This decision was made as IPI is increasingly being used for protein identification in high-throughput proteomics, due to its wider coverage of certain species. The databases were also expanded from just the human proteome, to cover several other organisms, including cow, chicken, rat, mouse and zebrafish. The protein sequence datasets were downloaded as FASTA files from the publicly available IPI repositories. The ratio databases can be updated monthly, in accordance with IPI’s latest releases. Ratio databases were also generated for *A. thaliana* and *C. reinhardtii*, as rubisco interactome studies of these are currently undergoing within the Klug research group. The Arabidopsis dataset was obtained...
Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

from IPI, whilst for Chlamydomonas the draft release of the genome from the Joint Genome Institute (JGI) [133] was used.

**Populating the databases**

Following from the peptide fingerprinting results (Section 3.3), the first database to be constructed was composed of \( AA / CH_2 \) ratios. However, once the \( CH_3 \) group had been identified as a suitable internal reference for proteins (Section 3.4), the second version of the database comprised \( AA / CH_3 \) ratios. Since glutamic acid (E) and aspartic acid (D) might be expected to have some EVV 2DIR vibrational features in common due to their carboxylate side chains, an extra ratio, \( D+E / CH_x \) was calculated. This is similarly the case for the amide side chains of asparagine (N) and glutamine (Q), so an \( N+Q / CH_x \) query option was also included.

Once a protein dataset had been downloaded, the total \( CH_x \) (\( CH_2 \) or \( CH_3 \)) groups and amino acid content of each protein were computed. These were then converted into \( AA / CH_x \) ratios, and the theoretical molecular weights and isoelectric points of each protein were also calculated. This was performed using the *pepstats* (Emboss 5.0) package [134], and the output files were parsed using custom Python scripts. The ratios, Mw and pI of each protein were used to populate a MySQL database on a Linux workstation.

**4.3.2: Database searching**

**Query submission**

A web-based query form was developed for searching the database. The search parameters are the twenty-two \( AA / CH_x \) compositions (including \( D+E / CH_x \) and
Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

$N +Q / CH_3$, molecular weight and isoelectric point. (It is envisaged that capillary electrophoresis based separation of proteins prior to EVV 2DIR analysis will provide protein mass and charge information – see Section 6.3.) Any number of experimentally determined ratios can be input and, to allow for any experimental error, a value for the level of precision on each measurement must also be included. This also applies if a Mw. or pI value is used in the query. When the database is searched, the output comprises proteins whose ratios (and Mw or pI) fall within the stipulated precision levels of the input values. The database hits from a protein query are calculated using custom SQL and Python scripts.

**Search output**

Currently the output from a database search is simply a list of hits, with no indication as to which of these proteins has the closest amino acid composition to the protein being identified. The present method can also exclude the unknown protein from the list of results, if only one of its measured ratios falls outside the precision level specified in the query. A method of ranking the protein hits, based on ExPASy’s *AACompIdent* tool, is therefore in the process of being implemented. Rather than only returning proteins whose ratios fall within a certain range of values, all database entries will be given a score based on the degree of difference between their ratios and the input ratios of the unknown protein. Every protein in the database will then be given a ranking, with the top-ranked entry representing the best match to the target protein. The progress of the implementation of the scoring and ranking system is covered in more detail in Section 6.3.
4.4: How many amino acid ratios to identify a protein?

To date, vibrational signatures have been identified and validated for three different amino acid side chains. Clearly, however, a protein cannot be identified using so few relative amino acid levels, and work is ongoing to establish EVV 2DIR cross peaks for more residues. Methods for measuring protein mass and charge using capillary electrophoresis are also under development [135] – this additional information about a protein will help to narrow down or confirm an identification.

Analyses of the ratios database were therefore devised to determine how many amino acid / CH\textsubscript{x} ratios need to be measured to achieve a unique protein identification, and also to discover which ratios are the most effective for narrowing down a search. The results presented in this section were obtained using the human protein database constructed using the Ensembl dataset (~ 33,000 proteins). Initially tests were performed using the AA / CH\textsubscript{2} ratios database. Following the designation of the CH\textsubscript{3} group as the most suitable internal reference for proteins, analysis continued with the AA / CH\textsubscript{3} ratios database.

4.4.1: Narrowing down a search

The average number of hits output from the human protein database when it is searched using just one amino acid ratio was calculated for each different residue. To achieve this, the actual value for a particular AA / CH\textsubscript{2} ratio for the first protein in the database was submitted as a query with an error range of, for example, ± 10%. All proteins whose ratio of the amino acid fell within ± 10% of that of the test protein were returned as hits. This procedure was then repeated for each of the ~33,000 protein entries in the database. The mean value of the number of proteins
returned is considered to be the average number of hits for that particular amino acid. This operation was carried out for each of the amino acid ratios in turn at four different accuracy levels (± 5%, ± 10%, ± 15% & ± 30%), with and without the molecular weight (± 10%) as a search parameter. Table 4-2 shows the average number of hits returned from the database for each of the twenty amino acid / CH$_2$ ratios when searching with the four different error ranges. Also shown are the numbers of hits returned when the molecular weight is additionally used as a search parameter (at ± 10%).

The fewest number of hits were returned from the database when the Cys / CH$_2$ ratio was used to perform a search; the most hits were returned when the Leu / CH$_2$ ratio was used. This indicates that proteins are most distinguishable from each other by their cysteine levels. To identify an unknown protein, it would therefore be more advantageous to be able to measure its Cys / CH$_2$ ratio than its Leu / CH$_2$ ratio. The ‘top six’ ratios for narrowing down a protein identification search are those of cysteine, tryptophan, tyrosine, proline, isoleucine and phenylalanine. The three aromatic amino acids of this list are those that can currently be measured using EVV 2DIR. The relative spectral decongestion around their side chain cross peaks have made them the easiest to find and validate spectral signatures for. As the hunt for more side chain vibrational signatures continues, the results of the database tests suggest that it would be wise to concentrate on identifying cross peaks of those amino acids in the top half of Table 4-2.

Although it was originally intended that the EVV 2DIR fingerprinting strategy will only require relative amino acid compositions to carry out protein identifications, it is likely that the mass of the proteins to be identified will be established during the CE-based separation step. [85] This additional information about the proteins will therefore be used as an extra aid in the identification process. The results in Table
4-2 show that having mass data for a protein will help significantly in its identification. When the protein molecular weight is also used as a search parameter, the number of average hits reduces by an order of magnitude; \( C / CH_2 (\pm 10\%) \) gives 2515 results, whereas a search with the same ratio and the molecular weight (\( \pm 10\% \)) returns only 201 hits.

Interestingly the most important amino acids for distinguishing proteins are amongst those that are least abundant. Figure 4-1 shows the amino acid composition across the entire UniProtKB / SwissProt protein database. The least abundant residues are tryptophan and cysteine at only 1.23% and 2.24% respectively, whilst the most abundant is leucine at 9.87%. Along with tryptophan, the other aromatic residues – tyrosine and phenylalanine – are also relatively scarce. The most abundant amino acids have aliphatic side chains, which contribute significantly to the \( CH_3 \) internal reference peak.

**Figure 4-1:** The average percentage amino acid compositions across all sequences of the UniProtKB / SwissProt protein database, Release 57.4. The colour coding indicates the nature of the side chains: grey = aliphatic, red = acidic, green = hydroxyl group, blue = basic, black = aromatic, white = amide, and yellow = sulphur. Reproduced from the UniProtKB website [31].
### Table 4-2: Average number of protein hits from the human protein database when it is queried using only a single amino acid / CH\(_2\) ratio with accuracy levels of ± 5%, ± 10%, ± 15% and ± 30%. Also shown is the number of hits returned when the molecular weight is included as a search parameter (also at ± 10%).

<table>
<thead>
<tr>
<th>AA/CH(_2) Ratio</th>
<th>AA/CH(_2) ± 5%</th>
<th>AA/CH(_2) ± 10%</th>
<th>AA/CH(_2) ± 15%</th>
<th>AA/CH(_2) ± 30%</th>
<th>AA/CH(_2) &amp; Mw. ± 10%</th>
<th>AA/CH(_2) &amp; Mw. ± 15%</th>
<th>AA/CH(_2) &amp; Mw. ± 30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys (C)</td>
<td>1339</td>
<td>104</td>
<td>2515</td>
<td>201</td>
<td>3910</td>
<td>296</td>
<td>7813</td>
</tr>
<tr>
<td>Trp (W)</td>
<td>1389</td>
<td>116</td>
<td>2624</td>
<td>205</td>
<td>3646</td>
<td>291</td>
<td>7088</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>1590</td>
<td>128</td>
<td>3053</td>
<td>241</td>
<td>4514</td>
<td>352</td>
<td>8861</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>1601</td>
<td>119</td>
<td>3202</td>
<td>236</td>
<td>4802</td>
<td>353</td>
<td>9594</td>
</tr>
<tr>
<td>Ile (I)</td>
<td>1684</td>
<td>123</td>
<td>3341</td>
<td>242</td>
<td>4993</td>
<td>361</td>
<td>9875</td>
</tr>
<tr>
<td>Phe (F)</td>
<td>1750</td>
<td>131</td>
<td>3475</td>
<td>259</td>
<td>5196</td>
<td>385</td>
<td>10284</td>
</tr>
<tr>
<td>Met (M)</td>
<td>1814</td>
<td>133</td>
<td>3627</td>
<td>265</td>
<td>5437</td>
<td>395</td>
<td>10812</td>
</tr>
<tr>
<td>Asn (N)</td>
<td>1834</td>
<td>139</td>
<td>3634</td>
<td>272</td>
<td>5423</td>
<td>403</td>
<td>10649</td>
</tr>
<tr>
<td>His (H)</td>
<td>1845</td>
<td>138</td>
<td>3659</td>
<td>272</td>
<td>5465</td>
<td>403</td>
<td>10781</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>1969</td>
<td>145</td>
<td>3933</td>
<td>289</td>
<td>5891</td>
<td>431</td>
<td>11668</td>
</tr>
<tr>
<td>Ala (A)</td>
<td>1997</td>
<td>146</td>
<td>3990</td>
<td>290</td>
<td>5975</td>
<td>434</td>
<td>11828</td>
</tr>
<tr>
<td>Lys (K)</td>
<td>2050</td>
<td>152</td>
<td>4083</td>
<td>301</td>
<td>6097</td>
<td>449</td>
<td>11897</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>2199</td>
<td>164</td>
<td>4386</td>
<td>325</td>
<td>6553</td>
<td>484</td>
<td>12802</td>
</tr>
<tr>
<td>Val (V)</td>
<td>2262</td>
<td>164</td>
<td>4514</td>
<td>326</td>
<td>6745</td>
<td>486</td>
<td>13194</td>
</tr>
<tr>
<td>Asp (D)</td>
<td>233</td>
<td>179</td>
<td>4735</td>
<td>354</td>
<td>7039</td>
<td>525</td>
<td>13460</td>
</tr>
<tr>
<td>Gln (Q)</td>
<td>2419</td>
<td>176</td>
<td>4821</td>
<td>350</td>
<td>7190</td>
<td>521</td>
<td>13924</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>2552</td>
<td>187</td>
<td>5085</td>
<td>372</td>
<td>7578</td>
<td>552</td>
<td>14606</td>
</tr>
<tr>
<td>Glu (E)</td>
<td>2569</td>
<td>191</td>
<td>5110</td>
<td>377</td>
<td>7597</td>
<td>559</td>
<td>14481</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>2572</td>
<td>190</td>
<td>5125</td>
<td>378</td>
<td>7644</td>
<td>562</td>
<td>14792</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>2689</td>
<td>194</td>
<td>5352</td>
<td>385</td>
<td>7966</td>
<td>571</td>
<td>15269</td>
</tr>
</tbody>
</table>
4.4.2: Combinations of amino acid ratios

Clearly the more information that can be acquired about a protein, the easier it will be to achieve an unambiguous identification. The question that needs to be addressed is how many amino acid ratios it is necessary to measure of an unknown protein to confidently identify it. The average numbers of hits returned from the database when different combinations of two amino acid ratios are used in a query are shown in Table 4-3. (These tests were performed in the same manner as for the single ratios shown in Table 4-2.) Even when using only two ratios in combination, the number of protein hits from a search is greatly reduced. When the $\text{Cys} / CH_2$ and $\text{Trp} / CH_2$ ratios are used independently with an accuracy interval of $\pm 10\%$ they each return ~ 2500 results, whereas in combination the number of hits is reduced ten-fold to 235.

<table>
<thead>
<tr>
<th>AA/CH$_2$ Ratio Pairs</th>
<th>Ratios $\pm 10%$</th>
<th>Ratios $\pm 10%$ &amp; Mw $\pm 10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W + C</td>
<td>235</td>
<td>24</td>
</tr>
<tr>
<td>L + S</td>
<td>876</td>
<td>69</td>
</tr>
<tr>
<td>W + F</td>
<td>291</td>
<td>27</td>
</tr>
<tr>
<td>W + Y</td>
<td>261</td>
<td>26</td>
</tr>
<tr>
<td>F + Y</td>
<td>416</td>
<td>37</td>
</tr>
</tbody>
</table>

Table 4-3: The average number of protein hits output when various pairs of amino acid ratios (with a $\pm 10\%$ error range) are used to search the human protein database. Shown are the results for the combinations of the two ‘best’ (W + C) and two ‘worst’ (L + S) ratios, taken from Table 4-2. The three possible pairs of the three amino acid ratios that can currently be measured with EVV 2DIR (F, Y and W) are also included.
Table 4-4 demonstrates how the average number of hits output from the protein database decreases as more amino acid ratios are used in combination to perform a search. Just by using four ratios at ± 10% (the three aromatic amino acids and histidine), the ~ 33,000 proteins are reduced on average to merely 7 possible candidates. If the mass is also input with a 10% accuracy range, then the average number of hits is reduced further to just 3. When the three aromatic residues currently measureable with EVV 2DIR are used, the > 30,000 proteins of the database are on average narrowed down to 40 candidates.

<table>
<thead>
<tr>
<th>AA/CH$_2$ Ratio Combinations</th>
<th>Ratios ± 10%</th>
<th>Ratios ± 10% &amp; Mw ± 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>2624</td>
<td>205</td>
</tr>
<tr>
<td>W + Y</td>
<td>261</td>
<td>26</td>
</tr>
<tr>
<td>W + Y + F</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>W + Y + F + H</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4-4: The average number of protein hits output when combinations of two, three and four amino acid ratios (with ± 10% error ranges) are used to search the human protein database. When the three aromatic residues currently measureable with EVV 2DIR are used, the average number of hits is reduced to 40.

The database analyses presented so far were performed using the AA / CH$_2$ database, constructed in response to the peptide fingerprinting results. Once the CH$_3$ peak had been identified in proteins as a suitable internal reference, work continued with a AA / CH$_3$ database instead. Nevertheless, both types of ratio are equally suitable and shown in Table 4-4 are the average numbers of hits output from the protein database when three and four ratios are used in combination for either the CH$_2$ or CH$_3$ ratios.
## Table 4-5: Average numbers of hits output from the CH\textsubscript{2} and CH\textsubscript{3} ratio databases when searching with three and four amino acid ratios in combination. The numbers of results are very similar for both the CH\textsubscript{2} and CH\textsubscript{3} databases, indicating that both types of ratios are equally suitable and either can be used depending on EVV 2DIR experimental requirements.

<table>
<thead>
<tr>
<th>AA/CH\textsubscript{2} Ratio Combinations</th>
<th>CH\textsubscript{2} Ratios ± 10%</th>
<th>CH\textsubscript{3} Ratios ± 10%</th>
<th>CH\textsubscript{2} Ratios ± 10%</th>
<th>CH\textsubscript{3} Ratios ± 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>W + Y + F</td>
<td>40</td>
<td>6</td>
<td>44</td>
<td>6</td>
</tr>
<tr>
<td>W + Y + F + H</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

### 4.4.3: Distribution of database hits

The database search results presented above discuss the average numbers of hits returned when performing searches with different combinations of ratios at varying accuracy levels. These results do not however give an indication of the distributions of the numbers of hits, nor how many times just one protein result is returned (i.e. only the protein itself is output when performing a search using its ratio values).

Figure 4-2 (a) shows the distribution of hits returned when the ratios of the three amino acids presently measureable with EVV 2DIR (F / CH\textsubscript{3}, Y / CH\textsubscript{3} and W / CH\textsubscript{3}) are input with a ± 10% accuracy interval for each protein in the human database. The first bar of (a) indicates that ~ 850 proteins give only one result – themselves – when these three ratios are used. If, however the molecular weight is also entered as a search parameter, Figure 4-2 (b), ~ 5,000 proteins are uniquely identified and ~ 10,500 (32% of the entries in the database) give only one or two hits. In fact, 99% of the proteins give fewer than 25 results when a search is performed using their three ratios and molecular weight with accuracy levels of ± 10%. (The bar in Figure 4-2 (a) showing that ~ 200 proteins each give ~ 200 hits is due to a group of proteins that have no F, Y or W residues and thus have zero values for their ratios; when a search is performed using any one of these proteins, 200 results are output.)
Figure 4-3 (a) shows the distribution of hits returned when a total of five amino acid ratios (the three currently measureable with EVV 2DIR and two others; H / CH$_3$ and C / CH$_3$) are input with ± 10% accuracy levels for each protein in the human database. The histidine and cysteine ratios are chosen because, with regards to the bioinformatics, these residues are good candidates for the fingerprinting strategy; Table 4-2 shows these both to be useful for limiting a protein search. Preliminary EVV 2DIR measurements on peptides, and calculations of the spectra of these residues, show that, with resolvable features at 1475 / 2650 cm$^{-1}$ and 1485 / 2560 cm$^{-1}$ respectively, these are both also experimentally favourable. The numbers of hits returned when the molecular weight at ± 10% is additionally used in a query with the five ratios is shown in Figure 4-3 (b).

The results of Figure 4-3 (a) demonstrate that if the relative levels of only these five amino acids are input with an accuracy range of ± 10%, then 44% of the proteins in the database (~ 15,000 of the ~ 33,000 database entries) give only one result, i.e. they are uniquely identifiable. 72% of the proteins (~ 25,000) give only one or two hits. If the molecular weight is also used for a search (with a ± 10% accuracy range), for 60% of the proteins the only result output is themselves, and more than 85% return only one or two hits. In fact, 99% of the proteins give fewer than 10 results when a search is performed using their five ratios and molecular weight with accuracy levels of ± 10%.

Although the database can only currently be searched in the manner used for these tests, with accuracy levels being used to query within certain ratio intervals, this will not be the case in the future. As previously discussed, this method is problematic – by only searching within a certain range of a ratio, the protein of interest can actually be excluded from the results. Additionally, the list of returned hits gives no indication as which of the proteins is the most closely matched to the unknown
sample. With the system currently being implemented, a list of hits will not be returned from a query, rather every protein in the database will be given a score and ranking indicating its level of similarity to the unknown protein’s measured data. Nonetheless, the database tests presented in this section do give a good indication of how many amino acid ratios will be required for an unknown protein to be distinguished from the other proteins in the database, and thus be ranked highly in the list of possible candidates. When the score of the protein ranked first is significantly lower than that of the second placed protein, there can be more confidence that this is the identity of the unknown protein. The number of amino acid ratios measured and then input for a search needs to be sufficient for this to be the case.

The analyses of the human protein database presented in this section are extremely encouraging. They indicate that for the EVV 2DIR fingerprinting strategy to be successful, a protein’s mass and the relative levels of upwards of five of its amino acids need be measured. Therefore the vibrational signatures of possibly only two or three more residues are required before protein identifications can be achieved. It is however recognised that the database tests were performed using the real values for the proteins’ ratios, and so are likely to produce more favourable results than when searching with measured ratios of unknown proteins. Of course, when more information is available about an unknown protein, the easier it will be to identify. The aim is therefore of course to be able to measure as many amino acid ratios, as well as other protein parameters such as the mass and charge, as possible (see Section 6.3).
Figure 4-2: (a) Distribution of hits that are returned when the ratios of three amino acids ($F/CH_3$, $Y/CH_3$ and $W/CH_3$) are input with a 10% accuracy interval for each entry in the human protein database of ~33,000 sequences. (b) Distribution of hits returned when, in addition to the three ratios, the molecular weight of each protein is also input at ±10%.
Figure 4-3: (a) Distribution of hits that are returned when the ratios of five amino acids (F / CH₃, Y / CH₃, W / CH₃, H / CH₃, and C / CH₃) are input with a 10% accuracy interval for each entry in the human protein database of ~ 33,000 sequences. (b) Distribution of hits returned when, in addition to the five ratios, the molecular weight of each protein is also input at ± 10%.
4.5: Database searches using EVV 2DIR data

To assess the EVV 2DIR data collected from the proteins used in Section 3.4 for the demonstration of the relative quantification of three amino acid levels, a couple of test database searches were performed. The first of these involved trialling a system for ranking protein results, and, for the second, ExPASy’s AACompIdent tool was used to test the EVV 2DIR data.

4.5.1: Ranking of protein results

The sequences of the ten proteins (details in Section 3.2.1) were used to construct a test database comprising their calculated AA / CH$_3$ number ratios. Each protein was then in turn treated as an unknown; using the calibration graphs produced from their EVV 2DIR data (Figure 3-12), their four measured peak intensity ratios ($\sqrt{I_{AA}} / \sqrt{I_{CH3}}$) were converted into the number ratios ($N_{AA} / N_{CH3}$) required for performing a search of the database. The data of the first ‘unknown’ protein was then used to carry out a search to discover its identity. Each of the ten proteins in the database was given a score based on the similarity of its AA / CH$_3$ ratios with the measured, input ratios of the unknown protein. The score was calculated as follows, where, for example, $Y_{Protein}$ is the $Y / CH3$ ratio of the protein in the database being scored, and $Y_{Unknown}$ is the experimental $Y / CH3$ ratio of the unknown protein. (F(1) and F(2) are the two phenylalanine ratios that can be measured using the different beam polarisations.) The individual scores were also scaled so that each ratio contributed equally to the overall score:
Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

Protein score \( = \sqrt{[(Y_{\text{Protein}} - Y_{\text{Unknown}})^2 + (W_{\text{Protein}} - W_{\text{Unknown}})^2 + (F(1)_{\text{Protein}} - F(1)_{\text{Unknown}})^2 + (F(2)_{\text{Protein}} - F(2)_{\text{Unknown}})^2]} \)

Equation 4-2

This procedure was repeated so that each of the ten proteins was treated as an unknown, with its measured ratios being used to perform an identification search. Five of the proteins achieved the first place in the rankings when their searches were performed, one protein was ranked second, three proteins came third in the rankings and chymotrypsin only achieved a fifth place. Shown in Table 4-6 are the results obtained when the measured ratios of lysozyme (a) and BSA (b) were used to perform searches. The scores for all of the ten proteins in the database are shown in each case. Lysozyme and BSA are clearly easily differentiated by their aromatic amino acid compositions, as they are each ranked last in the other’s search.

The measured ratios of each of the proteins were also used to perform searches of the database of ten proteins by submitting queries within a ± 20% error range (chosen because the average precision (and dispersion) of the four calibration graphs in Figure 3-12 is 20%). When the searches were performed using their ratio data, only two of the proteins were returned as hits. For the other eight proteins, one or more of their measured ratios fell outside of the ± 20% range of their actual values, and they were thus excluded from the list of possible results. Table 4-7 compares the results from the two different methods of database searching.
### Table 4-6: The similarity scores and rankings of the dataset of ten test proteins when searches were performed using EVV 2DIR measured ratios of (a) lysozyme and (b) BSA.

<table>
<thead>
<tr>
<th>Database protein</th>
<th>Score</th>
<th>Database protein</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(a) Lysozyme as ‘unknown’ protein</strong></td>
<td></td>
<td><strong>(b) BSA as ‘unknown’ protein</strong></td>
<td></td>
</tr>
<tr>
<td>Lysozyme</td>
<td>0.216</td>
<td>Pepsin</td>
<td>0.254</td>
</tr>
<tr>
<td>α-Lactalbumin</td>
<td>0.313</td>
<td>Concanavalin A</td>
<td>0.299</td>
</tr>
<tr>
<td>Alkaline Phosphatase</td>
<td>0.438</td>
<td>BSA</td>
<td>0.315</td>
</tr>
<tr>
<td>Aldolase</td>
<td>0.536</td>
<td>Albumin</td>
<td>0.441</td>
</tr>
<tr>
<td>β-Lactoglobulin B</td>
<td>0.617</td>
<td>Alkaline Phosphatase</td>
<td>0.444</td>
</tr>
<tr>
<td>Chymotrypsin</td>
<td>0.676</td>
<td>α-Lactalbumin</td>
<td>0.491</td>
</tr>
<tr>
<td>Concanavalin A</td>
<td>0.689</td>
<td>Aldolase</td>
<td>0.792</td>
</tr>
<tr>
<td>Pepsin</td>
<td>0.755</td>
<td>β-Lactoglobulin B</td>
<td>0.822</td>
</tr>
<tr>
<td>Albumin</td>
<td>0.836</td>
<td>Chymotrypsin</td>
<td>0.938</td>
</tr>
<tr>
<td>BSA</td>
<td>0.891</td>
<td>Lysozyme</td>
<td>0.964</td>
</tr>
</tbody>
</table>

### Table 4-7: A comparison of the two different methods of database searching. The first column shows that eight of the proteins did not appear in their lists of possible hits when queries were submitted using their measured ratios at ± 20%. Five of the ten proteins, however, achieved the top-ranking in their lists of results when a scoring system was used to rank the database proteins in order of ‘nearest fit’ to the input values.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Was the protein returned as a possible hit when a query was submitted of its measured ratios at ± 20%?</th>
<th>Ranking of the protein when the ten database entries were scored on the degree of difference between their ratio values and the input values.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>NO</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Aldolase</td>
<td>NO</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>BSA</td>
<td>NO</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Chymotrypsin</td>
<td>NO</td>
<td>5&lt;sup&gt;th&lt;/sup&gt;</td>
</tr>
<tr>
<td>Concanavalin A</td>
<td>NO</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>YES</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>Pepsin</td>
<td>NO</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>Alkaline Phosphatase</td>
<td>NO</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>α-Lactalbumin</td>
<td>NO</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>β-Lactoglobulin</td>
<td>YES</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
4.5.2: Using the \textit{AACompIdent} tool with EVV 2DIR data

ExPASy's \textit{AACompIdent} database analysis tool [29] is very similar to the one being developed for use as part of the EVV 2DIR protein identification strategy. As explained in Section 4.3, \textit{AACompIdent} is used to achieve protein identification by performing searches of the UniProtKB / SwissProt database with amino acid compositional data. The CH$_3$ ratios data measured using EVV 2DIR is not suitable for this program, since molar percentages of amino acid composition are required to perform a search. However, in order to test some of the EVV 2DIR data over a full database using this search facility, the measured amino acid ratios were converted into molar percentages using the known values for the total number of CH$_3$ groups in each protein.

