Evaluation of infrared QCL, Synchrotron and bench-top sources for cell imaging in aqueous media

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science & Engineering

2016

Zhe Zhang

School of Chemical Engineering and Analytical Science
## Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Abbreviations</td>
<td>5</td>
</tr>
<tr>
<td>List of Figure</td>
<td>6</td>
</tr>
<tr>
<td>Abstract</td>
<td>12</td>
</tr>
<tr>
<td>Declaration</td>
<td>13</td>
</tr>
<tr>
<td>Copyright Statement</td>
<td>14</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>15</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>16</td>
</tr>
<tr>
<td>2. Theories and instruments</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Spectroscopy</td>
<td>19</td>
</tr>
<tr>
<td>2.1.1 Infrared</td>
<td>19</td>
</tr>
<tr>
<td>2.1.2 Vibrations</td>
<td>19</td>
</tr>
<tr>
<td>2.1.3 Morse potential</td>
<td>20</td>
</tr>
<tr>
<td>2.1.4 Vibrational-Rotational coupling</td>
<td>21</td>
</tr>
<tr>
<td>2.1.5 Beer–Lambert law</td>
<td>22</td>
</tr>
<tr>
<td>2.2 Spectrometer</td>
<td>23</td>
</tr>
<tr>
<td>2.2.1 Dispersive infrared spectrometer</td>
<td>24</td>
</tr>
<tr>
<td>2.2.2 FT-IR spectrometer</td>
<td>24</td>
</tr>
<tr>
<td>2.2.3 Infrared microscopy</td>
<td>28</td>
</tr>
<tr>
<td>2.2.4 Synchrotron light source</td>
<td>36</td>
</tr>
<tr>
<td>2.2.5 Quantum cascade lasers</td>
<td>38</td>
</tr>
<tr>
<td>3. Literature Review</td>
<td>19</td>
</tr>
<tr>
<td>3.1 Cell and cell spectrum</td>
<td>45</td>
</tr>
<tr>
<td>3.2 Water barrier</td>
<td>50</td>
</tr>
<tr>
<td>3.3 Live cell imaging with benchtop FTIR spectrometer</td>
<td>52</td>
</tr>
<tr>
<td>3.4 Live cell imaging with SYNCHROTRON based IR microspectroscopy</td>
<td>59</td>
</tr>
<tr>
<td>3.5 Live cell imaging with QCL based microscope</td>
<td>63</td>
</tr>
<tr>
<td>4. Experimental setup</td>
<td>19</td>
</tr>
<tr>
<td>4.1 Sample Preparation</td>
<td>76</td>
</tr>
</tbody>
</table>
4.2 Modified compression cell ................................................................. 78
4.3 Liquid cell for synchrotron experiment ................................................ 80
4.4 Imaging with benchtop FTIR spectrometer ........................................... 80
4.5 Cell spectra with synchrotron-based FTIR spectrometer ....................... 82
4.6 Cell imaging with QCL microscope ..................................................... 83
5. Data analysis method ........................................................................... 87
   5.1 Least square fitting based on Beer-Lambert law ................................. 87
   5.2 Water correction method on different type of data .............................. 90
   5.3 Fringing removal ............................................................................. 92
   5.4 Cell extraction method for high throughput analysis ......................... 93
   5.5 Pre-processing method for multivariate analysis ............................... 95
   5.6 Multivariate analysis ........................................................................ 98
   5.7 Random forest classifier .................................................................. 100
6. Benchtop FTIR imaging of cell in aqueous medium .................................. 87
   6.1 Model of water correction ............................................................... 105
      6.1.1 Effects of fitting different spectral region ..................................... 105
      6.1.2 Choice of reference .................................................................... 109
      6.1.3 Validation .................................................................................. 111
   6.2 Spectra obtained in low mag setup .................................................... 113
      6.2.1 Cell imaging ............................................................................... 113
      6.2.2 Cell spectra ............................................................................... 114
   6.3 Attempted oscillation removal process .............................................. 120
   6.4 Multivariate analysis on ‘low mag’ spectra ........................................ 124
      6.4.1 PCA .......................................................................................... 124
      6.4.2 PC-LDA .................................................................................... 128
      6.4.3 Random Forest .......................................................................... 131
   6.5 Spectra obtained in high mag setup ................................................... 133
      6.5.1 Cell imaging ............................................................................... 133
      6.5.2 Fitting-average or average-fitting? .............................................. 134
List of Abbreviations

ATR  Attenuated Total Reflectance
AUC  Area Under Curve
EMSC Extended Multiplicative Signal Correction
FTIR Fourier Transform Infrared
FOV Field Of View
FPA Focal Plane Arrays
IR Infrared
LN2 Liquid Nitrogen
MCT Mercury Cadmium Telluride
NA Numerical Aperture
PC Principal Component
PCA Principal Component Analysis
PC-LDA Principal Component-Linear Discriminant Analysis
PBS Phosphate-buffered Saline
ROC Receiver Operating Characteristic
SNR Signal to Noise Ratio
SR-FTIR Synchrotron Radiation-based Fourier Transform Infrared
QCL Quantum Cascade Laser
List of Figure

FIGURE 2.1 SCHEMAT DIAGRAMS SHOWING THE VIBRATIONAL MODE OF -CH₂ GROUP .......................................................... 20
FIGURE 2.2 COMPARISON OF MORSE POTENTIAL AND HARMONIC OSCILLATOR. (D₀)DE ARE (ACTUAL) DISSOCIATION
ENERGY. ................................................................................................................................................................. 21
FIGURE 2.3 VIBRATIONAL-ROTATIONAL COUPLING ON SPECTRUM OF CO............................................................. 21
FIGURE 2.4 DIAGRAM OF DOUBLE-BEAM DISPERSIVE INFRARED SPECTROMETER ...................................................... 24
FIGURE 2.5 SCHEMAT DIAGRAM OF A BASIC FTIR INTERFEROMETER ........................................................................ 25
FIGURE 2.6 FOURIER TRANSFORM. THE WAVE FUNCTION IN THE TIME DOMAIN HAS BEEN DECOMPOSED IN TO A
COMBINATION OF MULTIPLE SINUSOIDAL WAVES WITH PARTICULAR AMPLITUDE AND FREQUENCY. EACH SINUSOIDAL
WAVE IS THEN PROJECTED INTO THE FREQUENCY DOMAIN ACCORDING TO ITS AMPLITUDE AND FREQUENCY ............ 26
FIGURE 2.7 SCHWARTZSCHILD OBJECTIVE. BLUE PART REPRESENTS THE MIRRORS AND THE YELLOW ARROWS STANDS FOR
THE LIGHT PATH....................................................................................................................................................... 29
FIGURE 2.8 COMPONENTS IN A SYNCHROTRON SITE ................................................................................................... 36
FIGURE 2.9 TWO TYPE OF RADIATION GENERATORS. BLUE LINE IS THE ELECTRON FLUX AND THE YELLOW BEAM SHAPE IS
THE EMITTED RADIATION....................................................................................................................................... 37
FIGURE 2.10 PIN LASER DIODE (LEFT) AND INTERBAND (RIGHT) ................................................................................. 38
FIGURE 2.11 QCL DIODE (LEFT) AND INTERSUBBAND (RIGHT) ....................................................................................... 39

FIGURE 3.1 SPECTRUM OF A SINGLE CELL (FIXED)........................................................................................................... 45
FIGURE 3.2 WATER SPECTRUM (BLUE) AND ITS INTERACTION WITH CELL SPECTRUM (GREEN) ............................. 51
FIGURE 3.3 LIQUID CELL WITH SANDWICH SETUP ....................................................................................................... 52
FIGURE 3.4 MODIFIED MOVEABLE ATR CRYSTAL WITH CELL CULTURE CHAMBER ON THE TOP .............................. 53
FIGURE 3.5 3D PRINTING MICROFLUIDIC FLOW CELL WITH EXTRA CAF₂ LENS, ASSEMBLED WITH A HOLDER ......... 54
FIGURE 3.6 (TOP) TOP VIEW AND (BOTTOM) CROSS SECTION OF THE DEVICE. THE SMALL CIRCLE IN TOP VIEW IS SAMPLE
CELLS WHICH ARE PRE CULTURED ON THE UPPER WINDOW. THE INLET AND OUTLET HOLE ARE EXPRESSED IN CROSS
SECTION WHICH FILLED WITH DASH LINE. THE HEATING CARTRIDGES ARE NOT INCLUDED IN THIS DIAGRAM. ............ 55
FIGURE 3.7 FULLY SEALED MICROFLUIDIC LIQUID CELL. CELLS WERE PLACED IN THE MIDDLE CHAMBER. A FLOW OF CULTURE
MEDIUM WAS PASSING THROUGH FROM THE TWO SECTIONAL CHAMBER. ............................................................ 56
FIGURE 3.8 CAF₂ WINDOW EMBEDDED POLYMER MICROFLUIDIC DEVICE FOR FTIR LIVE CELL IMAGING. THE CELLS WERE
INJECTED TO THE CELL CHAMBER AND CULTURED WITH SUPPLY OF A FLOW OF CULTURE MEDIUM AND CO₂ .......... 57
FIGURE 3.9 DESIGN OF OPEN CHANNEL IMAGING WITH FLOW OF MEDIUM. THE CELL WAS PLANTED ON GOLD COATED
SILICON ......................................................................................................................................................... 60
FIGURE 3.10 (A) TOP VIEW AND (B) CROSS SECTION OF THE DEVICE. THE T SLOT WAS ETCHED ON BOTTOM WINDOW TO
CREATE LAMINAR FLOW. TWO HOLES WERE DRILLED ON BOTTOM WINDOW AS INLET AND OUTLET FOR EXCHANGING
CULTURE MEDIUM............................................................................................................................................... 61
FIGURE 3.11 LITHOGRAPHED STATIC CELL ON CAF₂ WINDOW, CELLS SOLUTION WERE DROPPED INTO SAMPLE CHAMBER.
AIR CHAMBER IS USED TO TAKE BACKGROUND .................................................................................................. 62
FIGURE 3.12 CHANNEL DESIGN FOR CELL-DRUG INTERACTION. CELLS AND DRUGS ARE INJECTED INTO DIFFERENT CHAMBER
(UPO AND DOWN), THE FLUID IN THESE CHAMBER WILL EXCHANGE THROUGH VERY NARROW CHANNEL (5 MM)
SHOWED ON THE RIGHT ................................................................................................................................. 63

FIGURE 4.1 MODIFIED COMPRESSION CELL WITH CAF₂ WINDOWS PS SPHERE (BLUE DOT) AS SPACER................. 79
FIGURE 4. 2 LIQUID CELL WITH GOLD COATED SPACER ..................................................................................... 80
FIGURE 4. 3 DIAGRAM OF LIGHT PATH AND KEY COMPONENTS IN THE BEAMLINE ......................................................... 82

FIGURE 5. 1 WATER CORRECTION PROCESS. THE RAW SPECTRUM (TOP-LEFT) FIRSTLY BEEN FITTED BY TWO REFERENCE SPECTRA (RIGHT). SECOND, APPLY THE COEFFICIENT TO BUFFER REFERENCE SPECTRUM AND OBTAIN THE WATER SPECTRUM. THIRD, THE WATER SPECTRUM BEEN SUBTRACTED FROM THE RAW ..................................................................................... 88
FIGURE 5. 2 COMPARISON OF SPECTRA BEFORE AND AFTER WATER CORRECTION ........................................................................... 89

FIGURE 6. 1 SPECTRA OF 50 PC-3 CELLS IN PBS SOLUTION WITH A PATH LENGTH OF 5.5 MM ................................................. 105
FIGURE 6. 2 SPECTRA OF PC-3 CELLS OBTAINED AFTER WATER CORRECTION WITH DIFFERENT FITTING REGIONS. EACH SPECTRUM REPRESENTS THE AVERAGE OF 50 SPECTRA. ALL SPECTRA WERE NORMALIZED TO AMIDE I BAND. THE SPECTRUM OF MATRIGEL IS SHOWN FOR REFERENCE .......................................................................................................................... 106
FIGURE 6. 3 THE FINGERPRINT REGION OF SPECTRA IN FIGURE 6.2. RESULTS SEPARATED INTO TWO PARTS: (TOP) SPECTRA FITTED WITH AMIDE A BAND; (BOTTOM) SPECTRA FITTED WITHOUT AMIDE A BAND .......................................................................................................................... 106
FIGURE 6. 4 SPECTRUM OF MATRIGEL (BLUE) AND SPECTRUM OF AIR DRIED PC3 CELL (GREEN). VECTOR NORMALISED AND MEAN CENTRED.......................................................................................................................... 109
FIGURE 6. 5 A COMPARISON OF CORRECTED SPECTRA OF PC3 CELLS OBTAINED BASED ON DIFFERENT CELL REFERENCE SPECTRA AND FITTING RANGE. SPECTRA CUT INTO THE FINGERPRINT REGION, VECTOR NORMALISED AND MEAN CENTRED. (TOP) SPECTRA FITTED WITH ALL THREE PROTEIN BANDS (BOTTOM) SPECTRA FITTED WITHOUT AMIDE A BAND .......................................................................................................................... 110
FIGURE 6. 6 COMPARISON OF SPECTRA OF PC-3 CELLS WITH ETHANOL FIXATION (EF, BLUE) AND FORMALIN FIXATION (FF, GREEN) IN (TOP) LIPID BAND AND (BOTTOM AMIDE I AND II BAND) .......................................................................................................................... 112
FIGURE 6. 7 TWO EXAMPLES (ROWS) OF CELL IMAGING. (LEFT COLUMN) VISIBLE IMAGES OF CELLS IN PBS SOLUTION; (MIDDLE COLUMN) INTENSITY IMAGES OF AMIDE II PEAK CENTRE (1550 cm⁻¹) BEFORE WATER CORRECTION; (RIGHT COLUMN) INTENSITY IMAGES OF AMIDE II PEAK CENTRE AFTER WATER CORRECTION. IR IMAGE HAS A FOV OF 704x704mm² .......................................................................................................................... 113
FIGURE 6. 8 RAW SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS IN PBS SOLUTION. SPECTRA WERE VECTOR NORMALISED AND MEAN CENTRED .......................................................................................................................... 116
FIGURE 6. 9 CORRECTED SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS. .......................................................................................................................... 117
FIGURE 6. 10 CORRECTED SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS. THE SPECTRAL REGION ABOVE 3000cm⁻¹ WERE CUT OUT AND THE REMAINING PART WERE VECTOR NORMALISED AND MEAN CENTRED. .......................................................................................................................... 118
FIGURE 6. 11 MEAN SPECTRA OF PC-3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELLS. SPECTRA VECTOR NORMALISED BASED ON THE FULL SPECTRA REGION AND MEAN CENTRED. .......................................................................................................................... 119
FIGURE 6. 12 SPECTRA OF PC-3 CELLS WHICH EXHIBIT ‘FRINGING-LIKE’ FLUCTUATIONS. SPECTRA WERE WATER CORRECTED, VECTOR NORMALISED AND MEAN CENTRED. .......................................................................................................................... 120
FIGURE 6. 13 FLUCTUATION PATTERN EXTRACTED FROM THE SPECTRA IN THE FIGURE 6.12 .......................................................................................................................... 121
FIGURE 6. 14 (TOP) ABSOLUTE VALUE OF THE COMPLEX NUMBER OBTAINED FROM THE REVERSE FOURIER TRANSFORM. THE CURVE ARE SYMMETRIC. THE TIME STEP OF THE SPIKE WAS SHOWN AS x=0.093; (BOTTOM) SIMULATED SINUSOIDAL WAVE BASED ON THE VALUE OF x .......................................................................................................................... 122
FIGURE 6. 15 SPECTRA OF PC3 CELLS AFTER BEEN CORRECTED WITH THE OSCILLATION REMOVAL ALGORITHM. .......................................................................................................................... 123
FIGURE 6. 16 MEAN OF SELECTED SPECTRAL REGION (1200-1400cm⁻¹; 1700-1800cm⁻¹; 2800-3000cm⁻¹) IN RAW CORRECTED SPECTRA (TOP), 1ST DERIVATE SPECTRA (MIDDLE) AND 2ND DERIVATE SPECTRA (BOTTOM) OF PC3 (RED), LNCAP (BLUE) AND PNT2 (GREEN). SPECTRA WERE VECTOR NORMALISED AND MEAN CENTRED .......................................................................................................................... 125
FIGURE 6. 18 (LEFT) PCA PLOT FOR 1ST PC AGAINST 2ND PC OF 1ST DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE...
Figure 6. 17 (Left) PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCAP (blue dot) and PNT2 (green dot) cells with low mag setup. (Right) 1st and 2nd loadings correspond to each PC. .............................................................................................................................................................................................. 126

Figure 6. 19 (Left) PCA plot for 1st PC against 2nd PC of 2nd derivative spectra of PC3 (red dot), LNCAP (blue dot) and PNT2 (green dot) cells with low mag setup. (Right) 1st and 2nd loadings correspond to each PC. .............................................................................................................................................................. 126

Figure 6. 20 PC-LDA results based on the raw spectra (top), 1st derivative spectra (middle) and 2nd derivative spectra (bottom) of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines. ............................................ 127

Figure 6. 21 Importance of each wavenumber in the classifier of PC-3, LNCAP and PNT2 cell lines. The value of magnitude is averaged based on 100 times test.......................... 128

Figure 6. 22 Cell imaging based on the integrated lipid band (2800-3000 cm⁻¹) in the spectra of PC3 cell line with low mag setup (left) and high mag setup (right). The FOV for the images are 704x704 µm² (left) and 140x140 µm² (right) ........................................................................................................................................................................... 129

Figure 6. 23 PCA on spectra with different sequence of fitting process: the fitting-average process (blue circle) and the average-fitting process (red dot).................................................................................................................. 130

Figure 6. 24 Water corrected spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom) cells. The spectral region above 3000 cm⁻¹ were cut out and the remaining part were vector normalised and mean centred. .............................................................................................................................................. 131

Figure 6. 25 Mean spectra of PC-3 (red), LNCAP (blue) and PNT2 (green) cells. Spectra vector normalised based on the full spectra region and mean centred. ......................................................................................................................................... 132

Figure 6. 26 (Left) PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCAP (blue dot) and PNT2 (green dot) cells with high mag setup. (Right) 1st and 2nd loadings correspond to each PC. ........................................................................................................................................................................ 133

Figure 6. 27 (Left) PCA plot for 1st PC against 2nd PC of 2nd derivative spectra of PC3 (red dot), LNCAP (blue dot) and PNT2 (green dot) cells with high mag setup. (Right) 1st and 2nd loadings correspond to each PC. ........................................................................................................................................................................ 134

Figure 6. 28 PC-LDA results based on the raw spectra (top) and 2nd derivative spectra (bottom) of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines obtained from high mag setup.......................................................... 135

Figure 6. 29 ROC Curve of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines. ................................................. 136

Figure 6. 30 Mean of 2nd derivative spectra of PC3 taken from high mag (H) and low mag (L) setup .......... 136

Figure 7. 1 Raw spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom) cells in PBS solution taken from SR-FTIR spectrometer. Spectra were vector normalised and mean centred. .............................................................................................................................. 137

Figure 7. 2 Corrected spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom) cells taken from SR-FTIR spectrometer. Spectra were vector normalised and mean centred. Spectral region below 1000 cm⁻¹ was removed................................................................. 138

Figure 7. 3 Corrected spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom) cells with limited spectral region recorded from SR-FTIR spectrometer. Spectra were vector normalised and mean centred. 140

Figure 7. 4 Spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom) cells from different sample load taken from SR-FTIR spectrometer. Vector normalised and mean centred................................................................. 141

Figure 7. 5 Corrected spectra (top), 2nd derivative spectra (middle) and PCA results on spectra region of
1300-1500 cm\(^{-1}\) of PC3 cell line. .................................................................

Figure 7. 6 Corrected spectra (top), 2\(^{nd}\) derivate spectra (middle) and PCA results on spectra region of 1300-1500 cm\(^{-1}\) of LNCaP cell line. .................................................................

Figure 7. 7 Corrected spectra (top), 2\(^{nd}\) derivate spectra (middle) and PCA results on spectra region of 1300-1500 cm\(^{-1}\) of PNT2 cell line. .................................................................

Figure 7. 8 1\(^{st}\) loading of PCA results in Figure7.5-7.7 for PC3 (red), LNCaP (blue) and PNT2 (green) cells. .................................................................

Figure 7. 9 Selected spectral region in the mean 2\(^{nd}\) derivate spectra of PC3 (red), LNCaP (blue) and PNT2 (green) cells. Vector normalised and mean centred. .................................................................

Figure 7. 10 PCA plot for 2\(^{nd}\) derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1400 cm\(^{-1}\), 1700-1800 cm\(^{-1}\) and 2800-3000 cm\(^{-1}\). .................................................................

Figure 7. 11 PCA plot for 2\(^{nd}\) derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1380 cm\(^{-1}\), 1720-1800 cm\(^{-1}\) and 2800-3000 cm\(^{-1}\). .................................................................

Figure 7. 12 PC-LDA results based on 2\(^{nd}\) derivate spectra of PC3 (red), LNCaP (blue) and PNT2 (green) cell lines. .................................................................

Figure 7. 13 ROC Curve of PC3 (red), LNCaP (blue) and PNT2 (green) cell lines. .................................................................

Figure 7. 14 PCA plot for 2\(^{nd}\) derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1380 cm\(^{-1}\), 1720-1800 cm\(^{-1}\) and 2800-3000 cm\(^{-1}\). Raw spectra smoothed with 9 points smoothing function. .................................................................

Figure 7. 15 1\(^{st}\) loading in PCA results from Figure 7.11 and Figure 7.14. .................................................................

Figure 7. 16 PC-LDA results based on 2\(^{nd}\) derivate spectra of PC3 (red), LNCaP (blue) and PNT2 (green) cell lines. Raw spectra were smoothed with 9 points smoothing function. .................................................................

Figure 8. 1 Image (480x480 pixels with 1.33 micron pixel resolution) of cell in PBS solution with PMMA sphere taken from QCL microscope. (left) Image of integrated intensity. (right) Image of amide II peak intensity at 1550 cm\(^{-1}\). .................................................................

Figure 8. 2 Image of an extracted cell with surrounding region. The letters correspond to the position where a spectrum was taken. .................................................................

Figure 8. 3 Spectra taken from seven different position across the cell and the surrounding area in Figure 8.2. The colour of spectra represent the colour in the Figure 8.2. .................................................................

Figure 8. 4 Spectra of the PC3 cell (red), water (blue) and dried PMMA sphere (light blue) extracted from the QCL image. The spectra were vector normalised and mean centred. .................................................................

Figure 8. 5 Five different setup of windows. (A) two CAF\(_2\) square slides stack together (B) two BAF\(_2\) square slide stack together (C) a 25 mm CAF\(_2\) disc placed on a CAF\(_2\) slide (D) two 25 mm CAF\(_2\) disc stack together (E) a 13 mm CAF\(_2\) disc placed on a 25 mm CAF\(_2\) disc. .................................................................

Figure 8. 6 Spectra taken from different combination of IR windows. .................................................................

Figure 8. 7 Spectra of the same pixel in the image of same location in the same CAF\(_2\) window. .................................................................

Figure 8. 8 Plot of signal level received by the detector and the output power of laser source. The blue line and the light blue dots represent the target and actual power level of the laser. The red line stands for the target signal level. The green, black, purple and pink dots explain the signal level of four background image taken with 5 minutes interval. .................................................................

Figure 8. 9 (top) 200 spectra of PC3 cells in PBS solution taken from the QCL microscope. (bottom) 200 spectra of PC3 cells directly corrected with the new water removal method. .................................................................
FIGURE 8.10 200 SPECTRA OF PC3 CELLS SMOOTHED WITH 13 POINTS SAVITZKY-GOLAY FILTER. THE ORIGINAL SPECTRA WERE THE SPECTRA IN FIGURE 8.9 (BOTTOM). ................................................................. 178

FIGURE 8.11 SPECTRA OF WATER WITH 10MICRON (BLUE) AND 5MICRON (RED) PATH LENGTH RECORDED FROM QCL MICROSCOPE. THE NUMBER ABOVE EACH SPECTRUM IS THE STANDARD DEVIATION AND RANGE OF SPECTRA REGION FROM 1100 TO 1500CM^{-1}. ................................................................. 179

FIGURE 8.12 (TOP) SPECTRAL REGION OF 1000~1600CM^{-1} IN THE AVERAGED SPECTRA OF WATER ACROSS THE HORIZONTAL DIRECTION IN THE IMAGE. (BOTTOM) SPECTRAL REGION OF 1000~1600CM^{-1} IN THE AVERAGED SPECTRA OF WATER ACROSS THE VERTICAL DIRECTION IN THE IMAGE ........................................................................... 181

FIGURE 8.13 NOISE CORRECTION PROCESS. (TOP LEFT) THE SURROUNDING AREA OF THE TARGET CELL WAS CHOSEN. (TOP RIGHT) THE SPECTRA IN THIS REGION WAS FITTED WITH A NOISE FREE REFERENCE TO OBTAIN THE NOISE SPECTRUM (BOTTOM RIGHT). (BOTTOM LEFT) THE NOISE SPECTRUM WAS SUBTRACTED FROM THE SPECTRA OF CELL. ............ 182

FIGURE 8.14 RAW SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS IN PBS SOLUTION RECORDED BY QCL MICROSCOPE. SPECTRA WERE VECTOR NORMALISED AND MEAN CENTRED. .................................................. 184

FIGURE 8.15 NOISE CORRECTED SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS IN PBS SOLUTION RECORDED BY QCL MICROSCOPE. SPECTRA WERE VECTOR NORMALISED AND MEAN CENTRED. ............ 185

FIGURE 8.16 WATER CORRECTED SPECTRA OF PC-3 (TOP), LNCAP (MIDDLE) AND PNT2 (BOTTOM) CELLS IN PBS SOLUTION RECORDED BY QCL MICROSCOPE. SPECTRA WERE VECTOR NORMALISED AND MEAN CENTRED. ............ 186

FIGURE 8.17 (TOP) MEAN SPECTRA OF PC-3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELLS AFTER WATER CORRECTION. (BOTTOM) PCA PLOT FOR 1ST PC AGAINST 2ND PC OF THE ORIGINAL SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS ......................... 188

FIGURE 8.18 (TOP) MEAN 1ST DERIVATE SPECTRA OF PC-3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELLS AFTER WATER CORRECTION. (BOTTOM) PCA PLOT FOR 1ST PC AGAINST 2ND PC OF 1ST DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS ......................... 189

FIGURE 8.19 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF SPECTRAL LIMITED 1ST DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS ......................... 190

FIGURE 8.20 (TOP) MEAN 2ND DERIVATE SPECTRA OF PC-3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELLS AFTER WATER CORRECTION. (BOTTOM) PCA PLOT FOR 1ST PC AGAINST 2ND PC OF SPECTRAL LIMITED 2ND DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS. .... 191

FIGURE 8.21 PC-LDA RESULTS BASED ON THE 1ST DERIVATE (TOP) AND 2ND DERIVATE (BOTTOM) SPECTRA OF PC3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELL LINES. ................................................................. 192

FIGURE 8.22 AVERAGED IMPORTANCE OF EACH WAVENUMBER IN THE CLASSIFIER OF PC-3, LNCAP AND PNT2 CELL LINES BASED ON 100 TIMES TEST ................................................................. 194

FIGURE 8.23 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF 1ST DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS. THE ORIGINAL SPECTRA WERE SMOOTHED WITH 9 POINTS SMOOTHING FUNCTION. ........................................................................ 195

FIGURE 8.24 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF 2ND DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS WITH THE CORRESPONDING LOADINGS. THE ORIGINAL SPECTRA WERE SMOOTHED WITH 9 POINTS SMOOTHING FUNCTION. ........................................................................ 196

FIGURE 8.25 PC-LDA RESULTS BASED ON THE 1ST DERIVATE (TOP) AND 2ND DERIVATE (BOTTOM) SPECTRA OF PC3 (RED), LNCAP (BLUE) AND PNT2 (GREEN) CELL LINES. THE ORIGINAL SPECTRA WERE SMOOTHED WITH 9 POINTS SMOOTHING FUNCTION. ........................................................................ 197

FIGURE 8.26 AVERAGED IMPORTANCE OF EACH WAVENUMBER IN THE CLASSIFIER OF PC-3, LNCAP AND PNT2 CELL LINES BASED ON 100 TIMES TEST. THE ORIGINAL SPECTRA WERE SMOOTHED WITH 9 POINTS SMOOTHING FUNCTION. . 198
FIGURE 9.2 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF RAW SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS FROM FTIR HIGH MAG SETUP. (RIGHT) 1ST AND 2ND LOADINGS CORRESPOND TO EACH PC. ... 208

FIGURE 9.1 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF RAW SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS FROM FTIR LOW MAG SETUP. (RIGHT) 1ST AND 2ND LOADINGS CORRESPOND TO EACH PC. ... 208

FIGURE 9.3 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF RAW SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS OBTAINED FROM SR-FTIR SPECTROMETER. (RIGHT) 1ST AND 2ND LOADINGS CORRESPOND TO EACH PC. ........................................................... 209

FIGURE 9.4 PCA PLOT FOR 1ST PC AGAINST 2ND PC OF RAW SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS OBTAINED FROM QCL MICROSCOPE. (RIGHT) 1ST AND 2ND LOADINGS CORRESPOND TO EACH PC. ........................................................................................................ 209

FIGURE 9.5 PC-LDA RESULTS ON 2ND DERIVATE SPECTRA OF PC3 (RED DOT), LNCAP (BLUE DOT) AND PNT2 (GREEN DOT) CELLS OBTAINED FROM FTIR LOW MAG SETUP (TOP LEFT), FTIR HIGH MAG SETUP (TOP RIGHT), SR-FTIR SPECTROMETER (BOTTOM LEFT) AND QCL MICROSCOPE (BOTTOM RIGHT). ...................................................... 210

FIGURE 9.6 AVERAGED IMPORTANCE OF EACH WAVENUMBER IN THE CLASSIFIER BUILT FROM THE SPECTRA RECORDED BY FTIR SPECTROMETER WITH LOW MAG SETUP (ROW 1), FTIR SPECTROMETER WITH HIGH MAG SETUP (ROW 2), SR-FTIR SPECTROMETER (ROW 3) AND QCL MICROSCOPE (ROW 4). ............................................................... 213
Abstract

- University of Manchester
- Zhe Zhang
- Doctor of Philosophy in the Faculty of Science & Engineering
- “Evaluation of infrared QCL, Synchrotron and bench-top sources for cell imaging in aqueous media”
- 27th September 2106

Live cell imaging with FTIR spectroscopy offers a high throughput, non-damage and lab-free method to study the cells in vivo which has significant advantages in the field of cancer diagnosis and drug screening. However, due to the strong absorbance of water, using infrared spectroscopy in such field remains to be an underdeveloped topic. This project demonstrates a novel method to perform IR imaging of cells in solution. A novel water correction method, which avoids the using of water combination band, is proposed. A buffer reference and a cell reference spectra were introduced to fitting the contribution based on protein bands. This method was implemented on three types of IR spectrometers, namely conventional FTIR spectrometer, synchrotron-based FTIR spectrometer and quantum cascade laser (QCL) microscope. To date, most of the live cell imaging carried out with IR sources utilise synchrotron radiation. Recently, a new bench top system, QCL microscope, has been developed. It incorporates four tunable QCL laser sources covering the wavenumber range 900–1800 cm⁻¹ which are many orders of magnitude brighter than conventional sources. The proposed water correction method is, therefore, capable of processing the data recorded by all three types of IR spectrometers. Three prostate cancer cell lines were employed to evaluate the water correction method and the performance of three spectrometers on imaging of cell in solution. The obtained spectra was analysed with multivariate analysis, PCA and PC-LDA which shows good separation between cell lines. The data was also examined with Random Forest algorithm to establish a classifier and the diagnostic capability of the water corrected spectra was proven.