Each protein's measured peak intensity ratios were first converted into number ratios using the calibration graphs of Figure 3-12. These amino acids ratios were then changed into molar percentages using the known total amino acid and total CH$_3$ contents of the protein. Table 4-8 shows the real values of the AA / CH$_3$ ratios and AA molar percentages for lysozyme, along with the measured ratios and molar percentages simulated through the combination of experimental (the measured ratios) and known data (the total AA and CH$_3$ content) about the protein.

<table>
<thead>
<tr>
<th></th>
<th>Real AA/CH$_3$ ratio</th>
<th>Real AA molar %</th>
<th>Measured AA/CH$_3$ ratio</th>
<th>Simulated AA molar %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.04918</td>
<td>2.33</td>
<td>0.05809</td>
<td>2.75</td>
</tr>
<tr>
<td>W</td>
<td>0.09836</td>
<td>4.65</td>
<td>0.09179</td>
<td>4.34</td>
</tr>
<tr>
<td>F1</td>
<td>0.04918</td>
<td>2.33</td>
<td>0.05439</td>
<td>2.57</td>
</tr>
<tr>
<td>F2</td>
<td>0.04918</td>
<td>2.33</td>
<td>0.04123</td>
<td>1.95</td>
</tr>
</tbody>
</table>

\textit{Table 4-8:} The known amino acid ratios and molar percentages of lysozyme compared with the ratios measured with EVV 2DIR and the simulated molar percentages.
Chapter 4: Bioinformatic resources for EVV 2DIR protein fingerprinting

The simulated amino acid molar percentages for each of the proteins were used to search the UniProtKB / SwissProt database via the AACompIdent program. For each query, only the data of the compositions of the three amino acids (Phe, Tyr & Trp) and the species were submitted – no mass or pI data were used. The results were emailed back and comprised a list of the top twenty-ranked protein matches from the database.

Nine of the ten proteins did not feature in the top-twenty list when a search was performed using their amino acid compositions. However when a search was submitted to the *G. Gallus* database using lysozyme’s data, the protein ranked as the second closest match was indeed lysozyme. Shown in Table 4-9 is the list of twenty protein matches returned for the lysozyme data. It can be seen that lysozyme has the same score as the first-placed protein, which only has this spot by virtue of its mass being closest to the average value of the database; had mass data also been used in the query, lysozyme would have been ranked first. However, in this case lysozyme would not be considered as a confident match, since its score is neither below thirty, nor significantly different from that of the second-ranked protein. It is of interest to note that, although the twenty returned proteins have similar aromatic amino acid compositions, they have a wide spread of isoelectric points, molecular weights and functions.
The closest Swiss-Prot entries (in terms of AA composition) for the species GALLUS:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Protein</th>
<th>pI</th>
<th>Mw</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>EDF1_CHICK</td>
<td>9.87</td>
<td>16364</td>
<td>Endothelial differentiation-related factor</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>LYSC_CHICK</td>
<td>9.32</td>
<td>14313</td>
<td>Lysozyme C. /FTId=PRO_000018495.</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>CCHL_CHICK</td>
<td>6.08</td>
<td>31438</td>
<td>Cytochrome c-type heme lyase. /FTId=PRO_0</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>GBLP_CHICK</td>
<td>7.57</td>
<td>34946</td>
<td>Guanine nucleotide-binding protein subunit</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>SSPO_CHICK</td>
<td>5.56</td>
<td>558688</td>
<td>SCO-spondin. /FTId=PRO_0000245045.</td>
</tr>
<tr>
<td>6</td>
<td>88</td>
<td>SPON1_CHICK</td>
<td>5.90</td>
<td>87965</td>
<td>Spondin-1. /FTId=PRO_0000035868.</td>
</tr>
<tr>
<td>7</td>
<td>101</td>
<td>IRF1_CHICK</td>
<td>5.44</td>
<td>36010</td>
<td>Interferon regulatory factor 1. /FTId=PRO</td>
</tr>
<tr>
<td>8</td>
<td>125</td>
<td>RBBP7_CHICK</td>
<td>4.81</td>
<td>47703</td>
<td>Histone-binding protein RBBP7. /FTId=PRO</td>
</tr>
<tr>
<td>9</td>
<td>132</td>
<td>SUOX_CHICK</td>
<td>5.80</td>
<td>50205</td>
<td>Sulfite oxidase. /FTId=PRO_0000166076.</td>
</tr>
<tr>
<td>10</td>
<td>139</td>
<td>RM50_CHICK</td>
<td>9.12</td>
<td>17975</td>
<td>39S ribosomal protein L50, mitochondrial</td>
</tr>
<tr>
<td>11</td>
<td>141</td>
<td>VSX1_CHICK</td>
<td>9.22</td>
<td>37351</td>
<td>Visual system homeobox 1. /FTId=PRO_00000</td>
</tr>
<tr>
<td>12</td>
<td>168</td>
<td>DC12_CHICK</td>
<td>9.10</td>
<td>38221</td>
<td>UPF0361 protein DC12 homolog. /FTId=PRO</td>
</tr>
<tr>
<td>13</td>
<td>178</td>
<td>RSSA_CHICK</td>
<td>4.80</td>
<td>33021</td>
<td>40S ribosomal protein SA. /FTId=PRO_00001</td>
</tr>
<tr>
<td>14</td>
<td>186</td>
<td>CED043_CHICK</td>
<td>10.12</td>
<td>8709</td>
<td>UPF0542 protein C5orf43 homolog. /FTId=PR</td>
</tr>
<tr>
<td>15</td>
<td>201</td>
<td>P133_CHICK</td>
<td>10.02</td>
<td>28355</td>
<td>Protein FAM133. /FTId=PRO_0000287620.</td>
</tr>
<tr>
<td>16</td>
<td>201</td>
<td>GLL7_CHICK</td>
<td>9.49</td>
<td>4794</td>
<td>Gallinacin-7. /FTId=PRO_0000288567.</td>
</tr>
<tr>
<td>17</td>
<td>201</td>
<td>MK_CHICK</td>
<td>9.54</td>
<td>13436</td>
<td>Midkine. /FTId=PRO_000024665.</td>
</tr>
<tr>
<td>18</td>
<td>203</td>
<td>ASHWN_CHICK</td>
<td>9.83</td>
<td>26536</td>
<td>Ashwin. /FTId=PRO_0000268862.</td>
</tr>
<tr>
<td>20</td>
<td>203</td>
<td>CCKN_CHICK</td>
<td>3.80</td>
<td>949</td>
<td>Cholecystokinin-7 (By similarity). /FTId=PRO_0000035868.</td>
</tr>
</tbody>
</table>

Table 4-9: The list of twenty proteins from the G. Gallus UniProtKB / SwissProt database that were returned from an AACompIdent search [29] using simulated EVV 2DIR data of the protein lysozyme. All proteins of the database are given a score and ranking based on the degree of similarity of their amino acid compositions to those submitted of the protein of interest. The lysozyme entry in the database was ranked as being the second most likely match for the submitted data.
Chapter 5: Electronically resonant EVV 2DIR spectroscopy studies of bacteriorhodopsin

5.1: Electronically enhanced EVV 2DIR

Resonance electronic enhancement of EVV 2DIR signal can be achieved when a sample has a suitable electronic transition coupled to the vibrational states being probed by the infrared beams. By tuning the visible beam close to resonance with such an transition, one should get considerable enhancement of the signal size in the same way as occurs with resonance Raman scattering. Triply resonant four wave mixing signal, representing electron-vibration-vibration couplings, can thus be observed when the third input beam – the visible beam – probes such an electronic resonance and enhances the FWM signal (see Section 2.3.4).

5.1.1: Initial studies of dye molecules

Due to a number of instrumental constraints, the EVV 2DIR set-up used for the results presented here has the visible beam fixed at 790 nm. This means that the protein spectra shown in Chapter 3 were collected with a visible beam frequency far
from any electronic resonances of these samples (for example, ~ 280 nm for the aromatic side chains), and so the measured FWM signal was simply doubly vibrationally enhanced. P. M. Donaldson therefore made initial attempts to observe the triply resonant effect by measuring the signals emitted from red-absorbing dyes.

Various simple dye molecules – in the form of solutions, as well as thin films dried onto glass slides – with absorptions in the 600 to 780 nm region were tested for signs of electronic enhancement. The signal that was observed from some of these tended to decay with time, indicating bleaching of the dye molecules with the laser beams. With the visible beam energy reduced, however, electronically enhanced four wave mixing signals with single vibrational resonance enhancements were observed, an example of which is shown in Figure 5-1. Attempts to measure triply resonant signal (doubly vibrationally enhanced and singly electronically enhanced) from the dyes were however unsuccessful.

Figure 5-1: An example ER EVV 2DIR spectrum collected of a thin film of the red-absorbing dye, IR746. Structure is seen only along $\omega_\alpha$, indicating electronically enhanced four wave mixing signal with single vibrational resonance enhancements. Reproduced with kind permission of P. M. Donaldson. [98]
5.1.2: Preliminary bacteriorhodopsin studies

Triply resonant EVV 2DIR signal was first observed from the retinal chromophore that lies at the centre of the photosynthetic protein bacteriorhodopsin (bR). During the initial assessment of the aromatic spectral region (1350 – 1600 cm\(^{-1}\) / 2700 – 3300 cm\(^{-1}\)) for signatures of amino acid side chains, bacteriorhodopsin was amongst the set of test proteins used. A unique group of cross peaks was observed for bR in this region. These were not only shifted in frequency compared to the aromatic cross peaks of the other proteins (of the tyrosine, phenylalanine and tryptophan side chains), but their signal sizes were also significantly larger than those measured from the usually dominant CH\(_2\) and CH\(_3\) features. Additionally, the signal was seen to decay rapidly over time in the same manner as for the dye molecules. Together this indicated that the visible beam was possibly exciting an electronic transition of bacteriorhodopsin’s chromophore and enhancing the FWM signal of vibrational couplings.

Through consultation of various IR and Raman studies [136-139], as well as of DFT calculations performed by R. Guo, it was established that two vibrational modes of the retinal’s polyene backbone (Figure 5-2) were indeed being excited by the infrared beams, and that the doubly vibrationally enhanced signal was being further, electronically enhanced by the visible probe beam. While the visible absorption spectrum of bR shows it to absorb most strongly at 570 nm, it appears that it still absorbs sufficiently at 790 nm for significant enhancement to occur.

![Figure 5-2: The retinal chromophore located at the centre of bacteriorhodopsin.](image-url)
This result demonstrated for the first time an electron-vibration-vibration triple resonance experiment. It was also the first EVV 2DIR spectrum to be recorded of a protein active site; the resonance electronic enhancement provided by the visible beam allowed the small retinal molecule to be selectively probed whilst avoiding the signal from the large surrounding protein. The enhanced signal from only the single retinal molecule was considerably higher than that from the CH₄ groups of the 248-residue protein. This illustrated perfectly the ability of this triple resonance technique to selectively home-in on the structural changes occurring at the active sites of proteins that contain groups with appropriate electronic transitions, for example metal ions or chromophores.

This line of research emerged during the protein fingerprinting work and, following from the initial discovery of the enhanced retinal signal, more in-depth studies of bacteriorhodopsin using the electronically resonant EVV 2DIR technique commenced. The aromatic region was studied more fully, before a second spectral region was also investigated and the observed cross peaks interpreted and their vibrational modes assigned. Measurements of the level of visible enhancement of the retinal signal that can be achieved were also made – these were then compared with the non-electronically enhanced signal levels obtained from amino acid side chains. Finally, through humidity control of the samples, attempts were made to observe various structural intermediates of the retinal that are formed at different stages of the bacteriorhodopsin photocycle. Before these results are presented, therefore, the next section comprises a short review of the current understanding of the intermediates and pathways of bacteriorhodopsin’s photocycles.
5.2: Bacteriorhodopsin; structure, function and photocycles

5.2.1: Structure and function

Bacteriorhodopsin (bR) is a small (26 kDa, 248-residue) trans-membrane protein found in *Halobacterium salinarium* (*halobium*), a halophilic archaeabacteria. More specifically, bR is located in particular patches of the plasma membrane known as the purple membrane (PM). The PM contains 75% protein and 25% lipid by mass where bR, the sole protein present, is arranged in the plane of the membrane in a hexagonal crystalline array. The repeating element of this hexagonal lattice is composed of bR trimers; three identical protein chains, each rotated by 120° relative to the others. The PM from *H. salinarium* is the simplest photosynthetic membrane, for which bR acts as a light-driven proton pump. bR converts the energy of green light (500 – 600 nm) into an electrochemical proton gradient of up to 250 mV. [140] This gradient is subsequently used by a second protein, ATP synthase, to generate chemical energy in the form of ATP.

Although the basic function of bR as a photon-driven proton pump was established soon after its discovery in the 1960s [140-144], the exact details of its mechanisms that lead to proton transport across the membrane continue to be investigated and are only now being fully elucidated. For over 40 years bR has been the focus of hundreds of research groups – encompassing structural biology, biophysics and biochemistry – who have used a wealth of techniques to further the understanding of this important protein. bR has garnered such intense interest as it serves as a simple model for two specific types of proteins. First, it is very similar to G-protein-coupled seven-helix receptors; these cell-membrane receptors include most well-known drug targets in humans and probably operate by a switch mechanism similar to bR’s. [140] Second, it is also used as a model for membrane transporters – the
proteins involved in the movement of ions across membranes against an electrochemical potential.

![Figure 5-3: The seven trans-membrane helices of bacteriorhodopsin, with the retinal chromophore at their centre.](image)

Each 248 residue-long bR chain folds into seven trans-membrane alpha helices (A-G), linked by short, extra-membrane interconnecting loops (Figure 5-3). At the heart of the protein is its photoactive moiety, an all-trans retinal molecule. This chromophore is covalently bound via a protonated Schiff base to the ε-amino group of a conserved lysine (Lys-216) near the middle of helix G. The seven helices of the protein chain enclose a cavity that spans the width of the membrane. This cavity is divided by the transversely lying Schiff base into two sections – the extracellular and cytoplasmic half-channels. These are not open pores, but occluded pathways that define the route of protons during their transport across the membrane.

Within the extracellular half-channel there are many charged and polar residues which carry out functional roles, the most important of these being the anionic aspartates, Asp-85 and Asp-212, which act as counter-ions to the Schiff base. More
importantly, the Asp-85 also plays a role in the proton movement across the membrane, acting as the proton acceptor when the Schiff base is deprotonated. Late on in the transport process, the protonated Asp-96 donates a proton back to the Schiff base. This residue is contained in the cytoplasmic region, where the majority of the other residues are hydrophobic in nature. The active site of the protein is centred around a highly polarised water molecule (water 402) which is involved in hydrogen-bonding with the Asp-85 and Asp-212 residues, as well as with the protonated Schiff base. [145-148]

### 5.2.2: Light-adapted and dark-adapted bR

When continuously illuminated, the retinal in bR assumes an all-trans, 15-anti conformation which has an absorption maximum at 568 nm (bR\textsubscript{568}). In the dark, however, a thermal equilibrium between this isomer and a retinal conformation of 13-cis, 15-syn, is slowly established. (The two isomer conformations are shown in Figure 5-4). This resting form, in which ~ 34% of the retinal molecules are in an all-trans, 15-anti conformation and ~ 66% are 13-cis, 15-syn isomers, is called dark-adapted bR (DA-bR). [149, 150] Accordingly, when bR is illuminated and contains almost 100% bR\textsubscript{568}, it is described as light-adapted (LA-bR). The absorption maximum for bR comprising only the 13-cis, 15-syn isomer is found at 548 nm (bR\textsubscript{548}). Due to the 1:2 ratio of the bR\textsubscript{568} and bR\textsubscript{548} populations of DA-bR, its absorption maximum therefore occurs at 560 nm (bR\textsubscript{560}). The isomeric composition of DA-bR can also change depending on environmental conditions such as temperature, pH, pressure, and hydration level. Up to ~ 40°C, however, the ratio remains almost constant. [146, 149-154] The absorption spectra of bR\textsubscript{548}, bR\textsubscript{560} and bR\textsubscript{568} are shown in Figure 5-5, along with the isomeric compositions of light-adapted and dark-adapted bR.
Since the all-trans, 15-anti geometry is the lowest energy form of the isolated retinal in solution, the formation in the dark of the 13-cis, 15-syn isomer indicates that the protein preferentially binds this conformation. [137, 152] In the thermal bR\textsubscript{568} to bR\textsubscript{548} conversion, the retinal isomerises about both the C\textsubscript{13}=C\textsubscript{14} and C\textsubscript{15}=N double bonds, and it is due to this double isomerisation that both the isomers are nearly linear and are therefore accommodated by the binding pocket without undue strain.
X-ray crystallography indicates that the bR protein structure with the 13-cis, 15-syn isomer is very similar to that with the all-trans, 15-anti isomer – in particular, similar water-containing hydrogen-bonding networks are found at the Schiff base in both. [156]

In order to dark-adapt an aqueous suspension of PMs and achieve the 1:2 bR_{568}:bR_{548} equilibrium, it is usually sufficient to store the solution in the dark at room temperature for about 12 hours. [154, 157] The rate of dark adaptation increases with temperature, and so at higher temperatures shorter incubation times are required; at 1°C the process may take up to 1 week. [149, 150] Light adaptation of a sample is achieved through continuous illumination using light above ~ 490 nm. Typically a 150 – 500 W mercury- or tungsten-halogen lamp is used, with heat and cut-off filters (to halt the dark adaptation process and to ensure elimination of the actinic light below 500 nm, respectively) in place between it and the suspension. [153, 158-160] The light adaptation process can take from a couple of minutes with moderate illumination (e.g. 500 mW/cm² incident on the sample) [154] to 4 hours under weak illumination (e.g. 500 µW/cm² incident on the sample). [151]

5.2.3: The photocycles of bR_{568} and bR_{548}

Upon light excitation, both of the bR_{548} and bR_{568} isomers undergo distinct photocycles. The two isomers form different photoproducts and subsequent intermediates, but in both cases the cycles are initiated by photo-isomerisation involving rotation around the C_{13}=C_{14} double bond. The intermediates of the photocycle of LA-bR (bR_{568}) have been well characterised, but much less is known about those of the cycle of the 13-cis, 15-syn component (bR_{548}) of DA-bR.
During the light-adapted photocycle the primary photoproduct returns via dark reactions through a series of transients to replenish only the starting retinal isomer, all-trans, 15-anti. There are no dark reactions from any transient of this cycle into that of bR$_{548}$, nor does photo-conversion from bR$_{568}$ to bR$_{548}$ occur. However, due to a branching reaction from an intermediate of the bR$_{548}$ cycle, this photocycle leaks into that of the bR$_{568}$ form and the all-trans isomer becomes the major component in the light-adapted state. The two cycles also differ in that the photocycle of bR$_{548}$ does not involve a vectorial proton translocation; proton pumping is restricted to that of bR$_{568}$. Details of the two photocycles are provided in Figure 5-6.

![Diagram of the photocycles of bR$_{568}$ and bR$_{548}$](image)

**Figure 5-6:** The photocycles of bR$_{568}$ and bR$_{548}$ – the details from a variety of static and time-resolved studies have been combined to provide a full picture of both cycles. [142, 145, 148, 153, 155, 157-159, 161-187]

Over the past forty years, various spectroscopic techniques have been employed to identify and provide information about the intermediate states of the photocycles. Detection of the structural changes that occur in the retinal and its surrounding protein during the cycles has been carried out, either,
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

- by cryo-trapping the intermediates, determining their structures by static methods (NMR, neutron scattering, electron microscopy) and then comparing these with the ground state; or,
- directly, using time-resolved spectroscopies for measuring the transient changes in real time (ultra-fast laser spectroscopies, flash photolysis, resonance Raman, FTIR).

Step-by-step mechanisms have thus been proposed for the photocycles. Overtime, with the emergence of new techniques and data, these mechanisms have changed and developed extensively, and to date there are still many contradicting views regarding these.

**The bR\textsubscript{568} photocycle**

The photocycle of bacteriorhodopsin when it is in a light-adapted state is responsible for the generation of *H. salinarium*’s energy. During the bR\textsubscript{568} cycle a proton gradient is established and subsequently used by a second protein, ATP synthase, in the synthesis of the cell’s energy, ATP. The mechanism of proton transport involving the retinal’s Schiff base is explained more fully in Figure 5-7.

In the first step of the reaction cycle, the absorbed photons electronically excite the all-trans retinal and drive a rapid trans-to-cis isomerisation about its C\textsubscript{13}=C\textsubscript{14} double bond, resulting in the 13-cis, 15-anti conformation. This photo-isomerisation initiates a series of thermally-driven reaction steps in which the bR proceeds through a number of structurally distinct intermediates before returning to its initial all-trans, 15-anti state. The individual intermediates evolve on timescales of femto- to milli-seconds, and one complete cycle takes place in less than 15 ms. When in the bR and O states, the retinal has an all-trans, 15-anti conformation; all other
intermediates are the 13-cis, 15-anti isomer. Details of the intermediate conformations and the timescales over which they evolve are provided in Figure 5-8.

Figure 5-7: The molecular mechanism of proton pumping in bacteriorhodopsin, reproduced from Kuhlbrandt [140]. Red arrows represent proton movements and blue arrows movements by groups of atoms. The ‘paddle’ attached to helix F represents bulky side chains which move to open the cytoplasmic proton channel.

The intermediates of the bR568 photocycle have different protonation states of the Asp-85 and Asp-96 residues and the Schiff base. Up to the L intermediate, the initial state of the protein is maintained; the Asp-85 is unprotonated, whilst the Schiff base and Asp-96 are protonated. [147, 166] In the L to M1 transition, the Schiff base transfers its proton to Asp-85. As M2 is reached, the F and G helices open up to bring the retinal to Asp-96 in the cytoplasmic half-channel. In the M2 to N step, the deprotonated Schiff base then accepts a proton from the Asp-96. [182, 188, 189] During the N to O transition, the cis to trans reisomerisation is accompanied by the uptake of a proton from the cytoplasmic surface to reprotonate Asp-96. The O state is the only state in which the retinal, Asp-85 and Asp-96 are all protonated. Finally the O to bR transition sees the retinal return to its twisted all-trans conformation, whilst the Asp-85 releases it proton through a network of H-bonds and water molecules to the extracellular medium. [148, 190-194]
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

Figure 5-8: The bR568 photocycle providing details of the retinal’s intermediate conformations and the timescales over which they evolve. [145, 148, 158, 165-167, 169-187, 194]

Step 1, bR to J to K: In the primary reaction of the photocycle, absorption of light by the ground state all-trans, 15-anti retinal produces an excited state, bR*, which rapidly depopulates (~ 500 fs) to produce the 13-cis isomerised intermediate, J. Subsequent to this isomerisation around the C\textsubscript{13}=C\textsubscript{14} double bond, the J state relaxes to the K intermediate within ~ 3 ps. It has been suggested that J is a highly twisted and thermally excited form of K, and it is thus thought that the J to K transition occurs via vibrational cooling and torsional relaxation of the retinal.

Step 2, K to L: The K state, in which the retinal has the 13-cis, 15-anti conformation, has an absorption maximum of 590 nm. During the K to L transition, the hydrogen-bonding interaction between water 402, the protonated Schiff base and Asp-85 in the extracellular half-channel is strengthened. Within ~ 1 µs the L intermediate, whose absorption maximum is found at 550 nm, is formed.

Step 3, L to M\textsuperscript{1} to M\textsuperscript{2}: The M intermediate is split into two states, M\textsuperscript{1} and M\textsuperscript{2}, which, with absorption maxima at 412 nm, are yellow in colour. M\textsuperscript{1} is reached from the L intermediate within several microseconds, during which the first proton translocation step of the photocycle occurs as the Schiff base transfers its proton to the Asp-85 in the extracellular half-channel. During the M\textsuperscript{1} to M\textsuperscript{2} transition, taking up to ~ 100 µs, the retinal moves up towards the Asp-96 in the cytoplasmic half-channel.

Step 4, M\textsuperscript{2} to N: The second proton transfer step of the photocycle occurs during the transformation of M\textsuperscript{2} to the N intermediate. The proton transferred to the Schiff base does not originate directly from the cytoplasm; the Asp-96 residue in the cytoplasmic half-channel acts as a proton store for this task. The reprotonation of the Schiff base from Asp-96 takes ~ 3 ms.

Step 5, N to O to bR: The thermal isomerisation of the retinal from a 13-cis back to an all-trans conformation marks the transition from the N to the O intermediate. This takes a few milliseconds, and during this period the Asp-96 is also reprotonated from the cytoplasmic side of the membrane. In the O state the retinal has a considerably red-shifted absorption maximum of 640 nm. In the final, and strongly unidirectional step of the photocycle, the initial bR state is regenerated from the O intermediate and the retinal returns to its twisted all-trans conformation. Virtually coincident with this, the Asp-85 residue transfers its proton to the extracellular side of the membrane.
The bR$_{548}$ photocycle

Unlike that of LA-bR, the photocycle of the 13-cis, 15-syn component of DA-bR does not pump protons and so does not result in energy generation for the cell. It is probably for this reason that the intermediates and pathways of this photocycle have been far less well-characterised and defined than those of the bR$_{568}$ cycle. The photocycle proposed in Figure 5-9 is therefore a composite of the transients and pathways that have been suggested by various studies. Depending on the intensity of the excitation light, bR$_{548}$ can undergo two differing transitions. [153, 154, 158, 163] The first, which occurs at higher (and continuous) light intensities, involves the direct photo-conversion from bR$_{548}$ to bR$_{568}$. Alternatively, at lower light intensities, the bR$_{548}$ progresses through its photocycle which has distinct, resolvable intermediates.