Keyword: QCL; FTIR; SR-FTIR; cell; water; in vivo; PC-LDA; random forest; classifier;
Declaration

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

Zhe Zhang
Copyright Statement

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

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

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

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

First of all, I would like to thank my parents for both the financial and emotional support during my four-year study. Without your encouragement, I would not accomplish this honour.

Second, I would like to thank my supervisor Professor Peter Gardner, who guild me through the study and lead me to my PhD degree.

Third, I would like to give heartfelt thanks to people in my group: Michael Pilling and Caryn Hughes, who taught me the knowledge and shared their experience on Spectroscopy; Alex Henderson for the tutorial discussion on Matlab programming. In particular, Michael Pilling, who has been an inspiration to work with and has great teamwork mentality.

I would also like to thank Gianfelice Cinque for letting me use the facility at Diamond synchrotron site; Daylight Solutions (USA) for the lending of QCL microscope.

Finally, I would like to thank my lovely wife Yanjie Li for the company and love. I apologise for the grumpiness during the writing of the thesis.
1. Introduction

To date, most of the cancer studies in IR spectroscopy field were focused on fixed tissue biopsy. The produce of fixed tissue cut evolves both physical and chemical changes in the sample, such as ethanol dehydration, formalin flush and paraffin infiltration [1]. One way of avoiding such changes is performing single cell study, especially live cell imaging [2]. Prostate and bladder cancer cells were found in the urine at the early stage of diagnosis [3]. Detecting these cancer cells with IR spectroscopy could potentially establish a fast, accurate and pain-free diagnosis method.

Live cell imaging with IR microspectroscopy offers a high throughput, non-damage and lab-free method to study the cells in vivo which has significant advantages in the field of cancer diagnosis and drug screening. However, due to the strong absorbance of water, using infrared spectroscopy in such a field remains an underdeveloped topic. Different groups have developed multiple water correction methods to remove water contribution in the spectra [4-6]. However, each method has its limitations. Also, to obtain a signal through the water layer, the IR source needs to be strong enough to be detected. Thus, to date, the most suitable IR source for live cell imaging is synchrotron radiation. Recently, a new bench top quantum cascade laser (QCL) microscope system has been developed, which incorporates four tunable QCL laser sources covering the wavenumber range 900–1800 cm$^{-1}$, which are many orders of magnitude brighter than conventional sources [7]. Therefore, establishing a water correction method and couple with the new QCL microscope could potentially create a fast, extensive and even portable diagnosis method.

In this thesis, a novel water correction method which can process the spectra obtained from the three types of IR spectrometer, namely conventional Fourier Transform IR (FT-IR) spectrometer, synchrotron radiation FTIR (SR-FTIR) spectrometer and QCL microscope, will be proposed. The previous water correction methods computer the
water proportion by fitting the water combination band. However, this spectral region currently is not available in the QCL microscope. The new method avoids the requirement of water combination band and removes the water contribution based on the protein ratio. Therefore, this water correction method can be employed to evaluate the performance of imaging of cell in an aqueous environment with all three types of IR spectrometers. Three prostate cancer cell lines were measured to assess the capability of the new correction method. The obtained spectra were analysed with multivariate analysis method, such as PCA and random forest classifier algorithm.

In the beginning few chapters, first, IR spectroscopy and the instrumentation of different IR spectrometers will be discussed. Second, IR spectroscopy of cells and the water barrier to performing live cell imaging will be discussed. The previous studies on live cell imaging using different method and spectrometers will be reviewed. Third, the experimental methods for cell imaging on these three type of spectrometers and the data analysis method will be reported.

The results will be demonstrated in order of the development history of IR spectrometers. First, the conventional FTIR spectrometer will be used as a standard to establish and examine the new water correction method. The key parameter of the method will be discussed in detail. Second, the SR-FTIR spectrometer will be used to test the new method under an optimised light source condition as the theory behind these two spectrometers is the same. Third, the new water correction method will be used to analyse the spectra obtained from the newly developed QCL microscope. By comparing the results from all three spectrometers, the feasibility of the new water correction and the cell imaging performances can be revealed.
Reference


2. Theories and instruments

2.1 Spectroscopy

Spectroscopy is a subject which studies the interaction between matter and electromagnetic radiation. These interactions are classified by the energy levels involved and the mechanism by which energy gets transformed. Two methods that are growing in their popularity for the analysis of biological and biomedical samples are infrared and Raman spectroscopy. In this thesis, different instrumental and experimental methods of infrared spectroscopy will be used to study biological cells in aqueous media.

2.1.1 Infrared

Infrared spectroscopy studies the interaction specifically associated with light with the characteristic wavelength range from approximately 1 μm to 1 mm (or wavenumber from 10000 cm\(^{-1}\) to 10 cm\(^{-1}\)). The light within this region interacts with the vibrations of molecules and the associated change of the dipole moment. It is especially useful in detecting chemical information and specifically identifying the chemical compound. Importantly, infrared radiation contains much lower energy than other electromagnetic probes such as visible, UV or x-rays, and as such can be considered a non-destructive method of analysis particularly suitable for the study of sensitive biological systems.

2.1.2 Vibrations

A molecule can vibrate in a specified amount of modes depending on the number of atoms in the molecule. However, only part of these vibrations will interact with infrared light. If a vibration does not change the dipole moment of the molecule, it does not absorb any infrared light and therefore it is not infrared active. Figure 2.1 shows a list of vibrational modes for a -CH\(_2\) group, which generally can be detected by infrared spectroscopy.
2.1.3 Morse potential

In quantum physics, the energy of a system is not continuous; the system can only exist in certain allowed discrete energy states. For a diatomic molecule, which can be considered as two masses connected by a spring exhibiting simple harmonic motion, the solution of the Schrödinger equation shows that the only allowed energy states are given by

\[ E(\nu) = \left( \nu + \frac{1}{2} \right) \hbar \nu \]  

(Eq. 1)

Where \( E \), \( \nu \), \( \nu \) and \( h \) are the energy, energy level, frequency and Planck constant. This results in a series of equally spaced energy levels within the parabolic potential energy well. In a real system, however, it is easier to stretch the bond than compress it and indeed it is possible for the bond to break, leading to dissociation of the molecule. The vibrational motion of a real diatomic is, therefore, anharmonic leading to a gradual decrease in the separation of the energy levels at higher energies, as illustrated in Figure 2.2.

---

\[ \text{Symmetrical Stretch} \]
\[ \text{Asymmetrical Stretch} \]
\[ \text{Rocking} \]
\[ \text{Wagging} \]
\[ \text{Scissoring} \]
\[ \text{Twisting} \]

Figure 2.1 Schematic diagrams showing the vibrational mode of \(-\text{CH}_2\) group
2.1.4 Vibrational-Rotational coupling

The molecule does not only vibrate, but it can also rotate if it is unhindered in the gas phase. The rotations give the molecule some small moment of inertia which will affect the vibration in a process known as [1]. Since the rotational levels are also quantized, this coupling results in the appearance of absorbance lines on each side of the characteristic vibrational peak. Depending on the polarisation of the molecule, the vibrational-rotational coupling will lead to either just P (low wavenumber side) and R (high wavenumber side) or PQ and R branches, whereby the Q branch appears as a series of lines essentially superimposed on the centre position of the vibrational frequency. A vibrational-rotational spectrum of CO exhibiting P and R branches is shown in Figure 2.3. Although vibrational-rotational coupling is not observed in the liquid and solid phase and is therefore not relevant to most biological spectra ambient

![Figure 2.2 Comparison of Morse potential and harmonic oscillator. (Da)Dc are (actual) dissociation energy.](image)

![Figure 2.3 Vibrational-rotational coupling on spectrum of CO](image)
gas phase, water vapour and carbon dioxide are a persistent problem in any infrared experiment and strategies need to be employed to remove these contributions from the observed spectra. This normally means purging the system and possibly digital removal in a post-processing step [2].

2.1.5 **Beer–Lambert law**

When measuring a liquid sample in a cuvette, it is found that a proportion of the light is absorbed by the sample. The amount of light absorbed by the sample is directly proportional to the concentration of the sample, and the path length of the light through the sample. These two conditions are known collectively as the Beer-Lambert Law. For a homogeneous sample, the absorbance can be expressed by:

$$A = \epsilon cl$$  \hspace{1cm} (Eq. 2)

where $A$ is absorbance, $\epsilon$ is extinction coefficient, $c$ is molar concentration and $l$ is the pathlength of light. The extinction coefficient is a function of wavelength (or wavenumber). Therefore, the difference in extinction coefficient leads to the peaks in the measured spectrum. For a heterogeneous sample, the absorbance is defined as:

$$A = \sum_{i=1}^{N} \epsilon_i c_i l_i$$ \hspace{1cm} (Eq. 3)

where $N$ represents the number of species in the sample. Thus, the total absorbance is the sum of individual absorbance of each species.

The Beer-Lambert law strictly only holds in an ideal situation at low concentrations. In a real situation, there are many ways in which deviations from the Beer-Lambert law may be observed. This can occur at high concentrations, or where there are strong changes in refractive index leading to scattering within the sample. For analysis of single cells, these effects can be significant. So although the Beer-Lambert law may represent a good starting point for analysis, it should be used with caution.
2.2 Spectrometer

The instrument that is used to measure the difference in intensity of light as a function of the wavelength (or wavenumber) is known as a spectrometer. In the field of infrared spectroscopy, the spectrometer has a long history. The first commercial infrared spectrometer, the dispersive infrared spectrometer, came out on the 1940s [1]. In the 1960s, the more advanced Fourier transform infrared (FT-IR) spectrometer was developed, and this gained popularity in the 1980s with the advent of small dedicated computers that could rapidly carry out the Fourier transform of the measured interferogram [1]. Since then, there have been no major changes in FTIR spectrometer design, and it is this basic system that it still in use today. Although the spectrometer (based on the Michelson interferometer) has largely remained the same, the use of synchrotron radiation infrared sources coupled with dedicated infrared microscopes have facilitated bringing IR spectroscopy largely from the chemical domain to the biological and more specifically the biomedical domain [3]. The high brightness of the SR IR source enabled measurement of single biological cells for the first time.

Around 2010, another new infrared source was developed that has the potential to revolutionise the field of infrared spectroscopy [4]. The advent of tunable Quantum Cascade Lasers (QCLs) meant that a very high powered bench-top source was available to infrared spectroscopists. This tunable, high-powered, monochromatic source means that scanning through the frequency spectrum does not suffer the disadvantage of old dispersive instruments since no slit is required to define the measured wavelength, and thus there no subsequent loss of power. This also means that if you know the key peak or peaks of interest, you can selectively measure these without measuring the full spectrum; something that cannot be done with FTIR. In this thesis, these three techniques are compared.
### 2.2.1 Dispersive infrared spectrometer

![Diagram of double-beam dispersive infrared spectrometer](image)

Figure 2.4 shows a sketch of the double-beam dispersive infrared spectrometer. By selecting the beam with a grating mirror, the light can be dispersed into single wavelength component. The resolution is defined by a slit or series of slits which have the disadvantage of seriously reducing the signal intensity arriving at the detector. The radiation passes through a beam splitter and forms two beams; one of which interacts with the sample, and the other one does not. The second beam acts as a reference i.e. a beam that contains all of the instrument response function and any contributions from atmospheric water vapour and CO₂. These two beams periodically and alternately pass through the chopper and hit the detector. The detector then measures the intensity of chosen light and ratio of the two signals produces a transmission spectrum. This system is eventually out of date due to the slow measurements as it can only measure one wavenumber at a time. However, the capability of measuring the background and sample at the same time is still a significant advantage.

### 2.2.2 FT-IR spectrometer

The FT-IR spectrometer, currently the most commonly used type of IR spectrometer, is based on a Michelson interferometer. There are different types of FT-IR spectrometer, but the basic theories behind them are similar.
Michelson interferometer

The Michelson interferometer is a device which separates the beam and adds a path difference to the beams before recombining them. The path difference will induce interference into the beam and lead to a variation of intensity which can be measured by the detector.

Figure 2.5 shows a simplified Michelson interferometer. An interferometer consists of a beam splitter, a fixed mirror and a movable mirror. The incident light is divided into two beam by beam splitter. Those two beams of light meet the two mirrors and reflect back to the beam splitter where half is recombined and directed to the sample, and the other half is directed back to the source. By moving the movable mirror, a retardation (path length difference) will be created for the two separated beams, and the light at a particular wavelength will experience either constructive or destructive interference depending on the path length difference. In this way, an interferogram is produced and can be detected by the detector.
Fourier Transform

The interferogram generated by the detector is in the time domain. It need be translated into the frequency domain to be analysed further.

The mathematical tool used to perform this translation is known as Fourier transform. It breaks down the wave in the time domain (interferogram) into a series of sinusoidal waves and then projects the amplitude of those waves to the correspond frequency in the frequency domain. The transform can be expressed by [5]:

\[
\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x) e^{-i2\pi x \xi} \, dx
\]  
(Eq. 4)

where \( \xi \) represents the frequency and \( x \) stands for time. The inverse transform to compute from frequency domain to time domain is written by:

\[
f(x) = \int_{-\infty}^{+\infty} \hat{f}(\xi) e^{i2\pi x \xi} \, d\xi
\]  
(Eq. 5)
Light source for mid-infrared

To obtain the light with a particular frequency, a specific light source is required based on the desired frequency range. In this project, the light sources which produce radiation in the mid-infrared region will be discussed. For most of FT-IR spectrometers, a heated silicon carbide rod, known as a Globar, is used as a mid-infrared light source. Globar is a non-ideal thermal blackbody [1]. When the Globar is heated up to 1000~1500°C, it will start to transform the energy from heat into IR radiation. The radiation frequency of Globar can be as large as 50-6000 cm⁻¹. The output intensity has a distribution in which the highest intensity is located at around 2500 cm⁻¹. Advanced mid-infrared light sources, such as synchrotron radiation and semiconductor lasers will be discussed in chapter 2.2.4 and 2.2.5.

Detectors

There are two main types of IR detector, namely thermal detectors and quantum (photon) detectors. Both of them are commonly used in a FT-IR spectrometer.

The thermal detector senses the changes in temperature in the material by measuring the change in resistance of the sensing element. The detector has different names depend on the sensing elements, but is normally known as a bolometer or thermopile. The most commonly used detector is deuterated l-alanine-doped triglycine sulphate (DLATGS) detector. This type of detector is limited by two aspects. First, sensing the temperature changes requires a large amount of energy to obtain an appropriate signal to noise ratio. The intensity of a normal Globar source is weak and therefore cannot couple with the thermal detector well. Second, the sensing of temperature change results in a response time of few milliseconds and therefore a sampling frequency of less than 1000Hz [1]. Compared to a spectrometer with a higher sampling frequency of 2500Hz or 5000Hz, the time required for the measurement is significantly longer. On the other hand, the low manufacture cost and no operation cooling required still makes it a great IR detector [1].
The quantum detectors normally have two operation mode, namely photoconductive (PC) mode and photovoltaic (PV) mode. The PC mode detects current generated by electrical conductivity. The PV mode detects the current generated by the voltage difference across a p-n junction. The PC mode has a higher frequency response and signal to noise ratio compare to PV mode. For both modes, the generated current/voltage is proportional to the intensity of light shining on the detector. The *mercury cadmium telluride* (MCT, HgCdTe) detector is one of the quantum detector operating in PC mode.

Due to the low energy of infrared photons, the detected signal suffers interference from noise which generated by the sensing material itself from random thermal energy. The cooling system is introduced to prevent the detector generating this type of noise. The MCT detector is usually cooled to 77k by liquid nitrogen (LN₂) to obtain high sensitivity.

### 2.2.3 Infrared microscopy

An infrared spectrometer itself is limited by the spatial resolution. The recorded spectra can only present the average biochemical and structural information of the illuminated area. The large projected area of the IR radiation source also requires a large size sample or the intensity of the recorded spectrum is attenuated. The implementation of the microscope with IR spectrometer offers the capacity to measure the molecular vibrations in a restricted area. The first IR microscope was demonstrated in 1949[6]. However, for the following decades, the application of IR microscope was restricted by the slow scan rate. The situation changed after the development of FTIR spectrometer.

The FTIR spectrometer coupled with a microscope were used to examine biochemical samples including cells, such as blood cells[7] and stem cells[8], and tissues, such as breast[9], prostate[10] and lung[11]. Through these studies, FTIR microspectroscopy has shown its advantages in cancer research and clinical diagnosis field [3].
The use of all reflecting optics

Mid-infrared spectroscopy focuses on radiation with a wavelength from 2 to 10 μm. This results in a limitation of optical components. The normal refracting objectives contain glass which is not transparent in the region of interest.

IR microscopes have a similar structure as conventional visible light microscopes. The main difference is the microscope does not contain any glass components. The microscopes utilise Cassegrain-type reflecting objectives, which are known as Schwarzschild objectives, to establish the light path. Figure 2.7 shows a sketch of a Schwarzschild objective.

To measure a sample in transmission mode, two Cassegrain objectives are used as a pair, the second known as a condenser. The mirror in the objectives also reflects visible light so the sample can be visually located and focused by set the visible light source and IR light source in the same light path.

Focal-plane-array (FPA) detector

In the first quarter century history of IR spectrometers, only single element detectors were required, since the norm was to measure one spectrum per sample. The advent of the infrared microscope, however, changed the nature of the questions being asked. The microscope enabled spectral information to be coupled to spatial information.
meaning that chemically inhomogeneous samples could be analysed. In order to measure the spatial distribution, the sample needs to be mapped in a raster fashion point by point. Each point gives an average spectrum in a region defined by the size of the aperture, which will also affect the resolution and mapping speed [12]. This process is extremely time-consuming and detailed mapping of large areas is unfeasible using this method.

Developed by the military for use in heat seeking missiles, the first MCT multi-element detector was built in the 1970s. After years of development, (and partial declassification by the military), the first commercial FPA microspectrometer came out at 1995 [13].

Most of the FPAs consist of quantum sensors either indium antimonide (InSb) or MCT, which are usually cold-shielded and LN$_2$ cooled for better performance. These arrays are either linear or square with arrays ranging from 1x16 to 256x256 elements. The size of each element is of the order of microns, and therefore the array can be less than 1 mm$^2$ in size.

At the same time, other FPAs which employ thermal detectors have been developed as well [14]. These FPAs are known as microbolometers. The manufacturing cost of microbolometer is much lower compared to quantum-based devices, and the operation does not require cooling, thus reducing running costs. However, this type of FPA faces the same problem of the thermal detector which has a long response time and needs a powerful light source due to low sensitivity. For the same light source, the noise for microbolometer can be two orders of magnitude larger than a LN$_2$ cooled MCT FPA system [4].

The use of FPAs enables the simultaneous measurement of multiple points in the target area. This is known as hyperspectral imaging, and the resultant array of spectra is named a spectral-hypercube [15]. This is a much more efficient method of obtaining image compared with mapping samples with single point detector.
**Signal to Noise Ratio**

When performing experiments, there is always noise present in every measured spectrum. The source of the noise depends on the particular instrument used for the measurement. The most common method to evaluate the effects of the noise on different platforms is to compare the ratio between measured signal and noise, known as the signal to noise ratio (SNR) [1].

\[
SNR = \frac{P_{signal}}{P_{noise}} \quad (Eq. 6)
\]

where \(P\) stands for the power (intensity) of signal or noise. In an FTIR spectrometer, the power of signal or noise can be expressed by absorbance (amplitude) of the spectrum. Thus, the equation can be rewritten by:

\[
SNR = \left( \frac{RMS_{signal}}{RMS_{noise}} \right)^2 \quad (Eq. 7)
\]

where \(RMS\) is the root mean square of the amplitude.

The noise in an FTIR spectrometer is generated by different components and experimental setup factors. Most of the noise is random and can be removed by signal averaging. The SNR of FTIR spectrometer can be estimated. The relation between SNR, resolution and measurement time is known as spectroscopic trading rules [1]. If the resolution increases by a factor of \(N\), for the same measurement time, the SNR reduces by a factor of \(N\). The spectral resolution is proportional to the travel distance of interferometer mirror. Thus, if the resolution increases by a factor of \(N\), for the same measurement time, the SNR reduces by a factor of \(N\). For the same resolution, taking longer measurement time or increasing the number of interferograms co-added by a factor of \(N\), the SNR can be improved by \(\sqrt{N}\). Thus, to achieve the same SNR ratio for 2 times the resolution, the measurement time or number of scans need be 4 times more.

For an imaging system, the SNR is measured in each sensor and therefore affected by the number of sensors in the detector. For detectors with the same size, the more elements in a detector, the less the size of each single elements. Thus, if the number of
the sensor element in a detector is increased by a factor of \( N \), the SNR for each sensor is reduced by \( N \).

**Diffraction effects**

Diffraction happens when a wave, such as sound, water or electromagnetic wave, meets an obstacle. For modern IR microscopes, two types of diffraction need be considered.

*Diffraction of the aperture*

For a microscope using a single point detector, an aperture is used to confine the illumination region on the sample. However, the light with a wavelength larger than aperture size is strongly diffracted upon passing through the aperture [1]. This results in severe loss of signal as the aperture size approaches the diffraction limit.

*Diffraction limit of microscope*

The spatial resolution of a microscope system depends on the numerical aperture of the lens and the specific frequency of light. The diffraction limit \( d \) can be expressed by:

\[
  d = \frac{0.61\lambda}{NA}
\]

\((Eq. 8)\)

where \( d \) is the spatial resolution, NA is numerical aperture and \( \lambda \) is the wavelength of the light. Most of the objectives in IR microscope system have a numerical aperture about 0.6. Therefore, the achievable resolution is approximately the wavelength of the radiation. Some advanced lenses have NA of 0.81 which gives a spatial resolution smaller than the actual wavelength.

The microscope using an FPA is not affected by the aperture; generally, a large area of the sample is illuminated by the source when integrated with a microscope, the FPA pixel resolution can be from \(~20\) microns to less than one micron, depending on the objective used and coupling optics to the detector. With pixel resolution significantly greater than spatial resolution, oversampling can be achieved, meaning that the FPA detector is able to achieve the diffraction-limited spatial resolution.
It is important to note that the diffraction limit normally relates to the Rayleigh criterion which is valid for two points of the same spectral output. If the objects to be imaged and resolved have significantly different spectral output, then spatial resolution exceeding the diffraction limit can be achieved [16].

**Scattering effects**

When light passes through matter, the original trajectory may change due to the heterogeneity of the matter; this change in trajectory is known as scattering. There are three types of scattering which depend on the size of the particle which is causing the scattering. When the size of the particle is far smaller than the wavelength of light, the scattering can be described as Rayleigh scattering [17]. Conversely, the geometric scattering happens when the size of particle is far larger than the wavelength of light. When the size of particle is similar to the wavelength, the scattering is explained by Mie scattering.

Most of the biochemical samples consist of cells. The general size of cells varies from 8 to 30μm, and the organelles in the cell range from <1 to 10 μm. This size range is perfectly matching the wavelength of IR radiation and therefore causing intensive Mie scattering [18].

Mie scattering will reduce the light passing through the sample and subsequently reaching the detector, thus resulting in non-chemical apparent absorbance in the spectrum [19]. The broad undulating baseline in the spectra of the oral cell has been reported and related to Mie-type scattering by B.Mohlenhoff [20]. Later, M.Romeo demonstrated the large variance in the spectrum of the epithelial cells in the resting stage [21]. The cell should be relatively homogeneous, and therefore the variance is dominated by Mie-scattering.

Mie-scattering affects the spectrum of a biochemical sample in three different ways. First, the most commonly observed effect is the broad fluctuating baseline. The baseline in the region of hydrogen stretching bands is raised up while the region of fingerprint region is lower down. A bridge-like curved baseline in the non-biochemical
region from 1800 to 2800 cm\(^{-1}\) can be noticed as well. Second, a significant intensity drop was found at 1700 cm\(^{-1}\). This results in a derivative-like amide I band which known as ‘dispersion artefact’ [19]. These effects significantly alter the peak intensity of amide I and II bands, which present the secondary structure of proteins [22]. Third, the scattering leads to shifts in the peak position. All those effects are dominated by the physical property of the sample which makes the results concluded from these spectra inaccurate.

The scattering effects in the spectra can be overcome experimentally and mathematically. The scattering results from the change of refraction index. Therefore, the scattering can be avoided by embedding the sample with paraffin or placing in an aqueous environment to obtain a matching refraction index [19]. However, the wax itself has a huge absorbance peak in the lipid region which leads to complexity or sacrifice in the analysis. Water is also a strong absorber of IR light which overlaps with most of the biochemical peaks.

Different research groups have successfully created different mathematic approaches to remove this type of scattering. A.Kohler et al. demonstrated a scattering correction approach based on Extended Multiplicative Signal Correction (EMSC) algorithm[23]. The correction fits the target spectra with a large number of possible Mie-scattering curves obtained from different particle sizes and refractive indices. The corrected spectra show a normal baseline, but the derivative-like amide I band has not been removed. To explore the physics behind the Mie-scattering, P.Bassan et al. studied the polymethyl methacrylate (PMMA) microspheres with different size [19]. The sample is chemically homogeneous. Later, P.Bassan et al. reported a resonant Mie scattering correction method which modified the EMSC method by adding a scattering term [24]. By comparing the raw spectrum with a scattering-free reference spectrum, the real and imaginary part of the refractive index was estimated. This refractive index was then used to build up a database of 1000 possible scattering curves. After being decomposed by PCA analysis, the curves were then used to fit the target spectra. The corrected
spectra were then used as the reference spectra and compared with the raw spectrum. Multiple iterations of the above steps were carried out until the resultant spectra are constant. The correction method successfully eliminates all the scattering effects and, especially, reveal the true position of amide I peak.