At higher light intensities, the main route responsible for light adaptation of bR is a photo-induced conversion of bR$_{548}$, or its primary photoproduct, J, to a species in the bR$_{568}$ cycle (bR$_{568}$ itself, or possibly K$_{590}$). [158, 161, 163, 195] This transition competes with the bR$_{548}$ photocycle and the formation of its characteristic 610 nm intermediate. Various experiments using ultra-fast laser spectroscopy have demonstrated that, even at high pulse energies (for example, 2 – 10 mJ/cm$^2$), DA-bR is only partially converted to LA-bR; additional and continuous illumination of the sample is required to make complete light adaptation take place and to maintain it in a light-adapted state. At low excitation intensities, bR$_{548}$ undergoes its photocycle and, after initial formation of photoproducts J and K, forms the intermediate characteristic of this photocycle, L$_{610}$. [158]
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

Figure 5-9: The retinal intermediates of the br<sub>548</sub> photocycle. [142, 155, 157-159, 161-164, 178]

**Step 1. Formation of the primary photoproducts:**
- Following absorption of light, photo-isomerisation probably occurs at the C<sub>13</sub>=C<sub>14</sub> bond, leading from the 13-cis, 15-syn initial isomer to an all-trans, 15-syn intermediate. The absorption spectrum of this intermediate is very similar to that of the K intermediate in the br<sub>568</sub> photocycle, and has thus too been labelled K<sub>590</sub>. The K<sub>590</sub> of the br<sub>548</sub> photocycle however decays more slowly than that of the br<sub>568</sub> cycle.
- Although they are formed along different pathways, the initial photoproducts, bR*, J and K, of both the br<sub>568</sub> and br<sub>548</sub> cycles are formed with nearly identical time constants. It has been suggested that is due to similar separations between the protonated Schiff base of the retinal and its counter-ion in both cases.

**Step 2. The L<sub>610</sub>-intermediate and its decay pathways:**
- L<sub>610</sub> is formed via a thermal transition from the K<sub>590</sub> intermediate. Its decay pathway is split into two dark reactions, one leading to br<sub>548</sub> and the other to br<sub>568</sub>. The first of these thermal transitions results in the replenishment of the initial 13-cis, 15-syn isomer, and it is this pathway that predominates. In the alternate thermal transition, L<sub>610</sub> decays to the all-trans, 15-anti isomer. This branching reaction into the br<sub>568</sub> photocycle has a much lower yield (~10%) than the L<sub>610</sub> to br<sub>548</sub> transition, and so most molecules remain within the br<sub>548</sub> cycle. The fraction of molecules that however ‘leaks’ into the br<sub>568</sub> photo cycle is able to form M intermediates and then pump protons.
- Molecular dynamics simulations have shown that, following photo-isomerisation of the br<sub>548</sub> retinal, its Schiff base nitrogen points to the intracellular rather than to the extracellular side. In the L<sub>610</sub> intermediate, the Schiff base proton is pointing towards the cytoplasmic side and thus loses contact with Asp-85, the normal proton acceptor. This is possibly why the br<sub>548</sub> cycle is unable to pump protons. Conversely, an intact hydrogen-bonding network between the Schiff base proton and Asp-85 has been shown in the simulated L<sub>550</sub> intermediate of the br<sub>568</sub> photocycle; this should allow proton transfer from the Schiff base to Asp-85.
5.2.4: Effect of hydration level on bR

The intermediates and pathways of the bR photocycles described above are those that occur under physiological conditions, i.e. in aqueous suspensions at neutral pH. When a purple membrane sample is dried out, however, not all of the intermediates of the photocycles are formed, and those that do form do so at very different rates. In particular the process of light adaptation is considerably affected by the hydration level of the PMs. Thin films of dried membranes can be produced by simply leaving a PM suspension to dry on a slide or, alternatively, to achieve oriented PMs, an electrophoretic deposition technique can be used. [196, 197]

Dark and light adaptation

Purple membrane films at 95 to 100% relative humidity display the same spectroscopic and light-to-dark adaptation properties as PMs suspended in water. At lower hydration states of the PM, however, the cis-to-trans photo-isomerisation of bR$_{548}$ to bR$_{568}$ (light-adaptation) is increasingly hindered and the amount of bR$_{548}$ observed in LA-bR increases. In fact, below 10% relative humidity light-adapted and dark-adapted bR are spectroscopically indistinguishable. Whether a dry LA-bR sample is produced through irradiating a solution at above 95% relative humidity and then lowering its hydration level, or whether it is produced at low humidity and then illuminated, its absorption maximum shifts from 568 nm to ~ 560 nm. [159, 196, 198-200]

The thermally-activated dark adaptation process – the establishment of an equilibrium between bR$_{568}$ and bR$_{548}$ – is unaffected by drying of the sample. There is no difference between the isomer ratios reached in dried samples and aqueous suspensions of PMs that have been kept in the dark for 12 hours. [162] Nor is a
change observed in the absorption maximum of a previously DA-bR (bR\textsubscript{560}) sample when its humidity is lowered to below 10%. [198, 199, 201]

**The bR\textsubscript{568} and bR\textsubscript{548} photocycles**

At less than 95% relative humidity both the bR\textsubscript{568} and bR\textsubscript{548} photocycles are prolonged. The effect of dehydration on the bR\textsubscript{568} photocycle is to accelerate the first steps of proton movement but to prevent the last steps of the cycle. At relative humidities of < 95%, the L\textsubscript{550} intermediate is formed and, during its decay to the first M\textsubscript{412} intermediate, the Schiff base loses its proton to Asp-85. The transition of M\textsubscript{1} to M\textsubscript{2} is however drastically slowed and the O\textsubscript{640} intermediate is not observed. It has therefore been suggested that the recovery of bR\textsubscript{568} occurs through shunts directly from the M intermediates, thus by-passing the O intermediate. These proposed shunt pathways do not result in proton translocation, but rather in the return of the proton from the Asp-85 residue to the retinal. In such cases, the overall photocycle is prolonged from ~ 15 ms to ~ 100 ms, and accumulation occurs at the M state. At humidities below ~ 50%, the cycle is thought to halt completely at the M intermediates. In the bR\textsubscript{548} photocycle, a relative humidity below 90% results in the inhibition of the formation of the L\textsubscript{610} intermediate. [181, 197-199, 202]
5.3: Electronically resonant EVV 2DIR spectroscopy of bacteriorhodopsin’s retinal chromophore

5.3.1: Experimental methods and considerations

Lyophilised powder of bacteriorhodopsin as purple membranes was purchased from Actilor NanoBioSecurity. [203] This had been isolated from *H. Salinarium* by the standard procedure described by Oesterhelt. [143, 144] A 20 µM aqueous suspension of the PMs (in nanopure water) was used to prepare dry bacteriorhodopsin films on glass slides in the same manner as for the peptide and protein fingerprinting experiments (Section 3.2.1). 1 to 3 µl of the PM solution were deposited onto a glass microscope slide and left to dry at room temperature (Figure 5-10). The humidity at which the films were produced was not controlled, but during the 2DIR experiments the samples were in a sealed box purged with nitrogen at a humidity of < 1%.

![Figure 5-10: Photograph of dry purple membrane films deposited on a glass microscope slide.](image)

Visible absorption spectra of the PM films were measured and these gave absorption maxima, $\lambda_{\text{max}}$, of between 557 and 562 nm. This indicated the bR was in a dark-adapted state, DA-bR. These absorption maxima for the films were observed even if they were prepared from a light-adapted solution ($\lambda_{\text{max}}$ of 569 nm) and kept illuminated during the drying period. This was expected as dried LA-bR films are known to be spectroscopically indistinguishable from DA-bR aqueous samples; the
light-adaptation process is hindered by dehydration, whereas the reaching of a thermal equilibrium between bR_{548} and bR_{568} is unaffected. [196, 198-200]

The first issue with the electronically enhanced EVV 2DIR signal observed from the purple membrane films was that it was unstable and decayed rapidly, in the same manner as observed for the dye molecules (Section 5.1.1). Typically the FWM signal would halve in less than 10 seconds and, within 600 seconds, would drop and level off to less than 10% of its initial value. Each time the sample was moved to expose a fresh area to the laser beams, a large signal could once again be initially observed. Once the signal had decayed from a particular sample spot, it was not possible to achieve signal from the same area again, even if it was left unexposed to the laser beams for more than 2 hours. This indicated that the visible laser beam was photo-bleaching the retinal chromophore.

Due to sample area constraints, it was not possible to spin the PM films to keep exposing fresh bacteriorhodopsin to the beams. It was therefore necessary to use a low visible beam pulse energy to minimize the photo-damage to the retinal. A ten-fold reduction in the visible energy from 5 µJ (used for the protein spectra in Chapter 3) to 0.5 µJ resulted in more stable signal that, although still seen to decay slightly, would remain at > 50% of its initial value after 600 seconds. This then allowed for collection of complete 2D spectra. Once FWM signal had been found andoptimised in a particular spectral region (through adjustment of the beam pointing and longitudinal sample position), the sample was moved so that a fresh spot was exposed to the laser beams for the collection of data. The spectra were measured in transmission.

Initially, published Raman and IR spectra of bacteriorhodopsin were used to identify the possible locations of cross peaks in a particular spectral region, along with trial
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

and error. Later, theoretical 2DIR spectra of the retinal, calculated by R. Guo and I. R. Gould (Section 5.5), also provided guidance. Due to the experimental constraints, only two small spectral regions were explored for bR. The first region that was surveyed was that in which the tyrosine, phenylalanine and tryptophan side chain cross peaks had been found; $\omega_\alpha$ of 1350 – 1600 cm$^{-1}$ and $\omega_\beta$ of 2700 – 3300 cm$^{-1}$. As this is where the C=C double bond stretches of the retinal appear in the $\omega_\alpha$ direction, this is designated the ‘ethylenic region’. Following this, experiments were carried out in the ‘fingerprint region’ of 1200 – 1300 cm$^{-1}$ / 2600 – 2950 cm$^{-1}$, where the CC–H bending and C–C single bond stretches are located in the $\omega_\alpha$ direction. For the ethylenic region, delays of $T_{12} = 1$ ps, $T_{23} = 1$ ps were used. In the fingerprint region, the observed cross peak signal levels were lower than in the ethylenic region and here shorter delays of $T_{12} = 0.5$ ps, $T_{23} = 0.5$ ps were required.

5.3.2: The ethylenic region

Two unresolved cross peaks, centred around 1548 / 3060 cm$^{-1}$ and 1560 / 3070 cm$^{-1}$, were observed in the ethylenic region. Through consultation of various Raman and IR studies, the vibrational mode being probed by the $\omega_\alpha$ beam was assigned as a C=C double bond stretch, $\nu$(C=C), of the retinal’s polyene backbone. This was later confirmed by the DFT calculations of the retinal’s Raman, IR and preliminary EVV 2DIR spectra performed by R. Guo and I. R. Gould (Section 5.5). Further to this, the assignment of the vibrational state excited by the second IR beam, $\omega_\beta$, was identified as the first overtone of the C=C double bond stretch, $2\nu$(C=C). The observed cross peaks are therefore associated with the couplings between the C=C double bond stretch and its first overtone; $\nu$(C=C) / $2\nu$(C=C).

The relative level of signal from the two ethylenic cross peaks varied from sample to sample, and also over multiple spectra taken of the same sample spot. Attempts
were made to reproducibly observe the variations in their relative intensities, for example to see if the signal from either of the cross peaks increased or decreased over time with continued exposure to the laser beams or with increased visible beam energy. It was however not possible to control which of the two cross peaks was the more intense through adjustment of these parameters. Three example spectra of the ethylenic region are shown in Figure 5-11, in which the cross peaks are labelled with their mode assignments.

![Figure 5-11: Three example ethylenic region ER EVV 2DIR spectra measured of bacteriorhodopsin’s retinal chromophore. The cross peaks represent triply resonant FWM signal from two coupled vibrational modes of the retinal’s polyene backbone; ν(C=C) and 2ν(C=C). Two unresolved cross peaks can be observed, centred around 1548 / 3060 cm\(^{-1}\) and 1560 / 3070 cm\(^{-1}\). Spectrum (a) is an example in which the lower frequency cross peak predominates, whilst (b) is an example in which the higher frequency cross peak is the more intense. Spectrum (c) shows the same level of signal from the two cross peaks. In all three spectra, there is a strong dependency on the \(\omega_a\) frequency, with singly vibrationally enhanced signal observed along \(\omega_a\) above and below the cross peaks. The spectra were measured with a visible beam energy of 0.5 µJ, delays between the pulses of \(T_{12} = 1\) ps, \(T_{23} = 1\) ps, and with all three beams polarised in the plane of propagation (PPP).]
Previous Raman (both resonance and stimulated [137, 138, 141, 165, 204-208] ) and IR (FTIR difference and femtosecond time-resolved [42, 136, 139, 155, 209-216]) spectroscopy studies have shown the frequency of the C=C stretching mode, ν(C=C), to be extremely sensitive to the degree of π-electron delocalisation along the polyene backbone of the retinal. In fact, there is an inverse linear relationship between this C=C stretching frequency and the wavelength at which the retinal absorbs most strongly, λ_{max}. Hence any changes in the position of the ν(C=C) band can serve as an indirect gauge of changes in the absorption maximum of the retinal, and thus its conformation. [139] Depending on whether the bR’s retinal is in a 13-cis, 15-syn (bR_{548}) or all-trans, 15-anti (bR_{568}) conformation, this ethylenic band therefore shifts in frequency. In resonance Raman spectra (shown in Figure 5-12), the ν(C=C) band shifts by as many as 9 cm$^{-1}$, from 1536 cm$^{-1}$ for bR_{548} to 1527 cm$^{-1}$ for bR_{568}. Dark-adapted bR, which contains both the bR_{548} and bR_{568} isomers, has a broad band at 1532 cm$^{-1}$ encompassing both of these peaks. [138, 207, 217] In FTIR difference spectra, the ν(C=C) absorption is observed at 1536 cm$^{-1}$ for bR_{548} and at 1524 cm$^{-1}$ for the bR_{568} conformation. [42, 136, 213, 216]

![Figure 5-12: Resonance Raman spectra of bacteriorhodopsin reproduced from Smith et al. [137]
Spectrum B is of light-adapted bR (bR_{568}), C is of bR_{548}, whilst spectrum A, of dark-adapted bR (bR_{560}), was obtained by subtracting ~ 40% of spectrum B from spectrum A. The most intense ethylenic peak seen at ~ 1530 cm$^{-1}$ has contributions from all five C=C stretches of the retinal’s polyene backbone.](image-url)
The absorption maxima ($\lambda_{\text{max}} \sim 560 \text{ nm}$) of the PM films, measured prior to the collection of the ER EVV 2DIR spectra of Figure 5-11, indicated that they were in a dark-adapted state and thus initially composed of both the bR$_{548}$ and bR$_{568}$ isomers. The two cross peaks, which are separated by about 13 cm$^{-1}$ in the $\omega_\alpha$ direction, could therefore be tentatively assigned to each of these isomers. In this case, the lower frequency peak at $1548 / 3060 \text{ cm}^{-1}$ would be attributed to bR$_{568}$ (observed at 1527 cm$^{-1}$ in the above resonance Raman spectrum), and the higher frequency peak at $1560 / 3070 \text{ cm}^{-1}$ to bR$_{548}$ (seen at 1536 cm$^{-1}$ in the above Raman spectrum). The 2DIR cross peaks are shifted along the $\omega_\alpha$ axis by about 15 cm$^{-1}$ compared with Raman and IR spectra, but this is common to all of the EVV 2DIR spectra measured, including the protein spectra shown in Chapter 3.

The assignment of the observed $\nu$(C=C) / 2$\nu$(C=C) retinal cross peaks to the two isomeric components of DA-bR is however problematic. The intensity of the visible probe beam (13 mJ/cm$^2$) was sufficient to trigger both the bR$_{568}$ and bR$_{548}$ photocycles and produce their many intermediates. [158, 165, 218] The dehydrated nature of the PMs means that these cycles would have been substantially prolonged or, most probably, halted part of the way through. It is more likely therefore that the cross peaks represent a combination of both bR$_{568}$ and bR$_{548}$ and various trapped intermediates of their photocycles, in particular the M state. This is also supported by the irreproducible variation in the relative intensities of the two peaks. It is clearly not possible to assign these spectral features to specific conformations of the retinal; this would require pump-probe 2DIR experiments, to initiate the photocycles and then measure the spectra of particular intermediates after specific lengths of time. This would in any case be particularly difficult for such extremely dehydrated, dark-adapted samples.
5.3.3: The fingerprint region

In the second two-dimensional region to be explored – the ‘fingerprint region’ – two clearly resolved features were observed at 1225 / 2745 cm\(^{-1}\) and 1270 / 2790 cm\(^{-1}\). As had been the case for the cross peaks in the ethylenic region, the relative intensities of these varied from sample to sample and also for multiple spectra collected of the same sample spot. In the majority of spectra, however, far stronger signal was achieved from the lower frequency cross peak and in some cases the higher frequency cross peak was not observed at all. Sometimes this was simply due to the signal dropping off with time; the spectra were collected from lower to higher \(\omega_\alpha\) frequency and so the signal from this peak had more time to decay. Quick signal level readings on the cross peaks before spectra collection did however show the higher frequency cross peak to be significantly weaker.

Again through consultation of previous Raman and IR studies, and also of DFT calculations of Raman, IR and preliminary EVV 2DIR spectra, the cross peaks have been assigned to two coupled vibrational modes of the retinal polyene backbone – this time to a fundamental and a combination band. The vibrational states excited by the \(\omega_\alpha\) IR beam have been attributed to C–H bending modes, \(\delta(CCH)\), of different sites along the conjugated chain. These rocking motions also have significant contributions from the C–C stretching of the backbone. The CCH bending modes couple with a combination band of themselves with the C=C double bond stretch, \(\delta(CCH) + \nu(C=C)\), which is excited by the \(\omega_\beta\) IR beam. The electronic enhancement of these coupled modes, \(\delta(CCH) / \delta(CCH) + \nu(C=C)\), by the visible beam was lower than that observed for the ethylenic cross peaks. Two example fingerprint region spectra of the retinal are shown in Figure 5-13, in which the cross peaks are labelled with their mode assignments.
Both IR and resonance Raman studies have shown the CCH rocking modes to be crucial markers for retinal chromophore conformation. [42, 136-138, 155, 217] As for the C=C stretching mode, the frequencies of these vibrations shift as the retinal molecules progress through the intermediates of their photocycles. The resonance Raman spectra of Figure 5-12 show that both the bR\textsubscript{568} and bR\textsubscript{548} ground state isomers have four resolved vibrations from the CCH rocks (which also have strong contributions from the C–C stretching of the backbone). In both cases, the three most intense and overlapping peaks (centred at ~1201 cm\textsuperscript{-1} and ~1183 cm\textsuperscript{-1} for the bR\textsubscript{568} and bR\textsubscript{548} isomers respectively) are separated from a much weaker one by ~50 cm\textsuperscript{-1} (at 1255 cm\textsuperscript{-1} for bR\textsubscript{568}, and at 1234 cm\textsuperscript{-1} for bR\textsubscript{548}). The two cross peaks of the EVV 2DIR spectrum are too separated by ~50 cm\textsuperscript{-1} along \(\omega_\alpha\), with the more intense, broader cross peak at 1225 / 2745 cm\textsuperscript{-1} and the weaker one at 1270 / 2790 cm\textsuperscript{-1}. It is however, not yet possible to be certain of the exact assignments of these cross peaks; it is hoped that theoretical EVV 2DIR spectra will provide these in the future (see Section 5.5). Additionally, as for the cross peaks observed in the
ethylenic region, the fingerprint region 2DIR features cannot be attributed to particular bR isomers or intermediates of the photocycles. Due to the severely dehydrated states of the samples, as well as their continual irradiation with the visible beam, these cross peaks are most probably from a combination of bR$_{548}$, bR$_{568}$ and accumulated intermediates of their photocycles.

5.4: Studying the intermediates of bR$_{568}$’s photocycle using ER EVV 2DIR

The EVV 2DIR spectra discussed above in Section 5.3 were of severely dehydrated PM samples in which the bR was in a dark-adapted state. It was not possible to study particular retinal intermediates of the bR photocycles with such samples. Ideally it would be feasible to perform pump-probe EVV 2DIR experiments of light-adapted samples, i.e. to initiate the bR$_{568}$ cycle and then measure the spectra of different intermediates after specific lengths of time. (Alternatively, specific intermediates could be examined by trapping them cryogenically.) Although this may be realisable in the future with the new instrument, when the work presented here was completed an alternative method for studying specific intermediates had to be found. A viable solution was to attempt to observe different photo-accumulated intermediates of the bR$_{568}$ cycle through the use of humidity control.

In Section 5.2.4, it was explained that the hydration level of a PM sample has a significant effect on the length of the bR’s photocycle. It was therefore calculated that it should be possible to photo-accumulate two different intermediates of the bR$_{568}$ cycle – M$_{412}$ and O$_{640}$ – at different specific sample humidities. The collection of 2DIR spectra of the M and O intermediates was attempted only in the ethylenic region. As explained in Section 5.3, the position of the C=C stretch is an indication
of retinal conformation and can indicate whether a sample comprises the bR$_{548}$ or bR$_{568}$ isomers. Similarly, the C=C frequency shifts with the different photocycle intermediates due to their changing conformations and protonation states. [42, 136-139, 214, 215, 219]

5.4.1: Experimental methods and considerations

Purple membrane films at three different hydration levels were produced by equilibrating them at relative humidities (RH) of 35%, 65% and 100%. 20 µl of a 50 µM bR solution (10 mM Hepes, pH 7.4) was deposited onto the centre of a CaF$_2$ window and left overnight in a sealed glove box. The required humidities in the glove box (checked with a humidity probe) were produced by trays of saturated salt solutions; magnesium chloride for 35% RH, calcium nitrate for 65% RH and just water for 100% RH. One hour before spectra were to be measured, a sample was sealed into a cell within the glove box in order to maintain the humidity of the film. A 3 mm thick rubber ring, coated with vacuum grease, was placed onto the CaF$_2$ window around the deposited PM solution, and a second CaF$_2$ window was then clamped on top of this. All samples were produced, and their spectra measured, in a temperature-controlled room at 20 °C ± 0.5 °C. Figure 5-14 is a photograph of a disassembled cell, as left to equilibrate overnight in the glovebox.

Light-adaptation of the films (once sealed in a cell) was achieved using a projector lamp with a 150 W xenon bulb. A heat filter was placed between the sample and the lamp to prevent the acceleration of the dark-adaptation process; a 495 nm long-pass filter was also used to cut out light below 500 nm. After 30 minutes of exposure to the lamp, both the 65% and 100% relative humidity samples consistently had absorption maxima of 569 nm. This confirmed them to be fully light-adapted and consisting of solely the all-trans, 15-anti retinal isomer. Even with two hours of
irradiation with the lamp, the $\lambda_{\text{max}}$ of the 35% RH samples would remain at 553 – 560 nm. This indicated that at this hydration level the light-adaptation of the film was hindered, leaving more than 66% of the retinal molecules as the $\text{bR}_{548}$ isomer (13-cis, 15-syn).

Figure 5-14: An example sample cell used to maintain a PM film at a particular relative humidity whilst 2DIR spectra were measured. The cell comprises two $\text{CaF}_2$ windows separated by a rubber ring and clamped together by a stainless steel outer casing. Once sealed, the cell is able to maintain the sample at the required humidity.

The 2DIR spectra of the 35% and 65% RH films were measured with a visible beam energy of 0.05 $\mu$J. This was found to cause no damage to the samples, with the retinal signal levels maintained even after prolonged exposure to the laser beams. This was however not possible for the 100% RH samples – for signal to be obtained from these the visible beam needed to be at least 1.5 $\mu$J. Even though such an energy would damage the 35% and 65% RH samples within a few seconds, the retinal signal from the 100% RH sample was not observed to decay at all. Once FWM signal had been found and optimised at the required frequencies, all the samples were moved so that a fresh spot was exposed to the laser beams for the collection of data. The spectra were measured in transmission.

For this work, the spectra measured covered only the ethylenic region of 1350 – 1600 cm$^{-1}$ ($\omega_\alpha$) and 2700 – 3300 cm$^{-1}$ ($\omega_\beta$). Although the longest delays that could be used for the severely dehydrated films (< 1% RH, shown in Section 5.3) were $T_{12} = 1$ ps, $T_{23} = 1$ ps, longer delays could be used for the more hydrated samples. In fact
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

pulses delays of up to $T_{12} = 2$ ps, $T_{23} = 2$ ps could be used. This indicated that some degree of broadening of the IR pulses by the water vapour within the sample cell was occurring. The spectra were recorded with all three beams polarised in the plane of propagation (PPP).

In calculations to determine the degree of photo-accumulation of the M and O intermediates in the 65% and 100% RH samples, the quantum yield of the $\text{bR}_{568}$ ground state to initial K intermediate isomerisation was assumed to be 0.6. The quantum yield of this trans-to-cis isomerisation has experimentally been shown to be wavelength independent and insensitive to different light sources and methods. [220-224] This independence has however only be tested over the 500 – 670 nm region, and has not been investigated at 790 nm (the wavelength of the 2DIR visible beam).

5.4.2: ER EVV 2DIR studies of $\text{bR}_{568}$’s M and O intermediates

EVV 2DIR spectra were initially measured of the 35% relative humidity PM films, which were known to be in a dark-adapted state and thus comprising both the $\text{bR}_{568}$ and $\text{bR}_{548}$ isomers. The spectra had two ethylenic cross peaks, as observed for the dried films used in Section 5.3 whose 2DIR spectra are shown in Figure 5-11. This was to be expected from films of low humidity, and no further work with these samples was carried out.

The PM films at 65% and 100% relative humidity were used to attempt to observe the M and O intermediates respectively. Both of these samples could be successfully light-adapted, meaning that only the ground state $\text{bR}_{568}$ isomer (all-trans, 15-anti) was initially present. At > 95% RH, the $\text{bR}_{568}$ photocycle is triggered and proceeds through the same intermediates as PMs in solution, returning to the ground state in
the normal time of ~15 ms (Figure 5-7). Therefore 100% RH samples were used to attempt to observe the longest-lived intermediate of the photocycle, the O state, by its photo-accumulation from continued irradiation with the visible laser beam. Conversely, at < 90% RH, the O state is not observed at all due to a shunting pathway that occurs straight from the M intermediate back to the ground state. In this case accumulation of the M state occurs and the whole cycle is prolonged to ~100 ms. [181, 184, 196, 200, 202] The 65% RH samples were therefore used to photo-accumulate and observe the M intermediate.

Previous resonance Raman studies have demonstrated how the C=C stretch frequency changes with the differing conformations of the bR$_{568}$ photocycle. In particular, it shifts from 1526 cm$^{-1}$ for the ground state isomer, to 1566 cm$^{-1}$ when in the M$_{412}$ state and finally to 1509 cm$^{-1}$ for the O$_{640}$ retinal conformation. [137, 138, 217] This reflects the inverse linear relationship between the C=C stretching frequency and the wavelength at which the retinal absorbs most strongly, $\lambda_{\text{max}}$. IR difference spectra similarly display the same frequency shifts, from 1529 cm$^{-1}$ for the ground state to 1555 cm$^{-1}$ and 1506 cm$^{-1}$ for the M and O intermediates respectively. [136, 185, 186, 213, 214, 225]

The purpose of these experiments was therefore to observe any frequency shifts of the $\nu$(C=C) / 2$\nu$(C=C) cross peak between the ER EVV 2DIR spectra of the M and O intermediates. It was fortuitous that these two could viably be photo-accumulated under the constraints of the humidity control experiments, since the difference, observed with Raman and IR spectroscopies, in the frequency of their C=C stretching is > 45 cm$^{-1}$. With the 5 cm$^{-1}$ resolution of the EVV 2DIR spectra shown here, and also the broad nature of their cross peaks, smaller frequency shifts would have been more difficult to observe. In any case, the objective was not to give precise $\nu$(C=C) / 2$\nu$(C=C) frequencies for the M and O intermediates, but rather to
observe a significant shift of this cross peak from the two samples and thus indicate the presence of different retinal intermediates. Concurrently with this experimental work, the theoretical Raman and IR spectra of these isomers were also calculated by I. R. Gould and R. Guo for comparison. The ultimate aim of this strand of research is to compare DFT calculations of the EVV 2DIR spectra of the retinal intermediates with the experimental results.

**65% relative humidity bR – The M intermediate?**

To measure 2DIR spectra of the 65% RH samples, the visible beam energy was 0.05 µJ. This was sufficient to photo-convert ~ 22% of the retinal molecules within the beam diameter with each pulse. At this hydration level the bR<sub>568</sub> photocycle is prolonged, with the retinal molecules taking ~100 ms to return directly from the M intermediate to the ground state. Under these conditions, photo-accumulation of > 99% of the retinal molecules in the M state could thus be achieved.

The EVV 2DIR spectra measured of these samples reproducibly gave one clearly resolved cross peak at 1545 / 3055 cm<sup>-1</sup>. Two example spectra, measured with different delays between the infrared pulses, can be seen in Figure 5-15. These contrast well with the spectra of the dark-adapted 35% RH and dried films (shown in Figure 5-11); the unresolved double cross peaks observed from these had indicated the presence of more than one intermediate and / or both the bR<sub>548</sub> and bR<sub>568</sub> ground state conformations. The 65% RH samples however produced the first spectra in which only one symmetrical ethylenic cross peak was observed. This indicates the presence of just one retinal conformation – under the experimental conditions, the photo-accumulated M<sub>412</sub> intermediate. The position of this cross peak is close to that of the lower frequency peak (1548 / 3060 cm<sup>-1</sup>) of the ethylenic doublet seen in the spectra of the dehydrated DA-bR samples. As this cross peak
was unable to be attributed to a particular bR$_{568}$ photocycle intermediate, this neither confirms nor contradicts the assignment of the single peak of the 65% RH samples to the M$_{412}$ intermediate.

![Figure 5-15](image.png)

**Figure 5-15:** Two example ethylenic region ER EVV 2DIR spectra of bacteriorhodopsin’s retinal, measured of PM films at 65% RH. In both spectra, one clearly resolved cross peak is observed at 1545 / 3055 cm$^{-1}$; this represents triply resonant FWM signal of two coupled vibrational modes of the retinal’s polyene backbone, $\nu$(C=C) and $2\nu$(C=C). At this hydration level of the PMs, and under the 2DIR experimental conditions used, the peak can be attributed to the photo-accumulated M$_{412}$ intermediate of the bR$_{568}$ photocycle. The spectra were measured with a visible beam energy of 0.05 $\mu$J, delays between the pulses of (a) $T_{12} = 1.5$ ps, $T_{23} = 1$ ps and (b) $T_{12} = 2$ ps, $T_{23} = 2$ ps, and with all three beams polarised in the plane of propagation (PPP).

100% relative humidity bR$_{568}$ – The O intermediate?

To measure 2DIR spectra of the 100% RH samples, a visible beam energy of at least 1.5 $\mu$J was required. This was unexpected, since for the 35% RH, 65% RH and completely dehydrated samples an energy above 0.1 $\mu$J was observed to cause photo-damage to the retinal molecules. At 1.5 $\mu$J, however, no loss of retinal signal was observed from the 100% RH films. This highlighted the difficulty of using humidity control for studying photocycle intermediates. Nevertheless, the visible beam energy of 1.5 $\mu$J was calculated to be sufficient to photo-convert ~ 60% of the retinal molecules within the beam diameter with each pulse. At 100% RH, light-
adapted bacteriorhodopsin progresses through all the intermediates of the bR$_{568}$ photocycle on the normal timescales, with a total cycle taking ~ 15 ms to complete. In this case the longest-lived intermediate is O$_{640}$; the last step of the cycle from O to bR, in which the retinal returns to its twisted all-trans conformation and the Asp-85 residue transfers its proton to the extracellular side of the membrane, takes ~ 5 ms. For the particular sample used here and under the explained experimental conditions, photo-accumulation of > 99% of the retinal molecules in the O state should therefore have been achieved within 15 ms.

The EVV 2DIR spectra measured of the 100% RH samples were not as reproducible as those of the 65% RH samples. Even so, all again had only one ethylenic cross peak, though its position varied slightly from spectrum to spectrum in both the $\omega_\alpha$ and $\omega_\beta$ directions (by < 5 cm$^{-1}$). The position of this cross peak in the majority of the collected spectra was however at 1550 / 3055 cm$^{-1}$, and two examples, measured with two different sets of delays between the infrared pulses, can be seen in Figure 5-16.

The observation of only one ethylenic cross peak – compared with the two seen from the dehydrated samples – was again promising and indicated the presence of only one photocycle intermediate. On the other hand, the objective of using the 100% RH samples was to photo-accumulate and observe the O$_{640}$ intermediate in particular, and this did not appear to have been achieved. According to Raman and IR studies, the C=C stretching mode of the O intermediate is ~ 45 cm$^{-1}$ lower in frequency than that of the M intermediate. If indeed the spectra measured of the 65% RH samples (Figure 5-15) represent the photo-accumulated M intermediate, then the spectra of the 100% RH samples (Figure 5-16) appear not to be of O$_{640}$. In fact, not only are the ethylenic cross peaks of these not shifted to lower $\omega_\alpha$ frequency compared to those of the 65% RH samples, but are 5 cm$^{-1}$ higher in
frequency. The irreproducible nature of the experimental spectra, and also the difficulties experienced working at 100% RH indicate that the photo-accumulation of the O_{640} intermediate was not achieved.

![Figure 5-16](image)

**Figure 5-16**: Two example ethylenic region ER EVV 2DIR spectra of bacteriorhodopsin’s retinal, measured of PM films at 100% RH. In both spectra, one cross peak is observed at 1550 / 3055 cm\(^{-1}\); this represents triply resonant FWM signal of two coupled vibrational modes of the retinal’s polyene backbone, \(\nu(C=C)\) and \(2\nu(C=C)\). At this hydration level of the PMs, and under the 2DIR experimental conditions used, the aim was to photo-accumulate the O_{640} intermediate of the bR_{568} photocycle. The position of the cross peak, however, makes it unclear whether this was in fact achieved. The spectra were measured with a visible beam energy of 1.5 \(\mu\)J, delays between the pulses of (a) \(T_{12} = 1.5\) ps, \(T_{23} = 1\) ps and (b) \(T_{12} = 2\) ps, \(T_{23} = 2\) ps, and with all three beams polarised in the plane of propagation (PPP).

### 5.5: Theoretical studies of bR’s retinal

In parallel with the experimental work to measure the ER EVV 2DIR spectra of different bR photocycle retinal intermediates, complementary theoretical studies were carried out by R. Guo and I. R. Gould. A rigorous comparison between quantum mechanical simulations of a number of small molecules’ EVV 2DIR spectra and those measured experimentally had previously been performed by Donaldson and Guo. [89, 90, 94] Additionally, Guo has demonstrated the clear
correlation between theoretical spectra of amino acid side chain models and those measured of small peptides. [92] To aid the assignment of measured retinal cross peaks, and also as an extension of work within the group regarding comparisons of experimental and theoretical spectra, simulations of the retinal’s IR, Raman and EVV 2DIR spectra were performed.

Due to the binding of the retinal Schiff base to its surrounding protein through Lys-216 and to its varying protonation states and delocalised electronic structures throughout the bR photocycle – as well as to the manner in which it moves and twists in its ‘binding pocket’ as it interacts with its surrounding residues and participates in the transfer of protons across the PM – the frequencies and intensities of its vibrational modes are significantly different compared to those of the free retinal molecule. For example, a resonance Raman study of the bR$_{548}$ isomer has shown there to be an almost 50 cm$^{-1}$ difference between the C=C stretching mode of the free 13-cis, 15-syn retinal molecule at 1584 cm$^{-1}$ and the bound retinal of bR$_{548}$ at 1536 cm$^{-1}$. The frequencies of the protonated and unprotonated free Schiff base retinals fall between these at 1565 cm$^{-1}$ and 1583 cm$^{-1}$ respectively. [137]

If meaningful comparisons with experimental 2DIR spectra of bacteriorhodopsin’s retinal are therefore to be made, simulations of the spectra of the free retinal molecule are inadequate. Ideally, any calculations need to take into account the effect of the surrounding protein. This, however, is in no way a trivial task; even to perform DFT calculations of a molecule the size of the free retinal is very involved and requires a considerable amount of computing time due to the large number of its possible vibrational modes. When residues from the protein’s binding pocket are also included, the calculation becomes appreciably more complicated and time-consuming. In this case, the more tractable simulations of spectra of the free molecule can provide a sensible starting pointing. Calculations of two different sets
of theoretical Raman and IR spectra have therefore been undertaken; i) of the free retinal Schiff base, and ii) of a model in which the retinal is embedded within the residues of its surrounding protein with which it has the most significant interactions. Following from this, the full EVV 2DIR spectrum of one conformation of the free retinal Schiff base was calculated; to do this for the embedded system would take many weeks.

5.5.1: Theoretical methods and considerations

The first set of spectra to be completed (by R. Guo) were of the free retinal Schiff base. For these, the Lysine residue to which the retinal Schiff base is bound in bacteriorhodopsin was replaced with a methyl group. Even using this much-simplified model, the molecule still has 156 possible vibrational modes. The optimisations and spectral constants were calculated with Gaussian 03; DFT calculations of the Raman and IR spectra were performed at the B3LYP level using the 6-31+G(d,p) basis set, whilst the EVV 2DIR spectrum was calculated using the smaller 6-31G(d) basis set. The second set of Raman and IR spectra simulations are currently being performed by I. R. Gould. These are of a model bacteriorhodopsin system, essentially comprising the retinal molecule and its binding pocket taken from the protein’s crystal structure. This includes Asp-85, Asp-96, Asp-212, the bound Lys-216, all other residues within 8 Å of the retinal (each capped with a hydrogen atom) and the water 402 molecule.

To date, four different bR intermediates have been studied – the bR$_{568}$ and bR$_{548}$ ground states (GS) and, for comparison with the experimental spectra presented in Section 5.4, the M$_{412}$ and O$_{640}$ intermediates of the bR$_{568}$ photocycle. The conformations and protonation states of each of these are shown in Table 5-1, along with details of the two residues involved in proton transfer with the retinal, Asp-85.
and Asp-96. Although the bR\textsubscript{568} and O\textsubscript{640} states both have the protonated all-trans, 15-anti conformation, in bacteriorhodopsin their geometries are significantly different. When in the protein environment, the bR\textsubscript{568} GS is held in a twisted conformation and the nearby Asp-85 residue is unprotonated, whilst the retinal of the O\textsubscript{640} intermediate is flat and the Asp-85 protonated. This leads to the shifts in vibrational frequencies (and absorption maxima) observed experimentally between the two in bacteriorhodopsin samples.

Attempts to calculate spectra of the twisted, protonated all-trans, 15-anti conformation (as a model of the bR\textsubscript{568} GS) were unsuccessful for the free retinal molecule; when the initial energy optimisation was performed, it would always optimise back to the planar geometry of the O\textsubscript{640} intermediate. Although the studies using the embedded retinal molecule have not yet extended to the bR\textsubscript{568} ground state, initial results of those of bR\textsubscript{548}, M\textsubscript{412} and O\textsubscript{640} have indicated that the binding pocket residues included in the calculations are successfully providing a good model of the protein environment. Under these conditions, no problems were experienced when the energy optimisations of the different conformations were performed. No constraining of the different geometries was required. For example, the bent nature of the M\textsubscript{412} intermediate of the embedded retinal was retained by two nearby tyrosine residues.

<table>
<thead>
<tr>
<th>Retinal conformation and protonation state</th>
<th>bR\textsubscript{568} GS</th>
<th>bR\textsubscript{548} GS</th>
<th>M\textsubscript{412} state (bR\textsubscript{568})</th>
<th>O\textsubscript{640} state (bR\textsubscript{568})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retinal conformation and protonation state</td>
<td>all-trans, 15-anti Protonated</td>
<td>13-cis, 15-syn Protonated</td>
<td>13-cis, 15-anti Unprotonated</td>
<td>all-trans, 15-anti Protonated</td>
</tr>
<tr>
<td>Asp-85 protonation state</td>
<td>Unprotonated</td>
<td>Unprotonated</td>
<td>Protonated</td>
<td>Protonated</td>
</tr>
<tr>
<td>Asp-96 protonation state</td>
<td>Protonated</td>
<td>Protonated</td>
<td>Protonated</td>
<td>Protonated</td>
</tr>
</tbody>
</table>

Table 5-1: The conformations and protonation states of the four different retinal photocycle intermediates studied to date: the bR\textsubscript{568} and bR\textsubscript{548} isomers, and the M\textsubscript{412} and O\textsubscript{640} intermediates of the bR\textsubscript{568} photocycle. Also included are the protonation details of the two residues involved in proton transfer with the retinal, Asp-85 and Asp-96.
5.5.2: Comparison of theoretical and experimental spectra

A summary of the theoretical Raman and IR spectra completed to date is provided in Table 5-2. The calculated frequencies of the C=C stretches for the free retinal molecule are compared with those previously measured experimentally of bacteriorhodopsin using FTIR difference and resonance Raman and spectroscopies. [42, 136-138, 207, 213, 216, 219]

The large effect that the surrounding protein has on the stretching frequencies can be observed; although the positions of the ν(C=C) bands calculated of the free retinal follow the same pattern as though observed experimentally, their frequencies are significantly shifted. For example, the measured O$_{640}$ and bR$_{548}$ bands are ~ 30 cm$^{-1}$ apart in both IR and Raman spectra of bacteriorhodopsin, but in the theoretical spectra of the free molecule there is only a ~ 4 cm$^{-1}$ difference. Initial results from the calculations of the embedded retinal conformations are however very encouraging, and suggest that the included residues have largely been successful in providing the protein environment for the molecule. In particular the calculated IR frequencies seem to match those measured experimentally very closely. However, some of the Raman frequencies are significantly different. The choice of residues used to provide the protein environment is therefore currently being further assessed before this embedded retinal model is used to perform simulations of the EVV 2DIR spectra. The inclusion of more residues in order to more closely mimic the surrounding protein will probably be considered, though any increase in complexity of the system will increase the size of, and time required for, any future calculations.
To date, the theoretical EVV 2DIR spectrum of one retinal intermediate, O\textsubscript{640}, has been completed. This used the free retinal model, with the molecule in the all-trans, 15-anti protonated conformation and planar geometry. Originally this calculation was to be of the twisted bR\textsubscript{568} state but, as explained above, without any surrounding residues, the initial energy optimisation would always flip the molecule back to the flat geometry of the O\textsubscript{640} intermediate. Although the calculation only considered the solo retinal molecule, the number of vibrational modes involved was still large enough for it to take several weeks to complete.

The theoretical EVV 2DIR spectrum of the O\textsubscript{640} intermediate, calculated by R. Guo, is shown in Figure 5-17; as it is of only the free molecule, and was also calculated using a small basis set, it should be treated with caution. Nevertheless it is encouraging that ethylenic $\nu$(C=C) / 2 $\nu$(C=C) cross peaks can be observed around the 1525 cm\textsuperscript{-1} / 3050 cm\textsuperscript{-1} region. (That there are four in total does not indicate that

Table 5-2: Results of DFT calculations of the Raman and IR spectra of four different bacteriorhodopsin photocycle intermediates; bR\textsubscript{568}, bR\textsubscript{548}, M\textsubscript{412} and O\textsubscript{640}, compared with those observed experimentally. The calculations were performed using a model for bacteriorhodopsin’s Schiff base molecule, comprising just the free retinal molecule. The calculations were completed by R. Guo.

<table>
<thead>
<tr>
<th></th>
<th>bR\textsubscript{568} GS</th>
<th>bR\textsubscript{548} GS</th>
<th>M\textsubscript{412} state (bR\textsubscript{568})</th>
<th>O\textsubscript{640} state (bR\textsubscript{568})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raman $\nu$(C=C) frequencies (cm\textsuperscript{-1})</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Raman results. [137, 138, 207, 219]</td>
<td>1527</td>
<td>1536</td>
<td>1566</td>
<td>1509</td>
</tr>
<tr>
<td>Theoretical results of free retinal (R. Guo).</td>
<td>(1502)</td>
<td>1509</td>
<td>1555</td>
<td>1502</td>
</tr>
<tr>
<td><strong>Infrared $\nu$(C=C) frequencies (cm\textsuperscript{-1})</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental FTIR results. [42, 136, 213, 216]</td>
<td>1525</td>
<td>1536</td>
<td>1555</td>
<td>1506</td>
</tr>
<tr>
<td>Theoretical results of free retinal (R. Guo).</td>
<td>(1492)</td>
<td>1495</td>
<td>1555</td>
<td>1492</td>
</tr>
</tbody>
</table>
this number would be measured from bacteriorhodopsin samples. Many of the peaks observed in the calculated Raman and IR spectra of the free molecule were split into doublets, for example.) Also of note are the series of cross peaks along $\omega_\alpha \approx 1225 \text{ cm}^{-1}$. These have been attributed to the coupling of the $\delta$(CCH) mode with various others. In particular, the features at around 1225 / 2700 cm$^{-1}$ have been assigned to the $\delta$(CCH) / $\delta$(CCH) + $v$(C=C) cross peaks observed experimentally and shown in Figure 5-13. The most intense features in the theoretical spectrum appear at $\sim 1450 / 2900 \text{ cm}^{-1}$ and $\sim 1650 / 2950 \text{ cm}^{-1}$. For these, the modes involved have tentatively been assigned as $\delta$(CH$_3$) / 2$\delta$(CH$_3$) and $v$(C=N) / $v$(C=N) + $\delta$(CCH), respectively. The $\omega_\alpha$ region above 1600 cm$^{-1}$ has not yet been surveyed experimentally, though may be in the near future. However when the region around 1450 / 2900 cm$^{-1}$ was scanned, no cross peaks were observed.

![Figure 5-17: Theoretical EVV 2DIR spectrum of the planar all-trans, 15-anti retinal protonated retinal Schiff base molecule, which serves as a model for the O$_{640}$ intermediate of the bR$_{568}$ photocycle. The DFT calculations, carried out by R. Guo, were performed at the B3LYP level using the 6-31+G(d,p) basis set. As this spectrum is of model system comprising only the free retinal Schiff base, and with no effects of the protein environment accounted for, direct comparisons with experimental EVV 2DIR measured of bacteriorhodopsin cannot be made.](image-url)
5.6: Sensitivity of electronically resonant EVV 2DIR

In the previous sections of this chapter, the use of electronically-resonant EVV 2DIR to selectively observe structural elements of bacteriorhodopsin’s active site was demonstrated. It was possible to avoid signal from the surrounding protein due to the electronic enhancement of the retinal chromophore’s FWM signal afforded by the visible beam. For example, for an electronic chromophore with a linewidth of $\sim10^{14}$ rad s$^{-1}$ and transition dipole of 4 Debye, an improvement in signal level of $\sim10^3$ over non-electronically enhanced signal is predicted. [98] In this section, the level of enhancement that can be achieved in practice with this triply resonant technique is investigated.

The level of electronically resonant signal observed from bacteriorhodopsin’s retinal chromophore was compared with the non-electronically enhanced signal measured from the CH$_3$ groups of the protein concanavalin A (chosen for its similar size to bR). To achieve this, the absolute quantification procedure outlined in Section 3.5 was used to determine limits of detection for each system. The signal from the $\nu$(C=C) / 2$\nu$(C=C) cross peak (at 1545 / 3050 cm$^{-1}$) of the retinal molecule was mapped across a series of dried bacteriorhodopsin films of varying concentrations, and a calibration graph produced for determining the number of retinal molecules being probed by the lasers from the level of measured signal. The lowest detectable signal was considered to be two standard deviations (2$\sigma$) above the average background level measured from the stainless steel substrate. The number of molecules corresponding to this signal level was then extracted from the calibration curve as the limit of detection for the retinal molecule. The sensitivity of the CH$_3$ group of concanavalin A was determined in the same manner, using the $\delta$(CH) / 2$\delta$(CH) cross peak at 1485 / 2930 cm$^{-1}$. The results from the two systems were then compared.
5.6.1: Experimental methods and procedures

A series of five drop-coated films of concanavalin A were produced on a polished stainless steel substrate. Solutions with concentrations of 0.1, 0.2, 0.3, 0.4 and 0.5 mM were used, and the volume of each deposited droplet was 0.3 µl. This corresponded to approximately $2 \times 10^{13}$, $3.5 \times 10^{13}$, $5.5 \times 10^{13}$, $7 \times 10^{13}$ and $9 \times 10^{13}$ total protein molecules respectively in each of the five deposits. The procedure was repeated for bacteriorhodopsin, using PM solutions with concentrations of 2, 3, 4, 5, and 6 µM, corresponding to approximately $4 \times 10^{11}$, $5.5 \times 10^{11}$, $7 \times 10^{11}$, $9 \times 10^{11}$ and $1 \times 10^{12}$ total protein molecules respectively.

Due to the heterogeneity of the analyte density across the dried films, the EVV 2DIR signal was mapped across the whole of each of the five deposits for both sample systems. In the case of the concanavalin A samples, the doubly resonant EVV 2DIR signal from the CH$_3$ feature at 1485 / 2930 cm$^{-1}$ was measured, whilst for the bacteriorhodopsin the triply resonant signal from the retinal ethylenic cross peak at 1545 / 3050 cm$^{-1}$ was recorded. For both, the measurements were taken in reflection with delays of $T_{12} = 1$ ps and $T_{23} = 1$ ps and with a 100 µm step size. The visible beam energy was 1.5 µJ. Although this was known to photo-bleach the retinal molecule, an acquisition time of only 1 second per pixel was used. Each spot was therefore only irradiated with the visible beam for 1 second before the sample was moved to a fresh spot for the next signal acquisition. Images were acquired for four different sample sets for both concanavalin A and bacteriorhodopsin, and examples of these are shown in Figure 5-18.

Interestingly, the bacteriorhodopsin deposits did not produce films with the ‘coffee-ring’ effect, but rather the majority of the protein material was deposited at the centre of the films. This was possibly due to the lower concentrations of the
deposits, or more probably because purple membrane samples, containing 75% bacteriorhodopsin and 25% lipid by mass, were used. The formation of the membrane on the slide will have prevented the proteins from being carried to the outside of the drop during drying. As the experiments were performed at < 1% humidity, the observed retinal cross peak is attributable to various photo-accumulated and trapped intermediates of bR’s dark-adapted photocycle ($\lambda_{\text{max}}$ of the sample measured at 562 nm).

Figure 5-18: EVV 2DIR images of (a) concanavalin A and (b) bacteriorhodopsin drop-coated deposits. The deposits are each labelled with the total number of molecules they contained. The concanavalin A images were acquired by measuring the CH$_3$ signal level from the 1485 / 2930 cm$^{-1}$ cross peak across each drop. For the bacteriorhodopsin samples, the triply resonant ethylenic retinal signal at 1545 / 3050 cm$^{-1}$ was mapped to produce the images.

5.6.2: Limit of detection determination

The acquired concanavalin A and bacteriorhodopsin images were first corrected for background signal and any photon counting non-linearity. For each film, the integrated intensity of its square-rooted image was then calculated and plotted as a function of its known protein or retinal content. The calibration graphs demonstrating that the square-rooted integrated signal intensity of a deposit is proportional to the total number of molecules it contains – for concanavalin A, the
number of proteins, and for bR, the number of retinal chromophores – are shown in Figure 5-19 (a).