**Fringing**

When light travels through a thin layer with two or more parallel surfaces, depending on the angle of incidence, a path length difference may be generated by internal reflection within the layer. This pathlength difference leads to constructive and destructive interference of light, which leads to a small but measurable “second centre burst” in the interferogram [25]. This mini spike in the interferogram transforms into a sinusoidal oscillation in the frequency domain spectrum, and this is known as a fringing effect. The most common fringing occurs when measuring biochemical sample is from either the sample itself, when it is in the form of say a thin biofilm, or from the IR windows which support the sample.

As noted above, the fringing is generated from the path length difference which directly relates to the thickness of sample or window. The relation between number of oscillations and the thickness of parallel layer can be described by [1]:

$$d = \frac{10 \times N}{2 \times (\nu_1 - \nu_2)}$$

(Eq. 9)

Where $d$ is the thickness of layer, $\nu_1$ and $\nu_2$ are two selected wavenumbers, $N$ is the number of oscillations in the region between $\nu_1$ and $\nu_2$. Note, that this is strictly for measurements at normal incidence, but can be used to give an indication of the layer causing the problem even in a microscopy experiment.
2.2.4 Synchrotron light source

A synchrotron is the combination of a particle accelerator and an electron storage ring [26]. The electrons from an electron gun are accelerated, normally first in a linear accelerator and then a booster ring. Subsequently, the electrons are injected into a vacuumed storage ring. After being accelerated to near the speed of light, the relativistic electron will produce electromagnetic radiation when changing velocity.

In first and second generation storage rings, bending magnets were used to alter the trajectory of electrons and extract the light from storage ring. Subsequently, in third generation storage rings, insertion devices were used as the main way of extracting radiation. An insertion device (named as such because they are inserted into the straight sections of the ring) are, a set of dipole magnets was periodically arranged which force the electron to move sinusoidally, through the device.

There are two types of insertion device, a wiggler and an undulator. In a wiggler, the electron moves in a number of arcs and radiation is emitted at each one. As in the case of a bending magnet, the radiation is considered to be broad band covering a wide energy range. The intensity scales with the number of dipoles, so can be many times that of the bending magnet. In the case of an undulator, the arcs are much shallower.
than a wiggler and, as such, the apparent source size is small and, most importantly, the radiation emitted from one arc can interfere with that of another producing very high peak power over a narrow energy range [27].

The extracted synchrotron beam contains the light of a wide frequency from X-rays to microwaves. A water cooled beam stop is applied to block the X-rays and overcome the heat generated from X-rays. Before entering the collimation system, a diamond window is used to filter the light.

A synchrotron IR source is based exclusively on a bending magnet. The reason for this is that wigglers and undulators are extremely expensive and give little benefit in the IR since they are generally optimised for x-rays. Bending magnets, on the other hand, are necessary for the integrity of the ring, and since they are not in demand for X-rays, they are readily available for IR. The IR beam from a synchrotron bending magnet is still hundreds of times brighter than a Globar source [28]. Although the total power output is not much different, in the case of the synchrotron the IR radiation is emitted in a very narrow angular cone. The collimation angle of radiation depends on the energy (speed) of the electron, the higher the energy of the electron, the smaller the spread angle. A typical size of synchrotron IR beam varies from 10 to 20 microns, which means that all the intensity can be directed onto the sample even through a small aperture. The typical size of the focused Globar beam can be few hundred microns and therefore most of the beam is lost on passing through an aperture in the microscope.
This leads to significant improvements in SNR.

The improved SNR offers significant advantages. The number of interferograms required to achieve the same SNR for a synchrotron source is less compared to a Globar source. Second, the high flux of photons makes measuring thick samples easier. The confined beam also reduces the proportion of light sacrificed in the transflection mode.

When comparing single point mapping to array detector imaging, the situation becomes more complex. The brightness of synchrotron source comes from the confined beam. Thus, when the beam is projected onto a larger area, it loses its brightness depending on the splitting ratio. The SNR decreases with the increase of measuring area. For example, the SNR reduces 100 times when a 15 micron sized synchrotron beam is projected onto a 150-micron area. In other words, the synchrotron source loses its advantages when imaging large sample area.

2.2.5 Quantum cascade lasers

Quantum cascade lasers (QCL) are a new type of semiconductor laser. The regular semiconductor laser, such as PIN laser diode, contains three regions: two high bandgap regions N, P and a low band gap intrinsic region I (Figure 2.10). The electrons and holes are first pumped into N and P layers respectively. After reaching the threshold voltage, the electrons and holes move to the intrinsic layer and recombine. This
involves the release of energy which in this case is electromagnetic radiation. The frequency of the radiation depends on the energy difference between pumped level (conduction band) and ground level (valence band), which is the bandgap of the intrinsic layer (interband).

When the thickness of the intrinsic layer is small enough, a quantum well is formed [29]. In the quantum well, the energy change of the electron is not controlled by the conduction and valence energy levels, but the thickness of the layer.

A QCL is a semiconductor diode in which the intrinsic region, known as a superlattice, contains periodic thin layers of quantum wells (Figure 2.11). This will generate a series of virtual energy states which splits the permitted bandgap (interband) into small discrete electronic subbands, known as intersubband.

With the formation of intersubbands, the electrons can now remain in these particular states and be excited to emit radiations in between. By designing the thickness of the quantum well, the frequency of emission is tunable over a wide range [4].

In a traditional semiconductor diode, the electrons can only emit a photon once when they recombine with the hole. However, for a QCL, the electrons can emit photons every time they change energy state between subbands which give multiple photon emissions per electron [30]. This results in higher output power and a more intense laser beam. The energy difference for intersubband is smaller than the energy of visible
light which leads to laser frequency in the IR, especially mid-IR region. This makes QCL an excellent light source for an IR spectrometer.

Another advantage of QCLs comes from their capacity to generate light with discrete frequency. Through the development of IR applications, such as clinical diagnosis and drug screening, the application of chemometric classifiers such as Random Forest, has demonstrated that key information to a particular question is often contained within a small number of discrete frequencies across the spectral region[31, 32]. Therefore, measuring the full spectral range, which is time-consuming, becomes unnecessary. By the frequency of radiation, the QCL microscope can measure specified frequency in a short period of time. The generation of single frequency also benefits the intensity of light since all the output power is utilised to produce the radiation with single wavelength rather than light with a broad range of frequency.

It is noteworthy that the frequency range of the laser is based on the design of the laser chip. Each chip is tunable in a small frequency range, typically up to a few hundred wavenumbers. Thus, by integrating multiple lasers in one system, a wide spectral range of the system can be achieved. Systems with one[33], two [34] or four [32, 35] QCLs are used to study gas components [36], chemical process [37] and biosamples [31, 32, 35].

As described in the previous section, the thermal detector is limited by the frame rate (response time) and SNR (sensitivity of temperature change). The QCL offers significant advantages of coping with the thermal detector. The intense beam gives an excellent SNR, and the grating system does not require high frame rate. Hyperspectral imaging has been achieved by combining the QCL with a low-cost uncooled microbolometer FPA detector [33, 35]. This combination of laser and detector could potentially reduce the cost of the instrument and measurement operation compared with using an LN2 cooled MCT FPA detector.

There are also some potential disadvantages of using QCLs. The QCL with single frequency output faces the same problem as an old dispersive spectrometer. Measuring
a large wavenumber region with QCL microscope can take more time than the FTIR spectrometer. Moreover, the available spectral range depends on the laser chips integrated into the system. Currently, the largest frequency range available for practical use is limited to 5-10 microns. The high wavenumber region of the mid-IR is still not accessible. Manufacturing of high-frequency QCL requires material with higher bandgap. This will increase the cost of instruments, and the development can take a long time.

3. Literature Review

3.1 Cell and cell spectrum

There are numerous reasons why one might wish to study and analyse biological cells at the single cell level. Drug development will require at some point a study of the drug with known cells associated with a particular disease [1]. Stem cell analysis at the single-cell level is important particularly with a view to using them for therapeutic purposes, since it is important to know the differentiation state of the cells prior to treatment. The field of personalised and stratified medicine also relies on the analysis of single cells. In the case of some disease, diagnosis may also be made on the detection of say circulating tumour cells or cells obtained from a biopsy.

When studying biological samples, the size, shape and morphology of the sample will affect the answers which can be achieved. Large biological samples, such as biopsy, are ideal for fast clinical diagnosis. But, to understand the biochemical changes which can be attributed to a particular disease, especially cancer, the scale of the sample is often limited to a single cell. Cells contain not only genetic information but also functioning information, which are represented by, for example, the type and structure of proteins. An IR spectrum of a fixed PC3 cell is shown in Figure 3.1.

![Figure 3.1 spectrum of a single cell (fixed)]
There are four types of biomolecules in a cell, namely proteins, lipids, nucleus, and carbohydrates. All those components take 30% of the weight of a cell and the rest, 70%, is water. The behaviour of those four type of molecules in IR spectroscopy is discussed below.

**Proteins**

Protein is the main functional microstructure within the cell, and is made up from amino acids bound by a peptide bond. The order type of amino acids determines its function [2, 3]. The protein expresses itself in the infrared spectrum as four major peaks, amide A, I, II and III [4]. The amide A band is a broad and intense band which sits in the region from 3200 to 3600 cm\(^{-1}\). This band results from the N-H stretching vibrations. The amide I band is another intense vibration and is located in the region from 1600 to 1700 cm\(^{-1}\). It is mainly due to the C=O stretch (80%) coupled with the N-H bending mode. The width of the Amide I band, and its range of frequencies is as a result of the hydrogen bonding arrangement associated with the protein secondary structure. The amide II band sets in the region from 1480 to 1575 cm\(^{-1}\), and this is caused by two strong vibrations of C-N and N-H bond. The amide III band is relatively small and is observed in the region 1230 to 1350 cm\(^{-1}\). It is formed by the coupling of C-H and N-H bonds [2, 5].

**Lipids**

Lipids are organic molecules which are normally hydrophobic or amphiphilic. These molecules are multifunctional which include storing energy, signalling and forming cell membranes [6]. As a hydrophobic or amphiphilic molecule, the lipid has a long hydrocarbon chain which gives rise to vibrations of the \(-\text{CH}_2\) group at 1466, 2852 and 2874 cm\(^{-1}\) in the IR spectrum. Moreover, the stretching vibration of C=O band gives a small peak at 1738cm\(^{-1}\) [2].
**Nucleic acids**

There are two kinds of nucleic acid in the cell, known as DNA and RNA. The first nucleic acid contains the genetic information for organisms. This information is then extracted by a second nucleic acid and translated by ribosome to synthesise proteins. In a resting cell, most of the time the DNA are covered by proteins which make the DNA disappear from the spectrum [6]. However, when the cell is preparing to duplicate itself, the genome unpacked from the nucleus and results in a few bands in the region of, 1055, 1090, 1230 and 1620-1750 cm\(^{-1}\) which hide in the amide I and amide II band.

**Carbohydrates**

Most of the carbohydrates in a cell act just like the ‘fuel’ of cars. They consist of sugar, which stores the energy for keeping the cell alive [3]. Meanwhile, some carbohydrates are found coupled with proteins and form glycoproteins which contribute the adhesive property and recognise whether a molecule is an invader [2, 3]. The carbohydrate bands occur at 931, 996, 1016, 1043, 1078 and 1149 cm\(^{-1}\). These are contributed by stretching of C-O, C-C and C-O-H groups. Noteworthy, the glycogen in an animal cell is indicated by peaks at 1028, 1078 and 1151 cm\(^{-1}\) [2, 5].

In recent years, the focus has been put onto new analytical methods to study biological systems, such as mass spectroscopy [7], NMR spectroscopy [8] and magnetic resonance imaging [9]. However, none of those studies can explore all the biochemical information without damaging samples. The bio-samples go through a series of fixation and labelling process which certainly change the biochemical information they carry [10]. Thus, a non-damaging, label-free method which can measure samples under proper physiological condition (aqueous environment) is required.

Thanks to the low energy level of IR light, infrared microspectroscopy (IRMS) offers a non-invasive, label-free and sensitive method to study biological samples. The first study on individual cells with IRMS combined with synchrotron source was carried
out in 1998 [11]. The cell cycle and cell death were then observed by SR-IRMS [12], and the chemical changes of cell death were studied [13]. In the meantime, the cytotoxicity and heating effects from synchrotron source were evaluated [14]. Recently, a lot of work on cell-drug interaction with SR-IRMS has been reported from different groups, including the cell responses to multiple drugs [15], the staining effects and response to collagen [16, 17], kidney cancer [18] and breast cancer [19] drug treatment. Without the use of synchrotron source, the first multiple cell study with IRMS was performed on 2000 [20]. Several studies on single cells were reported later [21-23]. In 2004, the advantages and limitations of employing either SRFTIR spectroscopy or normal bench top spectroscopy to measure single cell spectra were discussed [24]. Studies on effects of cell culture process [25] and differentiating stem cells [26] were reported as well. Due to the limitation of spatial resolution and low SNR, the cell imaging on conventional FTIR spectrometer was focused on the spectra on the cellular level, until high-resolution imaging taken from the high numerical lens and FPA detector was demonstrated [27].

Due to the limitation of using aqueous media, the cells studied were mainly either dried or fixed [28]. These cells studied with IR spectroscopy are distorted by two effects. First, the chemical information of the cell is changed due to the fixation process [29]. Although there was a report that the fixation process is capable of retaining most of the biochemical information in the sample [10], the degree of variations and the missing information still remains unknown. Second, the Mie-scattering effects were reported on dry cell spectra and large distortion of the spectrum was observed [30] and subsequently a mathematical algorithm on removing Mie-scattering was demonstrated [31]. Later, however, in a series of papers, specifically investigating the so-called dispersion artefact, it was shown that resonant Mie scattering was also an important factor [32, 33] and an extension of the original algorithm was developed [34-36], but these extra step can still be time-consuming.

Live cell imaging offers a great opportunity to avoid both of those two sources of
distortion. No fixation process is required, and the matching of the refractive index of water and cells will significantly reduce Mie scattering effects to a minimum and acceptable level[37]. Measuring the live cell spectrum will not only provide a more realistic and accurate result but also give the potential of measuring biochemical changes in real time and in vivo. For example, prostate cancer cells can be found in urine. If such cells can be monitored while in flow, it might reduce the need for prostate-specific antigen (PSA) test or more importantly an invasive needle core biopsy. This will reduce the cost of surgery and reduce unnecessary patient discomfort.

However, for the past few decades, studying living cells was treated as a ‘forbidden zone’ for IR spectroscopy due to the strong water absorbance issue. This task was taken by Raman spectroscopy which is another branch in vibrational spectroscopy, since 2005 [38]. Degradation and photodamage were reported but solved by using NIR laser [39]. Studies on live cell imaging have been demonstrated by different groups, including effects of substrates [40], cell imaging on 3D collagen gel matrix [41, 42], combined confocal microscope [40, 43] and drug interaction on bio-molecules [44]. Raman live cell imaging has successfully been widely used in clinical applications [45]. Indeed, in some respect, Raman spectroscopy has advantages over IR spectroscopy, such as no strong water absorbance issue, fluorescent/visible image and higher spatial resolution. However, Raman requires a focused laser beam having a spot size of a micron or so and is therefore only good at measuring individual cells. A large percentage of variance in cell spectra comes from the different stage in the cell cycle. To measure the cell population, normally hundreds of cells need be measured. Using Raman to measure this amount of cells is time-consuming. Fortunately, IR microspectroscopy combined with multichannel FPA detector offers a high throughput measurement by which spectra of hundreds of cell can be recorded in a single image within few minutes. Thus, performing live cell imaging with IR spectroscopy is being developed.
3.2 Water barrier

In order to appreciate the challenge of live cell imaging with IR spectroscopy, it is important to understand the barrier associated with the strong absorption of water.

Water is a very strong absorber in the IR region of the electromagnetic spectrum and is, therefore, the main obstacle of performing live cell imaging with FTIR spectroscopy. Since the 2000s, a few groups of scientists have been exploring the possibility of solving the water issue to attempt to reveal live cell FTIR spectra.

The infrared absorption spectrum of water is very complex. To understand how water affects FTIR live cell imaging, it is important to identify the origins of the peaks in the water IR spectrum.

For an isolated water molecule (H$_2$O), there are three vibrational modes. These vibrations are symmetric ($\nu_1$) and asymmetric ($\nu_3$) stretching of O-H group and bending of H-O-H ($\nu_2$), located at 3450, 3600 and 1643 cm$^{-1}$ respectively. Meanwhile, the rotations of water molecule also affect the spectrum in the form of vibrational-rotational coupling. This will shift the vibration peak into higher or lower frequency. For water vapour, the spectrum is particularly complex consisting of both rotational and vibrational bands.

In liquid water, the situation becomes more complex due to the formation of intermolecular hydrogen bonds [46]. The rotations of water molecules are restricted to a certain degree by hydrogen bonds and form librations. Consequently, this results in the stretching bands shifting to a lower frequency and the bending one to a higher frequency [47]. The stretching bands also become more intense as a result of increasing of dipole moments. It is noteworthy that anything that affects hydrogen bonding has an impact on the water spectrum, such as temperature and salt effects [48, 49]. Furthermore, in the spectrum of liquid water, the absorption bands broaden and overlap which result in extremely broad bands.
The main vibrational bands of a water spectrum are shown in Figure 3.2. The intense broad band on the left results from overlap of the two stretching vibrations ($\nu_1, \nu_3$) and the overtone of bending vibration ($\nu_2$). This is further proved by Raman spectroscopy with the appearance of three peaks, 3210, 3450 and 3650 cm$^{-1}$ in the OH band region [50]. The narrow band on the right is due to the bending vibration ($\nu_2$). Between the two bands mentioned above, a small peak known as the water combination band is located at 2150 cm$^{-1}$. This band may be due to the third overtone of libration band (800-500 cm$^{-1}$) and the asymmetric bending vibration ($\nu_2$) of H-O-H caused by the second overtone of libration band [51, 52]. In the remaining fingerprint region, the L2 libration band is so broad that the tail of the band is combined with the tail of the bending band and generates a long plateau in the spectrum [53].

Comparing the water spectrum to a cell spectrum, reveals there is overlap in the stretching band (3000~3800 cm$^{-1}$), bending band region (1550~1800 cm$^{-1}$) and the plateau formed by the tail of bending band and L2 libration band (1000~1550 cm$^{-1}$). For most of the live cell spectra, the stretching bands cover the amide A bands. This is due to the fact that the extinction coefficient of stretching bands (around 100 L•mol$^{-1}$•
cm⁻¹ [54]) is twice as large as that of the protein bands. In the region of 1600~1700 cm⁻¹, the extinction coefficient of water is approximately 20 L•mol⁻¹•cm⁻¹ which is more than two orders of magnitude smaller than what of proteins. Therefore one could expect that it is still possible to observe the amide I band. However, the main absorbance peak observed on the spectrum is still contributed by water due to the higher concentration of water molecule compared to proteins. The cell spectrum in this region can be observed directly, but the peaks are affected which cannot be used for further analysis.

To overcome the water issue, the water contribution to the spectrum (at least fingerprint region) needs to be subtracted. The previous works of live cell imaging with IR spectroscopy will be reviewed in the next few sections. Each section will discuss the works based on the same type of IR spectrometer.

3.3 Live cell imaging with benchtop FTIR spectrometer

Due to the low intensity of Globar light source, live cell imaging with benchtop FTIR spectrometer is a challenge.

The first experiment on live cell imaging with FTIR spectrometer can be dated back to early 2004. M. Miljkovic investigated the feasibility of measuring the infrared spectra of live cells in aqueous medium [55]. In this experiment, human cancer cells were cultured on a CaF₂ slide or liquid cell and measured utilising both the reflection

![Figure 3. 3 liquid cell with sandwich setup](image-url)
and transmission modes. During the experiment, cells remained in growth medium throughout to maintain the viability. A benchtop FTIR microscope utilising a Globar source and a 16x1 detector array (6.25*6.25 μm² resolution) was used for cell mapping. The experimental setup is described in Figure 3.3.

After obtaining the IR imaging of both the cell and the culture medium, the spectra surrounding the biological cell was chosen as a reference and visually fitted with linear scaling to remove the water contribution from the whole spectra. This method is similar to previous research on water correction from proteins and did generate a normal looking amide band [49]. However, there are a number of issues which need to be resolved. First, the live cell used in the experiment has a diameter of 40μm which is much larger than the path length (6μm /10μm) of the IR cell. This results in the compression of the cell and may induce complexity to the system, such as band shift [56]. Second, the viability of cells is questionable as no environment control (temperature, CO₂) was performed. This could lead to the death of cells which gives changes in biochemical information, such as band shift and variation in peak intensity. Finally, the water contribution was manually removed which makes obtaining consistent result difficult.

In 2009, M. K. Kuimova et al. in London utilised Attenuated Total Reflectance (ATR) FTIR spectroscopic imaging to obtain the spectrum of single live cell in an aqueous environment [57]. Cells were seeded into a chamber on the diamond ATR accessory and cultured in the incubator (Figure 3.4). After the cell had attached to the ATR crystal,
the spectrum was measured in a time dependent manner. A 64 x 64 FPA detector and a Globar source were used to perform the imaging task.

The ATR imaging successfully avoids the excess amount of water. However, due to the small penetration depth, this approach only works with a small and thin cell layer that is directly attached to the crystal. Besides, the SNR for imaging in ATR mode is usually worse than in transmission mode, especially when combined with transflection setup in which half of the light beam has been blocked.

Later, in 2013, Chan et al. utilising advancements in 3D printing techniques, fabricated an IR compatible microfluidic device by printing wax on a CaF2 window and measured the cell spectrum in transmission mode [58]. Spacers (12µm) and a clamp holder were used to seal the device. Two CaF2 lenses (Figure 3.5) were used to reduce the chromatic aberration caused by the refraction angle difference of light with varying frequency. A cooled FPA detector with 128x128 elements was employed. The result clearly showed the inconsistencies which were caused by chromatic aberration and demonstrated that these could be removed by adding lenses. The acquired spectrum was fitted by the water combination band to eliminate the water contribution. Moreover, the spectra of live cells in multiphase flow in the microfluidic device were obtained.

![Figure 3.5 3D printing microfluidic flow cell with extra CaF2 lens, assembled with a holder.](image)
In 2010, E.J. Marcsisin et al. demonstrated a live cell imaging approach with a demountable flow cell (Figure 3.6) [37, 59]. A well of approximately 10µm thickness was etched on a CaF$_2$ window with cells cultured inside. Another CaF$_2$ window was drilled with two holes so the culture medium could be pumped in. These two windows were fixed on a temperature controlled stage which was then screwed together.

![Figure 3.6 (top) top view and (bottom) cross section of the device. The small circle in top view is sample cells which are pre cultured on the upper window. The inlet and outlet hole are expressed in cross section which filled with dash line. The heating cartridges are not included in this diagram.]

Two set of experiments were performed with this device. In the first experiment [59], the viability of cells in the flow cell was evaluated. The spectra were obtained in single point mode with an aperture size of 50x50µm$^2$ using a MCT detector. To remove the water contribution, a background water spectrum of the adjacent area was subtracted from the acquired cell spectrum. The overcompensation was then fitted by the water combination band. Despite the wrong peak position of water combination band labelled in the paper, the corrected cell spectra look normal. The viability of cell gone through multivariate analysis and no sign of significant biochemical change has been noted. However, the non-separate classifier analysis results may as well be a result of
In the second experiment [37], the cell was imaged with an 8x2 elements MCT FPA detector. The spectral region obtained is limited to 1000-3000 cm\(^{-1}\). By applying a principal component based noise reduction process, the SNR of spectra is significantly improved. In the meantime, the avoidance of Mie scattering with live cell imaging is proved by the coincided amide I peak position. After fitting the water combination band, the resultant cell spectra successfully show the biochemical difference between cells before and after drug treatment.

![Diagram of fully sealed microfluidic liquid cell](image)

*Figure 3.7 fully sealed microfluidic liquid cell. Cells were placed in the middle chamber. A flow of culture medium was passing through from the two sectional chamber.*

In 2012, L.Vaccari’s group demonstrated the first fully sealed infrared compatible microfluidic device (Figure 3.7) and studied the spectra variation induced by different fixation methods [10]. A round chamber was etched in the middle of the CaF\(_2\) window. Two sector chambers were then created on two sides of the middle chamber. Three chambers are connected by narrow channels. Cells were placed in the middle chamber, and the culture medium was passing through from and to the two sector chamber. The effects of fixation were investigated. The cells were fixed with three different fixation process, namely formalin fixation, ethanol fixation and air dried fixation. All fixed cells were measured in dry condition. The live cell spectra were then measured with the cell in the chamber and a supply of proper culture medium.
The spectra for four group of cells were compared to identify the chemical changes due to the fixation process. Before comparison, each group of spectra were averaged with 20 cell spectra to increase SNR. No scattering correction for dried cell spectra was performed. The changes in band shapes and band position were reported which matched the theoretical biochemical change in the fixation process.

Early in 2016, a cooperation of groups lead by G. Birarda and G. Grenci was carried out which focused on manufacturing a cheap IR compatible microfluidic device [60]. Two silicon moulds were made to produce the device by injection moulding in a fast and cost-effective way. This microfluidic device is made of two polymer layers (UV cured NOA) each embedded a CaF$_2$ window (Figure 3.8). One of the layers contains a narrow channel and a small chamber with the depth of 10µm for the cell to pass through. By matching the position of CaF$_2$ windows, the IR light will only pass through IR windows and avoid the interaction with the polymer. Two cell lines with small size (~10µm) were used to evaluate the function of the system. Cells were injected into the chamber and cultured onto the CaF$_2$ window with a flow of culture medium and appropriate temperature with a heater stage.

The spectra were acquired on a benchtop FTIR spectrometer with a Globar source and a 128x128 MCT FPA detector. Obtained spectra were fitted with the spectrum of
culture medium at water combination band.

The viability of cell was tested for more than 60 hours, and no cell death was found. By oversampling with 1.2x1.2µm² pixel resolution, high spatial resolution chemical imaging of cells was achieved. Spectra from a different part of the cell was then compared. In the meantime, a time-resolved mapping was carried out which further proves the capacity of the device.
3.4 Live cell imaging with SYNCHROTRON based IR microspectroscopy

In 2006, synchrotron radiation was adopted as a light source with FTIR microscope spectrometer for studying live cells [61]. This resulted in the first set of consistent and reproducible spectra of a live cell in culture medium, reported by Moss and co-workers. In this work, a number of different human cancer cell lines were studied (four colorectal cancer cell lines and two primary cultures). The cells were grown directly on a CaF$_2$ window (pre-treatment of the CaF$_2$ with polypeptides or proteins for better adhesiveness was explored in this study but was found not to be necessary). The experimental set-up was similar to the device M.Diem’s group, used in 2004 [55] (Figure 3.3) but with a different spacer of 15 µm thickness instead of 10µm. A flow of culture medium with a temperature of 35℃ was supplied to maintain the environmental conditions. The synchrotron light source was implemented to reach a high SNR. A single point MCT detector was employed with an aperture size of 20~30µm.

The spectra from regions that did not contain cells were selected as the reference and subtracted directly from the whole spectra and resulted in a normal looking cell spectrum consisting of protein, nucleic acid and lipid peaks. Overcompensation of the amide I band has been reported due to the amount of water around the cell being larger compared to the cell. Each cell was treated individually with a water background spectrum next to it due to the variation of path length in the device. The result also confirmed that infrared synchrotron radiation causes no damage to the cell.

Later in 2009, Holman and her colleagues studied the bacterial activity in a biofilm with SR-FTIR [62]. Cells were placed and cultured in an open channel device made of gold coated silicon and supplied with a thin laminar flow of medium (Figure 3.9). The experiments were performed under transflection mode. The chemical mapping was performed with a 5x5µm aperture.
After obtaining the spectra, the scattering, the variation of the thickness of biofilm and water vapour contribution was removed by mathematical fitting and linear baseline correction. The resultant spectra still contain a certain amount of water contribution which left an abnormal amide I/II peak ratio. The baseline correction also results in artefacts in the presented chemical information. Moreover, the transflection mode has been proved to induce spectral artefacts which distort band intensities [63, 64], although the ramifications of these effects are disputed[65-67].

In 2010, live cell imaging was studied with a combination of SR-FTIR and microfluidics. In Australia, M.J. Tobin and his group performed experiments on live cell imaging with a microfabricated static liquid cell to study cell-drug interactions [68]. The liquid cell consists of a polymer spacer (spin coated layer with 7.6 or 12.7µm thickness) fixed on a CaF₂ window, and the cell suspension was placed in the chamber formed by the spacer (Figure 3.10). Another CaF₂ window was added on the top to sandwich the samples. Two holes were drilled on the top window to let the culture medium pass through. The windows were held together by a microcompression cell. The experiments were carried out in transmission mode using the IR synchrotron beam line utilising an aperture of 7µm x 7µm.