The integrated intensities of the films were then converted into signal level per pixel by simply dividing them by the number of pixels (the interaction area of the beams and thus the selected step size of 100 µm by 100 µm) covered by each film. Similarly, the total number of molecules in each film was also divided by the total number of pixels it covered. Simple plots of \( \text{signal} / \text{pixel} \) as a function of \( \text{number of molecules} / \text{pixel} \) are shown in Figure 5-19 (b) for both concanavalin A and bR’s retinal. These calibration curves can be used to determine the number of molecules being probed by the laser beams from the level of detected FWM signal. To determine the limit of detection for each system, the number of molecules corresponding to a signal level of 27 photons / second (2\( \sigma \) above the average background signal of 15 photons /second) was taken from each of the calibration plots.

5.6.3: Level of signal enhancement achievable

From the concanavalin A measurements, a limit of detection of \( 3.5 \times 10^{10} \) protein molecules per pixel was established, where a pixel corresponds to the area of the beam interaction with the sample. The bacteriorhodopsin measurements gave a sensitivity limit of \( 6.6 \times 10^8 \) retinal molecules per pixel. The limit of detection achievable when probing bR’s chromophore using electronically enhanced EVV 2DIR is therefore \( \sim 50 \) times lower than that of proteins’ when measuring their CH\(_3\) cross peak. Moreover, considering that each concanavalin A molecule contains 139 CH\(_3\) groups, and that each bacteriorhodopsin molecule contains only one retinal chromophore, the CH\(_3\) group and retinal detection limits in terms of number of chemical moieties are \( 4.9 \times 10^{12} \) and \( 6.6 \times 10^8 \) respectively. Thus, by using triply
resonant EVV 2DIR, it is possible to achieve ~ $10^4$ times higher signal from retinal’s conjugated backbone than it is from a protein CH$_3$ group (summarised in Table 5-3).

**Figure 5-19:** (a) Square root of the EVV 2DIR signal level as a function of the number of protein molecules that were probed. The integrated EVV 2DIR intensity of the square-rooted image for each protein film is plotted against the total number of molecules present in each deposited droplet. For concanavalin A this corresponds to the number of protein molecules, and for bacteriorhodopsin to the number of retinal molecules. The error bars are standard deviations from repeats performed on four different sets of five samples, and the solid lines represent the linear fit. (b) The signal level per pixel of the films plotted as a function of the number of molecules per pixel for both concanavalin A and bR’s retinal. These provide effective calibration curves for determining the number of molecules corresponding to a particular level of FWM signal. The limit of detection for each sample is taken from these curves as the number of molecules corresponding to two standard deviations above the average background signal level (from the stainless steel substrate). The error bars are standard deviations from repeats performed on four different sets of five samples. The solid lines represent the fits, corresponding to a 2$^{nd}$ order polynomial function as expected from the 2DIR FWM theory.
Chapter 5: Electronically resonant EVV 2DIR studies of bacteriorhodopsin

<table>
<thead>
<tr>
<th>Bacteriorhodopsin molecules</th>
<th>Retinal (C≡C) backbone</th>
<th>Concanavalin A molecules</th>
<th>Concanavalin A CH₃ groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit of detection</td>
<td>6.6 × 10⁶</td>
<td>6.6 × 10⁸</td>
<td>3.5 × 10¹⁰</td>
</tr>
</tbody>
</table>

Table 5-3: Summary of sensitivity achievable with triply resonant EVV 2DIR, compared with that of the non-electronically enhanced case. The limit of detection of bR’s retinal molecule is compared with that of the CH₃ group of concanavalin A.

Significant improvement in EVV 2DIR sensitivity can therefore be achieved by making the electronically-polarising visible ‘read-out’ beam resonant with the molecular system being studied. It should be noted that the fixed wavelength 790 nm visible beam used in our experiment is pre-resonant in the case of bacteriorhodopsin’s retinal. Further enhancement could be achieved in this case by tuning the beam closer to resonance at 570 nm (the absorption maximum of the retinal).
Chapter 6: Conclusions and future work

The work presented in this thesis covers the initial steps achieved towards the realisation of EVV 2DIR spectroscopy as a practical analysis tool for addressing a range of biological problems. Specifically two bioanalytical applications were demonstrated: i) the use of EVV 2DIR for label-free, high-sensitivity proteomics, in particular high-throughput protein identification and absolute quantification, and ii) the potential of triply resonant EVV 2DIR for studying protein structures and mechanisms. The following two sections of this chapter summarise the achievements to date, while in the final section the outlook and future direction for this research is assessed.

6.1: Protein identification and absolute quantification using EVV 2DIR

In Chapters 3 and 4, proof of principle of the EVV 2DIR protein fingerprinting strategy – based on the identification of proteins through their spectroscopically determined amino acid compositions – was demonstrated. To this end, 2DIR cross peaks of three aromatic amino acid side chains (tyrosine, phenylalanine and tryptophan) were identified and their vibrational modes assigned. Importantly, spectral signatures arising from peptide and protein CH\textsubscript{2} and CH\textsubscript{3} groups were
found to be appropriate for use as internal references. Furthermore, the identified cross peaks were found to be insensitive to structural influences and thus suitable for the purposes of fingerprinting. Using peptides for initial studies, it was demonstrated that the square root of the intensity of an amino acid cross peak relative to that of a reference CH$_2$ feature is proportional to the amino acid / CH$_2$ ratio of peptide samples. This was then also applied to a set of ten known proteins, for which a CH$_3$ feature was instead used as the internal reference, and it was shown that the spectroscopically measured data of these proteins’ amino acid / CH$_3$ ratios could be successfully used to differentiate them.

The feasibility of this form of amino acid composition analysis for protein identification was then also addressed. Specialised databases comprising the amino acid / CH$_2$ and amino acid / CH$_3$ ratios of proteins were developed, and a method for searching these with 2DIR data implemented. Preliminary tests using the Ensembl database of ~33,000 human proteins showed that the amino acid / CH$_x$ ratios of less than half of the twenty different amino acids need to be known to uniquely identify a protein. In fact as few as six to nine ratios are required, meaning that the signal from only seven to ten 2DIR cross peaks needs to be measured to perform a successful identification.

Significantly, the potential of EVV 2DIR for performing simple absolute quantification of protein levels was demonstrated. This is a valuable property of this spectroscopy since it is an increasingly important issue in proteomics, but has proved difficult to achieve with other established technologies such as mass spectrometry. Quantification with EVV 2DIR was however shown to be a relatively uncomplicated procedure. By measuring the level of FWM signal from a particular cross peak, for example the CH$_3$ feature, it is possible to use calibration graphs to convert this to the number of protein molecules being probed by the laser beams.
EVV 2DIR protein analysis qualifies as a top-down proteomic tool. The research presented in Chapter 3 established how the technique is label-free, without any complicated preparation steps of the protein analytes required. The separated samples simply need to be presented to the spectrometer in the form of dry films. The method is non-destructive and protein samples can be kept and repeatedly re-examined; samples more than two years old have been shown to produce the same results as when freshly produced. Importantly, the high-throughput capability of the technique was demonstrated. The collection of full 2DIR spectra was shown to be unnecessary for measuring the amino acid / CH₃ cross peaks ratios. A simple signal reading on resonance for each amino acid cross peak was established as perfectly adequate for performing an identification. Therefore the spectrometer need only be programmed to take signal level readings from specific sets of frequencies. With the acquisition time for each peak likely to be less than 5 seconds, the data required for a protein identification could be acquired in less than 50 seconds. Furthermore, the current limit of detection was determined to be ~ 10¹⁰ protein molecules, making the sensitivity of the technique sub-picomolar. This is set to improve significantly in the near future with the development of a second EVV 2DIR spectrometer (Section 6.3). The potential speed and sensitivity at which protein identification and quantification will be achieved makes EVV 2DIR a viable tool for high-throughput protein analysis.

6.2: Electronically resonant EVV 2DIR spectroscopy of bacteriorhodopsin

In Chapter 5, the first example of a triply resonant EVV 2DIR experiment was presented. Electronically enhanced EVV 2DIR was used to selectively study the retinal chromophore located at the active site of the photosynthetic protein
bacteriorhodopsin. Cross peaks representing the couplings between vibrational modes of the retinal molecule were enhanced by the additional probing of an electronic transition by the visible beam, thus making their signal significantly more intense than that from the normally dominant protein CH$_x$ features. This made it possible to pick out the retinal signal, whilst avoiding that from the surrounding protein. In fact, by comparing the signals achievable from bR’s retinal and from proteins’ CH$_3$ groups, the level of enhancement possible with triply resonant EVV 2DIR was determined to be $\sim 10^4$ times over the non-electronically enhanced case. The current limit of detection for CH$_3$ was established to be $10^{12}$ groups, whilst for the retinal was found to be $10^8$ molecules. This area of research highlighted the potential of this technique for selectively homing-in on protein active sites that contain groups with appropriate electronic transitions, for example chromophores or metal ions, whilst avoiding their surrounding proteins.

Two different two-dimensional spectral areas were investigated for bacteriorhodopsin’s retinal – the fingerprint and ethylenic regions. In the fingerprint region, cross peaks attributed to the coupling of two vibrational modes of the retinal’s polyene backbone, $\delta$(CCH) / $\delta$(CCH) + $\nu$(C=C), were observed at 1225 / 2745 cm$^{-1}$ and 1270 / 2790 cm$^{-1}$. The cross peaks measured in the ethylenic region around 1555 / 3065 cm$^{-1}$ were assigned to the C=C double bond stretch, $\nu$(C=C), and its first overtone, 2$\nu$(C=C).

The electronically enhanced EVV 2DIR technique was also used to study different retinal intermediates of bacteriorhodopsin’s photocycle. Through the use of humidity control, the observation of various photo-accumulated intermediates was attempted, with the M$_{412}$ intermediate being successfully observed. Additionally, theoretical Raman, IR and EVV 2DIR spectra calculated by I. R. Gould and R. Guo were shown to compare favourably with the experimental spectra.
6.3: Outlook for EVV 2DIR in bioanalysis

The work presented in this thesis only represents a couple of the several potential biological and biomedical applications of EVV 2DIR. As for any bioanalytical technique, EVV 2DIR may be expected to excel for use in some areas, whilst be of limited use in others. The aim is for it to offer complementary methods to those already available. Firstly in this section therefore, the steps that need to be taken to make the approaches covered in this thesis realisable will be summarised. Following this, some of the other possible functions of EVV 2DIR in bioanalysis will be illustrated. These particularly include three areas which are already being developed within the Klug research group; the measurement of protein post-translational modifications, the detection of protein-protein interactions and also EVV 2DIR imaging.

6.3.1: EVV 2DIR for protein analysis

Experimental developments

There are two main areas that need to be addressed for EVV 2DIR to become a useful proteomic tool for protein identification and absolute quantification. Firstly, cross peaks of more amino acid side chains need to identified, as currently those of only three different residues can be used and this is not sufficient to perform protein identifications. Secondly, an improvement in the sensitivity of the technique must be achieved, with the long-term aim being single cell sensitivity.

With regards to the locating of additional amino acid cross peaks, the theoretical spectra of the residues’ side chains (Figure 3-5) will be used for guidance, in particular to identify those cross peaks which are isolated from others. The spectral
area that can now be potentially accessed is larger than that shown in Figure 3-5, with the possibility of reaching as low as 800 cm$^{-1}$ and stretching to 7000 cm$^{-1}$ in both $\omega_a$ and $\omega_b$. With the current experimental set-up, only small spectral regions of about 500 cm$^{-1}$ by 1000 cm$^{-1}$ can be covered at a time, and moving between each is not a trivial task. This is however expected to become much simpler with the improved EVV 2DIR spectrometer that is currently under development. With this, changing between regions will be more automated and thus hopefully easier and faster. Using peptide and protein known samples, the small spectral regions will therefore be systematically surveyed in the hunt for amino acid cross peaks – at least three more are required to make protein identification viable. Once identified, amino acid cross peaks will be assessed for their suitability as fingerprint features. Cross peaks that are unaffected by primary sequence and higher order structural elements are required. Although the CH$_2$ and CH$_3$ features are currently being used as internal references, if other suitable cross peaks are found these too could be used. There is no reason why, for example, a feature associated with a particular amino acid side chain could not be used; in such a case the database of proteins’ amino acid ratios would just need to be altered accordingly.

The limit of detection for proteins, determined using the CH$_3$ cross peak, is currently at the sub-picomole level (~ $10^{10}$ protein molecules). For the validation experiments of the tyrosine, tryptophan and phenylalanine cross peaks covered in Sections 3.3 and 3.4, however, typically $10^{12}$ protein molecules were in the lasers’ interaction area. These sensitivity limits were established using the group’s original experimental set-up, and many improvements are being implemented in the second EVV 2DIR spectrometer currently under construction and validation. These will include, amongst others, changing to a collinear set-up, higher OPA pulse energies, tighter IR beam focusing, an improved detector and the use of a multi-pass set-up, and are covered in detail in the PhD thesis of P. M. Donaldson. [98] Additionally,
Chapter 6: Conclusions and future work

the possibility of visible resonance enhancements have been demonstrated with bacteriorhodopsin and with a tuneable visible beam may also be achievable from, for example, the aromatic amino acid side chains. With these advancements it is expected that in the near future the technique will operate with at least femtomole sensitivity (~ $10^8$ protein molecules).

Protein databases and search tools

Work is to continue on development of the specialised protein databases and search tools for performing protein identifications using EVV 2DIR amino acid composition data. Currently these databases have been constructed for only a few model organisms, but will be expanded to cover more in the future. In addition, the implementation of a ranking system for protein results is presently underway. As explained in Chapter 4, the current method of searching the databases using amino acid ratios within a certain precision level can lead to the protein of interest being excluded from the list of possible results. Nor does this method indicate which protein from the list is the closest match to the data of the unknown protein.

The ranking system being developed is based on that of ExPASy’s AACompIdent tool, for which all proteins in the searched database are given a score and then ranking based on the level of similarity of their amino acid compositions with the input data of the unknown protein. This method has already been used to check the EVV 2DIR amino acid ratios data of the ten test proteins, covered in Section 4.5.1. In addition, the ranking system is currently being tested with simulated data for the yeast ribosomal proteins, another area of interest within the research group. [226] For each of the 131 protein sequences, EVV 2DIR ratios are being simulated for the three currently measureable amino acids to be within $\pm 20\%$ of their actual values. Each protein’s simulated data will in turn be used to search the sub-database of 131
proteins, with each database protein being scored on the similarity of its ratios to those input of the test protein. The ranking of each protein when the database is searched using its simulated data will be established. The number of proteins that are ranked first when their simulated ratios were used to perform the search will give an indication of how viable it will be to attempt to individually identify the yeast ribosomal proteins by measuring only three amino acid / CH$_2$ ratios.

It is likely that this ranking system will prove adequate, since it has been successfully used for years by researchers using standard amino acid composition analysis for protein identification. Nonetheless, other scoring and ranking methods for protein hits will also be investigated to assess how the procedure can be further improved. For example, ways of indicating the probability that a protein represents the best match to the unknown protein, such as those used for the analysis of mass spectrometry data, will be considered. Another method for increasing the confidence of an identification is the use of calibration proteins. If the amino acid composition of a control protein is measured prior to the collection of data of the unknown protein, it can be submitted with a search to compensate for any errors in the analysis procedure.

**Contamination issues**

One of the issues that will affect the confidence of protein identifications with EVV 2DIR is the level of contamination of the samples. To date all experiments have been performed on off-the-shelf, purified proteins. In reality, the separated protein samples presented to the EVV 2DIR spectrometer for identification will have some degree of contamination from proteins and other biomolecules. To assess just how much this will affect the identification procedure, tests will initially be performed bioinformatically. Protein sequences will be taken from the database and their amino
acid compositions altered by changing, for example, 10% of their composition to the average amino acid composition of the entire database. The ‘average number of hits’ tests, discussed in Chapter 4, will then be performed to ascertain how much of an effect 10% contamination will have on the results. This will also be carried out with other levels of alteration to the amino acid compositions (5%, 15% and 20%) to establish what degree of contamination of the sample can be tolerated by the EVV 2DIR protein fingerprinting method. On a smaller scale, contamination tests will also be performed experimentally by simply adding specific amounts of other proteins and biomolecules to a protein sample before submitting it for 2DIR identification.

6.3.2: EVV 2DIR as part of a high-throughput proteomic tool

Ultimately, the intention is for the EVV 2DIR apparatus to form part of a platform of technologies for label-free, high-throughput proteomics at the single cell level. Some sections of the Klug research group are concentrating on the development of the ‘front-end’ methods for cell sorting, trapping and lysis, whereas others are working on the separation techniques to be directly coupled with the 2DIR spectrometer. The aim is to perform two-dimensional separations, most probably high pressure liquid chromatography followed by capillary electrophoresis, of cell lysate prior to 2DIR analysis. Using an electrodeposition method developed by C. B. Loeffeld [85] the separated proteins eluted from the CE instrument will be written onto stainless steel plates, ready to be submitted for protein identification. These protein samples are not destroyed by the 2DIR procedure and can be reanalysed over and over again. There are currently a number of people working on the continued development of the CE ‘track-writing’ procedure, and in the near future it is hoped that the first protein track will be submitted for EVV 2DIR identification.
In particular, one researcher is currently optimising separations of yeast ribosomal proteins, ready for track-writing and then EVV 2DIR read-out. [226]

A significant advantage of the CE separation process is the potential for mass determination of the proteins. Proof of principle of this method is currently being verified by C. B. Loeffeld [135], and this extra information about a protein will aid in the identification process. Studies using standard amino acid composition analysis procedures have shown that a protein is more likely to be correctly identified with the AACompIdent tool if its molecular weight and / or isoelectric point are also used to perform the database search. [21, 22] Furthermore, the database tests covered in Chapter 4 showed the number of hits returned from a search was reduced by a factor of ten when mass was used along with the amino acid ratios as a search parameter.

**6.3.3: Future bacteriorhodopsin studies**

There are a number of ways in which the bacteriorhodopsin research covered in Chapter 5 may be continued, though it is unlikely that further work will be carried out on this system in the near future within the Klug research group. bR provided a good starting point for investigating the potential of the triply resonant EVV 2DIR technique, and the gained knowledge will be channelled into the proteomic area of research. For example, the signal from the aromatic side chains could possibly be electronically enhanced if the visible beam is tuned nearer to 280 nm. Alternatively the triply resonant technique could be used to investigate the active sites of other proteins, for example the binding of NADH to liver alcohol dehydrogenase. In both cases, however, the problems of photo-damage would have to be overcome. If studies of bacteriorhodopsin with ER EVV 2DIR are to be continued, the areas that would have to be addressed are summarised below.
Chapter 6: Conclusions and future work

Clearly the method of humidity control to study accumulated photocycle intermediates, reported in Section 5.4, was not reliable enough. In any case, this approach only had the potential for observing two of the intermediates, and for only showing significant spectral shifts between them. If work is to continue with studying the retinal intermediates, then clearly a more viable method will have to be developed. This was difficult to do when the results presented here were collected; there was only one experimental set-up, with priority for fingerprinting studies, to which major alterations could not be made. With the development of the second EVV 2DIR apparatus, there is the possibility of designing parts with bR studies in mind. For example, with four OPAs, there is the potential to carry out pump-probe experiments, i.e. initiating the photocycle with one pulse and then measuring the EVV 2DIR signal after a specific length of time. Alternatively, the cryo-trapping of particular intermediates would have to be considered. Another possibility offered by the extra OPAs, is a tuneable visible beam. The level of retinal signal enhancement provided by the visible beam has the potential to be further increased by tuning the beam closer to resonance; currently it is fixed at 790 nm, though the retinal absorbs most strongly at 570 nm. Finally, it would interesting to survey the $\omega_{\alpha}$ region above 1600 cm$^{-1}$ for the $\nu$(C=N) / $\nu$(C=N) + $\delta$(CCH) cross peak observed at 1650 / 2950 cm$^{-1}$ in the theoretical spectrum of Figure 5-17.

With regards to the continuation of the theoretical studies of bR’s retinal, it would indeed be useful to be able to make meaningful comparisons between experimental and simulated EVV 2DIR retinal spectra. In Chapter 5, it was demonstrated that theoretical spectra of the free retinal molecule are of limited use. That the calculated 2DIR spectrum of Figure 5-17 shows the expected ethylenic cross peaks to be present is promising and serves as a useful starting point, but ideally spectra of a system comprising the retinal and its protein environment are required. On-going studies of Raman and IR spectra of an embedded retinal molecule model are
seeming to offer a significant improvement over those of the free molecule, though there are still problems. This suggests that for future calculations a more substantial protein environment comprising more residues would need to be provided. This unfortunately makes the system larger and the calculations even longer to complete; a decision would need to be made as to whether this is necessary and also as to whether this is an area of research that should continue to be pursued. The calculation of EVV 2DIR spectra of the full embedded molecule would be a huge undertaking.

6.3.4: Measurement of post-translational modifications

One of the most significant areas of proteomics is the study of proteins that have undergone post-translational modifications (PTMs), the most common of which is phosphorylation. The reversible phosphorylation of proteins plays a major role in a wide range of complex cellular processes, including protein regulation and signalling, and cell division, growth and apoptosis. In fact many human diseases, particularly cancer, are caused by problems with phosphorylation-dependent signalling systems. Being able to measure protein phosphorylation, which most commonly occurs at serine, threonine and tyrosine residues, is therefore of great importance.

It has recently been demonstrated that tyrosine phosphorylation can be detected with EVV 2DIR. Two near-identical tyrosine-containing peptides were examined, the only difference between them being that one had its tyrosine residues phosphorylated. The EVV 2DIR spectra of these two peptides, LRRYYLYG and LRRpYpYL, can be seen in Figure 6-1. The spectrum of LRRYYLYG has the usual features, including the CH₂, CH₃ and two tyrosine aromatic cross peaks (covered in Chapter 3). The spectrum of the phosphorylated peptide is however missing one of
the tyrosine features. The cross peak that remains whether the tyrosine is phosphorylated or not (Tyr 1), albeit red shifted by ~12 cm\(^{-1}\) in the \(\omega_\beta\) direction and ~6 cm\(^{-1}\) in the \(\omega_\alpha\) direction, is attributed to only ring modes: \(\nu_{13} / \nu_{13} + \nu_{16}\). The cross peak that however disappears upon phosphorylation (Tyr 2) also has contributions from CH bending modes: \((\delta(CH_2), \nu_{13}) / (\delta(CH_2), \nu_{13}) + (\delta(CH_2), \nu_{16})\).

Figure 6-1: A comparison of the EVV 2DIR spectra of the peptide LRRYYLG with its phosphorylated version, LRRpYpYLG. The spectra were recorded with delays of \(T_{12} = 1\) ps and \(T_{23} = 1\) ps and with all three beams polarised in the plane of propagation. The spectra are plotted with the same intensity scale.

These preliminary results suggest that the intensity of the Tyr 2 cross peak is inversely proportional to the amount of tyrosines that are phosphorylated. To confirm this, tests will be performed using peptides of varying tyrosine-phosphorylation levels. A number of peptides based on the LRRYYYYLG sequence have already been synthesised for this task. It should be noted that, as phosphorylation does not influence the intensity of the Tyr 1 cross peak, the
fingerprinting procedure is not affected. Whilst being to measure tyrosine phosphorylation levels indirectly using one of the aromatic cross peaks is promising, it would be better if they could be measured directly. (Raman studies have also demonstrated changes in the structure of tyrosine peaks upon phosphorylation, particularly the aromatic deformations around 835 cm$^{-1}$. [44, 45]) Therefore the spectral region where the various phosphorous bond vibrations are found is to be surveyed for suitable cross peaks. For example, the P–OH stretches are found in the 2000 – 2800 cm$^{-1}$ region, and the C–OP and P=O stretches at 1100 – 1400 cm$^{-1}$.

### 6.3.5: Imaging with EVV 2DIR

The optical imaging of cells and tissues is an important area in the life-sciences. Dyes are often used to provide contrast, but these can alter the natural state of the molecules in cells. Imaging methods that can achieve chemical contrast without the need for labels are particularly attractive. A range of chemically sensitive, label-free approaches have therefore been developed including IR, Raman and CARS imaging. EVV 2DIR spectroscopy has the potential to provide enhanced contrast in imaging of tissue samples by decongesting the spectra to a much greater extent than is possible with those methods based on techniques which use one spectral dimension. Imaging with EVV 2DIR could therefore provide alternative information on the spatial distribution of particular chemical species.

The feasibility of imaging using EVV 2DIR is currently being evaluated in the group by F. Fournier. Preliminary tests have been performed on mouse kidney histological sections, stained with haematoxylin. Images were obtained by setting the infrared frequencies to a specific 2DIR cross peak of the haematoxylin molecule, and then spatially mapping the tissue by scanning the sample relative to the beam position. The procedure was then repeated, this time mapping the intensity of the
protein CH$_3$ cross peak. The obtained images can be seen in Figure 6-12. The images mapped using the different cross peaks show differential contrast and thus the different distribution of proteins and haematoxylin in the tissue. These distributions can be explained by the different tissue structures and their functions. [93]

![Figure 6-2: EVV 2DIR and visible microscope images of a histological mouse kidney section. To obtain the images, the IR beams were set to the frequencies of interest and then the signal mapped across the sample. Diagram reproduced with kind permission of F. Fournier. [93]](image)

### 6.3.6: Detection of protein-protein interactions

Another area of EVV 2DIR research within the group regards protein-protein and protein-ligand interactions. R. Guo has recently demonstrated that EVV 2DIR can be used to detect molecular complex formation and also to determine the geometry of the molecules that have formed the complex. [93, 94] Electrical interactions between molecular vibrations can be non-linear and thereby produce coupling even in the absence of a chemical bond. This fact was used to detect the formation of an intermolecular complex between phenylacetylene (PA) and benzonitrile (BN), and to determine the distance and angle between the two molecules. The calculated cross
peaks, formed by dipolar couplings, appeared with the predicted strength and at the predicted locations in the experimental spectrum.

This work is now being applied to a number of protein systems, including the interactions between various proteins and their inhibitors, as well as the therapeutically important interaction between the tumour suppressor p53 and its natural antagonist MDM2. Predictions of the interaction peaks for these systems have been made by R. Guo. M. Miele is carrying out the experimental side of the research and is currently attempting to observe the calculated cross peaks. This work highlights the future potential for interactome analysis with EVV 2DIR.
18. SEQUEST. Available from: http://fields.scripps.edu/sequest/.


86. *Bioinformatics Support Service*, Imperial College London.


References


113. Anaspec, San Jose, CA, USA.

114. Tate & Leatherbarrow Research Group, Department of Chemistry, Imperial College London.


116. EMBL Nucleotide Sequence Database. Available from: [www.ebi.ac.uk/embl](http://www.ebi.ac.uk/embl).

131. *IPI - International Protein Index*. Available from: www.ebi.ac.uk/IPI.


226. Goyder, M.S., A new top-down proteomic approach to study the eukaryotic ribosome. 2008, Department of Chemistry, Imperial College London.
Optical fingerprinting of peptides using two-dimensional infrared spectroscopy: Proof of principle

Frédéric Fournier a,⁎, Elizabeth M. Gardner a, Rui Guo a, Paul M. Donaldson a, Laura M.C. Barter a, D. Jason Palmer b, Chris J. Barnett a, Keith R. Willison c, Ian R. Gould a, David R. Klug a

a Molecular Dynamics Research Group, Department of Chemistry, Imperial College London, London SW7 2AZ, UK
b European Laboratory for Nonlinear Spectroscopy, University of Florence, I-50019 Sesto Fiorentino (Florence), Italy
c Institute of Cancer Research, Chester Beatty Laboratories, Cancer Research UK, Center of Cellular and Molecular Biology, London SW3 6JB, UK

Received 27 September 2007
Available online 13 November 2007

Abstract

We employ a particular form of two-dimensional infrared four-wave mixing (2DIR FWM) as a vibrational spectroscopic tool to quantify the amino acid content of a number of peptides. Vibrational features corresponding to ring modes of the aromatic groups of phenylalanine (Phe) and tyrosine (Tyr), as well as a methylene mode that is used as an internal reference, are identified. We show that the ratios of the integrated intensities, and the amplitudes, of the aromatic peaks of Phe and Tyr relative to the methylene integrated intensity, and amplitude, are proportional to the actual ratio of Phe and Tyr to CH2 in the samples within a precision of ±12.5%. This precision is shown to be sufficient to use this form of 2DIR spectroscopy as a possible proteins fingerprinting tool.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Two-dimensional infrared spectroscopy; 2DIR; Proteomics; Peptides; Phenylalanine; Tyrosine; Protein fingerprinting; Amino acid

We propose an alternative strategy for protein fingerprinting based on the identification of proteins' amino acid content using a two-dimensional infrared (2DIR) spectroscopy. The number of amino acids that one needs to quantify depends on the proteins to be identified and on the precision of the measurements. Converting such an analytical strategy into a practicable proteomic method requires the attainment of particular levels of sensitivity, resolving power, and high throughput.

Optical techniques have been used for protein and peptide analysis, but overall they have been used with less widespread application than have the classical nuclear magnetic resonance (NMR) and mass spectrometry (MS) techniques. Raman vibrational spectroscopy and IR absorption spectroscopy can give secondary structure information and, when used in difference spectroscopy mode, can identify individual bonds within a protein [1-3]. Vibrational spectra are sensitive to chemical environment, secondary structure, and conformation [4,5]. The drop coating deposition Raman (DCDR) technique has also been used recently in the study of proteins and peptides [6,7] and has been applied to tyrosine (Tyr) phosphorylation [8] and protein mixtures [9].
Multidimensional nonlinear optical methods in recent years have become a practical and realizable approach in optical spectroscopy [10]. Coherent multidimensional vibrational spectroscopy appears to power the decongestion of vibrational spectra and to allow the fingerprinting of proteins without the need for significant sample manipulation. The most widely used 2D vibrational spectroscopy is based on a three-pulse photon echo approach [11,12]. Examples of its application to proteins and peptides include the structure determination of peptides by studying the amide I band that is associated with the amino acid backbone [13–15], the study of vibrational dynamics [16,17], and the study of hydrogen bond dynamics [18]. This technique has also been used to probe the coupling of DNA vibrational modes and secondary structure [19], exciton dynamics [20], and hydrogen bonds dynamics in water [21].

Linear optical spectroscopies have, in principle, both the sensitivity and potential throughput to be used as a proteomic tool, but their resolving power generally is insufficient to unambiguously quantify enough amino acids in a protein. The choice of 2DIR spectroscopy is motivated by the power that this technique has to spread the spectral information over two or more dimensions in a manner highly analogous to that of 2DNMR. Successful fingerprinting of peptides and proteins by 2DIR methods requires accessing a wide range of disparate coupled vibrational states. To access these, we have adopted and adapted a particular version of 2DIR developed by Wright and coworkers [22–27]. In this approach, the frequencies of the excitation pulses are scanned independently across the vibrational spectrum. The coupling information is “read out” by a visible laser beam that scatters from the resulting polarization. The advantage of this particular variant of 2DIR is that it allows detection of photons in the visible optical spectrum. The ability to restrict the measurements to narrow band, or even a single point, of the 2DIR spectrum greatly improves the overall throughput. The method also allows the spectrum to be decongested even further than is possible with other 2DIR methods [28].

In this article, we apply our 2DIR approach to peptides and demonstrate its utility in the detection of aromatic amino acid side chains. We demonstrate that the first stages of this fingerprinting strategy (i.e., amino acid identification and quantification) are workable and show the practicality of using an internal reference to achieve relative quantification of the amino acid level. Most important, we demonstrate that this 2DIR technique potentially has the required sensitivity and high throughput to be a viable proteomic tool; neither sensitivity to primary structure nor sample preparation and presentation is an insurmountable barrier.

We present results from eight peptide sequences having variable levels of phenylalanine (Phe) and tyrosine (Tyr), and give an indication of the accuracy and precision with which this information can be recovered using the current approach.

Fig. 1. Three-level energy diagrams for FWM DOVE-IR and DOVE-Raman processes for two different IR pulse orderings. Depending on the time order of the IR pulses, one can select gg → ga → ba → ea → aa as the DOVE-IR pathway for e, or gg → bg → ba → ea → aa for D, arriving first (B). In the case shown in diagram B, another coherence pathway exists: gg → bg → ab → eg → gg, corresponding to the DOVE-Raman process. All pathways exist and interfere when the pulses are overlapped in time. Diagram A corresponds to the pulse ordering used in the current work.

Materials and methods

Principle of 2DIR four-wave mixing spectroscopy

A complete description of this spectroscopic method can be found in the various publications from Wright’s group [22–27] and in some specific comments on its utility for protein fingerprinting in a previous article of ours [28]. Briefly, 2DIR four-wave mixing (FWM) is a nonlinear coherent optical technique giving 2D vibrational spectra. FWM is a third-order process where three laser beams interact with the medium through the third-order susceptibility tensor. By using two IR beams (ωa and ωb) and a third visible beam (ωe), this technique becomes a powerful vibrational spectroscopic tool. When the IR beams’ wavelengths are scanned over vibrational resonances, the FWM signal shows enhancement for the IR frequencies that are in resonance with vibrational transitions.

The benefit of this technique is to show multiplicatively enhanced cross-peaks when the IR beams are in resonance with coupled vibrational modes. This multiplicative enhancement comes from the fact that the hyperpolarizability describing the molecular response at the third order [25] contains mechanical and electric anharmonicity contributions [27]. These terms are nonzero only if the vibrational modes excited by ωa and ωb are coupled by intramonomer interactions or if they are composed of a fundamental mode and its overtone or a fundamental mode and a combination band. Fig. 1 shows the energy diagrams and coherence pathways involved for both processes giving multiplicative enhancements: DOVE–IR and DOVE–Raman.

Sample preparation

Automated synthesis of the peptides was performed using an Advanced ChemTech Apex 396 multiple peptide synthesizer. The syntheses followed a standard FMoc/tBu (N-(9-fluorenyl)methoxycarbonyl/tert-butyl) peptide
synthesis strategy [29]. Peptides were purified to more than 98% homogeneity using reverse-phase high-performance liquid chromatography (HPLC) and then characterized by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS.

To avoid solvent/water contributions, the peptides were prepared in the form of dried films cast on glass slides. To prepare each film, a peptide solution of 50 mM concentration was made using nonbuffered ultrapure water. A 2-μl drop of the solution was deposited onto a standard, clear glass microscope slide 1.0 to 1.2 mm thick. The drop was left to dry in air at room temperature for approximately 30 min until it had formed a film. A second drop of 2 μl of the solution was then deposited on top of the film and also left to dry. The peptide films were further dried when put into the sample compartment of the experiment that was purged with nitrogen (<0.1% humidity). In several cases, the drops show a ring similar to the rings present in the DCDR technique [30]. We estimated an average of 10¹³ molecules present in the laser spots (typically 100 μm diameter).

**Optical setup**

A full description of the optical setup can be found elsewhere [28]. Briefly, to acquire 2DIR spectra, two IR beams (ω₄ and ω₅) and a visible beam (ω₀) are noncollinearly overlapped in time and space on the sample (Fig. 2B). The FWM signal at ω₆ = ω₄ + ω₅ − ω₀ is detected in the visible part of the electromagnetic spectrum with a photomultiplier, while the IR beam wavelengths are scanned over the spectral region of interest. For the experiment presented in this article, ω₀ is scanned from 1350 to 1570 cm⁻¹ and ω₅ is scanned from 2700 to 3200 cm⁻¹, both with 5 cm⁻¹ increments. The data obtained are 2D maps of the FWM signal level as a function of ω₀ and ω₅. Because atmospheric water absorption bands are present in the first spectral region, the optical path of the IR beam at ω₅ is purged with nitrogen to avoid any spectral absorption that would cause pulse distortions and then artifacts in the 2D spectra. The visible beam wavelength is 790 nm (12,658 cm⁻¹). The FWM signal is then detected in the spectral range from 13,788 cm⁻¹ (725 nm) to 14,508 cm⁻¹ (689 nm).

The time delays of the ω₅ IR pulse and visible pulse relative to the ω₀ IR pulse are controlled by two delay stages (the pulse ordering is shown in Fig. 2A). The following delays were found to give a satisfying level of FWM signal for the systems studied in this article without nonresonant background or other processes. τ₁₂ = 1.5 ps and τ₂₃ = 1 ps. We found that in this configuration the spectral features are not significantly distorted by interference effects and, thus, are easier to exploit.

**Quantification procedure**

Once specific features corresponding to certain amino acids have been identified, the aim is to be able to establish the quantity of a particular amino acid in a sequence using the variations of its corresponding peak intensity from one 2D spectrum to another. Two different ways of measuring are used. First, to demonstrate the high-throughput potential, we use a simple peak amplitude measurement corrected for any experimental offset. Second, as a reference method (interference free and with potential broadening effects taken into account), we calculate the integrated intensity for each feature of interest. To perform the latter type of measurement, the spectra are fitted with 2D Gaussian functions (Eqs. (1a) and (1b)).

\[
S_{\text{FWM,CH₃}} = \text{Background} + A_\text{peaks} \left( \frac{(2\pi\sigma_\text{FWHM})^{-2}}{(2\pi)^{1/2}} \right)^{1/2} \left( \frac{\left( \frac{2\pi\sigma_\text{FWHM}}{e} \right)^{1/2}}{(2\pi)^{1/2}} \right)^{1/2}.
\]

\[
S_{\text{FWM,n}} = \text{Background} + \sum A_\text{peaks} \left( \frac{(2\pi\sigma_\text{FWHM})^{-2}}{(2\pi)^{1/2}} \right)^{1/2} \left( \frac{\left( \frac{2\pi\sigma_\text{FWHM}}{e} \right)^{1/2}}{(2\pi)^{1/2}} \right)^{1/2}.
\]

Experimental line shape studies and calculations using time-dependent perturbation theory have shown that when the delays τ₁₂ and τ₂₃ increase, the peak line shape changes
from Lorentzian (for zero delays) to roughly Gaussian. The spectra used in this work are taken at $t_{12} = 1.5$ ps and $t_{23} = 1$ ps, delays for which the Gaussian line shape dominates [28].

$\omega_a$ and $\omega_b$ are the laser frequencies and are the variables of the fit. $\omega_{am}$ and $\omega_{pm}$ are the frequencies of the vibrational modes seen on the spectrum and are used as parameters of the fit. The subscript $n$ represents the peak, and the Greek letter represents either the $\omega_a$ direction or the $\omega_b$ direction. The parameter $A$ is the amplitude of the Gaussian function, $\varphi_n$ is the phase of the peak $n$, and $\Gamma$ is the full-width at half-maximum (FWHM) for each peak in both frequency coordinates. These are also parameters of the fit.

To speed up the fitting process, the spectra are cut into three regions of interest (Fig. 3): the aliphatic peak region, the low-frequency aromatic region, and the high-frequency aromatic region. Eq. (1a) is used to fit the aliphatic peak only; this mathematical form was chosen empirically and represents the diagonal behavior of this particular peak very well. For the two other spectral regions, we used Gaussian functions with relative phases (Eq. (1b)). This allows the possible interferences between peaks to be taken into account and allows interference-free amplitudes and widths of each peak to be extracted. We also included a background in the equation used for the fit to correct from any experimental offset. Fig. 3 shows the fit performed for the YQGF sample superimposed on the experimental spectrum.

Peak amplitudes and FWHM are extracted from the fits, allowing calculation of the integral of each peak independently from the others. Taking the modulus square of a 2D Gaussian peak having amplitude $A$ and FWHMs $\Gamma_a$ and $\Gamma_b$, the integral from $-\infty$ to $+\infty$ is easily calculated as

$$I_{\text{peak}} = \frac{A^2 \pi \Gamma_{\text{peak}} \Gamma_{mn}}{8 \ln 2}.$$  \hspace{1cm} (2a)

For the aliphatic peak, taking into account the mathematical form used for the fit, the integral is calculated as

$$I_{\text{CH}_2} = \frac{A^2 \Gamma_{\text{aliph}} \Gamma_{\text{CH}_2}}{16 \ln 2}.$$  \hspace{1cm} (2b)

With the FWM signal being proportional to the module square of the third order susceptibility, and then to the square of the number of molecules [31], the square root of the integral is calculated to give a quantity having a linear dependence with the number of molecules. The aliphatic peak at 1450/2850 (all of the peaks are noted by their coordinates in the frequency plane, $\omega_a/\omega_b$, in cm$^{-1}$) is taken as an internal reference, and the integral of the other peaks is taken relative to this one. For any peak $n$, the ratio of Eqs. (2a) and (2b) is calculated. With the square of the amplitude $A$ being proportional to the square of the number of molecules $N (A^2 = N^2)$, we obtain the following expression for the ratio (Eq. (3)):

$$\text{Ratio}_n = \frac{I_{\text{peak}}}{I_{\text{CH}_2}} = \frac{\frac{4 \pi \Gamma_{\text{peak}} \Gamma_{mn}}{A^2 \Gamma_{\text{aliph}} \Gamma_{\text{CH}_2}}}{\frac{2 \pi \Gamma_{\text{aliph}} \Gamma_{\text{CH}_2}}{A^2 \Gamma_{\text{aliph}} \Gamma_{\text{CH}_2}}} = \frac{N_a}{N_{\text{CH}_2}} \times \frac{\Gamma_{\text{CH}_2}}{\Gamma_{\text{aliph}} \Gamma_{\text{CH}_2}}.$$  \hspace{1cm} (3)

The number of CH$_2$ groups ($N_{\text{CH}_2}$) in each sequence is easily calculated, as are the numbers of phenylalanine and tyrosine groups ($N_{\text{Phe}}$ and $N_{\text{Tyr}}$). The experimental quantity, Ratio$_n$, is then plotted as a function of the ratio $N_a/N_{\text{CH}_2}$ for each Phe and Tyr peak and should show a linear variation.

**Results**

**2D spectra of peptides**

2D vibrational spectra of eight peptides are shown in Fig. 4. The spectra have in common a strong peak at around 1450/2850. This has been identified as corresponding to vibrational modes of CH$_2$ [32]; this feature is used as an internal reference for the calculation of the ratio of the integrated intensities. Other complex features appear around 1470/3050, 1520/3100, and 1480/2920.

For sequences in which Phe is the only aromatic amino acid present, two groups of peaks are measured in the upper part of the spectra (LRRFSLGF and LRRFFLG)
in Fig. 4). Three peaks are present along the vertical at \(\omega_x \approx 1515 \text{ cm}^{-1}\) and \(\omega_y = 3110, 3080, \text{ and } 3040 \text{ cm}^{-1}\); this part of the spectra is called the high-frequency aromatic region. Two peaks are also identified at 1470/3050 and 1480/3070; we call this spectral region the low-frequency aromatic region.

For sequences in which Tyr is the only aromatic amino acid present, only a very weak peak is measured in the low-frequency region at 1470/3040 and a stronger one is measured in the high-frequency region at 1530/3130 (LRRYSLGY and LRRYYLYG in Fig. 4).

The assignments of the modes of the aromatic peaks are confirmed by density functional theory (DFT) calculations of 2DIR FWM spectra of benzene derivatives and aromatic amino acids. The details of the DFT calculations, the detailed spectra calculations, and the assignments concerning the aromatic amino acids will be the subject of a future article, with some information already made available in a recent publication [28]. The identified peaks are depicted in Fig. 5. The spectral assignments are listed in Table 1. There are two vibrational ring stretching modes contributing to the aromatic peaks of Phe and Tyr: CC + HCC in-plane (\(\nu_{13}\)) and CC + HCC in-plane + CCC in-plane (\(\nu_{16}\)) [33].

We focus on the Phe's doublet, the most intense peak of the Phe's triplet, and the most intense Tyr peak. We extract the FWHM and amplitudes, as well as the frequencies, of the aliphatic Tyr and Phe peaks, and the relative amplitudes and integrated intensities are calculated using the procedure described in the previous section.

Fig. 4. 2DIR FWM spectra of the eight peptides studied. The amino acid sequence for each of the samples is indicated in each spectrum. For all spectra, the delays are \(t_{12} = 1.5 \text{ ps}\) and \(t_{23} = 1 \text{ ps}\). Both \(\omega_x\) and \(\omega_y\) are scanned with a 5 \text{ cm}^{-1} \text{ increment} (45 \times 101 \text{ points/spectrum}). The spectra are presented linearly scaled, and the peak line shapes are real. The color scales are linear, but not identical, from one spectrum to another to help identify the features. The peak at 1450/2850 common to all of the spectra is attributed to aliphatic CH\(_2\) vibrational modes and is taken as an internal reference. The high-frequency features are due to aromatic vibrational modes of Tyr and Phe.

Fig. 5. Identified spectral features in the peptide spectra. A spectral feature for leucine is identified, as is a CH\(_3\) peak that is present in all of the peptide spectra and used as an internal reference for the fingerprinting. The peak at 1460/2000 was left unassigned; it could be a weak contribution from glycerine. There are three features corresponding to Phe and one peak corresponding to Tyr. The mode assignments are detailed in Table 1.

Quantification of amino acid levels

Figs. 6 and 7 show plots of the normalized integrated intensities (\(\text{Ratio}_x\) in Eq. (3)) and the normalized peak amplitudes of the Phe and Tyr peaks against the known number of Phe and Tyr present in each peptide relative.
Table 1
Spectral assignments of the main peaks depicted in Fig. 5

<table>
<thead>
<tr>
<th>Peak position (cm⁻¹)</th>
<th>Group and amino acid</th>
<th>Mode assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1450/2880</td>
<td>CH₂, all</td>
<td>δ(CH₂)/2δCH₂ (first overtone)</td>
</tr>
<tr>
<td>1470/3070</td>
<td>Aromatic, Phe</td>
<td>δ(CH₂), v₁δ(CH₂) + (δCH₂, v₁₆)</td>
</tr>
<tr>
<td>1480/3070</td>
<td>Aromatic, Phe</td>
<td>δ(CH₂), v₁₂δ(CH₂, v₁₃) + v₁₆</td>
</tr>
<tr>
<td>1515/3040</td>
<td>Aromatic, Phe</td>
<td>Not assigned</td>
</tr>
<tr>
<td>1525/3080</td>
<td>Aromatic, Phe</td>
<td>v₁₂/v₁₆ + (δCH₂, v₁₆)</td>
</tr>
<tr>
<td>1515/3110</td>
<td>Aromatic, Phe</td>
<td>v₁₂/v₁₆ + v₁₆</td>
</tr>
<tr>
<td>1530/3130</td>
<td>Aromatic, Tyr</td>
<td>v₁₂/v₁₆ + v₁₆</td>
</tr>
</tbody>
</table>

Note: δ, deformation; v, stretch; +, combination band. Modes in parentheses are complex modes with two contributions.

Fig. 6. Correlation between the known Phe to CH₃ ratios and the ratios measured by 2DIR. The square root of the ratio of the integral (and amplitude) of the features relative to the aliphatic peak integral (and amplitude) is plotted as a function of the known number of Phe relative to the known number of CH₃ groups. Data points are shown by open circles (and open triangles) and are labeled by the amino acid sequence of the corresponding sample. The dashed red line is the linear fit constrained to go through the origin. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 7. Correlation between the known Tyr to CH₃ ratios and the ratios measured by 2DIR. The square root of the ratio of the integral (and amplitude) of the features relative to the aliphatic peak integral (and amplitude) is plotted as a function of the known number of Tyr relative to the known number of CH₃ groups. Data points are shown by open circles (and open triangles) and are labeled by the amino acid sequence of the corresponding sample. The dashed red line is the linear fit constrained to go through the origin. Note that the scales are smaller than those in Fig. 6, making the dispersion appear bigger than it really is (see Table 2).
LRRYSLGY) or Tyr (LRRFFLG and LRRFSLGF) does not show any features at the Phe or Tyr frequencies, and so these peak intensities are zero. The standard deviations for the dispersion of the data points for each feature for both the integral and amplitude data are shown in Table 2.

### Discussion

The linear correlation between the relative integrated intensities (or amplitudes) of specific vibrational features and the relative amount of the corresponding amino acid shown in Figs. 6 and 7 confirms the validity of Eq. (3). This demonstrates that relative quantification of the amino acid content by 2DIR is possible. The deviations from the linear behavior shown in Table 2 can have two different origins: structural effects (primary and secondary structures) and sample-to-sample variations.

The sample-to-sample variations include fluctuations in time of the intensity profiles of the lasers and pointing variations inducing changes in the phase matching conditions. It is not clear at this point whether the precision is limited by structural effects or instrumental/optical variations from sample to sample. Nevertheless, we have achieved sufficient precision for our purpose (see Table 2), as discussed at the end of this section.

The 2D spectra were acquired at 0.5 s per data point, with each spectrum comprising 4545 points (Fig. 4). The precisions resulting from the amplitude measurements compared with the integral measurements (Table 2) show that there is neither a need for a full 2DIR spectrum nor for a 2D fit of the entire spectrum for relative quantification. Once features are located, assigned, and identified, the peak strengths can be retrieved from one well-chosen data point. So long as a point on the spectrum that has relatively little interference from surrounding features (i.e., the spectral feature is resolved) can be found, the measurements of peak amplitudes are accurate enough. The data, therefore, suggest that it would be possible to measure a number \( N \) of amino acid quantity ratios (so \( N + 1 \) peak amplitudes) in a time of \( (N + 1)/2 \) s with sufficient precision. It is likely that with the current state of the art the initial throughput will be somewhat slower due to automation overheads, and so probably a time of 1 to 5 s per amino acid feature would be more realistic in the case of protein samples.

Given that quantification of amino acid levels using 2DIR seems feasible, the outstanding question is how many amino acid ratios one would need to measure to identify a protein unambiguously. To answer this question, a full bioinformatics study is needed and currently is under way (it will be presented in detail elsewhere). Nevertheless, we have constructed a database of the amino acid ratios of 48,000 human proteins (using the National Center for Biotechnology Information [NCBI] reference sequence collection) and have performed preliminary tests to investigate the number \( N \) of amino acid quantity ratios required. Table 3 summarizes our results.

The number of ratios required depends on the protein to be identified and on the type of amino acids that are effectively discovered and quantified by 2DIR FWM. That is why the number of ratios (peaks) in Table 3 is given as intervals. From this preliminary investigation, we can say that, with the current state of the experimental procedures (~12.5% precision), four to nine peaks would need to be measured to uniquely identify a protein. This means that protein identification should take between 4 and 45 s considering, as discussed previously, an acquisition time of 1 to 5 s per point. These acquisition times fulfill the requirement of high throughput necessary for the development of a protein fingerprinting technique.

For the data presented here, the average number of molecules in the laser spot is approximately \( 10^{13} \) (17 pmol), and good fingerprints can be obtained at approximately \( 10^{11} \) to \( 10^{12} \) molecules. With small modifications to the existing apparatus, but using the same general approach, we would expect \( 10^{10} \) molecules in the beam to be achievable; however, this has yet to be demonstrated, and we do not yet know the absolute lower limit for sensitivity. There are other variants of nonlinear optical spectroscopy that we plan to assess for viability as fingerprinting methods, but neither their utility nor the possibilities of analyzing post-translational modifications have yet been established.

### Conclusion

We have demonstrated the basis of an alternative analytical approach to protein identification based on 2DIR spectroscopy by measuring 2D spectra of peptide side chains. Specific features corresponding to the aromatic vibrational modes of Tyr and Phe, as well as an aliphatic mode, were identified. Using the spectra obtained for various amino acid sequences containing different amounts of Phe and Tyr, it was shown that the relative integrated signal strengths and the relative amplitudes are propor-
tional to the number of Phe or Tyr in the peptide relative to the number of methylene groups. This proportionality is achieved with a precision of approximately ±12.5%, which in principle would be enough for the purpose of fingerprinting. Our data did not show any evidence of structural effects that would affect the precision of the data.