The experiments successfully revealed the main features of a cell spectrum such as lipids and protein bands. It is also notable that reducing the path length to a certain value, which depends on each cell line, will result in a measurable amide II peak. The

Figure 3.9 Design of open channel imaging with flow of medium. The cell was planted on gold coated silicon.
water contribution was subtracted based on the adjacent area of cells, which led to overcompensation and abnormal amide I/II peak ratio. The thickness of the CaF$_2$ windows was proven to affect the quality of spectra. The large slope of baseline which was suspected to be due to scattering effects was discussed but not proven. This can also be the artefacts from the synchrotron light source, as no scattering effects have been reported in the similar setup elsewhere.

Figure 3.10 (a) top view and (b) cross section of the device. The T slot was etched on bottom window to create laminar flow. Two holes were drilled on bottom window as inlet and outlet for exchanging culture medium.
G. Birarda et al. in Italy, also reported two prototypes of IR compatible microfluidic device in the same year. These two prototypes were made by the photolithography technique. A path length of less than 10 µm can be easily achieved with photolithography, and the resulted gap distribution is more uniform.

The first prototype is a demountable static cell which used to study the response of live cells under mechanical compression [56]. A layer of polymer were coated on a square CaF$_2$ window, and four round chambers were etched out. Another CaF$_2$ window was placed on the top of the lithographed window. Two windows were mechanically assembled together (Figure 3.11). The chambers were designed in thickness of 9, 5 and 3µm to compress the cell. The cells in buffer solution were dropped into two chambers and left the other two chambers were filled with pure buffer solution to take background. The cell mapping was performed with an 8x8µm$^2$ aperture with a step length of 4µm to oversample and increase the resolution.

The resultant spectra were fitted with the water spectrum of the adjacent area at the water combination band. After water correction, the spectra of cells under a different level of compression were compared by peak ratios. It was found that, for cells with diameter around 10µm, the chamber thickness small than 5µm resulted in significant variations in obtained spectra. Due to the lack of nutrition, the origin of changes in the cell spectra cannot be concluded.
The second prototype was manufactured by two round CaF$_2$ windows with two inlets and two outlets hole in the upper window (Figure 3.12) [69]. A large chamber was made with some narrow channels in the middle which separates the chamber into two parts. Cells were placed in the chamber on one side, and the drugs were introduced into the other chamber to study the cell-drug interaction. In this setup, the culture medium was able to flow around cells and give a suitable environment for cells to live with. The manufacturing of those two IR compatible microfluidic devices proved to be an excellent guide for the future design.

3.5 Live cell imaging with QCL based microscope

Tunable IR QCLs are still an emerging technology, and the development of a dedicated QCL microscope is more recent still[70]. However, QCL microscopy techniques have demonstrated excellent performance exploiting the high-intensity light source and fast measurements. However, live cell imaging has been barely performed with this system due to the limited time such systems have been in existence. To understand the advantages and disadvantages of QCLs, different QCL microscope systems which were used to measure biosamples will be reviewed.
Other laser-based systems

A QCL microscope is not the only laser-based IR spectrometer used to study biosamples. Recently, H. Amrania and the group in ICL London demonstrated a fast IR imaging system with a laser-based light source [71]. The source used here is known as Optical Parametric Generator (OPG).

The OPG is a tunable source which generally consists of two laser gain media and an angle-tuned nonlinear media (birefringent crystal). When two beams of high-intensity light pass through a nonlinear crystal, the light can interact with each other, and one of the beams can be pumped up to a higher frequency and the other one down to a lower frequency. This character made the frequency of output light tunable. The detail of principle of OPG is beyond the scope of this thesis and therefore will not be discussed here [72].

With the OPG laser, live cell imaging was performed with a bench top IR spectrometer. The spectra of cell in heavy water (D₂O) has been successfully obtained in the spectral region of 1160-1550 cm⁻¹. The available spectral region of the OPG system can be extended by using different nonlinear crystals. The spectra were then compared with the spectra of the same sample taken from FTIR spectrometer. By oversampling, the spatial resolution is diffraction limited. The spectral resolution, however, due to the broad spectral linewidth of output laser, was limited to 12-20cm⁻¹. This work proves the capability of laser-based IR spectrometer to performing single cell imaging in aqueous environments.

QCL system

The first mid-IR QCL light source was demonstrated in 2007 [73]. A set of distributed feedback (DFB) QCLs and external cavity (EC) QCLs were combined to achieve a widely tunable frequency. Later, the possibility of using QCL laser based spectrometer as a replacement for the FTIR spectrometer was discussed [70]. By comparing the
obtained image from QCL and FTIR spectrometers, the advantages of QCL and limitations due to the coherent laser source were explored [74].

Employing QCLs to study biosamples only has a history of few years. In 2013, a group in Princeton lead by C.F.Gmachl utilised a QCL system to measure the concentration of glucose in biological fluids [75]. This system contains a single laser which is tunable from 1000-1200cm\(^{-1}\) and a LN\(_2\) cooled MCT detector. A solution of glucose was put into a liquid cell with a pathlength of 100 µm. Spectra of glucose in aqueous medium were taken in transmission mode. A mathematic model based on the transmission of light in glucose band were developed to predict the concentration. The mode was successfully predicted the concentration of glucose in the biological fluid to clinical accuracy. It also confirmed the capacity of QCL in measuring samples in aqueous environments with large pathlength.

Later on, this group created a fibre probe with QCL to measure the glucose level of human in vivo [76]. The QCL light was shining on the human palm between the thumb and index finger through an optical fibre. The scattered light was collected by another optical fibre with 6 channels. The collected light was then detected by a LN\(_2\) cooled MCT detector. The resultant spectra show that the prediction of glucose concentration is feasible in a particular concentration range. By comparing the predicted concentration with the one taken from a commercial glucose detector, only 2% error was found. This work successfully demonstrated the advantages of high-intensity QCL light source in measuring biochemical information in vivo.

In 2014, a number of groups of scientists employed the QCL to study biological tissues. In the UK, a commercialised QCL microscope system (SPERO, Daylight solutions, USA) was used to measure both tissue sections and bio-fluids. The tissue imaging of breast tissue microarray (TMA) was demonstrated by P. Bassan et al. [77]. The SPERO system contains a quad-QCL module covering the spectral range from 900 to 1800 cm\(^{-1}\). An uncooled 480x480 elements microbolometer detector was employed. A 4x/12.5x magnification lens with NA of 0.15/0.7 was used to achieve a pixel resolution
of 4.25/1.36 µm and spatial resolution of 24/5µm, respectively. A tissue imaging at a single wavenumber of a whole TMA with the size of 20x24 mm² was acquired within 9 min. The speed is 126 times faster than FTIR spectrometer, which suggests that the single mode QCL can be widely used for detecting known biomarkers.

Later, Clemens et al. also reported tissue imaging acquired with SPERO system [78]. The full range spectra of a TMA core was obtained and compared with the spectra taken from FTIR spectrometer. It is found that, under similar measurement time, the QCL microscope can offer approximately four times better pixel resolution than FTIR spectrometer with the same spectral resolution and good SNR. Urine and serum were also studied with the QCL using discrete wavenumber. The results clearly showed that spectral data at 11 discrete wavenumbers was enough to perform multivariate analysis and identify differences in the brain cancer serum.

In Germany, N.Kröger et al. demonstrated a QCL microscope which contains a single QCL module and an uncooled microbolometer detector [79]. The single QCL in the system is tunable from 1030 to 1090 cm⁻¹ while the FPA detector contains 640x480 elements. The pixel resolution is approximately 7µm, which gives a field of view of 3.1x2.8 mm². A sample of mouse tissue with a thickness of 8µm was placed on 2mm thick BaF₂ slide. After taking the IR chemical imaging, the sample was stained using H&E to compare the result.

The obtained spectra were spectrally averaged every 2.4 cm⁻¹ to compare with the spectra taken from FTIR spectrometer (Globar source, LN₂ cooled MCT detector). The resultant chemical imaging from QCL and FTIR were compared with visible imaging of H&E stained section. After applying a colour code scheme, the imaging from both QCL and FTIR spectrometer identify the epithelium from other tissue and are highly coincident with visible imaging.

Later on, the same group integrated two QCLs in the system to study tissue sections[80]. The system uses the same set up as before except the laser tunable range changes to 1027-1087cm⁻¹ and 1167-1319cm⁻¹ (second QCL).
By digital staining, the resultant single frequency chemical imaging successfully distinguish different type of tissue in the sample. The full range spectra were then analysed by random forest classification method to explore the detailed chemical information. A colour scheme was developed based on the peak intensity and peak ratio to label the different type of tissue. The random classified imaging was proved to be nearly identical to the H&E stained or unstained tissue imaging. This suggests that the increase of available spectral range gives more rich information which gives a more accurate result.

In 2016, the same system was employed to study samples in vivo [81]. Two sets of experiment were carried out which the first one studied the fermentation process and the second one studied the microorganisms, both in the aqueous environment.

In the first experiment, a temperature controlled liquid cell with two BaF$_2$ windows and a set of spacers with varied thickness was developed. Only one QCL laser was used, which gives a spectral region of 1027-1087 cm$^{-1}$. Before performing the experiments on fermentation, a quantitative study on the limitation of SNR on different thickness of water layer was carried out. The liquid cell was filled with water at a thickness from 6µm to 100µm. The results suggest that the achievable SNR is optimised at 20µm and can be still higher than 10 at water layer thickness of above 50µm. A suggestion that the SNR noise is laser source limited was made. This confirmed the advantages of QCL on measuring samples in the aqueous environment.

To evaluate the capacity of QCL on measuring chemical reaction process in aqueous, a set of glucose, ethanol, yeast and LB medium was put into the liquid cell which was maintained at 35ºC. Spectra of those four compounds were taken before the experiments. By comparing the spectra of the mixed solution with spectra of each single compound, the concentration of different component was then estimated. A set of images of the fermentation process was obtained to monitor the volume of generated CO$_2$. 


In the second experiment, a single-cell microorganism and a multi-cellular organism were placed in the liquid cell with the appropriate aqueous medium. A spacer of 25µm was used to generate a suitable path length, and the experiment was performed at room temperature. Another QCL in the system was utilised to investigate the behaviour of these organisms which gives a spectral region of 1160-1320 cm\(^{-1}\).

By applying a k-means clustering algorithm, a chemical image of single-cell microorganism in aqueous medium was classified into three clusters. The results suggest that the real chemical difference between different types of sub-cellular structure can be revealed by only a few discrete wavenumbers.

Later, the movement of a multi-cellular organism was studied in real time measurement. By comparing the absorbance of a particular frequency (discrete wavenumber), a video with a frame rate of 50Hz was obtained. It is found that the SNR is well above 10 in each frame and can be improved up to 18 by averaging 20 frames. The potential impact of heating was also studied which suggest that the heating from high brightness laser source is negligible and will not affect the biochemical information of bio-samples.

All these results confirmed the potential improvement of utilising a QCL based light source to measure bio-sample in aqueous environments.

3.6 Summary

Live cell imaging with IR spectroscopy has been studied in all three type of IR measurement system. However, there are still problems to be solved.

The elucidation of biochemical information requires the removal of the water contribution from cell spectra. The water in the live cell consists of free water and bonded water. Therefore, the amount of water which ‘should’ be removed cannot be calculated in any method and therefore the true live cell spectra is not achievable. However, it is still possible to create a method which consistently extracts most of the
information from cell spectra and can be used to address a particular biological problem.

Two mathematical water removal methods have been widely applied in the past. The first method uses a water spectrum of the same path length as the cell as a background and directly ratio the obtained cell spectrum to the background. This will cause overcorrection due to the excess water contribution be removed. The second method is based on the fitting of water combination band. By assuming the peak centred at 2150cm\(^{-1}\) is only due to the vibration of water, the cell spectra were fitted with a pure water spectrum to flatten the baseline in water combination band region. This method proves to be able to acquire consistent and normal looking fingerprint region. However, the assumption made is dangerous since baseline shifts, scattering and other possible effects were not taken into account. Moreover, the water combination band region was not available in any QCL microscope system. In the studies on live cell imaging with a QCL microscope, the progress was still held in the ‘observation’ step. Due to the lack of water removal algorithm, the cell spectra obtained from QCL system cannot be used for multivariate analysis.

Studies using heavy water [71, 82] did avoid the broad absorbance peak in protein region. However, due to the cytotoxicity of heavy water, the cells behave differently, for example, not undergoing mitosis, which will induce biochemical differences in the resulting spectra.

Thus, the first problem that needs be addressed is to find a more appropriate water correction method which has theoretical support, can consistently uncover desired information and needs to be compatible with QCL system.

The second problem is most of the live cell work done to date has focussed on detecting the chemical difference of a single cell line before and after treatment, such as fixation and drug interaction. The chemical differences between multiple cell lines have not been studied in live cell form. Thus, the second problem to deal with is evaluating the performance of new water correction algorithm on multiple cell lines.
Thirdly, most of the QCL studies has been used to detect homogeneous biofluid samples or multi-cellular tissue samples. The studies on the single cell utilised cells with a size larger than few hundreds of micron. For a small cell, only imaging at discrete wavenumber is obtained. The available spectral region was also limited by the two QCL system. No spectra of a single cell with full available spectral range, especially in solution was obtained and analysed. Thus, the third task is to evaluate the performance of measuring single cell in aqueous solution with a full available spectral range in the quad-QCL system.
Reference


4. Experimental setup

4.1 Sample Preparation

In this project, three prostate cancer cell lines will be used to explore the protocol for live cell imaging. These three cell lines are PC-3, LNCaP and PNT2. PC-3 is a prostate cancer cell line derived from prostate cancer cells found in bone marrow obtained from a 62 year old male in the 1970s [1]. LNCaP is a prostate cancer cell line derived from the lymph node of a 50 years old male in 1977 [2]. PNT2 is derived from originally healthy epithelial prostate cells which been immortalised [3], the transfected PNT2 cells behave similar to cancer cells and therefore can be treated as a model of a cancer cell. The average size of PC-3, LNCaP and PNT2 cell (based on a thousand of each type) measured by the microscope are around 30 µm, 25 µm and 20 µm, respectively.

Cell culture

All three cell lines were stored in liquid nitrogen at -196°C with 0.5% of DMSO. After thawing the frozen sample, the cells was then transferred into the flask with culture medium and placed in an incubator.

The PC-3 cell line was cultured in Ham's F-12 Medium with 10% v/v fetal bovine serum (FBS), 1% v/v L-glutamine and 0.5% v/v penicillin-streptomycin. The LNCaP and PNT2 cell lines were cultured in RPMI-1640 medium with 10% v/v fetal bovine serum (FBS), 1% v/v L-glutamine and 0.5% v/v penicillin-streptomycin. The culture medium without penicillin was controversial, but the cells can be easily contaminated in the lab. The effect of penicillin on the resultant spectra is found to be negligible and will not be discussed in this project.

The cells are cultured in plastic flasks. After cells have been suspended in the medium, they start to attach to the bottom surface of the flask, spread out and proliferate. The
flasks were placed in the incubator at 37°C and 5% CO₂. The percentage of surface been occupied by the cell is known as confluence. Once the confluence is above 80 percent, the cells are ready to harvest. As all three cell lines adopted are adherent, the cells, therefore, need to be detached from the bottom surface by adding trypsin. Trypsin will be deactivated in culture medium and therefore the flask needs to be washed with phosphate-buffered saline (PBS) solution.

**Cell fixation**

As discussed in the previous chapter, this study focuses on the protocol to perform live cell imaging instead of actually creating a live cell environment due to the lack of funding to purchase appropriate device. Two fixation processes are employed in this project.

*Formalin fixation*

The formalin fixation process uses 10% formalin solution which is 4% formaldehyde in water. The formalin will create chemical bonds to proteins in the sample and hold all the cell components in the cell [4]. No carbohydrates are preserved by formalin solution. Thus, most of the cell structure including protein, lipids and DNA will be well preserved, and the morphology of cell is kept.

After being harvested from the flask, cells were centrifuged into a small pellet, and the solution was decanted out. Then the cells were washed with PBS solution twice to remove the remaining culture medium. The cells are washed by the following process: the cell pellet was mixed with PBS solution and centrifuged down to pellet form, and the solution was removed. Finally, 10% formalin solution was added and mixed with the cells. The cells in formalin were stored in a cold room at 5°C for at least 12 hours before use. It is worth mentioning that the formalin fixation is reversible [5], and therefore all the formalin fixed cell was measured within 12 hours after suspension into PBS solution.
Alcohol fixation

The cells were alcohol fixed to examine the established water correction method. The alcohol fixes the cell by removal of water from proteins, which also involves breaking of hydrogen bonding and altering the protein structure [6]. It also fixes carbohydrates by eliminating the adjacent water. As lipids are soluble in alcohols, the damage of cell membrane can be expected [7].

The cultured cells were washed with PBS twice and then added into 70% ethanol for 30 mins. The cells were then centrifuged down and suspended in the PBS solution and stored in the cold room at 5°C.

4.2 Modified compression cell

As no suitable flow cell was available in the group, a microcompression cell was modified to perform experiments using the FTIR and QCL microscopes. The compression cell consisted of one holder, one cap, two rubber ring and two CaF$_2$ windows (Figure 4.1). The thickness and diameter of IR windows used for compression cell are 0.5 mm and 13 mm, respectively. By reducing the thickness of window, the chromatic aberration is limited to a reasonable level [8], and the amount of light passing through is increased. Due to the problem in sourcing a suitable spacer, small PS/PMMA non-dispersed microspheres were employed to create the designed pathlength. The size of the sphere is averaged around 10.8 µm. The pressure applied to the compression cell is not high, since the windows are fragile, and therefore the rigid spheres should not significantly deform which suggest that the path length should be around 10 µm.

The appropriate path length should not only be able to obtain an unsaturated fingerprint region [9] but also leave enough room for the cell to remain alive and relatively un stressed[10]. Agreement on a path length less than 10 µm is not clearly stated but widely used in previous studies [9, 11-14].
The attenuation coefficient of liquid water at 1600 cm$^{-1}$ is approximately 1400 cm$^{-1}$ [15]. Thus, the absorbance can be calculated to be 1.4. This value may vary depending on the temperature, actual instrument and setup used, but should be within the trustable range for the detector. In the FTIR spectrometer and QCL microscope systems, with 10 µm pathlength, no saturated O-H bending band was observed.

As these initial measurements are on fixed cells, they should not show the obvious chemical difference on compression as long as the cells maintain their physical structure. No breakage/damage of any of the cells was noticed for all three type of cell line during the experiments.

The formalin fixed cell in formalin solution needs to be washed twice with PBS solution before the experiment in order to completely remove any traces of the formalin. The ethanol fixed cell can be used directly. The fixed cells were suspended in PBS solution and mixed with PS spheres. The mixed solution was then dropped on the bottom window in the compression cell and then sandwiched by the upper window and sealed by the cap. As the compression cell is made for solids/mulls and it is not designed to prevent leakage, thus the vaporisation of water can be foreseen. Therefore the measurement time for each sample loading is limited to maximum one hour.

![Figure 4.1](image_url)
4.3 Liquid cell for synchrotron experiment

The laboratory associated with the MIRIAM infrared beamline at the synchrotron facility (Diamond Light Source, Didcot Oxford) has a well-designed liquid cell which has been tested and is fully capable of performing cell measurements. The design of this cell is similar to the modified micro-compression cell. It consists of a holder, a cap, two rubber ring, two CaF₂ windows and a set of spacers. The spacer used in the experiment is made of gold-coated rubber and has a thickness of 10μm. Two large CaF₂ window with a diameter of 25mm and thickness of 1mm offer a large sampling area. As the CaF₂ window is brittle, the thickness of 25mm CaF₂ window cannot be smaller than 1mm or the window will easily break under compression. The cap and holder are fixed by screws. The specially designed cell well, rubber ring and the spacers are able to prevent leakage of water for at least 4 hours.

4.4 Imaging with benchtop FTIR spectrometer

The experiments on the FTIR spectrometer were performed on an Agilent (now Cary) 670 benchtop spectrometer with an Agilent (Cary) 620 imaging infrared microscope. A Globar source with an output power of 150 mW was employed in the spectrometer. The microscope consists of a 15x magnification Cassegrain objective with a numerical
aperture (NA) of 0.62, an additional 5x magnification ‘High-mag lens’, a visible camera and a 128x128 elements LN2 cooled MCT FPA detector. The High-mag lens can be used in addition to the IR objective which gives a total magnification of 75x. Therefore, the achievable pixel resolutions for this system are 5.5\(\mu\)m and 1.1\(\mu\)m with and without the 5x lens. The field of view (FOV) under these setups are 704x704 \(\mu\)m\(^2\) and 141x141 \(\mu\)m\(^2\). However, as the 15x objective has a NA of 0.62, the spatial resolution of the system is diffraction limited at equal to the wavelength. For example, at a wavenumber of 1650 cm\(^{-1}\), the spatial resolution is approximately 6 \(\mu\)m. The available spectral range of the spectrometer is from 950 to 3975 cm\(^{-1}\) due to a low pass optical filter (undersampling of 4). This limits the possibility of acquiring high-frequency information and can cause a problem in the subsequent data analysis (see fringing removal). The system was purged with dried air and a relative humidity of less than 4% needed to be achieved before any measurements were performed. A minimum of 20 min is required for a FPA detector to stabilise after topping up with LN2.

Cell imaging with the lab based FTIR spectrometer was carried out using two different setups. The first set of experiments utilised the 15x objective on its own without the additional High-mag option. (This is referred to as the low-mag configuration). This gives a large FOV to observe a large number of cells in a single image. The cells are compressed slightly in the sample holder which leads to a diameter of approximately 25 to 30 \(\mu\)m. Thus, the 5.5 \(\mu\)m pixel resolution is enough to identify a single cell. Under this setup, a high throughput automatic spectra extraction method can be used.

The second set of experiments employed the 15x objective in combination with the High-mag lens to obtain a pixel resolution of 1.1 \(\mu\)m. The High-mag option is used in order to establish whether the water correction protocol established in the first experiments is still valid for the higher resolution data. Both experiments were performed in transmission mode.

The background spectra were taken with only air in the compression cell. When the compression cell is assembled, a clear visible light diffraction pattern can be observed
which suggest the gap between two windows CaF2 is less than 1µm. The number of co-scans for the background and the sample is 256 and 128 respectively. The spectral resolution is set at 4 cm⁻¹ with zero filling of one. For each cell line, 500 cell spectra were obtained in the low-mag configuration, and 100 spectra obtained using the High-mag set-up.

4.5 Cell spectra with synchrotron-based FTIR spectrometer

The cell spectral measurements with the synchrotron infrared spectrometer were carried out at the B22 MIRIAM infrared beamline at Diamond light source, Didcot, Oxford, UK. The beamline utilises the radiation generated at the bending magnet (Figure 4.3). The radiation induced from the storage ring consists of light with a wide range of frequency. The x-rays occupy the centre of the beam as the energy of radiation decreases from the centre the outside layer. A hollowed flat mirror was used to filter the X-rays. The X-rays were then dumped by a water cooled photon absorber due to the excessive heat generated from the absorption. The rest of the light was reflected by the mirrors and equally divided into two beams by two flat mirrors. The beam before entering the spectrometer is in ultra high vacuum (UHV), and a diamond window was employed at the end of the vacuum tube allow all the whole IR range passing through. The final beam at the sample has a full width at half maximum (FWHM) size of 15 µm. The average photon flux at the sample for light with wavelength less than 10 µm
Spectra were measured with a Bruker Vertex 80V vacuum spectrometer coupled to a Hyperion 3000 infrared microscope. The microscope consists of a 36x magnification objective with NA of 0.6 and a LN$_2$ cooled MCT single element detector. Thus the spectra were obtained in single point mode. In single point mode, the spectrum obtained is an average of the sample area illuminated through the aperture (the area is therefore directly related to the aperture size). The effects of averaging will be discussed in an experiment with the High-mag benchtop FTIR spectrometer. An aperture size of 15x15 µm$^2$ was used. Although the size of the aperture is smaller than the size of the cell, this area should be able to cover most centre part of the cell. Four blade slits were used to restrict the microscope to fit the size of the aperture. It is noteworthy that no purge system is employed in this experiment. A stabilising time of half hour is required after topping the LN$_2$.

The experiments using the synchrotron IR system employed a liquid cell. The fixed cells in PBS solution are dropped into the liquid cell with a precise volume which guaranteed the formation of an air bubble in the field of view after assembling the liquid cell. The existence of air bubble allows the measurement of a background with exactly same pathlength as the sample. Thus the background is taken after loading the sample into liquid cell. The background and sample are measured at 4 cm$^{-1}$ spectral resolution with 512 scans and 256 scans respectively. For each cell line, 200 spectra of cells were taken in transmission mode.

### 4.6 Cell imaging with QCL microscope

The QCL microscope employed in the experiment is the SPERO from Daylight solution, San Diego USA. This system has demonstrated its capability of measuring the different type of biochemical samples [16-19]. The SPERO system consists of four external cavity quantum cascade lasers (EcQCL) which collectively cover the
wavenumber range from 900 to 1800 cm\(^{-1}\). A low (4x, 0.15 NA) and high (12.5x, 0.7 NA) magnification lens were equipped with the system which gives a FOV of 2x2 mm\(^2\) and 0.65x0.65 mm\(^2\). Due to the high brightness of laser source, the system uses a 480x480 element uncooled microbolometer instead of LN\(_2\) cooled MCT detector. This large FPA detector coupled with the objectives will result in a pixel resolution of 4.25 µm and 1.36 µm on each side. The achievable spatial resolutions of the system at 1650 cm\(^{-1}\) are 24 µm and 5 µm for the low and high mag lens, respectively. The stage of the microscope is equipped with a modulator. This modulator is turned on in high mag setup to overcome the chromatic aberration [8] and coherent effects.

Cell imaging was conducted with a 12.5x lens with 4 cm\(^{-1}\) spatial resolution and full spectral range. Due to the unknown difference between the samples, the experiments using discrete wavenumber was not performed in this project. The protocol to measure the background and sample is the same as the benchtop FTIR spectrometer experiment. 100 cell spectra were taken for each cell line.
Reference


5. Data analysis method

5.1 Least square fitting based on Beer-Lambert law

According to the Beer-lambert law, the absorbance of any heterogeneous sample can be treated as total absorbance of the different homogeneous samples. This rule does have the limitation when non-Beer-Lambert behaviour is occurring, such as Mie-scattering [1]. However, as discussed in the previous chapter, studies have proven that the refraction index matching of water and cell reduces the scattering effects to a negligible level [2].

Assumption 1: Beer-Lambert law validity

Least squares is a linear regression analysis method. The algorithm minimises the sum of the squares of errors to obtain an optimum solution. It is used widely in data fitting and will be the main algorithm in this project. The ordinary least square (OLS) fitting on spectra data can be expressed by the following equation:

\[ A_{total} = \sum_{i=1}^{N} X_i A_i \]  \hspace{1cm} (Eq. 10)

where the \( A_{total} \) stands for the total absorbance of the sample, \( X_i \) and \( A_i \) represent the coefficient and absorbance of spices \( i \). Assuming \( A_{total} \) and \( A_i \) are constant, the result of OLS fitting is a set of coefficient \( X_i \).

Water correction method based on protein ratio

The water correction in the previous studies uses the water combination band at 2150cm\(^{-1}\) as standard to obtain a flat baseline [3-6]. However, the flatness of the baseline is relatively subjective. Thus a new standard which is based on real chemical information without personal judgment is required. In this project, a water correction
method based on protein ratio is presented.

There are three broad protein bands in the cell spectra, namely amide A, amide I and amide II band. Those bands result from different vibration and function group of protein. However, the intensity of those bands should all be proportional to the number of peptide molecules in the cell. Thus, the ratio of peak area for those three band should be constant.

**Assumption 2: the ratio of peak area for protein bands are constant.**

The spectra of cell in PBS solution can be treated as a combination of a cell spectrum and a spectrum of PBS solution based on assumption 1:

\[ A_{total} = X_w A_w + X_c A_c \]  \hspace{1cm} (Eq. 11)

where the subscript \( w \) and \( c \) correspond to water and cell.

Two reference spectra were employed to figure out the contribution of cell and PBS solution. By inducing a cell reference, the result can be fitted based on actual cell features which relies on Assumption 2, the ratio of protein peaks. After obtaining the coefficient \( X_w \), the water contribution was then removed by:

![Diagram of water correction process.](image)

*Figure 5.1* water correction process. The raw spectrum (top-left) firstly been fitted by two reference spectra (right). Second, apply the coefficient to buffer reference spectrum and obtain the water spectrum. Third, the water spectrum been subtracted from the raw
The remaining part is assumed to be the spectra of cells. While computing Eq. 2, only the absorbance of the selected protein region will be taken into the calculation. After fitting, the coefficient of water was then applied to the full spectra range to obtain the water spectrum which need be removed from the raw spectrum.

Choose of cell reference

The cell reference should have the appropriate protein information to fit the results. Two different cell references have been tested in the project. The fitting of results of those two references will be compared, to understand the reference dependency. The first reference is a spectrum of an air dried cell obtained from the PC-3 cell line and the second is a spectrum of Matrigel [7]. Matrigel is a general biology sample which contains most of the key information of a biomaterial. Although the Matrigel does not have information of nucleus, the protein band region is still available. Matrigel has been widely used to mathematically present biosamples. For all water correction methods on a different type of data, the cell reference remains unchanged to ensure consistent fitting results.
**Effect of fitting range**

According to assumption 2, the fitting process will be based on the peak ratio of protein bands. However, there are three protein bands and therefore six combinations of fitting range. The tested combinations and corresponding spectral regions are listed in Table 1. The fitting results based on the different spectral regions will be compared to understand the best fitting range.

<table>
<thead>
<tr>
<th>Amide Peak combination</th>
<th>A</th>
<th>A&amp;I</th>
<th>A&amp;II</th>
<th>A&amp;I&amp;II</th>
<th>II</th>
<th>I&amp;II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral region cm⁻¹</td>
<td>3100-3550</td>
<td>1600-1700 &amp; 3100-3550</td>
<td>1500-1600 &amp; 3100-3550</td>
<td>1500-1700 &amp; 3100-3550</td>
<td>1500-1600</td>
<td>1500-1700</td>
</tr>
</tbody>
</table>

**Universal fitting range**

As mentioned in review chapter, the new water correction algorithm should be able to apply for all three types of spectrometer. Thus, the final fitting range relies on the available frequency range of the spectrometers. As two FTIR spectrometers both can cover the frequency range from 950 to 4000 cm⁻¹, the determining factor will be the QCL microscope. The amide A band, is not achievable with QCL microscope and therefore the universal fitting range is limited to the amide I and II band, which is 1500-1700 cm⁻¹.

**5.2 Water correction method on different type of data**

As a result of different experimental setup and system configuration, the water correction method applied to the data taken from various spectrometers needs to be specified.
Water correction method for benchtop FTIR spectrometer

For the cell imaging obtained in the low mag setup, a PBS spectrum was selected from the non-cell region in the target image as water reference. This water reference was then combined with a cell reference to every single spectrum in the imaging. The spectra of each cell were then extracted and averaged as the corrected cell spectrum.

For cell imaging taken with high mag setup, the water reference is selected from the imaging as well. However, the fitting process is separated into two protocols. The first method undergoes a ‘fitting-average’ process which extracts the spectra of a single cell, fits all the spectra extracted and then averages the fitted spectra to obtain the corrected cell spectrum. The second method undergoes an ‘average-fitting’ process which extracts the spectra of a single cell, averages all the extracted spectra and then fits with two references to get the corrected cell spectrum.

There are two reasons to implement those two methods. First, the average-fitting process can be represented as a single point model to evaluate the comparability of imaging from benchtop FTIR spectrometer and single spectrum from SR-FTIR spectrometer. Second, at high pixel resolution, the protein distribution in the cell can be localised. Whether the new water correction algorithm is capable of removing the water contribution in a fine cell structure needs to be tested.

Water correction method for SR-FTIR spectrometer

The spectra of the sample as measured by the SR-FTIR spectrometer are recorded in single point mode. For each cell spectra, a buffer spectrum is taken next to the observed cell. Thus, every spectrum of a cell in solution is fitted with its own buffer reference and the same cell reference.
Water correction method for QCL microscope

The cell images obtained from the QCL microscope have a FOV of 650x650μm² which is similar to the FOV of the low mag setup in the benchtop FTIR spectrometer. In theory, the same correction protocol should be implemented. However, when taking cell images with the QCL microscope, strong and correlated noise is observed which hold the water correction method back. The detailed study of this noise is described in chapter 8.1. The water correction method, when applied to remove the water contribution from spectra taken from the QCL microscope, is more like a combination of the methods for the high mag setup in benchtop FTIR spectrometer and SR-FTIR spectrometer. The cells were extracted one by one, and their surrounding water spectra were taken as buffer reference. The spectra of the cell were then fitted with the obtained buffer reference and a cell reference.

5.3 Fringing removal

When performing the initial experiment with the low mag setup in benchtop FTIR spectrometer, a consistent periodical fluctuation is found in the baseline of acquired spectra. The amplitude of the fluctuation is comparable to the peak intensity in the fingerprint region and therefore it needs be removed. The origin of the fluctuation is suspected to be the fringing occurring in the optical path.

A mathematical algorithm based on the extended multiplicative signal correction (EMSC) method for removing fringes resulting from thin film interference was demonstrated by T.Konevskikh’s group [8]. The fringes in the spectra were treated as a combination of a sine wave and a cosine wave.

By measuring the high-frequency spectral region from 4000 to 6000 cm⁻¹, a clear fluctuating baseline was obtained. This spectrum was then Fourier transformed back into the time domain by discrete Fourier transformation (DFT), described as:

\[ \tilde{A}(k\Delta x) = \sum_{n=0}^{N-1} A(\tilde{b}_n)\exp\left(-i\frac{2\pi nk}{N}\right) \]  

(Eq. 12)
where $A(\tilde{\nu}_n)$ is the absorbance at wavenumber $\tilde{\nu}_n$, $N$ is the number of wavenumber points, $k = 0 \ldots N - 1$. The $\tilde{A}(k\Delta x)$ is a complex number whose its amplitude can be used to identify the period of fringes, presented by:

$$|\tilde{A}(k\Delta x)| = \sqrt{[Re(\tilde{A}(k\Delta x))]^2 + [Im(\tilde{A}(k\Delta x))]^2}$$  \hspace{1cm} (Eq. 13)

where $Re$, $Im$ stands for the real and imaginary part of the complex refractive index. By plotting the amplitude in the Fourier domain, the period of fringes can be observed based on the $x$ value of the point with highest amplitude. Finally, the spectra were fitted by two sinusoidal terms with period of $2\pi/x$ to remove the fringes.

The 4000 to 6000cm$^{-1}$ spectral region is not available in the benchtop FTIR spectrometer employed. Thus, in this project, the region between the CO$_2$ band and C-H stretching band was selected to define the period of fringes at 2500-2700cm$^{-1}$. The 1800-2300cm$^{-1}$ region is not available due to the water combination band. After obtaining the period of fringes, two sinusoidal terms were added into the least square fitting equation:

$$X_c A_c = A_{total} - X_w A_w - X_{f_s} \sin(x\tilde{\nu}) - X_{f_c} \cos(x\tilde{\nu})$$ \hspace{1cm} (Eq. 14)

No other EMSC term will be added into the equation. To ensure the least square fitting for fringes is not strongly affected by the huge protein band, the fringes were fitted with the same spectral region to find the period of fringes. Then the spectra were computed with protein bands to remove the water contribution.

### 5.4 Cell extraction method for high throughput analysis

Under the low mag setup in the FTIR spectrometer, the number of cells in one image can be more than a hundred. Manually selecting those cell are time-consuming. Therefore, in this project, a semi-automatic cell extraction process was implemented.

The morphological analysis based on spectral information has been demonstrated by J.Filik [9]. In that study, the cells were cultured and fixed on the CaF2 slides. The
natural growth resulted in clear boundaries between cells so that the cells can be located and separated by the absorbance of the amide I peak. However, when dealing with cell suspensions in a compression cell, the gathering and overlapping of the cell cannot be avoided. In the meantime, due to the high absorbance of the O-H bending band, the intensity of the amide I peak is blurred even after water correction.

In this project, the cell extraction method will be altered by using lipidic information instead of using protein to identify the boundary of cells. The process for this ‘cell finder’ method is stated below.

First, the absorbance of the lipid band in the 2800-3000 cm\(^{-1}\) has been integrated to form a two-dimensional grey scale mapping. This mapping may contain not only cell but also debris and therefore needs to be filtered. A threshold for the intensity of each pixel is manually set to remove the debris resulting in a clean background. The median filter is used to remove the debris further and smooth the jagged boundary. By computing the gradient of intensity, a mask with gradient boundaries for cells was created. The number and the centre location of the cell were then calculated based on the regional maxima of intensity in the gradient boundary. So far, the separated individual cell is identified, but the cell clumps are not broken into single cells yet.

Second, to break a large clump of cells into the small individual cell, a watershed function is employed. The value of the intensity of each pixel is inverted into a negative value, whereby the maxima become the minima. Consider each minimum represents the deepest point of a lake. If one keeps filling water into those lakes, water in each lake will eventually meet the water come from other surrounding lakes. The place where the water meets is defined as the border of the lake. In another words, the boundary of the cell. This is known as a watershed function [10]. To this point, all individual cells have been separated from each other.

Third, there is a possibility that some of the cells are still unresolved, and few cell have been recognised as one. To avoid those cells, the size of each cell need be filtered. As the shape of a computed cell can be random, the size of cell is described by an equivalent circular diameter (ECD):
\[ ECD = 2 \times \sqrt{\frac{S_c}{\pi}} \quad (Eq. 15) \]

where \( S_c \) is the actual area of the cell.

After calculating the ECD for all the cells, the distribution of ECD was created. The first and fifth 5-quantiles of the ECD is then evaluated to obtain the lower and upper limit of the size filter. Cells with size below or above the lower and upper bound are removed from the list of spectra.

It is arguable whether the size control is necessary. The estimated cell size is not ideal, and the information in the filtered cells was lost. However, to ensure the quality of spectra of cells and be confidence that each spectrum is only the presenting one cell, the size control is implemented in this project.

5.5 Pre-processing method for multivariate analysis

Before performing multivariate analysis of the cell spectra, some pre-processing is required to minimise the non-chemical effects in the spectra. A list of pre-processing methods is described below. The choice of pre-processing method depends on the target data and will be stated in the respective results chapter.

Quality control

Due to experimental errors, mechanical failure, computing errors, sometimes the obtained results contain spectra of non-sample or with artefacts. To remove those ‘bad’ spectra, a quality control needs to be carried out. In this project, considering the thickness of the sample, the absorbance of amide I (1650cm\(^{-1}\)) and II (1550cm\(^{-1}\)) peaks are limited to 0.2 to 1.6. Any spectrum which has absorbance beyond this value range will be removed from the data set.

Noise reduction

The random noise in the spectra increases the variance of the dataset and may affect
the further analysis to some degree. To remove random noise, a noise reduction based on principal component analysis was implemented. The details of principal component analysis are described later in section 5.6. For the results, a certain percentage of data variance is assumed to be random noise. Thus, after reconstructing the data set, the principal components representing that part of variance is removed from the dataset.

Rubber band correction

Rubber band correction is a robust baseline correction method. It treats any effects on baseline shift linearly and does not consider the origins. To perform rubber band correction, a number of wavenumber points are selected and the absorbance at those points is directly dragged down to a flat baseline. The trapezoidal integration between those points was removed so all the data in between was shifted as well. Thus, the rubber band correction should be carefully used.

Derivative

The variance between spectra sometimes can be subtle. The variance which a sample group sometimes can be even larger than the variance between different sample groups. To elucidate the subtle changes in the dataset, the first or second derivative may be required. However, the derivative not only reveals the changes sought but also enhances the noise in the data. Thus, after derivatising the data, a smoothing function is usually applied [11].

The second derivative may be performed instead of the first derivative in two situations. First, the first derivative does not always find the subtle changes, as the changes in gradient can be small. Performing second derivative enlarges the difference further. Second, the negative peak of second derivative and the peak of original spectra share the same location and therefore the changes can be directly linked to the particular peak. It also needs to be taken into account that the second derivative requires more smoothing than the first derivative.
Savitzky-Golay smoothing

The Savitzky-Golay algorithm is a well-known smoothing function that improves the SNR without significantly distorting the data. The smoothing function fitting the data to a low-degree polynomial curve by least square fitting. Before performing smoothing, a frame with a size of N and a polynomial degree p is required. The size of frame decides the number of points applied in the fitting process and the degree of polynomial defines the shape of the fitting curve. For a frame size of N with a degree of p, each data point is fitted by (N-1)/2 points on each side and therefore the N can only be an odd number and larger than 3. The degree of polynomial p need be smaller than N-1.

It is worth noting that the implement action of smoothing has a significant possibility of removing spectral features from the data set. The larger the frame size, the more spectral features are missing. In the meantime, the size of the frame also affects the starting and ending points of the dataset. As for the first and last (N-1)/2 points, there are not enough data to perform the same smoothing level. Thus, the use of smoothing function need be considered carefully.

Mean centre and vector normalisation

The mean centre and vector normalisation are two methods that remove the variance due to the baseline shift and sample thickness from the spectra. The mean centre process calculates the average absorbance of each spectrum and brings them up or down to the same level.

The vector normalisation process normalises the data based on a unit vector. The unit vector is the square root of the sum of square of data:

\[ \text{Unit Vector} = \sqrt{\sum [A(\vec{v_n})]^2} \]  

(Eq. 16)

After obtaining the unit vector, all the data was divided by the unit vector to remove the effects of thickness.
5.6 Multivariate analysis

Generally, a spectrum of a cell, recorded with an FTIR spectrometer at 4cm⁻¹ spectral resolution will contain approximately four thousand data points. Each data point represents a dimension of the spectrum. When comparing hundreds or thousands of spectra, the high dimensional dataset generates thousands of millions of possible data features analysis of which, one by one is not feasible. Multivariate analysis offers a way to simplify the dataset, reduce its dimensions to only a few numbers, and can still represent all the data without sacrifice.

Depending on the whether the class of data is known or not before performing the analysis, there are two type of multivariate analysis, supervised and unsupervised analysis. In this project, an unsupervised principal component analysis (PCA)[12], a supervised principal component based linear discriminate analysis (PC-LDA) [13] and a supervised random forest analysis [14] will be employed.

**Principal component analysis (PCA)**

Three groups of spectra of different cell lines were integrated as one data matrix to perform the same pre-processing steps. As an unsupervised method, the algorithm does not need to know which spectra belongs to which group. Therefore, the PCA is able to explore the independent variance based on the dataset without manual intervention.

To perform PCA, the dataset after preprocessing is in the format of a $m \times n$ matrix, where $m$ is the number of spectra and $n$ is the number of data points in each spectrum. After applying the PCA algorithm, this matrix will be reconstructed into a $m \times l$ matrix and a $l \times n$ matrix. The production of those two matrix is the original dataset. In this manner, the $n$ dimensional dataset was reduced into a $l$ dimensional
The $m \times l$ matrix is named the score matrix and the $l \times n$ matrix is known as the loading matrix. These two matrices are defined as principal components (PCs), and the number of $l$ represents the number of principal components in the system. Each PC represents a certain percentage of variance in the dataset. The PCA algorithm automatically order the PCs based on how much variance is explained by each PC. Therefore, the first PC always explain the largest variance in the dataset, followed by the second and third PCs.

The explained variance recorded by each PC is represented by the loadings. A plot of loading against the corresponding wavenumber of data points will directly reveal the spectral features that cause the largest variance. The score matrix stands for the degree of how the actual data varies based on the loadings. By plotting the score against score for each PC, whether the explained variance in those PCs is capable of distinguishing the spectra taken from different cell line can be evaluated. It is noteworthy that the PC that discriminates between spectra is not necessarily the first PC which stands for the largest variance in the dataset.

**Principal component based linear discriminate analysis (PC-LDA)**

Unsupervised analysis is good to study independent data. However, the variance between the spectra of the same cell line and the difference between the spectra of a different cell line are equally weighted. When the variance within a sample group is larger than the variance between different groups, unsupervised analysis is, therefore, unable to distinguish different sample groups. Supervised analysis then offers the
possibility to maximise the variance between the various sample groups while minimising the difference in data within the same group.

Linear discriminate analysis (LDA) is a supervised multivariate analysis which finds the best linear hyperplane between different sample groups on each dimension. In order to do so, the class of data is reported to the algorithm beforehand.

As LDA compares the dataset in every dimension, when the number of dimensions is significantly larger than the number of samples, over-fitting may occur [15]. To address this problem, the dimension of the dataset can be first reduced by the PCA algorithm. The score of PCs is then input into the LDA algorithm instead of the original spectra matrix. This method is known as principal component linear discriminate analysis (PC-LDA). Still, when performing PC-LDA, the number of dimensions (PCs) that is input into LDA algorithm needs to be carefully controlled.

5.7 Random forest classifier

PC-LDA analysis indicates the possibility to separate the sample based on its chemical information. However, creating an analysis tool to detect an unknown sample with the features obtained from PC-LDA analysis can be complex and time-consuming. In this project, random forest algorithm was implemented to establish a classifier to test unknown sample.

Random forest is a machine training algorithm that generates a classifier based on an input dataset and employs the classifier to another dataset if required. This method creates a particular number of decision trees and obtains results depends on those trees. Assume a $m \times n$ dataset matrix has been input into the random forest algorithm with group information. The algorithm first creates one decision tree and randomly pick up $m$ number of data with replacement as the sample to train itself. Thus, some of the data may not be chosen and some may be chosen multiple times. Then the decision tree will randomly select $j$ number of features (data points) from the ‘training’ data.
The number $j$ is predefined by the user and should be far less than the number $n$, which is the number of dimensions of the datasets. After comparing the $j$ number of features, this decision tree picks the best feature, which can separate the data taken from a different group, and splits into two decision trees. Each of those two new trees performs the same operation as the first decision tree. When the new tree picks the same feature as its father tree, the new tree stops splitting. The process will continue until the selected feature can fully separate the data from different groups, or the number of trees reaches a predefined number. In this manner, the number of decision trees grows up and the trees form a forest. As for each tree, the sample and the feature are selected randomly, the forest is named a random forest. The random forest approach is good at giving high accuracy, handling high dimensional datasets and reporting important features.

As described above, the two key parameters in the random forest algorithm are, first the number of trees, second, the number of features picked by each tree. The number of trees, in theory, should be as high as possible to obtain the best result. However, in practice, the accuracy of a classifier does not increase after a certain number of trees and the more trees, the longer the calculation time. Thus a suitable number of trees is required. For each tree, some of the data are not selected by the tree after picking up the target data and are labelled as out of bag (OOB). Those OOB data are used to monitor the error rate of the tree and the accuracy of the final results. By a function that calculates the OOB error rate versus the number of trees, the best number of the trees can be found. For the number of features picked by each tree, it should be small but not too small, in case of over-fitting. Assuming only one feature is picked by each tree, the depth of the forest will be enormous, as any repeated feature will stop the split of the tree. This will give a perfect classification result as every single feature is taken into account. However, this classifier cannot be used to detect another dataset as it won’t be exactly the same as the training dataset. Assuming all the data points have been picked by each tree, the tree will simply give the same best feature and stop splitting which gives a poor classification result.
The random forest reports the importance of each feature depending on the number of times that feature is selected by the tree. The dimension of the dataset can be easily reduced by this function as all the non-important features can be filtered out. If only a few data points are required, the data measurement, such as using QCL microscope, and the data process speed can be very much faster.

In this project, the random forest method will be used to build a classifier to detect cells of different types of cell line. The spectra of cells taken from three cell lines will be randomly divided into two datasets, namely training set and validation set. The training set will be used to train the machine to build a classifier. The number of trees will be calculated using the OOB error rate function. The number of features selected is defined as the square root of the total number of the features in the data. After being trained, the classifier will be tested by the validation set to calculate the classification accuracy. Each spectrum in the validation set will be judged by all the trees in the random forest. The spectrum will be classified as the group which most of the trees think it is. By setting a threshold for the number of trees required to perform a valid result, the sensitivity and specificity of the result can be manipulated (see receiver operating characteristic). By comparing the result from the classifier to the known group information, the accuracy of the classifier can be calculated. This process will be repeated 100 times to find the lowest, mean and highest of accuracy.

**Receiver operating characteristic**

The receiver operating characteristic (ROC) is a curve which estimates the performance of a classifier (usually binary). The curve plots the true positive rate (sensitivity) against false positive rate (1-specificity). A simple explanation for sensitivity and specificity is: for a binary system with two group of sample A and B, the sensitivity is the percentage of A been classified as A; the specificity is the proportion of B been classified as B. For a system with more than two groups, the specificity will be the percentage of other groups being classified as other groups. The accuracy between each other group is not taken into account. The reason that this rate
can vary is because of the change of judgment threshold in the classifier.

For the random forest classifier mentioned above, each data will be voted by all the trees in the forest. The judgment threshold is the percentage of trees required to decide which class the spectrum should belong to. Again, assume a system with sample A and B, if the threshold for A is 100%, then the A will only be classified as A, if all the trees are agreed. Therefore, compared to a threshold with 50%, the sensitivity is reduced while the specificity is increased. If the threshold for A is 1%, then as long as a small number of tree vote for A, the data will be classified as A. Under this circumstance, the sensitivity is higher, and the specificity is lower than a threshold of 50%. By changing the threshold, the sensitivity against 1-specificity can be obtained, and the curve can be drawn. The smoothness of the ROC curve depends on the number of samples in the dataset and the number of the threshold used to generate the curve. After obtaining the curve, the area under the curve (AUC) is used to estimate the performance of the classifier. The value of AUC is usually varied from 0.5 to 1. For the same dataset, the larger the AUC, the better the performance of the classifier.


6. Benchtop FTIR imaging of cell in aqueous medium

6.1 Model of water correction

6.1.1 Effects of fitting different spectral region

The new model of water correction, as described in data analysis chapter 5.1, is based on fitting the protein peaks. However, the protein molecules form three broad peaks in the mid-IR region. Fitting with different individual peak or combination of peaks may lead to different results. Thus, the effect of the fitting range is evaluated first.

In order to obtain an unsaturated OH stretching band, the thickness of water layer is limited to 5.5 μm. A total of 50 cells (PC3) were measured with the benchtop FTIR spectrometer with the low magnification setup. The normalised spectra of PC3 cells in PBS are shown in Figure 6.1. The spectra were vector normalised and mean centred. Most of the biochemical features of the cells can be observed due to the thin water layer.

![Figure 6.1 Spectra of 50 PC-3 cells in PBS solution with a path length of 5.5 μm](image)
Figure 6. 2 Spectra of PC-3 cells obtained after water correction with different fitting regions. Each spectrum represents the average of 50 spectra. All spectra were normalized to amide I band. The spectrum of Matrigel is shown for reference.

Figure 6. 3 The fingerprint region of spectra in Figure 6.2. Results separated into two parts: (top) spectra fitted with amide A band; (bottom) spectra fitted without amide A band.
Figure 6.2 shows the mean spectra of water-removed cells obtained from the water correction algorithm having been applied using the different fitting regions. Each spectrum results from an average of 50 spectra. A spectrum of Matrigel is used as cell reference in the algorithm and therefore is also shown. Spectra were normalised to the amide I band for comparison. The spectral region of 3000-3800 cm\(^{-1}\) clearly indicates the overcompensation of water when the fitting range is limited to the amide I and II bands. A baseline shift between spectra fitted with or without amide A band can also be observed. By comparing the resultant spectra with a spectrum of Matrigel, the difference in lipids (2800-3000 cm\(^{-1}\) and 1450-1500 cm\(^{-1}\)) and nucleus region (1090 cm\(^{-1}\) and 1230 cm\(^{-1}\)) indicates the biochemical information in the non-protein spectral region will not be affected by the selected reference.

A zoomed fingerprint region for five combinations of the fitting parameter is shown in Figure 6.3. The spectra are highly consistent when the amide A band is included in the fitting range. However, when the amide A region is not used, the resultant spectra exhibit a different amide I/II peak ratio depending on the fitting range.

This is because the least squares fitting algorithm is biased on the absolute value of the dataset. Thus, when fitting with the broad amide A band, the algorithm is more likely to give a positive ‘residual’ in the OH stretching region in the corrected spectrum. When fitting with the amide I and II bands, the result will not take the amide A region into account and only focus the best fitting in the amide I/II region, which tends to give a clear amide I/II band. The size of the fitting range also affects the results. The spectrum fitted with both amide I and II band is overcorrected, but the level of overcompensation is less than the spectrum fitted with amide II only (Figure 6.2).
In theory, the best fitting range should be all three protein bands. However, in this project, the spectral range for the water correction fitting process is the amide I and II bands. There are two reasons: First, obtaining a credible OH stretching band is difficult. To achieve an absorbance of OH stretching band less than 2 requires a pathlength far less than 10 m. When measuring the cell with a 10 μm spacer, the OH stretching band is saturated and therefore cannot be used for the fitting process. This limits the possibility of measuring large cells and therefore reduces the application range of this method. Second, the OH stretching band region is not accessible with the QCL spectrometer. Table 2 shows the available range of the three spectrometers used. Obviously, when applying the water correction method to the spectra recorded from QCL microscope, the only protein bands available are amide I and II. Thus, to have a universal water correction method for all three spectrometers, the spectral region of 1500-1700 cm\(^{-1}\) was selected for the fitting process.

Although overcompensation can be expected, the resultant spectrum should still be able to reveal most of the biochemical information and most importantly allow a valid comparison between data sets that have all been treated using the same methodology.

<table>
<thead>
<tr>
<th>Available spectral region/cm(^{-1})</th>
<th>FTIR</th>
<th>SR-FTIR</th>
<th>QCL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>950-4000</td>
<td>500-6000</td>
<td>1000-1800</td>
</tr>
</tbody>
</table>
6.1.2 Choice of reference

The water correction method utilised a reference to represent the cell contribution. Therefore, the effects of different references should be evaluated. In this section, the water-corrected cell spectra obtained using two different reference spectra will be compared.

Figure 6.4 shows the spectra of Matrigel and air dried PC-3 cell. The difference in these spectra is evident. The amide A bands of both spectra have the same centre position, but the shape of the band is altered. The spectrum of the dried cell has much sharper lipid bands. The amide I/II peak ratio in Matrigel is higher than the ratio in cell spectrum. Difference in remaining fingerprint region can be observed as well.

To evaluate the effects of the reference, 50 PC-3 cell spectra (same as the previous section) were fitted with both reference spectra in the spectral region with or without amide A band. Figure 6.5 (top) indicates the fingerprint region of mean corrected spectra obtained from fitting with both references in all protein regions. The resultant spectra are almost identical to each other. Both of the corrected spectra present most of the key absorbance peaks and therefore it can be concluded that they are essentially

Figure 6.4 spectrum of Matrigel (blue) and spectrum of air dried PC3 cell (green). Vector normalised and mean centred.
independent of the reference used. However, when fitting with amide I and II band only, the resultant spectra exhibit a reference dependency, as shown in Figure 6.5 (bottom). The spectrum fitted with Matrigel gives a higher amide I/II peak ratio compared to the spectrum fitted with cell spectrum. This suggests that the amount of water removed by fitting with cell spectrum is significantly larger than fitting with Matrigel. This also can be concluded by the much clearer small features in the remaining fingerprint region. However, overcompensation has already occurred when fitting with the Matrigel spectrum at amide I and II bands. Utilising a cell spectrum as a reference will remove extra water and lead to a much worse overcompensation.

Two aspects were considered when selecting the fitting reference for the water correction method. First, the level of overcompensation of water needs be controlled
to some extent. Employing the Matrigel spectrum as cell reference will result in less water being removed. Second, to fit the spectra of the cells in water with the spectrum of the cell from the same cell line, the spectrum of the dried cell needs be obtained beforehand. This requires and additional experimental process and will be time-consuming when measuring samples from a large number cell lines. For the above reason, the Matrigel will be used as cell reference for all the data analysis in this project.

6.1.3 Validation

To evaluate the real performance of the water correction method and justify whether this approach is able to extract differences in spectra, preliminary validation experiments were carried out.

The PC-3 cells were fixed using two different fixation processes, namely ethanol fixation and formalin fixation. 50 cells of each kind were measured with a path length of 10 μm. Recorded spectra were quality controlled, vector normalised and mean centred.

Figure 6.6 demonstrates the comparison of selected features in the spectra of two samples. The lipid band (2800-3000 cm⁻¹) of two mean spectra of fixed cells is shown in Figure 6.6 (top). It is evident that the –CH₂ and –CH₃ stretching bands have been fully resolved after water correction. For the spectrum of cells fixed with ethanol, the loss in symmetric and asymmetric stretching band of –CH₂ is due to the fact that alcohol solubilises lipids [1]. Lipids, as polar molecules, will interact with ethanol and the van der Waals interaction between the acyl chains will be broken. The lipids are also soluble in high concentration ethanol solution. On the other hand, the formalin solution forms strong bonds with lipids and preserves the content and structure of lipids [2].

Figure 6.6 (bottom) indicates the amide I and II band of the spectra of the cells. The broadened amide I peak is due to the denaturation of protein. The ethanol reduces the
α-helix segments (1650 cm\(^{-1}\) and 1550 cm\(^{-1}\)) and leads to the formation of β-sheet (1625 cm\(^{-1}\)) and random aggregations (1695 cm\(^{-1}\)) [3]. Due to the complexity of the protein system, the secondary structure of the protein will not be discussed in detail.