It should be noted that the number of amino acid peaks to be measured depends a priori on three factors: the search pattern of amino acids chosen, the protein to be identified, and the precision of the 2D IR measurements. With the precision demonstrated here (±12.5%), the number of peaks required would be, in principle, between four and nine. Moreover, it was shown that simple measures of the 2D IR signal amplitude at the peak positions is sufficient for this purpose and that there is no need to measure full 2D spectra. This suggests that protein identification could take between 4 and 45 s.

Acknowledgments

We acknowledge Dr. Darek Kedra of the Bioinformatics Support Service at the Imperial College London. This work was financially supported by the Engineering and Physical Sciences Research Council (EPSRC), the Royal Society, and the Chemical Biology Centre Doctoral Training Centre.

References

Protein identification and quantification by two-dimensional infrared spectroscopy: Implications for an all-optical proteomic platform


*Department of Chemistry and ‡Chemical Biology Centre, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom; †Bioinformatics Support Service, Division of Molecular Biosciences, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom; and ‡Institute of Cancer Research, Chester Beatty Laboratories, Cancer Research UK, Centre of Cellular and Molecular Biology, London SW3 6JB, United Kingdom

Edited by Robin M. Hochstrasser, University of Pennsylvania, Philadelphia, PA, and approved August 8, 2008 (received for review May 27, 2008)

Electron-vibration-vibration two-dimensional coherent spectroscopy, a variant of 2DIR, is shown to be a useful tool to differentiate a set of 10 proteins based on their amino acid content. Two-dimensional vibrational signatures of amino acid side chains are identified and the corresponding signal strengths used to quantify their levels by using a methyl vibrational feature as an internal reference. With the current apparatus, effective differentiation can be achieved in four to five minutes per protein, and our results suggest that this can be reduced to <1 min per protein by using the same technology. Finally, we show that absolute quantification of protein levels is relatively straightforward to achieve and discuss the potential of an all-optical high-throughput proteomic platform based on two-dimensional infrared spectroscopic measurements.

The potential of proteomic tools ranges from biomarker discovery and clinical diagnostics to the provision of data for systems biology and fundamental biological research (1–6). This broad range of applications is one of the drivers for the development of protein analysis tools with greater capability. Optical spectroscopies appear to have significant potential for protein analysis (7–9), but conventional approaches suffer from overcongested spectra, which makes feature assignment and quantification highly problematic. Multidimensional coherent infrared spectroscopic techniques, commonly referred to as two-dimensional infrared (2DIR) spectroscopies, might be expected to be able to relieve the congestion of infrared spectra sufficiently to allow such assignment and quantification to take place. Indeed, we recently demonstrated how picosecond electron-vibration-vibration (EVV) four-wave mixing experiments can deconvect 2DIR spectra to an even greater extent (10, 11), and showed how such an EVV approach can be applied to the analysis of peptides (12). In this article we take the approach further to show that it can be used to differentiate and identify proteins and to measure absolute protein quantities. We also demonstrate that the sensitivity and throughput of our EVV 2DIR apparatus is sufficient for this method to be considered for use as a real proteomic tool.

Although we have previously shown that it is possible to quantify relative amino acid levels for short peptides (12), there is always the possibility that primary, secondary, or tertiary structural effects would prevent such measurements on proteins. In this article we demonstrate that these structural sensitivities are not limiting factors either for differentiation/identification or for absolute quantification of protein levels.

The key proposition of this article is that protein identification can be performed by using spectrosopically determined amino acid content, relative to an internal reference. Amino acid composition analysis is an approach that has been used to determine protein relatedness (13) and protein structural classes (14, 15). Although it is known that compositional ratios of amino acids can also be used to identify proteins (for instance, by using the AAcMolIdent [http://www.expasy.org/tools/aacomp/) or MultIdent (http://www.expasy.org/tools/multident/) tools], this method of protein identification is not widely used. The experimental methods historically used for composition determination require hydrolysis of the protein substrate, followed by separation and derivatization of its amino acids before quantification can occur (16–18). In contrast, the identification strategy outlined here requires no chemical or biochemical preparation steps and achieves quantification by measuring spectroscopic features of the proteins.

In this particular study we use methyl groups (CH₃) as an internal reference, and the fingerprint of a protein in this case is the distribution of its relative amino acid quantities. We have identified spectral features corresponding to the CH₃ reference and to three different amino acids: tyrosine (Tyr), phenylalanine (Phe), and tryptophan (Trp). We use the peak ratios of these three amino acids to the CH₃ internal reference, with the Phe measured with two different polarizations. This gives a total of four amino acid cross-peaks to be monitored, as well as the CH₃ cross-peak that is measured for both polarization schemes.

Identification is achieved by comparing these spectroscopically determined amino acid ratios of a protein to the contents of a database. To this end we have constructed searchable protein databases for a number of model organisms; these are composed of the amino acid/CH₃ ratios for each of their proteins. Fig. 1 shows the distribution of hits that are returned when the CH₃ ratios of the three spectrally identified residues are input with 10% precision for each protein in our human database [taken from ENSEMBL release 44 (19)]. To demonstrate how this identification strategy scales, we also show the results for when another two amino acids, in this case histidine (His) and cysteine (Cys), are included, along with the three residues used experimentally in this article. The relative amounts of His and Cys residues, as well as of Trp, Tyr, and Phe, can vary significantly from one protein to another, and they are therefore good candidates for our protein identification strategy. Preliminary EVV 2DIR measurements on peptides, and calculations of the EVV 2DIR spectra of these species, show that histidine and cysteine, with resolvable features at 1475/2650 cm⁻¹ and 1485/
Fig. 1. Histograms demonstrating the feasibility of identifying proteins by using their amino acid/CH₃ ratios. Tests were performed over our human proteome database; amino acid/CH₃ ratios and their precisions were used as search parameters for each protein in the database. The horizontal axes correspond to the number of protein outputs from the database when a search was performed for a protein. The vertical axes represent the frequency with which a particular number of hits were output when the search was performed for each protein of the database (~33,000) in turn. (A) Shown is the number of hits output from the database when the three amino acid/CH₃ ratios studied experimentally for this article (TyrCH₃, TrpCH₃, and PheCH₃) were input with 10% precision for each protein. (B) Shown is the number of protein hits returned when the database search was extended to using five amino acid ratios (by also using the His/CH₃ and Cys/CH₃ ratios). (Inset) Histograms show the results for when the molecular weight (10% precision) was also included as a search parameter. The first bar of B shows that ~15,000 proteins of the ~33,000 present in the database (~44%) gave only one hit and thus were uniquely identifiable. The second bar shows that ~9,000 proteins gave two protein hits and so could be one of only two possible database candidates. When the molecular weight was also used as a parameter, ~20,000 (~60%) of the proteins were unambiguously identified.

2,560 cm⁻¹, respectively (data not shown), are useful residues to include in our identification strategy.

A preliminary bioinformatics analysis shows that identifying the relative levels of only five amino acids would allow ~44% of the proteins in the ENSEMBL human proteome database to be uniquely identified and ~72% to be one of only two proteins.

Results

EVV 2DIR spectroscopy (also known as Doubly-Vibrationally Enhanced Four-Wave Mixing spectroscopy [20–25]) requires the overlap of two picosecond infrared (IR) beams and a picosecond visible beam on the sample. A nonlinear visible signal generated by the induced polarizations is detected. The signal intensity is measured as a function of both IR frequencies, and the spectra are presented as two-dimensional intensity maps. Cross-peaks appear only at the IR frequencies corresponding to vibrational states that are coupled. The delays between the pulses, as well as the orientation of the visible electric field, can be varied, and we show below how alteration of these parameters helps to further decongest the spectra. The polarization states are denoted by the usual S and P notation; this describes the orientation of the electric fields relative to the plane of propagation. T₁ and T₂ are the delays between the first IR pulse (frequency ω₁) and the second IR pulse (frequency ω₂), and the second IR pulse and the visible pulse, respectively. Because this particular version of 2DIR is a homodyne spectroscopy, the total signal is proportional to the square of the number of molecules in the beam.

Spectral Signatures of Amino Acids and Use of Multiple Polarizations.

The first step of our protein differentiation/identification procedure is to identify spectral features that are amino acid specific. Spectral congestion has to be minimized to ensure that each cross-peak corresponds to a unique residue. To be exploitable, the features must also be free of interferences that could affect the amplitudes of the cross-peaks. The delays allow selection of the coherence pathways corresponding to the EVV 2DIR process, minimizing other nonlinear processes and also reducing the electronic resonance background (10, 11). The polarizations select which components of the susceptibility tensor will be probed (26) and therefore can be used to extinguish certain vibrational modes, thus further decreasing the spectral congestion.

All of the spectra presented here were measured at T₁ = 2 ps and T₂ = 1 ps and for two different sets of polarizations (PPP or PPS when all beams are polarized in the plane of propagation and PPS when the electric field of the visible is perpendicular to the plane of propagation). The EVV 2DIR spectra of peptides obtained in our previous studies were an aid in the identification of the features measured in the protein spectra (12), which were found to be very similar.

Fig. 2 shows typical examples of protein EVV 2DIR spectra.

Vibrational features of the amino acid side chains are present in the spectral regions of 1485/3070 and 1525/3120 for phenylalanine, and of 1545/5150 for tyrosine. Their assignments are presented more fully elsewhere (12).

In brief, the higher-frequency cross-peaks (labeled “Tyr” and “Phe 1”) for tyrosine and phenylalanine, respectively, arise from the coupling of an aromatic stretching mode with a combination band involving aromatic stretching modes and a CH₃ deformation. We estimate the full width at half-maximum (FWHM) of these cross-peaks to be 10–15 cm⁻¹. Although these Tyr and Phe 1 features are 20–30 cm⁻¹ apart and appear not fully resolved, the cross-contamination is sufficiently small to reliably quantify the amount of each residue through the cross-peak intensity on resonance.

The lower-frequency phenylalanine cross-peak (labeled “Phe 2(P)” and “Phe 2(S)” for the PPP and PPS schemes, respectively) arises from the coupling of a mode involving aromatic stretching and CH₃ deformation with a combination band that also involves aromatic stretching modes and a CH₃ deformation.
Changing the polarization of the visible beam from PPP to PPS helps to reduce the remaining congestion around the Phe 2 cross-peak, but renders the Phe 1 and Tyr aromatic modes very weak. Nevertheless, both polarization schemes produce data that can be used to assist in the differentiation of the proteins. Independent measurements of the phenylalanine levels at both polarizations contribute to the overall dataset used here.

Decongesting the spectral region around the Phe 2 feature also reveals a clearly resolved tryptophan peak at 1490/3020 (labeled “Trp”) in Fig. 3. This decongestion effect can be seen in Fig. 3 where the spectra of pepsin and a tryptophan-rich protein, α-chymotrypsin, are compared. We have also observed this tryptophan peak in tryptophan-containing peptides.

To summarize, the intensity of four amino acid features corresponding to three amino acid residues are monitored to perform the relative quantification procedure: the phenylalanine “Phe 1” and tyrosine “Tyr” cross-peaks in the PPP polarization scheme, and the phenylalanine “Phe 2(S)” and tryptophan “Trp” cross-peaks in the PPS polarization scheme. The “Phe 2(PP)” feature is considered congested and therefore not exploited. The internal reference “CH₃” is measured for both polarization schemes.

CH₃ and CH₃ Internal Standard. The structured peak at ~1475/2920 contains CH₂ and CH₃ contributions. These arise from a combination of fundamental modes, overtones of the CH stretch, and Fermi resonance between the CH stretch and the first overtone of the CH deformation.

In summary, peaks from four different CH modes are known to be present: CH₃ symmetrical (s) stretch at 2,865 cm⁻¹, CH₃ symmetrical stretch at 2,865 cm⁻¹, CH₂ asymmetrical (as) stretches at 2,920 cm⁻¹ and 2,930 cm⁻¹, and CH₃ asymmetrical stretches at 2,960 cm⁻¹ and 2,984 cm⁻¹ (27).

We found that the feature at 1,480/2,960 works well as an internal standard. A complementary experiment (data not shown) showed that the peak at 1,460/2,920 can also be used as a reference; a priori any or all of these features would probably be suitable for this particular role.

Amino Acid Quantification. The square roots of the intensities of the Tyr, Phe 1, Trp, and Phe 2(S) cross-peaks relative to the square root of the intensity of the CH₃ reference peak were calculated and plotted as a function of the actual amino acid to CH₃ ratios for each protein (Fig. 4). A more detailed procedure can be found in Methods and supporting information (SI) Text. The reproducibility of the measurements was estimated by repeating this procedure four times at different positions on each sample.

Protein Differentiation. The purpose of proteomic techniques is the identification of proteins and absolute quantification of their
amount in a given sample. For proteins to be identified, they need to be distinguishable from each other. We use a mathematical definition of distinguishability based on a multidimensional overlap integral comprising the overlap integrals for each amino acid peak (see Methods and SI Text). One protein can be distinguished from another protein if any amino acid is clearly present at different levels in the two proteins. Distinguishability in this case would mean that the overlap of the distributions, which peak at the expected amino acid ratio value and have a width of the standard deviation of the measurement, is much less than the difference in the expected amino acid ratios of the proteins being compared. The total overlap integral is the product of the integrals of two proteins for each amino acid. If the overlap integral has a value of 1, then the two proteins are wholly indistinguishable. If the overlap integral has a value of zero, then the proteins are wholly distinguishable.

The results of these calculations are presented as dimensional maps, in which the intensity corresponds to the value of the normalized integral from 0 (black) to 1 (white) and reflects the probability of two proteins being identical (i.e., their probability of identity). Figs. 5 and 6 show such maps for single amino acid peaks and combinations of amino acid peaks, respectively.

Fig. 5 shows that the tryptophan peak, “Trp,” alone provides quite a good basis for differentiating between many of these proteins. It does not, however, allow differentiation between all pairs of proteins, for example, alkaline phosphatase (protein 8) from BSA (protein 3) or β-lactoglobulin B (protein 10) from albumin (protein 1) and aldolase (protein 2).

As one would expect, the distinguishability increases when a combination of several amino acid peaks is used (Fig. 6). The cumulative number of pairs of discernible proteins as a function of the probability of identity can be deduced from each differentiation map (Fig. 7A). For example, if one accepts a maximum of 10% probability of identity between two proteins, the four cross-peak scheme (labeled (d) in Fig. 7A) gives 42 pairs of discernible proteins out of a total of 45 pairs. Instead, a two cross-peak scheme (phenylalanine and tyrosine PPP peaks, scheme (b) in Fig. 7A), gives only 12 pairs of discernible proteins out of 45 (again accepting a 10% probability of identity). The cross-peak intensity is measured in such a way that the protein differentiation efficiency for different regimes of acquisition times can also be assessed. The cumulative number of pairs of discernible proteins as a function of the probability of identity can be determined for different time-averaging regimes (data not shown). We deduced that, for a 2- to 4-min measurement time per protein, the four cross-peak scheme with an acceptance of 10% probability of identity gives a good result of 39 pairs of discernible proteins out of a total of 45 (Fig. 7B). More signal averaging increases the precision with which each amino acid is quantified; it also increases a priori the number of pairs of discernible proteins for cases of probabilities of identity <50–60% (data not shown).

**Discussion**

We have shown that protein differentiation with a maximum of 10% probability of identity can be achieved for 39 pairs of proteins out of 45 in 2 to 4 min of measurement time per protein (Fig. 7B).

In some cases, it appears there is a limitation on the precision with which the content of particular amino acids can be determined. This can be seen by the spread of the data points around
the straight-line fit in Fig. 4 for tryptophan. This spread, or dispersion, is greater than predicted from experimental precision alone, which suggests that there is some small residual structural sensitivity for this particular cross-peak. For cases where this limitation is reached, improved identification capability can only be achieved by finding spectral features for more amino acids, rather than by further signal averaging for that particular protein. However, the use of different polarizations for phenylalanine shows that higher precisions can be achieved by taking more than one peak per amino acid if desired.

Differentiation within a limited set of proteins is easier than absolute identification from an entire protein database of an organism. Given the limitations discussed above, we estimate that it will take between six and nine amino acids to unambiguously identify >90% of the proteins in our database of human proteins (~33,000 proteins taken from ENSEMBL). We also estimate that, with the current technology, the signal-averaging time per amino acid peak can be shortened to between 1 and 10 s. This gives a realistic protein identification time of somewhere between 10 s and 2 min.

An important property of EVV 2DIR as a proteomic technique is its potential for simple absolute quantification of protein levels. Absolute quantification is increasingly an important issue in many proteomic applications, but one that is relatively difficult to achieve with mass spectrometry. Quantification with EVV 2DIR is, however, relatively easy to achieve. This is because the average oscillator strength for the CH$_3$ cross-peaks (used as the internal standard in this study) appears to be the same for all proteins so far studied. This presumably reflects the fact that the oscillator strength is an average over many CH$_3$ groups in each protein, and that these groups are relatively insensitive to structural effects. The acquisition and treatment procedures for performing absolute quantification are described in Methods. As predicted by the theory, we found that the integrated intensity of the square-rooted signal of a dried drop of protein solution is proportional to the total number of protein molecules it contains (Fig. 8). From these measurements we have established that the practical sensitivity limit with our apparatus is ~10$^{12}$ protein molecules (~1.5 pmol).

An additional advantage of EVV 2DIR as a proteomic tool is that it is nondestructive, such that the samples can be retained for further and more detailed investigations. Preliminary results from peptides also strongly suggest that EVV 2DIR has the potential to monitor levels of posttranslational modifications such as phosphorylation.

Methods

EVV 2DIR Spectroscopy. A full description of the laser setup and the principles of this technique can be found elsewhere (11, 12). In brief, a commercial picosecond regenerative amplifier and two IR optical parametric amplifiers were used to provide a visible beam at 800 nm and two frequency-scannable IR beams. The three beams were overlapped on the sample, and the visible four-wave mixing signal produced at $\omega_4$ ($\omega_4 = 2\omega_2 + \omega_1 - \omega_0$, where $\omega_0$ and $\omega_1$ are the frequencies of the IR beams, and $\omega_2$ is the frequency of the incident visible beam) was detected in transmission with a photomultiplier. The detected signal was plotted as a function of both IR frequencies to produce two-dimensional spectra. For the results presented here, the photomultiplier was used in photon-counting mode. This allowed the data to be collected by using longer time delays between the laser pulses, thus producing less congested protein spectra. The congestion issue was also addressed by changing the polarization of the visible beam; we discovered that this helped to further decongest the spectra, making phenylalanine and tryptophan peaks exploitable.

Bioinformatics. Human protein sequences were obtained as fasta file from ENSEMBL release 44 (19). Amino acid compositions and protein molecular weights were then computed by using pepstats from Emboss 5.0 package (28). The output file was parsed by using custom Python scripts, and the data were stored in MySQL database on a Linux workstation. The number of database hits to a protein query was calculated by counting all entries within a given accuracy interval by using SQL and Python scripts.

Sample Preparation. Proteins were prepared in the form of dried films, cast onto glass slides. The following 10 proteins were used (all were bought from Sigma-Aldrich): albumin from bovine serum, albumin from chicken egg white, aldolase from rabbit muscle, alkaline phosphatase from bovine intestinal mucosa, $\alpha$-chymotrypsin from bovine pancreas, concanavalin A from Jack Bean, $\beta$-lactoglobulin B from bovine milk, lysozyme from chicken egg white, and pepsin from porcine gastric mucosa. The protein solutions had concentrations in the range from 10 to 50 mg/ml, and volumes of 1.5 $\mu$L were deposited (see SI Text).

Procedure for the Peak Intensity Measurements. Once two-dimensional spectral features have been identified and associated with amino acids, there is no need for two-dimensional spectra to be recorded for each protein. The signal intensity can be measured for the pairs of IR frequencies corresponding to the peaks of interest. For the measurements presented here, the delays were set at $\Delta T_1 = 2$ ps and $\Delta T_2 = 1$ ps, which was discovered to be a compromise between signal strength and decongestion. The intensity on each peak was recorded for 5 s over a period of 120 s, and the amino acid/CH$_3$ ratios (corrected from the nonresonant background) were calculated. To assess the reproducibility, the measurements were repeated four times for the full set of 10 proteins. The details of the acquisition procedure and ratio calculations are presented in the SI Text.

Protein Differentiation Overlap Integrals: Definition of Distinguishability. A normal distribution of amino acid/CH$_3$ ratios was constructed for each protein by using the measured ratios, the standard deviations, and the linear fit. To compare two proteins, the overlap integral of their ratio distributions was calculated to give a mathematical definition of distinguishability (see SI Text). If a protein is compared with itself, then the overlap integral is maximum and the normalized integral is one. A comparison of proteins with at least one orthogonal subintegral gives a null integral.

Absolute Quantification of Proteins Levels. BSA solutions of five different concentrations were deposited and left to dry on a microscope cover slide. Concentrations of 0.5, 0.4, 0.3, 0.2, and 0.1 mM were used, which correspond to $9 \times 10^{-5}$, $7 \times 10^{-5}$, $5.5 \times 10^{-5}$, $3.5 \times 10^{-5}$, and $2 \times 10^{-5}$ total protein molecules respectively in each of the 0.3-$\mu$L drops. Because of the spatial heterogeneity of the dried films (thickness, material density), the EVV 2DIR signal on the CH$_3$ peak (at 1.4852,390, delayed set at $\Delta T_1 = 1.5$ ps and $\Delta T_2 = 1$ ps) was mapped across all five dried drops. The images were measured with 1 s of acquisition time per point in photon-counting mode. They were corrected for any photon-counting nonlinearity and then background corrected and square rooted. The integrated intensity (of the square-rooted image) associated with each drop reflects the total number of protein molecules in the deposited volume. Plotting this against the known protein content of each film gives a straight line and an effective calibration curve for quantifying the protein levels.

Acknowledgments. We thank C. J. Barnett for technical support. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) and the Chemical Biology Centre Doctoral Training Centre (CBC DTC).
Biological and Biomedical Applications of Two-Dimensional Vibrational Spectroscopy: Proteomics, Imaging, and Structural Analysis

FRÉDÉRIC FOURNIER,*† RUÍ GUO,*† ELIZABETH M. GARDNER,*† PAUL M. DONALDSON,*† CHRISTIAN LOEFFELD,*† IAN R. GOULD,*† KEITH R. WILLISON,*† AND DAVID R. KLUG†*‡

†Department of Chemistry and Chemical Biology Centre, Imperial College London, Exhibition Road, London SW7 2AZ, U.K.; ‡Institute of Cancer Research, Chester Beatty Laboratories, Cancer Research U.K., Centre of Cellular and Molecular Biology, London SW3 6JB, U.K.

RECEIVED ON MARCH 18, 2009

CONCEPTUS

In the last 10 years, several forms of two-dimensional infrared (2DIR) spectroscopy have been developed, such as IR pump–probe spectroscopy and photon-echo techniques. In this Account, we describe a doubly vibrationally enhanced four-wave mixing method, in which a third-order nonlinear signal is generated from the interaction of two independently tunable IR beams and an electron-polarizing visible beam at 790 nm. When the IR beams are independently in resonance with coupled vibrational transitions, the signal is enhanced and cross-peaks appear in the spectrum. This method is known as either DOVE (doubly vibrationally enhanced) four-wave mixing or EVV (electron–vibration–vibration) 2DIR spectroscopy.

We begin by discussing the basis and properties of EVV 2DIR. We then discuss several biological and potential biomedical applications. These include protein identification and quantification, as well as the potential of this label-free spectroscopy for protein and peptide structural analysis. In proteomics, we also show how post-translational modifications in peptides (tyrosine phosphorylation) can be detected by EVV 2DIR spectroscopy.

The feasibility of EVV 2DIR spectroscopy for tissue imaging is also evaluated. Preliminary results were obtained on a mouse kidney histological section that was stained with hematoxylin (a small organic molecule). We obtained images by setting the IR frequencies to a specific cross-peak (the strongest for hematoxylin was obtained from its analysis in isolation; a general CH₂ cross-peak for proteins was also used) and then spatially mapping as a function of the beam position relative to the sample. Protein and hematoxylin distribution in the tissue were measured and show differential contrast, which can be entirely explained by the different tissue structures and their functions.

The possibility of triply resonant EVV 2DIR spectroscopy was investigated on the retinal chromophore at the centre of the photosynthetic protein bacteriorhodopsin (bR). By putting the visible third beam in resonance with an electronic transition, we were able to enhance the signal and increase the sensitivity of the method by several orders of magnitude. This increase in sensitivity is of great importance for biological applications, in which the number of proteins, metabolites, or drug molecules to be detected is low (typically pico- to femtomoles). Finally, we present theoretical investigations for using EVV 2DIR spectroscopy as a structural analysis tool for inter- and intramolecular interaction geometries.

1. Introduction

There are a number of forms of coherent multidimensional spectroscopy that have emerged as useful tools in a variety of research contexts. Each of these methods has particular strengths and capabilities that make them more or less suitable for addressing particular scientific questions. We have been exploring the utility of one family of 2DIR methods for a range of biological and bio-
medical applications. The approach, originally demonstrated by John Wright and colleagues,\textsuperscript{1–6} is known as electronic–vibration–vibration two-dimensional infrared spectroscopy (EVV 2DIR) or as DOVE-IR (doubly-vibrationally enhanced infrared spectroscopy). The purpose of this Account is to give a broad overview of what is currently possible and indicate what might realistically become possible, using this particular form of 2DIR with respect to biological and biomedical applications. It is entirely possible that other forms of spectroscopy including other forms of 2DIR may be able to do as well or better than EVV 2DIR with respect to some of these applications. But in this Account, we restrict ourselves to discussing this one class of methods.

2. The Basis of EVV 2DIR

EVV 2DIR has some significantly different properties compared with other published forms of 2DIR spectroscopy. This leads to particular strengths and weaknesses and also has implications for the practical implementation of EVV 2DIR experiments. This section is intended to highlight those qualities that differ from other methods.