The validation results give clear evidence of the feasibility of the water correction method. The resultant spectra were able to reveal important biochemical information which been reported by a previous study [4]. The non-protein region was well resolved, and the protein bands still retain the shape of secondary structure, even under the overcorrection of water.

In summary, the new water correction with compromised fitting parameter is capable of removing the water contribution in the spectrum of cells in aqueous media. The cell
spectra obtained are able to distinguish the chemical difference induced by two types of fixation process. The overcompensation of water can be expected but will not significantly alter the biochemical information. However, in order to minimise the effects of overcompensation, the protein bands (1500-1700 cm\(^{-1}\)) should only be used with an element of caution regarding over interpretation.

6.2 Spectra obtained in low mag setup

6.2.1 Cell imaging

Figure 6.7 Two examples (rows) of cell imaging. (left column) visible images of cells in PBS solution; (middle column) intensity images of amide II peak centre (1550 cm\(^{-1}\)) before water correction; (right column) intensity images of amide II peak centre after water correction. IR image has a FOV of 704x704μm\(^2\)

One of the key advantages of measuring live cells with an IR spectrometer coupled to an imaging microscope is the high throughput of the measurement, i.e. the capacity of measuring a large number of cells in a single image. Figure 6.7 shows an image of cells in aqueous solution. For each row, the IR image on the right corresponds to the visible image on the left. The two IR images in each row presents the absorbance value
The amide II peak (1550 cm⁻¹) before (middle) and after (right) water correction. 

The visible image demonstrates the number of the cells that can be captured within one single image. By comparing the visible image with the water corrected image, it is found that the IR camera actually produces a larger field of view than the visible camera. As the spectrum of water reference is obtained from the IR image, a large enough non-cell region should be captured for every image.

The IR image before water correction shows nothing but a nearly homogeneous block. After removing the water contribution, the OH bending band in the non-cell region is removed which increases the contrast of cell/water and exposes the location of cells.

Due to the low spatial resolution, the cell is blurred in the IR image. When clumping together, the cells cannot be separated directly from visual inspection of the IR image. Manually identifying the cells based on a visible image is difficult and time-consuming. Thus, a cell extraction algorithm was applied to obtain the spectrum of each single cell automatically. After removing the water contribution, the cell was separated based on the lipid bands, since these bands appear in the shorter wavelength region of the spectrum and thus will give a better spatial resolution. The spectra of pixels in each cell region in the image were then averaged to obtain the spectrum of that cell.

6.2.2 Cell spectra

The spectra of cells extracted from the IR image are shown in Figure 6.8. Three sub figures illustrate the spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom). Each figure contains the spectra taken from at least 10 measurements from more than 3 times of sample loadings. Spectra were vector normalised and mean centred.

The saturation of the OH stretching band can be clearly observed in the form of spikes and the altered peak shape in the spectra. The saturation makes the information in the spectral region above 3000 cm⁻¹ unusable. Due to the thick water layer, most of the cell features are hidden, and only the amide II band is notable in the spectra. The large variance in the amide I band resulted from the vector normalisation due to the saturated
OH stretching band.

To reveal the biochemical information of cells, the spectra were water corrected, and the results are shown in Figure 6.9. It is evident that the overcompensation of water has been controlled at a reasonable level. The spikes in the amide A band remain in the spectra and therefore it is not suitable to be analysed further. The significant fluctuation in the amide A band also indicates the correction method, which works on a cell to cell basis, will result in variations due to the different amount of water being removed from the cells.

The amide I and II bands and the C-H stretching bands can be observed in all the spectra. The maxima of amide I band in all three set of spectra are at an absorbance of approximately 0.2. The variance of those spectra results from the different cell thickness, baseline shift and water correction. The different height of the amide I band is again due to the scale to the spikes in the amide A band region.

After cutting the spectral region above 3000 cm\(^{-1}\), vector normalisation and mean centring, the spectra are shown in Figure 6.10. The amide I and II band are well resolved and the rest of the cell features are clearly illustrated. The peak ratios of amide I to amide II band in all spectra are reasonably consistent. The lipid region observed in spectra of LNCaP cells have sharper –CH\(_3\) stretching bands than the spectra of the other two cells. The large variance in the cell spectra results from different effects. First, the amount of water which has been removed from each spectrum is different. This leads to a baseline shift in lipid bands (2800-3000 cm\(^{-1}\)) and the tail of the spectra (1000-1200 cm\(^{-1}\)). Second, the low mag setup cannot clearly resolve the cell boundary, and the cell extraction algorithm is robust. This results in the surrounding non-cell part being taken into account when averaging the pixels.

Small oscillations can be found in the spectral region between 1800 cm\(^{-1}\) to 2800 cm\(^{-1}\) in the spectra. This periodical oscillation is suspected to be the fringing effect as described in chapter 2. The fringes are correlated and cannot be removed by averaging.
Figure 6. 8 raw spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells in PBS solution. Spectra were vector normalised and mean centred.
Figure 6. 9 corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells.
Figure 6. 10 corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells. The spectral region above 3000 cm$^{-1}$ were cut out and the remaining part were vector normalised and mean centred.
This type of variance may distort the result of further analysis and therefore needs to be eliminated. Before applying multivariate analysis on the resultant dataset, a mathematical method to remove the fringing pattern will be implemented. It is noteworthy that the fluctuations are not consistent across the whole wavenumber range and do not appear in all of the spectra. To compare the results from different cell lines, the corrected spectra were mean centred, vector normalised and averaged. The average spectra of three cell lines are shown in Figure 6.11. Apart from the amide A band, the water correction results in spectra containing essentially all of the peaks associated with the cell components. The amide A band region suggests that more water has been removed for the spectra of PNT2 and PC3 cells than LNCaP cells. The positive amide A band in the spectrum of LNCaP leads to large unit vector in vector normalise process. Therefore, the spectrum of LNCaP has a lower baseline in the fingerprint region. The low-frequency tail of the spectra shows an abnormal increasing trend which could result from the combination of water layer and the cut off of the CaF₂ windows.

Figure 6.11 mean spectra of PC-3 (red), LNCaP (blue) and PNT2 (green) cells. Spectra vector normalised based on the full spectra region and mean centred.
6.3 Attempted oscillation removal process

Figure 6.12 shows the spectra of PC3 cells with ‘fringes’. All spectra were obtained from one single imaging experiment. Spectra after the water correction process were vector normalised and mean centred. The clear periodical fluctuations can be observed in the spectral region of 1800 to 2800 cm\(^{-1}\). The amplitude of this fluctuation is approximately 0.02 in absorbance which is similar to the absorbance of the fingerprint region from 1000 to 1500 cm\(^{-1}\). If the fluctuation is due to the fringing effects, the spectra will not be credible for further analysis. As the periodical oscillation has a significant degree of similarity to the fringing effects, a mathematical algorithm which removes fringes is employed.

Fringes exist in the spectra in the form of periodical oscillation. This oscillation can be expressed by the combination of a sine wave and a cosine wave with the same period of the oscillation. The spectral region of 2500-2700 cm\(^{-1}\) was extracted out to calculate the period of the fluctuation.

Figure 6.13 shows the fluctuation extracted from a cell spectrum. The path length leads
to this fringing pattern can be calculated based on Equation 9. Three periods of oscillation are found in a range of 200 wavenumbers which gives a path length of 75μm. However, there is nothing in the experimental setup which gives a 75μm path length. The CaF$_2$ windows have a thickness of 500μm each. The thickness of water layer is approximately 10μm.

Although the path length suggests that the fringing is not from the sample, removal of such oscillation will still benefit the future analysis. By converting the spectrum from frequency domain to time domain, the spectrum will transfer to a series of complex numbers with the same number of data point as the spectrum. The amplitude of the spectrum in time domain was then calculated and plotted out in Figure 6.14.

The spectrum in Figure 6.14 (top) can be treated as an interferogram. The sinusoidal waves with different amplitude and frequency were added up and formed a peak next to the time zero in the time domain. The location of this peak represents the period ($2\pi/x$) of the oscillation in the frequency domain. After obtaining the value of $x$ of 0.093, two sinusoidal waves were generated as $\sin(x\tilde{u})$ and $\cos(x\tilde{u})$. The simulated wave of $\sin(x\tilde{u})$ is shown in Figure 6.14 (bottom). By fitting two sinusoidal wave to the spectra, the oscillation should be removed. However, the algorithm not only failed to eliminate the oscillation from the spectra but also induced extra fluctuation into the
spectra. The fitting result is shown in Figure 6.15.

First, the fluctuations in the spectra become more obvious, and the amplitude is increased. Second, the other regions of the spectrum are significantly altered by this oscillation and artificial peaks in the fingerprint region are observed. Thus, these
results cannot be used for further analysis.

The algorithm may have failed to remove the oscillation due to the following reasons. First, the fluctuations may not be generated from the sample. The path length for the measured fringes does not match any thickness in the sample or sample holder. If the fringes are from the spectrometer, the background spectrum should also contain the fringes, and therefore it should be removed in the final recorded the spectrum. Second, not all the spectra contain the oscillations, and the amplitude of fluctuation varies across the wavenumber region. Thus, fitting with two sinusoidal waves with constant amplitude may be inappropriate.

Despite the failure of removing the fringes, the fingerprint region in the resulted spectra exhibit features nearly identical to the area in the spectrum of the dried cell. This suggests that the fingerprint region is not affected by the oscillation, or the interference is negligible. Moreover, the strong oscillation in the baseline is only observed in the initial stage of the experiment. For most of the spectra analysed in this project, the periodical fluctuation cannot be observed, or the amplitude is too small to be noticed. Anyway, to avoid the variance from correlated oscillation, the spectral region of 1800 to 2800 cm\(^{-1}\) will not be used for further analysis.
6.4 Multivariate analysis on ‘low mag’ spectra

As the aim of this project, the water correction method should enable different cell lines to be distinguished, based on the corrected spectra. To evaluate the performance of the algorithm, the resultant spectra were analysed with multivariate analysis. The result of PCA, PC-LDA and random forest classifiers for spectra of cells obtained in the FTIR spectrometer with low mag setup will be reported in the following section.

6.4.1 PCA

Before performing the PCA method, some pre-processing steps were applied to the spectra. First, depending on the situation, the spectra may be transferred into the 1\textsuperscript{st} or 2\textsuperscript{nd} derivate. 11 points and 19 points smoothing function were applied for 1\textsuperscript{st} and 2\textsuperscript{nd} derivate, respectively. Second, due to the variance introduced by the water correction method and the limitation of the experiment, only the spectral regions of 1200-1400 cm\textsuperscript{-1}, 1700-1800 cm\textsuperscript{-1} and 2800-3000 cm\textsuperscript{-1} in the spectra were kept for the analysis. Third, the spectra were then vector normalised and mean centred based on the whole selected spectral region after been cut. In this sequence, the influence from other parts of the spectrum in the normalisation process can be avoided.

Figure 6.16 shows the average of the original corrected spectra, 1\textsuperscript{st} derivate spectra and 2\textsuperscript{nd} derivate spectra for the three cell lines. Significant baseline variance can be observed in the spectral region of 1200-1400 cm\textsuperscript{-1} in the raw spectra. This may result from the water correction algorithm and can be removed by the derivatization of the spectra. In the 1\textsuperscript{st} derivate spectra, the spectra of LNCaP exhibit a more drastic change in the lipid band at 1730 cm\textsuperscript{-1}. The difference in the mean of 2\textsuperscript{nd} derivate spectra was subtle.

The result of PCA analysis for cell spectra without derivation is shown in Figure 6.17. It is obvious that there is no separation between spectra from different cell lines. From
Figure 6. Mean of selected spectral region (1200-1400 cm⁻¹; 1700-1800 cm⁻¹; 2800-3000 cm⁻¹) in raw corrected spectra (top), 1st derivate spectra (middle) and 2nd derivate spectra (bottom) of PC3 (red), LNCaP (blue) and PNT2 (green). Spectra were vector normalised and mean centred.
Figure 6. 18 (left) PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with low mag setup. (right) 1st and 2nd loadings correspond to each PC.

Figure 6. 17 (left) PCA plot for 1st PC against 2nd PC of 1st derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with low mag setup. (right) 1st and 2nd loadings correspond to each PC.
the first PC, the largest variance is located in the lipids band at 2800 to 3000 cm\(^{-1}\). Over 90% of data variance has been explained by the first two PC. Therefore, it suggests that the difference in the spectra from the same cell line is larger than the difference between the different cell lines.

The tail of OH stretching bands will affect the height of the CH stretching bands. Thus for each cell line, the spectra of cells, which have different water contribution been removed, will result in lipids bands with similar shape but varies the height of baseline.

After removing the variance from the baseline and enlarge the variance from biochemical difference, the PCA results for the 1st derivate spectra were shown in Figure 6.18.

The LNCaP cell line was separated from the other two cell lines by the 1st derivate spectra. In the 1st PC, the largest variance in the data set was found in the tail of the amide I band at 1700 cm\(^{-1}\). This can result from the difference in the amide I band due to the water correction method. No separation can be found based on 1st PC.

A separation between LNCaP cells and the other two cell lines can be observed.

---

Figure 6.19 (left) PCA plot for 1st PC against 2nd PC of 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with low mag setup. (right) 1st and 2nd loadings correspond to each PC.
according to the second PC. The loading of 2\textsuperscript{nd} PC suggests that the LNCaP cells are different from the other two cell types from the lipid bands. However, the variance between spectra of PC-3 and PNT2 cells was still not significant enough to separate them.

To elucidate the subtle difference between those three cell lines, the 2\textsuperscript{nd} derivate of the spectra was obtained, and the result of PCA analysis is shown in Figure 6.19. The LNCaP cells were clearly separated from PC-3 and PNT2 cells in both 1\textsuperscript{st} and 2\textsuperscript{nd} PCs. Both the 1\textsuperscript{st} and 2\textsuperscript{nd} loadings suggest that the LNCaP was separated from PC-3 and PNT2 due to the difference in lipids. The difference in nucleus acid is also observed in 1200-1400 cm\textsuperscript{-1} in 2\textsuperscript{nd} PC loading. The results further confirm the separation of the cell lines is based on the chemical difference which agreed with the previous study carried out by T.J.Harvey \textit{et al.} [5, 6]. However, the PC-3 and PNT2 cells are still mixed together.

Through the PCA analysis, the LNCaP cell was successfully separated from the PC-3 and PNT2 cell line due to the significant variation in the lipid content. However, the PCA as an unsupervised analysis method is not able to distinguish between PC3 and PNT2 due to the variance between spectra of cells from same cell line is still larger than the variance between different cell lines. These three cell lines were studied in the dried fixed state, and PC-3 and PNT2 were also not separated by PCA [5, 6].

\textbf{6.4.2 PC-LDA}

As the unsupervised PCA analysis is unable to make a distinction of the variance between the same cell line and different cell lines, a supervised PC-LDA algorithm is employed. The data set used for PC-LDA is the scores obtained from the PCA analysis. Thus, three PC-LDA results corresponding to the original spectra, 1\textsuperscript{st} and 2\textsuperscript{nd} derivate were carried out. It is noteworthy that the number of PCs used for PC-LDA is much smaller than the total number of PCs. This is because the first 20 or 30 PCs already
explained more than 99% of the variance. The number of PCs input into the PC-LDA needs to be as small as possible to explain as much variance as possible. Thus, in this analysis, once the number of PCs is enough to explain 100% of the variance, the remaining PCs will be removed from the dataset.

Figure 6.20 indicates the PC-LDA results of the raw spectra, 1st derivate spectra and 2nd derivate spectra for the three cell lines. From Figure 6.20 (top), all the data points, which represents the raw spectra, were mixed together. The variance between different cell lines in the PCs is too small to separate the spectra. The round curve in the bottom of the figure suggests there is systematic variance in all of the spectra. This can result from the water correction process.

The PC-LDA results for the 1st derivate spectra are shown in Figure 6.20 (middle). Although small overlapping can be observed between the outlier of PC3 cells and PNT2 cells, the three cell line were successfully separated. However, this separation may result from the water correction algorithm as the largest variance were found in the tail of the protein band at 1700 cm⁻¹.

Figure 6.20 (bottom) shows the PC-LDA result for the 2nd derivate spectra. The three cell lines were clearly separated, and the overlap between PC3 and PNT2 cell lines was minimised. The 2nd derivate spectra should not be affected by any baseline difference which suggests the separation is truly dependent on the chemical difference between the three cell lines.

The results have proven that the spectra can be separated by a supervised analysis method. The difference between LNCaP and the other two cell lines is much larger than the difference between PC-3 and PNT2 cells. The same trend of separation was reported in a dried cell study [5, 6]. These three cell lines were also studied with Raman spectroscopy and the separation was demonstrated [7]. This gives the opportunity to build up the classifier to test unknown cell samples.
Figure 6. 20 PC-LDA results based on the raw spectra (top), 1st derivate spectra (middle) and 2nd derivate spectra (bottom) of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines.
6.4.3 Random Forest

The random forest algorithm offers a convenient way to create a classifier and is able to select the most important features that can be used to distinguish the data.

As the best result from PC-LDA are based on the spectra in the form of 2\textsuperscript{nd} derivative, only the 2\textsuperscript{nd} derivative spectra will be studied with the random forest method. The derivative spectra were cut to the spectral region of 1200-1400, 1700-1800 and 2800-3000 cm\textsuperscript{-1}. Vector normalisation and mean centreing were also employed on the spectra. 500 spectra from each cell line were selected and integrated as the input dataset.

The dataset was then divided into two sets. 60\% of the data for each group was randomly chosen as the training set, and the remaining spectra were used as the validation set. After establishing the classifier with the training set, the validation set is subjected to the classifier to test the performance. This process was repeated 100 times to obtain the minimum, mean and maximum accuracy of the classifier.

The accuracy of the classifier for each cell line is summarised in Table 3. The average accuracies of all three cell lines are above 99.5\%, which is nearly perfect. The minimum accuracy is also above 98\%. This suggests that the classifier based on the 2\textsuperscript{nd} derivate of spectra can successfully verify the unknown sample with high accuracy. The ROC curve is not presented, as the performance of the classifier is good enough.

<table>
<thead>
<tr>
<th>MIN</th>
<th>AVG</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.1%</td>
<td>99.8%</td>
<td>100%</td>
</tr>
<tr>
<td>99.3%</td>
<td>99.7%</td>
<td>100%</td>
</tr>
<tr>
<td>99.3%</td>
<td>99.8%</td>
<td>100%</td>
</tr>
</tbody>
</table>
One of the advantages of the random forest algorithm is the feature selection function. After building up the classifier, an index of importance can be generated. Figure 6.21 illustrates the importance of the feature in the classifier. The magnitude corresponds to the performance using that feature to separate data. By selecting the feature with high importance, the dimension of the dataset can be reduced further to less than 20 discrete wavenumbers. The 20 most important wavenumbers for this classifier have been listed in Table 4. As the dataset is the 2nd derivate of spectra, the feature location should correspond to the same wavenumber in the original spectrum. Thus, the biochemical bands for each feature are included in the list as well.

Table 4 20 most important feature for classifier of PC3, LNCAP and PNT2 with correspond vibration

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>vibration</th>
<th>Wavenumber</th>
<th>vibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1223</td>
<td>( \nu_{as}(PO_2^-) )</td>
<td>1750</td>
<td>( \nu(C = O) )</td>
</tr>
<tr>
<td>1235</td>
<td></td>
<td>2853</td>
<td>( \nu_s(-CH_2) )</td>
</tr>
<tr>
<td>1237</td>
<td></td>
<td>2872</td>
<td>( \nu_s(-CH_3) )</td>
</tr>
<tr>
<td>1312</td>
<td>( \delta CH )</td>
<td>2874</td>
<td></td>
</tr>
<tr>
<td>1375</td>
<td>( \delta CH_3 )</td>
<td>2886</td>
<td>( \nu_{as}(-CH_2) )</td>
</tr>
<tr>
<td>1395</td>
<td></td>
<td>2957</td>
<td></td>
</tr>
<tr>
<td>1397</td>
<td>( \nu(CO_2^-) )</td>
<td>2959</td>
<td></td>
</tr>
<tr>
<td>1399</td>
<td>( CH - NH )</td>
<td>2961</td>
<td>( \nu_{as}(-CH_3) )</td>
</tr>
<tr>
<td>1400</td>
<td></td>
<td>2963</td>
<td></td>
</tr>
<tr>
<td>1748</td>
<td>( \nu(C = O) )</td>
<td>2965</td>
<td></td>
</tr>
</tbody>
</table>
6.5 Spectra obtained in high mag setup

6.5.1 Cell imaging

The cell images based on the integrated intensity of lipid bands from 2800 to 3000 cm\(^{-1}\) in the spectra of PC3 cells are shown in Figure 6.22. The most straightforward difference between the cell images taken from the high mag and low mag setups are the number of cells and the pixel size. The high pixel resolution and the small field of view lead to a more homogenous non-cell background and a clearer cell boundary.

The spatial resolution for the cell images varies from 3~10μm for the objective with NA of 0.61. The resolution at the –CH stretching band region (3.3μm) is smaller than the pixel resolution at low mag setup, which is 5.5μm. Thus, there is a trade-off when identifying the boundary of a cell in the low mag image. The average diameter of the three cell lines is around 30μm, which is 6~7 pixels. When dealing with the edge of the cells, if all the pixels in a cell are taken, some proportion of the non-cell background will be taken as well. This will lead to the distortion in the spectra due to the diffraction limit. If all the edge pixels are rejected, then a large proportion of information is lost.
For the high mag setup, the pixel resolution is increased 5 times to 1.1μm which can clearly identify the boundary of the cells, as shown in Figure 6.22 (right). The equivalent average pixel diameter of a cell is 27 pixels. By rejecting a thin layer of one or two pixels, the distortion from the non-cell region can be avoided and most of the cell information is retained.

The image taken from the high mag setup has a high noise level. This is due to the amount of light received by the detector which is reduced 25 times while the field of view been enlarged. The low light intensity leads to the low SNR compared to the low mag setup. However, the number of pixels for each cell was increased by 25 times and therefore the averaged cell spectra should have a similar SNR as the spectra taken from the low mag setup.

### 6.5.2 Fitting-average or average-fitting?

As mentioned in the data processing section, different sequences of fitting process for data from high mag setup were tested. It was hypothesised that doing a water correction (fitting) on each pixel of the cell image and then averaging the pixels

---

![Figure 6. 23 PCA on spectra with different sequence of fitting process: the fitting-average process (blue circle) and the average-fitting process (red dot)](image-url)

---
making up the image of the cell might give a different result to averaging all the pixels to create a single spectrum of the cell and doing a single water correction. Computationally, the latter is faster, but it is important to know how the resultant spectra might change. After vector normalisation and mean centring, the resultant spectra from the fitting-average and the average-fitting process was analysed with PCA. The PCA result for those two groups of spectra is shown in Figure 6.23.

The plot of 1\textsuperscript{st} and 2\textsuperscript{nd} PCs clearly indicates that the cell spectra resulting from two different sequences of the fitting process are identical to each other. The first two PCs represent 71.2 percentage of variance. The rest of the PCs are also examined, and the data perfectly overlap (results not shown).

The main difference between the SR-FTIR and benchtop FTIR spectrometer in this project are the light source and the detector. The spectrum obtained from a single point detector can be treated as the average of spectra obtained from the multichannel detector with the same field of view. The result above firmly proved that the averaging of spectra does not affect the water correction results. Therefore, the difference in the spectra obtained from SR-FTIR and benchtop FTIR spectrometer will not result from the detector.

6.5.3 Cell spectra

The cell spectra recorded from the high mag setup after water correction are shown in Figure 6.24. 100 cell spectra were obtained from at least 15 images with more than 5 samples for each cell line. The spectra were cut to the spectral region of 1000-3000 cm\textsuperscript{-1}. Vector normalisation and mean centring were applied as well.

The spectra from the high mag setup evidently have advantages compared with the spectra taken from low mag setup. First, all the key peaks are nicely resolved and clearer. The variance between the spectra of the same cell line is minimised. Second, the SNR are
Figure 6. 24 water corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells. The spectral region above 3000 cm\(^{-1}\) were cut out and the remaining part were vector normalised and mean centred.
reasonable, and the spectral region of the lipid bands and nuclear bands are consistent. The periodical oscillation cannot be observed in the spectra.

The better spectral quality may result from two aspects. First, the much clearer cell boundary avoids the distortion in the spectra from the non-cell region. For the low mag setup, the cell was extracted by a mathematical method in which the resultant cell boundary is not ideal. The low spatial resolution also induces the distortion from the edge into the spectra. Second, the average intensity for pixel varies across the image in the low mag setup (see Figure 6.22 left). Thus, fitting the whole image with the same buffer spectrum may lead to large variance in the resultant spectra. For the high mag setup, due to the small field of view, each cell spectrum was fitted with the buffer spectrum taken next to the cell. Therefore, the fitting process was more accurate which will give a more consistent fitting result.

The mean spectra of the three cell line are shown in Figure 6.25. The variance between the three cell lines is minimised, but there are still some differences that can be observed. For example, the LNCaP cell has a sharper -CH₃ stretching band in the region of 2800-3000 cm⁻¹. A higher amide I/II peak ratio is notable in spectra of LNCaP.
as well. Unlike the spectra obtained from the low mag setup, no increasing absorbance in the tail of spectra at 1000 cm$^{-1}$ can be observed in spectra from high mag setup. A small oscillation can be found in the baseline, but the amplitude is too low to significantly affect any biochemical features. However, for consistency, the spectra will be cut into the same spectral region as the low mag setup.

6.5.4 PCA

To explore the difference between the cell lines, the spectra were analysed with the PCA algorithm. The results in Figure 6.17 suggest that the 1st derivate spectra were not able to eliminate the changes in the tail of amide I band which resulted from the water correction method. Thus, for the data recorded from high mag setup, only the original spectra and the 2nd derivate spectra were analysed with the PCA algorithm. The spectra were cut into the same spectral regions of 1200-1400 cm$^{-1}$, 1700-1800 cm$^{-1}$ and 2800-3000 cm$^{-1}$ as the low mag setup for consistency. The spectra were vector normalised and mean centred after being cut.

The PCA results from the original spectra are shown in Figure 6.26. Although the variance between the spectra of the same cell line is minimised, there is still no separation to be found in the plot. The variances expressed by both 1st and 2nd PC loadings suggest that the largest variance is from the slope of the baseline. The two loadings show a reversed feature in the lipid bands. This also indicates that the variance introduced by the water correction method was diminished but cannot be avoided by increasing the spatial resolution and number of spectra.

Figure 6.27 indicates the results from the 2nd derivative spectra. Surprisingly, not only the LNCaP is spotted out but also the PC3 and PNT2 cells were separated. The LNCaP cells were separated based on the 1st PC while the PNT2 cells were separated based on the 2nd PC. The separation between PC3 and PNT2 may not be completely reliable as the number of spectra is small and the large variance in the spectra of PC3 cells spread
Figure 6. 26 (left) PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with high mag setup. (right) 1st and 2nd loadings correspond to each PC.

Figure 6. 27 (left) PCA plot for 1st PC against 2nd PC of 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with high mag setup. (right) 1st and 2nd loadings correspond to each PC.
the data out. From both loadings, the main variance is confirmed as lipid band. A strong
difference in the nucleus region at 1235 cm\(^{-1}\) can be observed in 2\(^{nd}\) PC. This peak can
also be found in Figure 6.19. However, a clear difference at 1317 cm\(^{-1}\) can be noted
between the 2\(^{nd}\) loadings in Figure 6.19 and Figure 6.27. Whether this feature is
important for cell classification will be discussed later.

6.5.5 PC-LDA

As the 1\(^{st}\) derivate spectra were not obtained in PCA analysis, only the original spectra
and 2\(^{nd}\) derivate spectra will be analysed through PC-LDA method. The smallest
number of PCs which explained 100\% of variance were used as raw data. The results
were shown in Figure 6.28.

The LNCaP and PNT2 cell line were separated based on the original spectra as shown
in Figure 6.28 (top). The LNCaP cells were separated from PNT2 cells. The variance
in the PC3 cell is comparably larger which causes the data to be widely spread in the
score plot. Figure 6.28 (bottom) shows the result from the 2\(^{nd}\) derivate of spectra. The
three cell lines are separated with some overlapping and outlier features. As a
supervised analysis method, the PC-LDA requires a certain number of data to minimise
the variance in the spectra of cells in the same sample group. The 100 cells for each
cell line might not cover enough variance from the cell cycle stage which leads to the
imperfect separation.