2.1. The Practicability of EVV 2DIR Experiments

In EVV 2DIR, three independently tunable and independently timed picosecond pulsed laser beams are brought together on a sample in a phase-matched configuration. Two of these beams are in the infrared and are used to excite molecular vibrations, while the third is a visible beam and is essentially used to probe the polarization generated by the two IR beams, such that the coupling strength of the two vibrations can be "read-out" by the detection of visible photons. The signal is detected at the frequency \( \omega_b = \omega_a + \omega_g - \omega_b \) (with \( \omega_a \) and \( \omega_g \) representing the IR frequencies and \( \omega_b \) the visible beam frequency). The final "read-out" step is essentially half of a conventional Raman scattering event, but coherent. One of the consequences of this is that the electronic properties of the molecular system also affect the signal and that the experiment can be made triply resonant if the visible "probe" beam is tuned toward an electronic resonance.

The IR beams are independently scanned in frequency and when they are in resonance with coupled vibrational modes the detected signal is multiplicatively enhanced. The spectra represent the level of signal detected at \( \omega_b \) as a function of the IR frequencies; they show cross-peaks at the specific IR frequencies corresponding to coupled vibrational modes. Two delay stages control the timing between the pulses: when not further discussed, \( T_{12} \) and \( T_{23} \) denote the delay of the IR pulse at \( \omega_g \) relative to \( \omega_b \) and of the 800 nm pulse relative to the IR at \( \omega_g \), respectively.

2.2. Properties of EVV 2DIR

One of the most obvious and unusual aspects of EVV 2DIR is that it involves not only the excitation of molecular vibrations with infrared laser pulses but also the direct polarization of electrons via a "half-Raman" scattering step. This makes EVV 2DIR effectively a hybrid Raman–IR method where the final signal strength depends on both IR and Raman processes. This is perhaps most clearly demonstrated when EVV 2DIR spectra are calculated from first principles.\textsuperscript{3,7}

Another important property is that all of the EVV 2DIR experiments performed to date have involved direct excitation of vibrational combination bands, usually in the near-infrared. Vibrational combination bands are created by the anharmonic coupling of two vibrational modes. In general, two types of anharmonicities will contribute: mechanical anharmonicity, which measures the deviation of the molecular potential energy surface from simple harmonic potential, and electrical anharmonicity, which is the nonlinearity of the molecular dipole moment along vibrational coordinates.\textsuperscript{8} This means that EVV 2DIR is sensitive to couplings that arise from purely mechanical anharmonicity, from purely electrical anharmonicity, or from both. The processes that contribute to the total EVV signal are shown in Figure 1.

It can be seen that pathway 3 is essentially the same as coherent anti-Stokes Raman spectroscopy (CARS), but with the involvement of real excited states rather than the purely virtual states of a CARS experiment. Pathways 2 and 3 are only accessible if the near-infrared pulse that excites the combination band comes first. Thus it is possible to switch pathways on and off by changing the pulse ordering, and this additional
flexibility conveys some additional utility to the EVV method as discussed below.

2.3. Breaking the Fixed Fourier Relationship in 2DIR: Enhanced Decongestion of 2DIR Spectra. The independent timing of the infrared and visible pulses has an important consequence, namely, that it is possible to break the fixed transform relationship between wavelength and time that is used in the current pump-probe or echo variants of 2DIR. This means that pathways that are always present in other forms of 2DIR can be switched off in EVV 2DIR by manipulation of the pulse timings. This is seen clearly in the situation where Fermi resonances are present as in the case of benzene. EVV 2DIR has the advantage of offering enhanced spectral decongestion with other parameters such as delays between pulses and beam polarizations providing yet additional control of spectral congestion.

2.4. Sensitivity to Pure Electrical Anharmonicity. One of the more unusual properties of EVV 2DIR is that it is not a difference spectroscopy. As a consequence of this, cross-coupling can be observed even in the case where the mechanical anharmonicity is either too small to shift ground and excited vibrational states significantly or entirely absent. The corollary of this is that the other 2DIR techniques demonstrated to date are not able to measure couplings that are purely electrical unless the electrical coupling is sufficiently strong to cause secondary mechanical anharmonicity.

A good example of such a case is the methyl and methylene stretch or deformation overtone, coupled with the scissors deformation mode. Both methyl and methylene features appear as strong and highly reproducible features in EVV 2DIR spectra of peptides and proteins, and we have studied them extensively. Calculations suggest that these modes have either very weak or zero mechanical anharmonic coupling but significant electrical anharmonicity.

In order to demonstrate that these features are indeed the result of pure electrical anharmonicity, we use the independent pulse timings available in EVV 2DIR to extinguish the signal via interference of pathways 2 and 3 (Figure 1). If the time ordering of the IR pulses is reversed (E1 first) from the timing usually used (E2 first), then pathway 1 no longer contributes, and the signal is the sum of the contributions of pathways 2 and 3. Pathways 2 and 3 have opposite sign, and therefore if the combination band is unshifted by the interaction of modes a and b, as is the case when there is zero mechanical anharmonicity, then the signals will totally cancel. The spectra shown in Figure 2 demonstrate this effect: when only pathway 1 is opened (E2 first), CH2 and CH3 cross-peaks are present in the spectrum, whereas when the IR pulses order is reversed (E1 first), these two cross-peaks disappear. If there is a level shift, then splitting of the cross-peaks is observed as pathways 2 and 3 are no longer degenerate.

The cancellation of pathways 2 and 3 in the case of pure electrical anharmonicity (absence of mechanical anharmonicity) appears to be more or less total with less than 3% of the signal remaining for the CH2 cross-peak.

3. Biological and Biomedical Applications of EVV 2DIR

3.1. Proteomics. The field of proteomics is the identification and characterization of proteins on a large scale and at high throughput to provide data for the discovery and analysis of protein networks and to obtain a global view of cellular processes at the protein level, the so-called ‘proteome’. The proteome describes the number of each protein in a specific cell, the interactions of the proteins, and post-translational modifications of the proteins and allows one in principle to monitor any changes in the cellular pathways due to disease, drug action, or other biological perturbations. Determination of a proteome is an intensive task requiring protein identification, quantification, and analysis of post-translational modification and protein complexation. In so-called ‘bottom-up’ proteomics, the proteins to be identified are digested by an enzymatic or chemical process and the resulting peptides analyzed, usually by mass spectrometry. In the second approach, the ‘top-down’ strategy, the intact protein system is studied. The bottom-up approach is widely used and commercial instruments exploiting this strategy are available. Top-down approaches have the advantage of requiring no
additional chemical–biochemical preparative steps associated with producing fragments and, as the bottom-up approach, allowing the study of native proteins and native protein properties.

We are in the process of developing EVV 2DIR spectroscopy as an analysis tool for proteomics. The initial goal is to be able to identify proteins and in the future analyze the identified proteins in greater detail, such as assessing the extent and type of their post-translational modifications. Our protein identification strategy is based on using EVV 2DIR to quantify the amino acid content of a protein. EVV 2DIR is shown to be able to perform absolute quantification, something of major importance in the field of proteomics but rather difficult and time-consuming to achieve with mass spectrometry. Our technique can be qualified as a top-down label-free method; it does not require intensive sample preparation, the proteins are intact when analyzed, and it does not have any mass restriction on the proteins to be analyzed. Moreover, EVV 2DIR is a nondestructive technique; the samples can be kept for reanalysis in the light of further information.

3.1.1. Identifying and Counting Proteins. Known small peptides were used to identify tyrosine, phenylalanine, and CH3 specific vibrational signatures in the EVV 2DIR spectra. We showed that the square root of the relative intensity of the cross-peaks is proportional to the amount of amino acid relative to the CH3 internal reference cross peak.14 When applied to a set of 10 known proteins, these same amino acid signatures are detected, as well as a tryptophan cross-peak and a CH3 feature used as internal reference (Figure 3).

Figure 4 shows that there is indeed a direct correlation between the EVV 2DIR features signal level for each amino acid and the known amount of that amino acid in ten test proteins. This direct proportionality suggests that these cross-peaks are not structurally sensitive and that this approach can be applied efficiently to protein identification.

The curves of relative amino acid quantification (Figure 4) are the starting point of a protein differentiation procedure based on probability of identity.19 Briefly, each amino acid/CH3 ratio is described by a normal distribution with a width determined by the standard deviation of the data shown in Figure 4. The fingerprint of one protein is defined as a product of N orthogonal normal distributions corresponding to N amino acid number ratios, N being the number of identified amino acid cross peaks or the number of amino acids used for the differentiation strategy. The comparison of two proteins is performed by calculating the overlap integrals of their numerical fingerprints: if the distributions are exactly overlapped, the integral is maximum and the proteins are indistinguishable; on the contrary, if the distributions are totally different, the integral is zero and the proteins are totally distinguishable. These overlap integrals, when normalized to the value obtained from comparing one protein with itself, define a probability of identity of two proteins. Protein differentiation maps represent the probability of identity as a two-dimensional graph where the gray level of each square reflects the probability of two proteins to be identical (Figure 5).

Bioinformatics studies suggest that identifying the relative levels of only five amino acids with 15% precision would allow 44% of the proteins in the ENSEMBL human protein
FIGURE 5. Differentiation maps of the ten proteins studied using the tyrosine peak only and the full set of amino acid peaks (Tyr, Phe1, Phe2(S) and Trp cross peaks as noted in Figure 3). While squares correspond to normalized overlap integral values of 1 (completely indistinguishable) and black to 0 (completely distinguishable). The gray level of each square reflects the probability that the two proteins being compared are the same. The diagonal compares a protein with itself and so is white. The relative darkness of an entire map shows the relative contribution of each amino acid scheme in distinguishing the proteins.

Database (~33,000 proteins) to be uniquely identified and 72% to be one of only two proteins; 60% would be unambiguously identified if the molecular weight of the proteins is used as a parameter as well.

The speed of this process is high enough to consider EVV 2DIR as a real proteomics tool, because each amino acid peak can be quantified to sufficient precision with around 10–60 s of signal averaging giving around 1–4 min per protein depending on the search pattern used.

Finally there is the issue of absolute quantification. Because of spatial heterogeneity of the protein sample (a dried drop of protein solution placed on a substrate), it is better to map the signal across the whole sample than to compare the CH₂ cross peak intensity on local spots on each dried film. It is then possible to determine the total integrated intensity and check for linearity (Figure 6).

3.1.2. Post-translational Modifications. The phosphorylation of amino acids is of major importance in regulating protein activity. To see whether we could identify phosphorylation of tyrosine, we made identical peptides containing two tyrosines and two phosphorylated tyrosines. These were diluted in water at the same concentrations and deposited on a stainless steel substrate. EVV 2DIR spectra were then measured in reflection (Figure 7).

The nonphosphorylated peptide (labeled LRRYYLG) shows the usual pattern consisting of CH₂ and CH₃ cross peaks and two aromatic cross peaks of tyrosines previously identified in our peptides studies (Tyr 1 and Tyr 2). For the phosphorylated peptide (labeled LRRpYpYLG) the \( \Delta CCH₂(^{13}C) / [\Delta CCH₂(^{13}C) + \Delta C(\mu)H₂(\mu)C(\mu)H₂] \) cross peak (Tyr 2) totally disappears whereas the \( \nu_{13} / \nu_{13} + \nu_{18} \) cross peak (Tyr 1) is red-shifted by \( \sim 12 \text{ cm}^-1 \) in the \( e_{\alpha} \) direction and \( \sim 6 \text{ cm}^-1 \) in the \( e_\beta \) direction.

These preliminary results suggest that the intensity of the Tyr 2 cross peak is inversely proportional to the amount of phosphorylated tyrosines present in the peptide.
3.2. Imaging. Optical imaging of cells or tissue samples is an important area in the life sciences. Chemically sensitive label-free imaging methods are particularly attractive in principle, and a range of approaches have been developed including IR, Raman, and CARS imaging. In all of these cases, chemical sensitivity is possible but due to the extensive con-gestion of the vibrational spectra in these samples, this is usually restricted to classes of molecules, such as proteins, fats, and nucleic acids, or to particular molecules present at very high concentration. EVV 2DIR spectroscopy has the potential to provide enhanced contrast in imaging of tissue samples by decongesting the spectra to a much greater extent than is possible with imaging methods based on techniques that use one spectral dimension. EVV 2DIR imaging could therefore be useful for providing alternative information on the spatial distribution of particular chemical species. We therefore show here a simple pilot experiment that demonstrates that EVV 2DIR can be used in such a manner.

To perform the tissue imaging, the IR frequencies are set on a cross-peak specific to a chemical group, the sample is spatially scanned and the signal level, detected by transmission, is mapped as a function of the beam position on the sample.

The validation experiments shown here are performed on histological sections of mouse kidney stained with hematoxylin. The ~5 μm thick tissue sections are deposited on 1.5 mm thick microscope glass slides. The purpose is to show that the distributions of hematoxylin can be determined within different tissue types of the kidney section by using a vibrational cross-peak as the contrast-providing feature.

Prior to the sample imaging, hematoxylin solution was deposited and allowed to dry on a microscope glass slide in order to measure its EVV 2DIR signatures (Figure 8).

The hematoxylin signal at the strongest cross peak (1530/3040 in Figure 8) was easily detected in the kidney section. From previous studies, we know that EVV 2DIR spectra of proteins show a CH\textsubscript{3} cross peak at around 1490/2940, and this signal level was mapped as well.

Images were acquired for the two pairs of IR frequencies corresponding to the hematoxylin and the protein CH\textsubscript{3} (Figure 9).

The hematoxylin image reflects the distribution of cell nuclei and also ribonucleic acids since it stains mainly those entities. Differential contrast is clearly seen between the hematoxylin and the protein CH\textsubscript{3} images.

A simple histological interpretation of the kidney section is presented in Figure 10; a general anatomic description of the kidney structure can be found elsewhere. Hematoxylin stains nuclei and also ribonucleic acids necessary for protein synthesis. High levels of hematoxylin are measured in the cortical and medulla areas and a more dispersed signal in the collecting area. This is consistent with the presence of cells with nuclei in these areas. Instead, the blood clot contains cells without nuclei (red blood cells and platelets), which do not fix hematoxylin, hence the lack of signal.

Hemoglobin and fibrin present in large amounts in the blood clot give rise to a strong protein signal. Protein signal is high as well in the cortical area where cells are involved in urine formation and thus contain energetic machinery and synthesize a large amount of proteins. Weaker signals are measured in the medulla and collecting areas where cells are mainly involved in exchange of water and ions and structure and transport, respectively.

3.3. Enzyme Mechanisms. One field in which time-resolved spectroscopy has a significant impact is the study of enzyme mechanisms. This has largely been restricted to reactions that are naturally light-triggered, such as photosynthetic or visual processes, but has also included cases of direct photolysis and the release of caged compounds by photolysis.

Mechanistic studies have up to now used one-dimensional spectroscopy to study the formation of intermediate states during enzyme-catalyzed reactions, but in principle the information content of a multidimensional spectroscopy should be greater. In this section, we demonstrate the use of electronically resonant EVV 2DIR to selectively observe structural elements of a particular protein's active site, namely the retinal chromophore at the center of the photosynthetic protein bacteriorhodopsin (bR), while avoiding signal from the surrounding protein. By tuning the visible beam close to resonance with an electronic transition during an EVV 2DIR experiment, one should get considerable enhancement of the signal size in the same way as occurs with resonance Raman scattering.
Biological Applications of 2D Vibrational Spectroscopy

**FIGURE 9.** EVV 2DIR images and a visible microscope image of the kidney section area investigated. For EVV 2DIR, the signal is mapped across the sample for $T_{12} = 1$ ps and $T_{23} = 1$ ps for two pairs of IR frequencies corresponding to the hematoxylin and the protein CH$_3$ cross peaks. The EVV 2DIR images are recorded point by point (1 s acquisition per point) with 25 $\mu$m step size. The sample is scanned horizontally in the same direction for every vertical position to create the $3 \times 3$ mm field. The level of EVV 2DIR signal increases from blue to green and red; both EVV 2DIR images are shown with the same intensity scale. The microscope image is made out of ten narrow field images stitched together to represent the area scanned by EVV 2DIR.

**FIGURE 10.** Interpretation of the kidney histological section showing the different kinds of tissue present in the area scanned by EVV 2DIR.

Triply resonant signal from bacteriorhodopsin membranes (isolated from *Halobacterium salinarum*) was observed at 1545/3050 cm$^{-1}$ (inset of Figure 11b). The assignments of the coupled vibrational states giving rise to this cross peak were confirmed by DFT calculations. The modes being probed by the infrared pulses are a C=C stretch of the retinal’s conjugated backbone, $\nu$(C=C) (1545 cm$^{-1}$), and its first overtone, 2$\nu$(C=C), (3050 cm$^{-1}$).

Numerous infrared and Raman studies have used the shifting in frequency of this C=C stretch for observing the intermediates of bacteriorhodopsin’s photocycle. Because the experiments here were carried out on samples at <1% humidity, the observed cross peak is attributable to various photoaccumulated and trapped intermediates of br’s dark-adapted photocycle ($\Delta_{\text{max}}$, of the sample measured at 562 nm).

An example of the level of signal enhancement possible using this triply resonant technique can be seen by comparison of the signal level from the methyl group of a protein with that from bacteriorhodopsin’s retinal. At the end of section 3.1.1, we demonstrated how absolute quantification of proteins can be achieved with EVV 2DIR by imaging a sample on the protein CH$_3$ cross peak at 1485/2930 cm$^{-1}$. Here we show that the same quantification procedure can be performed for bacteriorhodopsin using the retinal feature at 1545/3050 cm$^{-1}$. To this end, a series of bacteriorhodopsin samples of varying concentrations is imaged on the retinal cross peak. Figure 11 compares calibration curves obtained for concanavalin A (chosen for its similar size to br) imaged at 1485/2930 cm$^{-1}$, and for bacteriorhodopsin imaged on the 1545/3050 cm$^{-1}$ retinal cross peak.

From the concanavalin A measurements, a sensitivity limit of $9 \times 10^{12}$ protein molecules ($\approx 15$ pmol) in the lasers interaction area was established. The bacteriorhodopsin measurements, however, give a sensitivity limit of $6 \times 10^{10}$ molecules (0.1 pmol) in the laser interaction area. Considering an interaction area with a diameter of 80 $\mu$m and a probed sample thickness of 10 $\mu$m, these sensitivity limits would correspond to solutions of 0.3 mol/L for concanavalin A and 0.002 mol/L for br. In terms of numbers of molecules therefore, the limit of detection it is possible to achieve for br when probing its chromophore is 100 times lower than that for a protein when measuring its CH$_3$ cross peak. Moreover, considering that each concanavalin A molecule contains 139 CH$_3$ groups and that each bacteriorhodopsin molecule contains only one retinal chromophore, the CH$_3$ group and retinal detection limits in terms of number of chemical moieties are $1 \times 10^{10}$ and $6 \times 10^{10}$.
FIGURE 11. Square root of the EVV 2DIR signal level as a function of the number of concanavalin A (a) or bacteriorhodopsin (b) molecules in the sample. Insets of panels a and b are EVV 2DIR spectra showing the protein CH$_3$ and the retinal ν(C=C) cross peaks, respectively. For concanavalin A, the signal level of the CH$_3$ peak at 1485/2930 was mapped across five deposited films with concentrations of 0.1, 0.2, 0.3, 0.4, and 0.5 mM and a drop volume of 0.3 µL. For bacteriorhodopsin, the signal level of the retinal peak at 1545/3050 was mapped across films with concentrations of 2, 3, 4, 5, and 6 µM and a drop volume of 0.3 µL. For both concanavalin A and bacteriorhodopsin, the measurements were taken in reflection with delays of $T_{12} = 1$ ps and $T_{23} = 1$ ps, a 1 s acquisition time per pixel and a 100 µm step size. The integrated EVV 2DIR intensity of the square-rooted image for each dried drop is plotted against the total number of protein molecules (NConA or NbR). The error bars are standard deviations from four repeats performed on four different sets of five protein samples. The solid lines represent the linear fits.

$10^{10}$, respectively. Thus by using triply resonant EVV 2DIR, it is possible to achieve $10^4$ times higher signal from retinal’s conjugated backbone than it is from a protein CH$_3$ group. Therefore significant improvement in EVV 2DIR can be achieved by making the electronically polarizing visible “read out” beam preresonant with the molecular systems being studied. Further enhancement could be achieved in this case by tuning the visible beam (currently at 790 nm) closer to the 560 nm absorption maximum of the Bt.

3.4. Structural Analysis: Measuring Structural Parameters via Electrical Anharmonicity in the Absence of Mechanical Anharmonicity. There have been a number of studies regarding the use of 2DIR spectroscopy to decipher the structures of small peptides.26–29 This is a particularly challenging problem and one that will require a good deal more effort to solve. There are however many other categories of molecular structural analysis of great importance in the biological and biomedical sciences. A particular class of these is the study of the geometry of Interaction between biomolecules that do not form a covalent bond or the geometry of interaction of chemical groups within large biomolecules such as proteins, which are often the interactions that convey various forms of biological function. In this section, we focus on this class of interactions and note that although our primary interest is in biomolecular interactions, the approach outlined here could potentially be used in other contexts where molecular species of interest are juxtaposed.

As discussed in section 2.4, one of the unusual features of EVV 2DIR spectroscopy is that it is sensitive to pure electrical coupling between molecular vibrations, that is, pure electrical anharmonicity. Good examples of this are the cases of the CH$_3$ and CH$_2$ modes shown in Figure 2 in section 2.4. These cross-peaks arise from pure electrical anharmonicity, a conclusion supported by both QM calculation and experimental evidence of the total cancellation of these features when pathways 2 and 3 are enabled by appropriate pulse ordering.

One of the useful properties of pure electrical anharmonicity is that it obeys simple physical rules for the distance and angle dependence of the coupling. It is already known that intermolecular electrical couplings between two groups should be inversely proportional to the third power of their distance as first proposed by Okumura et al.30 and Cho31 and later confirmed in CO–HCl dimer by Hahn et al.32 By a theoretical survey of a series of molecules with formulas such as H=C=O(CH$_2$)$_n$=C=N ($n = 2–4$), we were able to confirm that this relationship also exists for intramolecular electrical coupling as shown in Figure 12 between C=C and C=N groups, which largely couple with each other electrically.

Apart from this sensitivity to distance, the relative orientations of the interacting groups are almost as crucial. This can be illustrated by a theoretical examination of EVV 2DIR spectra of two conformers of 4-cyano-1-butyne (HC=C–(CH$_2$)$_2$–C=N), the anti and gauche conformers. Theory predicts that only the anti conformer exhibits two distinctive coupling cross peaks between the C=C and C=N stretching modes and none for the gauche one (Figure 13).

Since the distance between CC and CN groups is shorter in the gauche conformer (3.8 Å compared with 4.9 Å), this
should have given a greater signal by generating an electrical anharmonicity larger in magnitude. However the relative orientation of the two interacting groups in the gauche conformer means its resultant electrical anharmonicity lies at an angle from the directions of transition dipoles and transition polarizabilities of CC and CN groups, and this nonoptimal alignment makes the electrical anharmonicity unable to contribute to 2DIR signals. On the other hand, although the distance between the two groups is larger in the anti conformer, the alignment between the two groups is optimal, thus enabling the full contribution of the interacting electrical anharmonicity to 2DIR signals.

Although it would be difficult to verify these theoretical predictions experimentally, this preliminary theoretical study clearly showed the structural sensitivity of EVV 2DIR spectroscopy in terms of distances and relative orientations between two interacting groups. This hints at the possibility of developing EVV 2DIR into a structural analysis tool for intermolecular complex formation, in particular the binding of small molecules to proteins, protein–protein interactions, and the interactions of protein side chains within one other protein. In fact, the results with 4-cyano-1-butyne indicate that even for certain intramolecular coupling EVV 2DIR spectroscopy might have some use as a structural analysis tool as long as the coupling is largely electrical.

4. Conclusion

There are now many published examples of several coherent multidimensional spectroscopy techniques using different optical pulse sequences and combinations. These sequences and combinations may be used in different systems depending upon the context of the experiments and the details of implementation. We have sought to illustrate some applications of EVV 2DIR in the biological and biomedical sciences in particular with respect to proteomics and imaging and give a hint of what might be possible in the future for structural analysis of proteins in particular functional states.

BIOGRAPHICAL INFORMATION

Frederic Fournier obtained his Ph.D. in Physics from University Paris XI-Orsay in 2003. Since 2006, he has been a Postdoctoral Research Associate in the Chemical Biology Centre at Imperial College London, focusing on the development of nonlinear laser spectroscopy methods and their applications to biological problems.

Rui Guo received his Ph.D. in Physical Chemistry in 2000 from Peking University. He is currently a Postdoctoral Research Associate at Imperial College London working on theoretical simulations of EVV 2DIR spectroscopy, in particular the structural—spectral correlation from first-principle calculations.

Elizabeth M. Gardner obtained her M.Res. in Biomolecular Sciences in 2005 from Imperial College London. She is currently completing her Ph.D. with the Chemical Biology Centre based at the same institution. Her research interests include the use of optical spectroscopies for protein analysis, in particular, the development of EVV 2DIR spectroscopy as a protein fingerprinting tool.

Paul M. Donaldson obtained his Ph.D. from Imperial College London in 2008. He is currently a Postdoctoral Research Associate at the University of Zurich working on applying 2DIR methods to the study of energy transport in proteins.

Christian Loeffler received his M.Sc. from Technical University of Munich and National University of Singapore in 2006. Since 2007, he has been a Ph.D. student at the Chemical Biology Centre at Imperial College London investigating electokinetic phenomena and applying 2D-IR methods for proteomics-related issues.

Ian R. Gould is a Senior Lecturer in the Department of Chemistry at Imperial College London. His research interests include ab initio chemistry, molecular modeling, molecular dynamics, protein structure–function relationships, sugar and lipid force field development, and nucleic acids.

Keith R. Willison is Professor of Molecular Cell Biology at the Institute of Cancer Research. His research interests include bio-
chemical, kinetic, and structural studies on protein folding, molecular chaperones, and the actin cytoskeleton.

David R. Klug is Professor of Chemical Biophysics in the Department of Chemistry at Imperial College London and Chair of the Chemical Biology Centre. His research interests include natural and artificial solar energy conversion, protein structure-function relationships, and quantitative proteomics and protein networks.

REFERENCES