The PC-LDA results suggest that the water correction is capable of revealing enough
biochemical features to classify three different cell line under high pixel resolution.
The separation is based on the similar feature as the results taken from the low mag
setup. The separation is evident despite the outliers and overlapping due to the small
number of cells. However, whether the number of measured spectra is enough to build
up a high-performance classifier requires further analysis.
Figure 6. 28 PC-LDA results based on the raw spectra (top) and 2nd derivate spectra (bottom) of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines obtained from high mag setup.
6.5.6 Random forest

The 2nd derivative of the spectra was used to establish the classifier. The spectra, for consistency, were limited to 1200-1400 cm⁻¹, 1700-1800 cm⁻¹ and 2800-3000 cm⁻¹. Vector normalisation and mean centre are applied as well. 100 spectra each cell line were integrated as the dataset. The dataset was split into the training set and validation set with a ratio of 6:4. The results from 100-time calculations were obtained, and the range of accuracy was shown in Table.5.

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>80.56</td>
<td>81.25</td>
<td>82</td>
</tr>
<tr>
<td>AVG</td>
<td>92.83</td>
<td>93.63</td>
<td>91.07</td>
</tr>
<tr>
<td>MAX</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5 Accuracy (sensitivity) of classifier for each cell line

The performance of the classifier is lower than the classifier used for the spectra from the low mag setup. The main reason for lower accuracy may be due to the small sample amount. The number of spectra used for classifier from the high mag is 5 times less than the number of spectra used from the low mag. The number of spectra in the dataset significantly affects the feature selected by the decision tree and therefore, the small data set, the poorer the performance of the classifier. In the meantime, for 100 spectra for each cell line, the variance may not be able to cover the biochemical changes due to the different cell cycle stage. However, the average accuracies for three cell lines are still above 90%, and the minimum accuracy was higher than 80%.

The receiver operating characteristic (ROC) curve of the classifier is shown in Figure 6.29. The area under the curve (AUC) for PC3, LNCAP and PNT2 were calculated to be 0.982, 0.985 and 0.980. The value is close to 1, which stands for a perfect classifier.
The sensitivity of each cell line can be achieved above 95% with only 10%–20% sacrifice of the accuracy of other two cell lines if required.

6.5.7 Cross validation between dataset from high and low mag setup

According to the PCA analysis, the variance between data from high and low mag setup is expected. Thus, the two datasets were cross-validated by building a classifier with one dataset to test the other dataset. The spectra were underwent the same pre-processing to ensure consistency. The classifier was repeated 100 times and the mean accuracies were reported.

The accuracy of the classifier established with the high mag results to test low mag

| Table 6 mean accuracy of classifier build with high mag data to test low mag data |
|--------------------------------|---------|--------|
| PC3  | LNCAP   | PNT2   |
| Accuracy | 78      | 91.7   | 62.6   |

| Table 7 mean accuracy of classifier build with low mag data to test high mag data |
|--------------------------------|---------|--------|
| PC3  | LNCAP | PNT2  |
| Accuracy | 28      | 78    | 82    |
results is shown in Table 6. The PC3 and LNCaP cell line are still reasonable, but the accuracy for PNT2 was as lower as only 62%. Conversely, when using the classifier of low mag data to test high mag data, the accuracy of PC3 is far worse than 30%, as shown in Table 7. Therefore, it can be concluded that the spectra obtained from the low mag setup contain the features required by the classifier build by the spectra from the high mag setup. But the high mag setup does not have enough features to meet the requirement of the classifier constructed by the spectra from low mag setup. This suggests there are significant differences in the spectra of PC3 obtained from low and high mag setup. In both classifiers, the accuracy of LNCaP is relatively stable. This matches the PCA analysis which the LNCaP cell line can be easily separated but the PC3 and PNT2 are not.

To examine the difference in the spectra of PC3 obtained from different setup, the 2nd derivate spectra were compared and shown in Figure 6.30. Variances can be observed in the spectral region of 1200-1320 cm\(^{-1}\) and 1700-1800 cm\(^{-1}\). The difference may result from two aspects. First, the recorded sample amount for low mag setup is 5 times larger than high mag setup. The spectra taken from the high mag setup may not be able to present the full variance from the different stages of cell cycle. Second, the cell
spectra were extracted by two different processes. The spectra from the low mag setup was extracted through an automatic cell finder algorithm. This will introduce other variance from the distortion in the edge of the cell. For high mag setup, the cell was manually selected and the edges of the cells were avoided. Moreover, the buffer spectrum used to fit the cell spectra were different in the two setups. The high mag setup benefits from a smaller field of view which avoid the variance introduced by fitting the cell spectra with a buffer spectrum far from the cell region. These two types of variance may result in the overfitting which leads to the poor performance of the classifier for cross-validation.

6.6 Summary

In this chapter, the determining factors of the water correction, including fitting parameter and reference dependency were discussed. The algorithm was then validated with the PC-3 cells fixed with ethanol and formalin.

The new water removal algorithm successfully reveals all the key biochemical information in the corrected cell spectra. The protein bands, which were used to fit the water contribution in the corrected cell spectra, were distorted by the algorithm which removes different amount of water. The automatic cell extraction method may also induce distortions from the edge of the cell. However, the water corrected spectra still contain enough information that can be used to build a high-performance classifier to identify PC3, LNCaP and PNT2 cell lines.

The cells were measured under two different pixel resolutions. Both results show great performance in the classification analysis. The results with high resolution benefit from the clear cell boundary which leads to higher consistency in the spectra. The cross-validation between the spectra taken from the high mag and low mag setup were implemented, but the performance was restricted by the different processing steps. The algorithm requires further improvement to support cros-platform measurements.
Reference
7. SR-FTIR single point spectra of cell in aqueous medium

The infrared synchrotron radiation, as a high brightness IR light source, offers SNR which is a hundred to a thousand times better than a traditional Globar source when measuring a small sample area. In this chapter, the spectra of three prostate cancer cell lines in an aqueous environment will be measured using synchrotron-based FTIR spectroscopy. The spectra will be analysed by multivariate analysis and then used to build up a classifier for testing unknown samples.

Altered fitting process

There are two key differences between the spectra obtained with SR-FTIR spectrometer and conventional benchtop FTIR spectrometer, namely the light source and the spatial resolution. To achieve the best SNR from SR-FTIR spectrometer, the cells were measured with a single point detector with a 15x15 μm² aperture. This leads to the pixel resolution of the size of the aperture which is nearly three times less than the resolution of the FTIR spectrometer with low mag setup.

The difference in the resolution results in the difference in the water correction algorithm. Since the average size of the cells varies from 20 to 30μm [1], only one spectrum was taken for each cell. With the sampling area of 15x15 μm², and the aperture placed directly above the centre of the cell most of the cell will be measured, and the surrounding water area largely avoided. The recorded spectrum can be treated similarly to the averaged spectrum of 186 spectra taken with a FPA detector with 1.1x1.1 μm² pixel resolution. The original water correction algorithm first fits the spectra from each pixel then average the fitted spectra to obtain the cell spectrum. However, for the spectra obtained from SR-FTIR spectrometer, only one spectrum can be fitted. Therefore, the fitting process was altered to average the spectra first, and then
fit the mean spectrum with the reference spectra.

The effects of altering the fitting sequence were discussed in the previous chapter with spectra taken from FTIR spectrometer with high mag setup. Here, the consequence of altering the water correction steps was assumed to be negligible.

7.1 Water correction result

The raw spectra of three cell lines taken from SR-FTIR spectrometer were shown in Figure 7.1. For each cell line, 200 spectra were measured in three (PC3 and LNCaP) or four (PNT2) sample loadings. All spectra were vector normalised based on the whole spectral range and mean centred.

The most obvious features of spectra in Figure 7.1 are the saturated –OH stretching band and the spectral region below 1000 cm⁻¹. The saturation in the low wavenumber region is due to the cut-off of CaF₂ windows. Thanks to the better SNR, the shoulder of the OH stretching band in the spectra was much clear than the spectra obtained from FTIR spectrometer. Both the evident water combination band and the obscure amide II band suggests that the path length of the liquid cell used with SR-FTIR spectrometer are larger than the path length of the compression cell used with FTIR spectrometer. The window thickness in these two sample holders is also different. However, although the liquid cell consists of CaF₂ windows with 1 mm thickness which is twice as thick than the CaF₂ windows in the compression cell, the amount of light passing through the water layer with the same thickness in SR-FTIR spectrometer should still larger than in FTIR spectrometer due to the high brightness light source [2]. The OH bending band in the spectra are nearly saturated and even saturated in one of LNCaP cell spectra which suggests the pathlength is larger than 10μm [3]. Some variances in the baseline can be found in water combination band and fingerprint region in spectra of all three cell lines.

The corrected cell spectra after water correction were shown in Figure 7.2. Each cell spectrum was fitted with a matching buffer spectrum taken next to the target cell.
Figure 7. Raw spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells in PBS solution taken from SR-FTIR spectrometer. Spectra were vector normalised and mean centred.
Figure 7.2  Corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells taken from SR-FTIR spectrometer. Spectra were vector normalised and mean centred. Spectral region below 1000 cm⁻¹ was removed.
Figure 7. 3 corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells with limited spectral region recorded from SR-FTIR spectrometer. Spectra were vector normalised and mean centred.
The water correction algorithm successfully removed the water contribution and dragged the saturation plateau down to the baseline. However, this also confirmed the overcompensation of water. The saturation in OH bending band remained in the spectra which can be problematic.

To explore the corrected cell spectra further, the spectral region above 3000 cm\(^{-1}\) was removed, and the normalised spectra were shown in Figure 7.3. The corrected spectra clearly illustrate most of the key features of the cells and a straight baseline in the water combination band region. However, there are few issues that need be discussed. First, the amide I band is clearly affected by the spikes from the saturated OH bending band. Therefore this band cannot be used for further analysis. Second, the spectral region below 1200 cm\(^{-1}\) exhibits a high noise level. This is mainly due to the cut-off from the CaF\(_2\) windows. For CaF\(_2\) windows with 2 mm thickness in total, the cut-off was expected below 1250 cm\(^{-1}\)\[4\]. Using windows with a small thickness, such as 0.5 mm, will improve this situation. However, as the CaF\(_2\) windows are too brittle, 0.5 mm windows are not practical for the liquid cell which requires a window diameter of 25 mm. The third problem in the spectra was the distorted baseline in fingerprint region. The large variance in baseline level can be found around 1300 cm\(^{-1}\), and different slopes were observed at 1200 cm\(^{-1}\). This variance increases the difference between the spectra of the cells from same cell line and may lead to the result cannot be separated through the multivariate analysis. Therefore, the origin and the method to remove this distortion need be studied.

**7.2 Repeatability in SR-FTIR spectrometer**

From the Figure 7.3, the distorted baseline in the fingerprint region is consistent with a certain amount of spectra. Therefore, it is possible that such variance results from different sample loadings. The spectra of cells obtained from different load for each cell line were shown in Figure 7.4. 50 spectra were selected and shown for each load.
The Figure 7.4 evidently confirms that the distorted baseline in fingerprint region varies from different sample load. The spectra from same sample load show the same baseline shape which suggests that the variance resulted from the experimental errors. There are a few possible explanations for these differences. First, the liquid cell used in the experiments is sealed by three screws. There is no standard on how the screw should be fixed onto the holder. Therefore, the degree of parallelism of two CaF$_2$ windows is controlled by eye. This leads to an irreproducible pathlength and angel between two windows. Second, the focus and condenser were manually adjusted every time for each sample load. It is hard to keep the focus and condenser position consistent through the experiments. Third, as the beam size of synchrotron radiation is small, after adjusting the focus and condenser, the incident light needs to be recalibrated by moving a set of mirrors which are controlled by a computer system. The signal count from the detector can vary up to $\pm 15\%$ after the calibration process. All these unreproducible parameters will result in variances in the results which can be the origin of the baseline shift in the spectra.

To explore the effects of the distorted baseline in the further analysis, the corrected spectra, 2nd derivate spectra and the PCA analysis results on the 2nd derivate spectra of cells from different sample load for each cell line is shown in Figure 7.5 to 7.7. As the buffer spectrum used to fit the water contribution was taken next to the target cell, it should contain the same shape of baseline in the fingerprint region. Therefore, the water correction process should be able to remove a particular proportion of the distortion in the water contribution. This has been proved by the distorted baseline in the corrected spectra from Figure 7.5 to 7.7 (top) and a similar shape of the baseline between the corrected and raw spectra.

The remaining distortion in the baseline is still too large which will lead to the unreal separation in the multivariate analysis. To eliminate the effects of the distorted baseline, the 2nd derivate spectra were calculated out (Figure 7.5 to 7.7 (middle)). It is evident that the variance due to the distortion was successfully removed by the 2nd derivation.
Figure 7.4  spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells from different sample load taken from SR-FTIR spectrometer. Vector normalised and mean centred.
Figure 7. 5 corrected spectra (top), 2nd derivate spectra (middle) and PCA results on spectra region of 1300-1500cm$^{-1}$ of PC3 cell line.
Figure 7. 6 corrected spectra (top), 2nd deravite spectra (middle) and PCA results on spectra region of 1300-1500cm$^{-1}$ of LNCaP cell line.
Figure 7. Corrected spectra (top), 2nd derivative spectra (middle) and PCA results on spectra region of 1300-1500 cm\(^{-1}\) of PNT2 cell line.
However, a strong noise level can be observed in the amide I band and in the corrected spectra. This makes the amide I band unsuitable for further analysis. The noise in the low wavenumber region is due to the cut-off from the CaF$_2$ windows and the size of the aperture [2]. The raw spectra already contain high-level noise in this region which also affects the 2$^{nd}$ derivate spectra. Thus, the spectral region below 1300 cm$^{-1}$ cannot be used for further analysis.

The 2$^{nd}$ derivate spectra for cells from different sample load were undergoing the PCA analysis to evaluate the remaining effects of the baseline distortion. Due to the noise in the spectra, only the spectral region of 1300-1500 cm$^{-1}$ was employed. It is evident that the 2$^{nd}$ derivate spectra still contain particular variances which lead to the separation between different sample loads. From the loading of 1$^{st}$ PC which separates the data in all three figures, the largest variance can be located in the spectral region of 1380-1420 cm$^{-1}$. The 1$^{st}$ loadings for each cell line were plotted together and shown in Figure 7.8. The loading for PC3 cells was reversed upside down. It is notable that the variance pattern in the 1$^{st}$ loadings is similar between different cell lines. This suggests that the most of the variances are due to the cells. Therefore, although the distorted baseline still affects the 2$^{nd}$ derivate spectra, the introduced variance should not lead to separation between different cell lines.

![Figure 7.8](image.png)

Figure 7.8  1$^{st}$ loading of PCA results in Figure7.5-7.7 for PC3 (red), LNCAP (blue) and PNT2 (green) cells.
7.3 Multivariate analysis

7.3.1 PCA

As the original corrected cell spectra contain obvious distortion in the baseline, only the 2
nd derivative spectra were used to obtain the PCA results. To avoid the variance from the saturation band and noise, the derivative spectra were cut into few spectral regions. Only the spectral regions of 1300-1400 cm\(^{-1}\), 1700-1800 cm\(^{-1}\) and 2800-3000 cm\(^{-1}\) were kept for the analysis. The selected regions of spectra were then vector normalised and mean centred.

Figure 7.9 shows the averaged spectrum of cells from three cell lines. The spectra are relatively consistent. The largest variance can be observed at 1400 cm\(^{-1}\) and 1700 cm\(^{-1}\). The corresponding PCA results were shown in Figure 7.10. The PCA algorithm evidently picks up the large variance at 1400 cm\(^{-1}\) and 1700 cm\(^{-1}\) which is shown in the loading for 1st PC. However, the separation between LNCaP and the other two cell lines exists but is not that clear. The variance at 1700 cm\(^{-1}\) may result from the Savitzky-Golay smoothing when performing the 2nd derivation. The smoothing method takes the following 9 data points into account which may be affected by the
Figure 7.10  PCA plot for 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1400cm$^{-1}$, 1700-1800cm$^{-1}$ and 2800-3000cm$^{-1}$

Figure 7.11  PCA plot for 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1380cm$^{-1}$, 1720-1800cm$^{-1}$ and 2800-3000cm$^{-1}$
saturation peaks in the raw spectra. The difference at 1400 cm⁻¹ is suspicious to be from the distorted baseline as the largest variance which separates the spectra from same cell line was located in the region of 1380-1420 cm⁻¹.

To remove the doubtful variances, the spectra were further limited to the spectral region of 1300-1380 cm⁻¹, 1720-1800 cm⁻¹ and 2800-3000 cm⁻¹. The PCA results were shown in Figure 7.11. The separation between LNCaP and other two cell lines can still be observed with a small part of overlap. The largest variances between those cell lines were then located in the lipid bands which agree with the results obtained from FTIR spectrometer.

7.3.2 PC-LDA

As the corrected cell spectra contain significant variance from different sample loads, the unsupervised analysis method cannot tell the difference between spectra from different cell lines. The PCA scores based on the further limited 2nd derivate spectra obtained in the above section were employed to perform PC-LDA. Only the least
number of PCs which explains 100% of the variance were taken as the target data. The PC-LDA results were shown in Figure 7.12.

From the Figure 7.12, the separation between the three cell lines can be observed. However, there is still a small amount of data overlapping together. This suggests the limited spectra may not contain enough features to distinguish between different cell lines or the variance in the spectra of same cell line were still significant than the difference between the various cell lines [5, 6].

7.3.3 Random forest classifier

The further limited 2nd derivate spectra were also tested by the random forest method. 200 spectra were randomly selected for each cell line and combined into one dataset to build the classifier. The dataset was then divided into the training set and validation set with a ratio of 6:4. The classifier process has been executed 100 times to obtain a range of accuracy, which is shown in Table 8.

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>73.2</td>
<td>78.5</td>
<td>83.1</td>
</tr>
<tr>
<td>AVG</td>
<td>86.3</td>
<td>88.2</td>
<td>92.0</td>
</tr>
<tr>
<td>MAX</td>
<td>96.9</td>
<td>98.4</td>
<td>97.1</td>
</tr>
</tbody>
</table>

The performance of the classifier varies in a broad range. The average performance is reasonable, but the minimum accuracies for PC3 and LNCaP are problematic. This poor performance may result from the smaller spectral regions which do not contain enough features. The variance in the accuracy range results from the randomly selected training set. For example, if the training data used to build the classifier is located in
the overlapping area in Figure 7.12, the algorithm will not be able to select enough important features to separate the data which leads to poor performance in the classifier.

The ROC curve for the classifier was shown in Figure 7.13. The AUC for PC3, LNCaP and PNT2 were calculated at 0.957, 0.968 and 0.987. This suggests the performance of the classifier is reasonable. By sacrificing 10% of the specificity, the achievable accuracy of classifying each cell line was easily above 90%.

7.4 Improved results based on Savitzky-Golay smoothing

The spectra taken from the SR-FTIR spectrometer were affected by the high noise level in the fingerprint region. Therefore, a certain percentage of the variance resulted from the noise which can be removed by smoothing.

The water corrected spectra were smoothed by 9 points Savitzky-Golay algorithm. The smoothed spectra were then transferred to the 2nd derivate spectra and limited to the spectral region of 1300-1380 cm\(^{-1}\), 1720-1800 cm\(^{-1}\) and 2800-3000 cm\(^{-1}\). The spectra
were then vector normalised and mean centred for PCA analysis. The PCA results were shown in Figure 7.14.

It is obvious that the smoothed data gives a better separation between the LNCaP and the other two cell lines. The variance in the spectra of LNCaP and PC3 cells are reduced by the smoothing, and therefore the data points are aggregated. However, the PNT2 cells are still widely spread across the 2nd PC. The PC3 and PNT2 cells are still overlapping together as expected.

As the smoothing algorithm does change the spectra, the variance that leads to the separation may change as well. To confirm whether the smoothing will alter the features in the cell spectra, the 1st loading from the PCA results of both raw spectra and smoothed spectra were compared, as shown in Figure 7.15.

From Figure 7.15, small amplitude changes can be found at 2848 cm\(^{-1}\) and 2916 cm\(^{-1}\). This suggests the separation is less dominated by the \(-\text{CH}_2\) group. However, it is evident that the most of the variances in the data are remain unchanged. Therefore, it can be concluded that the smoothing algorithm does not introduce non-chemical

---

**Figure 7.14** PCA plot for 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with spectral region of 1300-1380cm\(^{-1}\), 1720-1800cm\(^{-1}\) and 2800-3000cm\(^{-1}\). Raw spectra smoothed with 9 points smoothing function.
The reduced variance in the spectra of the same cell line should give better results in the further analysis.

**PC-LDA**

Figure 7. 15 1st loading in PCA results from Figure 7.11 and Figure 7.14.

Figure 7. 16 PC-LDA results based on 2nd derivate spectra of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines. Raw spectra were smoothed with 9 points smoothing function.
To enlarge the difference between the spectra of three cell lines, the data was undergoing the PC-LDA method. The PCs contains 100% variance were employed, and the results are shown in Figure 7.16.

From Figure 7.16, the data points represent the spectra of three cell lines were clearly separated from each other. No overlapping can be observed in the plot. Comparing to the Figure 7.12, the results suggest that the smoothing method successfully reduced the variance between the spectra from the same cell line and kept the key features which are enough to distinguish between different cell lines. The more distinct separation should also result in a better performance in the classifier.

**Random forest classifier**

The 2\textsuperscript{nd} derivate spectra calculated from the smoothed cell spectra were used to build a random forest classifier. For the consistency, the spectral region was limited, and the same processing parameters as unsmoothed spectra were implemented. The results were shown in Table.9.

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MIN</strong></td>
<td>91.4</td>
<td>93.0</td>
<td>95.0</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td>98.2</td>
<td>98.6</td>
<td>99.0</td>
</tr>
<tr>
<td><strong>MAX</strong></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The performance of classifier significantly improved with the smoothed spectra. The minimum accuracy for three cell lines is all above 90%. The average accuracy for 100 runs is all above 98%. The results are close to the performance of the classifier built with spectra obtained in FTIR spectrometer with low mag setup. As the classifier has great accuracy for all three cell lines, the ROC curve is not necessary here.
Summary

In this chapter, the water correction algorithm was manipulated to remove the water contribution in the cell spectra taken from SR-FTIR spectrometer. The correction process successfully and consistently resolved most of the biochemical peaks. The distorted baseline in the spectra was observed and discussed. The water correction method was not able to remove all the distortion, but the resulting spectra consist of stable baseline shape which confirms the consistency of the algorithm. The distortion effects were minimised by the 2nd derivate spectra, and the questionable spectral region was identified.

The spectra obtained from the SR-FTIR spectrometer, in theory, should benefit from much better SNR. However, due to the limitation of the experimental setup, the results are a little below the expectation. The separation of three cell lines can be observed in the multivariate analysis. However, due to the unrepeatability of the experiments, noise and shorter spectral region, a large proportion of data from different cell lines was overlapping together. The random forest classifier built by the spectra shows a reasonable average performance, but the minimum accuracy was poor. The results were then improved by the Savitzky-Golay smoothing algorithm. The smoothed spectra exhibit clear separation with no overlap in the PC-LDA results. The performance of the random forest classifier using the smoothed spectra is also improved and close to the performance of classifier built with spectra obtained from FTIR spectrometer.
Reference


8. QCL imaging of cell in aqueous environment

8.1 Cell imaging and noise analysis

The images of cells in PBS solution taken from the QCL microscope are shown in Figure 8.1. The left image represents the total intensity, and the right one stands for the intensity at the wavenumber of 1550 cm\(^{-1}\). The images consist of 480x480 pixel with 1.33 \(\mu\)m pixel resolution. With large FOV, more than a hundred cells can be captured in one image. Moreover, thanks to the high pixel resolution, although the cells were clumped tighter, the boundary of the cell was still clear, especially in the image of the amide II peak.

Comparing the two pictures in Figure 8.1, a clear contrast difference can be observed. The number of high intensity (red) spots in the left picture is significantly reduced in the right image. Those dots are the PMMA spheres used as a spacer. The PMMA sphere
has a strong average absorbance across the wavenumber region. However, it does not have peaks at 1550 cm\(^{-1}\) (see Figure 8.4). The huge contrast between cell and water also suggests that the amide II band has been well resolved.

A diffraction-like ‘ring effect’ can be found in both images in Figure 8.1. Those rings exist around both cells and PMMA spheres and therefore should not result from the chemical property of samples. To explore this effect, a small region containing a cell and a ‘ring’ around it was extracted from the image and the total intensity of the image is shown in Figure 8.2. For this small area, the variance in the path length is negligible. Therefore, the intensity of water should be consistent. However, it is clear that the inner ring region close to the cell has less average intensity than the region away from the cell.

To explore the origin of the ring, spectra from 7 different positions labelled in Figure 8.2 have been taken and are shown in Figure 8.3. First, the spectra obtained in the cell region (C, D, E) have higher average absorbance and baseline in the fingerprint region. This is due to the high extinction coefficient of cell in the fingerprint region. The spectra taken from the FTIR spectrometer contains similar features. Second, the spectra extracted from the outside ring have a higher average absorbance than the
spectra extracted from the inner ring. The low absorbance means there is more light passing through this particular region. As the ring effects directly correspond to the size of the cell or PMMA sphere, the extra light may come from the scattering in the sample. Although the refractive index of water and the cell are similar, there is still some difference which can cause non-obvious scattering. Moreover, this effect has not been seen in any FTIR spectrometer. Thus this diffraction-like pattern should directly link to the property of QCL. Due to the time limit of the project, the physics behind the effects will not be addressed here.

Another notable feature is the noise. In theory, the QCLs should give much better SNR than the Globar source. However, noticeable noise can be found in the spectra extracted from the image. The noise covers the whole spectral region, and it becomes less obvious when there is a strong peak. Also, the noise in the spectra appears to be correlated as the same noise pattern can be found in all 7 spectra in Figure 8.3. This type of noise has not been observed in the spectra taken from FTIR spectrometer. To understand the origin of the noise, a series of studies were carried out.

![Figure 8.3 spectra taken from seven different position across the cell and the surrounding area in Figure 8.2. The colour of spectra represent the colour in the Figure 8.2.](image)
To explore the relation between the noise and the measured sample, the spectra of a PC3 cell, the water solution and a dried PMMA sphere are shown in Figure 8.4. The cell in water and the dried PMMA sphere are all placed in the compression cell. It is evident that all three spectra contain some observable noise. The dried PMMA also contains the noise which means the noise is not from the water layer. Thus, the noise should not result from the interaction between the laser source and the sample.

The above conclusion leads to the question that if the sample holder, the material and setup of windows to be specific, will interact with the laser and cause the noise. To evaluate the contribution from the windows to the noise, four different types of IR windows, namely a CaF$_2$ slide, a BaF$_2$ slide, small CaF$_2$ disc (φ13mm) and large CaF$_2$ disc (φ25mm) were measured with the QCL microscope in five different setups, as shown in Figure 8.5. All the windows have the thickness of 1 mm. The bottom window was placed on the microscope stage, and the upper window was fixed onto it with blue tack. The gap between windows was not controlled by any spacer. However, visual inspection of the slides revealed that a visible light (wavelength of 400~700nm)
interference pattern was observed for each setup which suggests that the air gap is less than 1 μm. At this path length, the fringing effects from the air gap should be minimised. The background scan was taken when nothing has been put in the light path.

The measured spectra from the five setups above are shown in Figure 8.6. The noise exists in all the spectra. The amplitude of the noise is similar between different configurations. A relatively large fluctuation at 1200 cm⁻¹ can be observed in the spectra of setup using two CaF₂ slides. The remaining spectral region in that spectrum does not share this significant oscillation. Some of the noise exhibits periodicity. However, it should not be related to the fringing effects from the windows for two reasons. First, the calculated path length responsible for this particular frequency would be around 140 μm. Second, the fringes are sinusoidal waves, but the noise is irregular. It is noteworthy that the BaF₂ window has a lower cutoff and therefore the spectrum (dark green) exhibits a flat tail at low wavenumber region.

To quantitatively discuss the noise level, the average standard deviation and range in the spectral region from 1100 cm⁻¹ to 1800 cm⁻¹ of 500 spectra from each setup were
shown in Table.10. The spectral region below 1100 cm$^{-1}$ was not analysed due to the cutoff of CaF$_2$ windows. From Table.10, the standard deviation between different window materials is similar. A decrease in the standard deviation and range was noticed with the round window. All the windows were made with IR grade material. Therefore, in theory, the shape of the window should not cause any difference in the

<table>
<thead>
<tr>
<th>Setups</th>
<th>Standard Deviation</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two CaF$_2$ slides</td>
<td>0.023</td>
<td>0.12</td>
</tr>
<tr>
<td>Two BaF$_2$ slides</td>
<td>0.022</td>
<td>0.10</td>
</tr>
<tr>
<td>CaF$_2$ slides + CaF$_2$ discs (large)</td>
<td>0.018</td>
<td>0.08</td>
</tr>
<tr>
<td>Two CaF$_2$ discs (large)</td>
<td>0.014</td>
<td>0.08</td>
</tr>
<tr>
<td>Two CaF$_2$ discs (large/small)</td>
<td>0.014</td>
<td>0.06</td>
</tr>
<tr>
<td>Two BaF$_2$ slides</td>
<td>0.022</td>
<td>0.10</td>
</tr>
</tbody>
</table>
signal. However, those windows were manufactured by different companies which may lead to differences in quality. Those differences were not obvious when performing experiments with FTIR spectrometer. This may suggest that the QCL microscope is sensitive to subtle changes in the environment.

To further confirm whether the noise is generated by the windows, two images were taken from the same position in the same round CaF$_2$ window, as shown in Figure 8.7. The results clearly demonstrate the instability of the noise which means the noise is not generated from the windows. Therefore, the noise can only result from the QLC microscope itself.

To explore the instability in the noise, 4 background images, with nothing in the light path, were taken with a time interval of 5 minutes. Due to the QCL microscope employed in this project being borrowed from Daylight Solutions and in demonstration mode, the actual background file is not available through the software. The measured average signal can be observed in the plot, but the value is not accessible. By photo editing, the signal from the background images is shown in Figure 8.8.
Figure 8.8 consists of two type of information. The blue curve with light blue dot represents the target (line) and actual (dot) laser power when measuring the background. The red line stands for the target signal level. The dots with four different colours attached on the red curve corresponds to average signal level measured four times. For the ideal situation, the actual signal should sit on the target curve and be consistent through multiple experiments. However, significant fluctuations can be observed in the signal level between four measurements. The laser power was driven exactly the same for the four measurements. Therefore, the variance in the measured signal level confirmed the instability in the output power of laser source. Later on, the correlated noise results from the laser source were confirmed by the Daylight solution. Ideally, the noise of 0.01~0.02 absorbance can be expected in the measured spectra.

From the water corrected spectra of cells from three cell lines, the absorbance of amide I peak are around 0.4~0.8. For the non-protein peaks, such as lipids bands and nuclear
bands, the absorbance can be less than 0.1. Therefore, even the ideal noise of 0.02 absorbance will lead to significant variance in the spectra of a single cell.

To evaluate the noise effects in the cell spectra, 200 spectra of PC3 cells were recorded and corrected with the water correction algorithm. The results are shown in Figure 8.9. For the spectra of cells in water, the noise is not obvious and does not affect the shape of the large peaks. However, after water correction, the noise in the spectra of cells was enlarged and significantly distort the shape of the spectra. Abnormal amide peak shoulders and strong noise peaks can be observed in the spectra. The spectra with this
type of noise cannot be used in the further analysis. Hence, a method that removes the noise is required.

**Noise removal**

The common noise removal methods employed for spectra of biochemical samples are PCA noise reduction and Savitzky-Golay smoothing. However, both of the methods have an excellent performance with random noise but cannot properly deal with correlated noise. For PCA noise reduction, the spectra are subjected to PCA. The spectra can then be reconstructed from the addition of the first x number of PCs that explain all the biological features. The remaining PC should explain the variance associated with the noise and so can be omitted. However, when the noise is correlated, and the variance of the noise will be recorded in the first few PCs which makes remove it impossible.

The Savitzky-Golay smoothing method uses the nearby points to estimate the value of the target data point. The smoothing performance depends on the order of fitting curve and the size of the fitting frame. When using a small frame, the noise will not be fully removed. However, if a large frame is employed, the shape of the spectra will be distorted. For the corrected cell spectra shown in Figure 8.9 (bottom), obvious

![Figure 8. 10 200 spectra of PC3 cells smoothed with 13 points Savitzky-Golay filter. The original spectra were the spectra in Figure 8.9 (bottom).](image-url)
distortion was found when the fitting frame is larger than 13 points. The 13 points smoothing results are illustrated in figure 8.10.

From Figure 8.10, the noise in the spectra was significantly smoothed, and the variance between the spectra was minimised. However, visible artefacts can be observed in the resulted spectra. Artefact peaks and bumps can be found at 1304, 1330, 1420, 1585 cm\(^{-1}\). The original peaks at 1167 and 1317 cm\(^{-1}\) disappeared. The shape of amide III peak at 1400 cm\(^{-1}\) was distorted in the left shoulder. All these features are noise related rather than chemical related. Therefore, the spectra fit with the smoothing method cannot be used in the further analysis. A new method that removes the correlated noise in the spectra is required.

Before establishing the new method, the character of the noise was studied. As mentioned above, the expected noise has a fixed absorbance value rather than SNR. Therefore, the change in the noise level for different sample thickness was evaluated by two water layers with different thickness. The PMMA spheres of 5.5 μm and 10.8 μm were used as a spacer to create two different path lengths between the CaF\(_2\)
windows. The CaF$_2$ windows with 0.5 mm thickness were employed to minimise the effects on the noise signal. The resultant water spectra were shown in Figure 8.11.

From Figure 8.11, the drastic absorbance difference can be observed between two water spectra. The absorbance of OH bending band in the spectrum of water with the 10 μm gap is more than two times larger than the water with 5 μm gap. This may result from the baseline shift. To evaluate the noise, the standard deviation and range of spectral region from 1100 to 1500 cm$^{-1}$ was calculated. Both spectra consist of similar standard deviation and range which confirm that the noise is independent of the thickness of the sample. However, the absorbance range of the noise in both spectra was calculated as 0.06–0.07 which is three times more than the expected level. This gives a reference value of the possible noise level in the cell spectra which raise the concern even further.

As the spectra exist in the format of the image (hypercube), it is important to understand how the noise varies across the pixels. A 100x100 pixel area in the image of water layer with 10 μm gap was extracted out to evaluate the difference of noise across a small region. The spectra of each row and column in the 100x100 matrix were averaged, which gives two 1x100 spectra matrices. By comparing the spectra in each matrix, the change of noise in the horizontal and vertical direction can be observed. The resultant spectra are shown in Figure 8.12.

From the Figure 8.12, it is evident that the noise in most of the spectral region is consistent across the 100 pixels. Variances can be observed at the wavenumber of 1335 and 1400 cm$^{-1}$ in the spectra of horizontal spectra comparison. In the vertical spectra comparison, the variances appear in the spectral region of 1100-1200 cm$^{-1}$ and 1400-1500 cm$^{-1}$. This suggests that the noise changes differently across different direction. Due to the instability of the laser, the noise is changing between two measurements. Therefore, the relation between the orientation of the window and the noise cannot be evaluated.

It is noteworthy that the changes in the noise only affect the absorbance at particular
wavenumber but the noise pattern does not change across the whole wavenumber range. The variance is also linearly dependent on the position in the area. Thus, it can be concluded that the difference in the noise is negligible across a small area in the image.

To sum up, the noise has three key properties. First, as shown in Figure 8.3, the noise is independent of the sample and therefore consistent in both spectra of cells and water. Second, the noise is independent of the thickness of the sample. Third, the noise is consistent across a small imaging area. Based on these properties, a new noise reduction method based on adjacent pixels was proposed.
8.2 Noise reduction based on adjacent pixels

The spectral pattern of noise was shown to be localised, and the change of noise from pixel to pixel is tiny. Therefore, the assumption was made that the correlated noise in QCL system is consistent in a small region with 5% of the total imaging area.

Based on the assumption, the average of noise within a cell should equal to the mean of the noise in a confined surrounding area of that cell. Therefore, if a spectrum of noise can be obtained, the spectra of the cell can be corrected. The spectra of surrounding region of the target cell are selected to get the noise spectrum. The process of noise reduction is shown in Figure 8.13.

The water region around the target cell was chosen, and the spectra in that region were averaged. The mean spectrum was then fitted with a noise free water spectrum premeasured from the FTIR spectrometer. The difference between the recorded water spectrum and the fitting result will be the noise spectrum. This noise spectrum was then subtracted from the spectra of the target cell.

Figure 8.13 noise correction process. (top left) The surrounding area of the target cell was chosen. (top right) the spectra in this region was fitted with a noise free reference to obtain the noise spectrum (bottom right). (bottom left) the noise spectrum was subtracted from the spectra of cell.
The raw spectra of cells in PBS solution are shown in Figure 8.14. Three sub figures illustrate the spectra of PC-3 (top), LNCAP (middle) and PNT2 (bottom). 200 spectra were extracted for each cell line. The spectra of each cell line were taken from at least 5 different images and three batches of the sample. The shown spectra were vector normalised and mean centred. As shown in Figure 88, obvious noise can be observed in the spectra of all three cell lines. The noise in the spectra of the PC3 cell line is less obvious than the other two cell lines. This is due to the normalisation process. The PC3 cells are much larger than the other two cells. Therefore, when measuring the spectra, the PC3 always gives a higher absorbance in the amide I and II bands which lead to a higher SNR. The absolute noise in the spectra was still expected to be at the same level of the noise in the spectra of other two cell lines. The noise level in the spectra of the same cell line was not consistent as well. This is due to the irreproducible path length in the compression cell which leads to a difference in the total intensity of the spectra. For the chemical features, an evident amide II peak can be found in the spectra. The remaining fingerprint region exhibits subtle peaks, but those peaks were blurred by the noise.

The spectra after noise reduction are shown in Figure 8.15. The fluctuating noise has been successfully removed from the spectra of all three cell lines, which leads to a smooth spectral curve in the figure. Although the noise level in the raw spectra is different, the correction method was capable of removing the appropriate noise from the spectra. The lipid, amide III and nucleus bands are clearly resolved in the spectra. A small amount of noise can still be found in the spectral region from 1200 to 1500 cm$^{-1}$. However, compared to the noise in the raw spectra, this level of noise is negligible.

The noise reduction method based on adjacent pixels correctly removes the noise from the spectra of cells from different cell lines. This correction method will not alter the shape of the spectra as the Savitzky-Golay smoothing method does. Therefore, the resultant spectra should be suitable for the further analysis.
Figure 8. 14 raw spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells in PBS solution recorded by QCL microscope. Spectra were vector normalised and mean centred.
Figure 8. 15 noise corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells in PBS solution recorded by QCL microscope. Spectra were vector normalised and mean centred.
Figure 8. 16 water corrected spectra of PC-3 (top), LNCaP (middle) and PNT2 (bottom) cells in PBS solution recorded by QCL microscope. Spectra were vector normalised and mean centred.
8.3 Water correction result

After removing the correlated noise, the spectra were corrected with the water correction method. The reference buffer spectrum was the noise corrected water spectrum from the noise reduction process. The cell reference was the Matrigel spectrum. All the spectra in the cell region were fitted first, and the resultant spectra were averaged to obtain the final cell spectrum.

The water corrected cell spectra are shown in Figure 8.16. The spectra were vector normalised and mean centred. The water correction method successfully reduced the water contribution from the cell spectra and therefore the cell features in the fingerprint region have been elucidated. A clear lipid peak at 1740 cm\(^{-1}\) can be observed in the spectra of LNCAP cells. The low-level noise can be observed in the spectra. However, the noise level in the spectra was better than the noise level in the spectra smoothed with Savitzky-Golay filter. The cell peaks were not distorted, and no artefact peaks can be found in the spectra. The peak ratio of amide I and II band in the spectra are more consistent and significantly larger than the ratio in the spectra taken from the FTIR spectrometer, which suggests the overcompensation of water was improved. As the OH stretching band is not available with QCL microscope, the degree of overcompensation cannot be evaluated. It is possible that there is still water contribution in the corrected spectra. The quality and consistency of the spectra were also better than the FTIR spectra taken with the low mag setup.

8.4 Multivariate analysis

8.4.1 PCA

The corrected cell spectra were analysed by the PCA algorithm. The original spectra, 1\(^{st}\) and 2\(^{nd}\) derivative spectra were studied. The derivative spectra were smoothed by 11 or 19 points Savitzky-Golay filtering. As the spectra of cells are consistent, the full spectral range was examined.
Figure 8.17 indicates the mean spectra of three cell type and the PCA results based on the corrected cell spectra. The average spectra of cells were nearly identical in most of the spectral region. Only two obvious variances can be observed, in the lipid band at 1740 cm\(^{-1}\) and the nucleus band at 1085 cm\(^{-1}\). The three cell lines were not separated in the PCA plot. The largest variance was found in the amide bands according to the 1\(^{st}\) loading and variance in the slope of baseline can be observed in the 2\(^{nd}\) loadings. The slope issue can be overcome by taking the derivative spectra.
The averaged 1st derivative spectra and the corresponding PCA results are shown in Figure 8.18. The derivative spectra enhanced the difference in the lipid band and minimised the variance in the nuclear band, as seen in Figure 8.17. However, the three cell lines were still not separated by PCA. Correlated variance in the LNCAP cells was noted in the 2nd PC. From the 2nd loading, the significant variance was found in the
spectral region from 1200 to 1400 cm\(^{-1}\). This variance is regular and therefore may result from the remaining noise in the spectra. The noise after been translated into the 1\(^{st}\) derivative spectra was smoothed with the 4\(^{th}\) polynomial order and 11 points smoothing method. As the original noise is correlated, it is possible that the smoothing turns the noise into fluctuation curves instead of removing the noise. Therefore, this region was cut off to eliminate the variance from the noise. The PCA results for the cut 1\(^{st}\) derivative spectra are shown in Figure 8.19. The LNCAP cells were successfully separated from the other two cell lines according to the 2\(^{nd}\) PC, with a small region of overlapping in the plot. Based on the 2\(^{nd}\) loading, the separation occurs due to the lipid bands which agrees with the results obtained from the FTIR spectrometer.

To further explore the variance in the spectra, the mean 2\(^{nd}\) derivative spectra and the PCA results were shown in Figure 8.20. The only obvious difference in the 2\(^{nd}\) derivative spectra between different cell lines is still the lipid band. The spectral region of 1200 to 1400 cm\(^{-1}\) was cut from the 2\(^{nd}\) derivative spectra as well. In the PCA plot, the separation between LNCAP and other two cell lines can still be observed. However, the percentage of data overlapping with other cells was increased. This may result from
the fact that the 2nd derivative spectra significantly increases the noise in the spectra. The 1st derivative spectra have already shown that the noise can be problematic after smoothing. This situation can be improved by smoothing the water corrected cell spectra. The results will be shown in section 8.5.
8.4.2 PC-LDA

The PC-LDA analysis based on the 1st and 2nd derivative spectra was performed, and the results are shown in Figure 8.21. The spectral region from 1200 to 1400 cm\(^{-1}\) was removed to avoid the over-fitting of noise. The separation of the three cell lines can be observed in both plots. However, the results are different to the results from FTIR spectrometer. In the previous study, the LNCAP cells were far away from the other two cell lines. The difference, with the overlapping of LNCAP and PNT2 cells and the separation of PC3 cells may be caused by the lack of features in the spectra. The spectra do not contain any information from the CH stretching band, and the broad spectral region was removed due to the noise. The results based on the 2nd derivative spectra show a similar distribution pattern as the 1st derivative spectra. This suggests that the features used to separate the cell lines are the same in both spectra. Therefore, the

Figure 8.21 PC-LDA results based on the 1st derivative (top) and 2nd derivative (bottom) spectra of PC3 (red), LNCAP (blue) and PNT2 (green) cell lines.
separation of the three cell lines should not result from the different noise level in the raw spectra.

8.4.3 Random forest classifier

Separation can be obtained from both the 1st and 2nd derivative spectra. Therefore, these two types of spectra were both evaluated by the random forest algorithm. The full spectral region, except from 1200 to 1400 cm⁻¹, was used to build the classifier. 200 spectra of each cell line were combined as the dataset and randomly divided into the training set and validation set with the ratio of 6:4. The classifier process was performed for 100 times to identify the range of the performance.

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MIN</strong></td>
<td>72.1%</td>
<td>82.1%</td>
<td>70.0%</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td>85.6%</td>
<td>92.2%</td>
<td>84.4%</td>
</tr>
<tr>
<td><strong>MAX</strong></td>
<td>95.0%</td>
<td>100%</td>
<td>96.6%</td>
</tr>
</tbody>
</table>

The accuracy of the classifier built with the 1st derivative spectra is shown in Table.11. The performance of the classifier exhibits strong instability as the accuracy varies over a broad range. The significant variance in the accuracy may result from the over-fitting of noise in the spectra. Although the PC3 cells were well separated in the PC-LDA plot, the performance is much lower for the LNCAP cells. This suggests the PC-LDA results were over fitted. The average performance of the classifier is not significant but reasonable.

The performance of the classifier built with the 2nd derivative spectra is shown in Table.12. The accuracy of the classifier is worse than the classifier from the 1st
derivative spectra. This may result from over-fitting of the noise in the 2nd derivative spectra.

To identify the features selected to classify the three cell lines, the importance of each wavenumber is shown in Figure 8.22. The importance was calculated from the classifier built with the 2nd derivative spectra. It is evident that the most important features were located at the lipid band around 1740 cm\(^{-1}\) and nuclear bands at 1085 cm\(^{-1}\) and 1108 cm\(^{-1}\).

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>70.8%</td>
<td>71.2%</td>
<td>68.9%</td>
</tr>
<tr>
<td>AVG</td>
<td>83.9%</td>
<td>91.4%</td>
<td>80.9%</td>
</tr>
<tr>
<td>MAX</td>
<td>92.5%</td>
<td>98.5%</td>
<td>93.1%</td>
</tr>
</tbody>
</table>

Table 12: Accuracy (sensitivity) of classifier built with 2nd derivative spectra for each cell line

Figure 8.22 averaged importance of each wavenumber in the classifier of PC-3, LNCaP and PNT2 cell lines based on 100 times test.
8.5 Results from smoothed cell spectra

From the results above, the remaining correlated noise in the cell spectra may still lead to poor performance in the multivariate analysis. Thus, the water corrected cell spectra was smoothed with 9 points Savitzky-Golay filter. The smoothed spectra were then transferred into the 1st and 2nd derivative spectra to perform the multivariate analysis.

PCA

Figure 8.23 illustrates the PCA results obtained from the 1st derivative spectra. A clear separation can be observed across the 2nd PC in the plot. A small number of LNCAP data points were spread out and overlapped with the other two cell lines. The fluctuation feature in the spectral region of 1200-1400 cm⁻¹ can still be observed in the 2nd loading. However, the contribution of this variance is much smaller than the lipid band which suggests that the noise was reduced by the smoothing algorithm.

Figure 8.23 PCA plot for 1st PC against 2nd PC of 1st derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells with the corresponding loadings. The original spectra were smoothed with 9 points smoothing function.
The PCA results of 2nd derivative spectra are shown in Figure 8.24. The separation between LNCAP and the other two spectra is evident. The variance between the spectra of LNCAP cells was reduced in the 2nd derivative spectra as all the data points are clearly separated from the other two cell lines. From the 2nd loading, the variance that separates the LNCAP cell was still dominated by the lipid band at 1745 cm\(^{-1}\).

**PC-LDA**

As the influence of noise in the spectra was minimised, the full spectral region of the derivative spectra was used to obtain the PC-LDA results. The performance from the 1st and 2nd derivative spectra is shown in Figure 8.25. The separation of three cell line can be observed in both plots. The LNCAP points were far away from the other two cell line, and the PC3 and PNT2 points were separated but relatively close. This result agrees with the conclusion from the FTIR results. No overlapping can be observed and therefore a better classification result can be expected.
Random forest classifier

The 2nd derivative spectra give the best separation in the PCA results. Therefore, the random forest classifier will be built on the 2nd derivative spectra only. The full spectral region was employed to evaluate whether the noise still affects the results. The training set and validation set was randomly selected from the dataset which contains the spectra of 200 cells for each cell line. The accuracy range of 100 processes is shown in Table. 13. The performance of the classifier was significantly improved. The minimum accuracy was found in the classifier for PC3 cells at above 90%. The average performance is close to 100% which is achievable under the particular arrangement of datasets.
The smoothing of corrected spectra before derivation successfully increased the performance of the classifier. However, it is possible that the separation was due to the artefacts created during the smoothing process. To verify the feature selected to distinguish the three cell lines, the importance of each wavenumber in the classifier is shown in Figure 8.26.

From Figure 8.26, it is evident that the most important feature is still located at the lipid band at 1740 cm\(^{-1}\) and nucleus band at 1085 cm\(^{-1}\) and 1108 cm\(^{-1}\). The order of importance changes between those features, but there is no extra feature being taken into account. The results prove that the smoothing of the water corrected spectra does not create significant non-chemical features in the spectra nor remove the important feature from the spectra.

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>90.5%</td>
<td>98.0%</td>
<td>92.0%</td>
</tr>
<tr>
<td>AVG</td>
<td>99.6%</td>
<td>99.3%</td>
<td>99.3%</td>
</tr>
<tr>
<td>MAX</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The smoothing of corrected spectra before derivation successfully increased the performance of the classifier. However, it is possible that the separation was due to the artefacts created during the smoothing process. To verify the feature selected to distinguish the three cell lines, the importance of each wavenumber in the classifier is shown in Figure 8.26.

From Figure 8.26, it is evident that the most important feature is still located at the lipid band at 1740 cm\(^{-1}\) and nucleus band at 1085 cm\(^{-1}\) and 1108 cm\(^{-1}\). The order of importance changes between those features, but there is no extra feature being taken into account. The results prove that the smoothing of the water corrected spectra does not create significant non-chemical features in the spectra nor remove the important feature from the spectra.

Table 13 Accuracy (sensitivity) of classifier built with 2nd derivate spectra for each cell line. The original spectra were smoothed with 9 points smoothing function.

![Figure 8.26 averaged importance of each wavenumber in the classifier of PC-3, LNCaP and PNT2 cell lines based on 100 times test. The original spectra were smoothed with 9 points smoothing function.](image-url)
Summary

In this chapter, the image of cells in aqueous solution taken from QCL microscope has been studied and discussed. The QCL laser sources were found to produce correlated noise at a fixed absorbance value in the spectra. This noise has been analysed, and a method of removing such noise based on fitting the spectra from adjacent pixel has been proposed.

The water correction successfully reduces the water contribution from the cell spectra. The corrected spectra exhibit large amide I/II peak ratio which suggests that the overcompensation of water has been improved. Variance between the spectra of cells from same cell line is minimised which results in high consistency in the spectra. However, due to the fact that the noise was not completely removed by the noise reduction algorithm, the performance of the built classifier was poor.

The water corrected spectra were then smoothed to minimise the noise in both the original and derivative spectra. It is found that the smoothing decreases the noise in the spectra to an acceptable level which does not affect the unsupervised PCA results. The performance of the classifier was significantly improved as well. The effects of smoothing were evaluated, and the results suggest that the smoothing will only minimise the noise and not affect the chemical features in the spectra.
9. Performance comparison between three spectrometers

The imaging of cell in aqueous solution has been evaluated using three spectrometers. Excellent performance of all the classifier built with the spectra recorded from each spectrometer was obtained. However, due to the experimental and instrumental limitation, the measurements and data analysis methods are significantly modified for each spectrometer which leads to difficulties in the cross-platform application. In this chapter, the effects of differences in the instruments on the recorded spectra will be discussed.

9.1 Comparison of instrumentation

The comparison of instrumentation is limited by the availability. The possible results on other platforms or experimental setup will not be discussed. The different spectrometers will be discussed following the order of the chapters. The difference between spectrometers will be summed up in Table.14.

Benchtop FTIR spectrometer with Globar source

Two configurations were employed in the experiments performed with the conventional FTIR spectrometer, namely the low magnification and high magnification setup. In the low magnification setup, the spectrometer was able to record an image with large FOV of 704x704 μm² using the liquid cooled MCT FPA detector. The spatial resolution is limited by both the pixel resolution and diffraction limit depending on the wavelength of light. The SNR is limited by the intensity of the Globar source and detector, which can be improved by measuring and averaging multiple interferograms. The full mid-IR spectral range from 1000 to 4000cm⁻¹ is available in the spectrometer. The measurement time for each image with 128 interferograms is around 10min. The number of cells measured in one image

200
significantly depends on the concentration of cell solution and can be more than a hundred.

For the high magnification setup, the FOV of a single image was significantly reduced, 25 times, which gives $105 \times 105 \, \mu m^2$. The high resolution of 1.1 $\mu m$ increases the spatial resolution, which is only diffraction limited, and the amount of spectra in a fixed sampling area. The SNR of each pixel is lower than the SNR in the low magnification configuration. This problem can be overcome by averaging spectra from more pixels. The time take to obtain one image with 128 interferograms is still around 10 minutes. The amount of cells captured in a single image is dominated by the small size of FOV rather than the concentration of the cell solution.

The sample holder used in the experiments with the conventional FTIR spectrometer is modified from the compression cell which is designed for solid samples. The leakage of the sample holder leads to a short measurement time of 30 minutes for each sample loading. Due to the lack of proper spacer, the PMMA sphere is used to generate the gap between windows. The diameter of the PMMA sphere is relatively constant but can still vary, which will lead to the irreproducible path length and angle between the windows. The path length may also vary due to the different pressure applied to the sphere from the sample holder. CaF$_2$ windows with 0.5mm thickness are employed to increase the throughput of light.

*Synchrotron-based FTIR spectrometer*

The FPA detector in the SR-FTIR spectrometer was not available at the time of experiment. Therefore, cell spectra could only be recorded with a single point liquid cooled MCT detector. The spatial resolution is controlled by a single aperture with a size of $15 \times 15 \, \mu m^2$. Thanks to the high brightness of the synchrotron radiation, the SNR was only limited by the detector. The full mid-IR spectral region is achievable in the SR-FTIR spectrometer. The measurement time for a single spectrum is only 20 seconds. However, the electrons in the storage ring need be topped-up every 10 minutes, and a 1 minute stabilisation time is required. The average amount of spectra
taken in an hour is around a hundred. The utilisation of synchrotron radiation in FTIR spectrometer is still under development. The optimised signal requires a lot of calibration of the system. This may reduce the reproducibility of the results.

The sample holder used with the SR-FTIR spectrometer is a well-designed liquid cell. The use of proper spacer ring results in a sealed environment which prevents the leakage. At least 4 hours of experiment time can be expected from one sample loading. The liquid cell employed two 1mm CaF₂ windows which lead to a higher wavenumber cutoff in the spectra. The cap of the liquid cell was manually fixed to the holder with three screws. This will lead to the irreproducibility of the path length and the angle between the two windows.

**QCL microscope**

The QCL microscope benefits from the high intensity of the laser source and the spectra were recorded with an uncooled microbolometer FPA detector. This detector contains 480x480 sensor elements which is 14 times more than the MCT FPA detector in the FTIR spectrometer. With the high magnification objective, the detector is able to provide a FOV of 640x640 μm², which is close to the FOV in the FTIR spectrometer with low magnification objective. It also offers a high pixel resolution of 1.3 μm which is close to the resolution in FTIR spectrometer with high magnification objective. The pixel resolution is also diffraction limited. The SNR, in theory, should benefit from the high-intensity laser. However, correlated noise with the amplitude of a fixed absorbance was generated due to the coherent property of laser. Therefore, the SNR of QCL microscope is dominated by the laser source rather than the detector. The currently available spectral range in the QCL microscope is from 910 cm⁻¹ to 1800 cm⁻¹. The measurement time of one image at one wavenumber only takes few seconds. However, to measure the full spectral range, the microscope need scan the wavenumber one by one which leads to a total time of 8 minutes. The amount of cell in one image is similar to the FTIR spectrometer with low magnification objective due to the similar FOV. The experiment with QCL microscope employed the same
A compression cell as the sample holder.

Table 14: Specification of the three spectrometers with the particular configuration.

<table>
<thead>
<tr>
<th></th>
<th>FTIR Low</th>
<th>FTIR High</th>
<th>SR-FTIR</th>
<th>QCL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Light source</strong></td>
<td>Globar</td>
<td>Synchrotron radiation</td>
<td>Laser</td>
<td></td>
</tr>
<tr>
<td><strong>Detector</strong></td>
<td>MCT FPA</td>
<td>MCT single point</td>
<td>Microbolometer FPA</td>
<td></td>
</tr>
<tr>
<td><strong>FOV (μm²)</strong></td>
<td>704x704</td>
<td>145x145</td>
<td>15x15</td>
<td>640x640</td>
</tr>
<tr>
<td><strong>Pixel resolution</strong></td>
<td>5.5μm</td>
<td>1.1μm</td>
<td>15μm</td>
<td>1.33μm</td>
</tr>
<tr>
<td><strong>Spatial resolution</strong></td>
<td>5.5μm (λ&gt;5.5μm)</td>
<td>λ</td>
<td>15μm</td>
<td>λ</td>
</tr>
<tr>
<td><strong>SNR</strong></td>
<td>Detector limited</td>
<td>Light source limited</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Spectral resolution</strong></td>
<td>2 cm⁻¹</td>
<td></td>
<td>4 cm⁻¹</td>
<td></td>
</tr>
<tr>
<td><strong>Available region</strong></td>
<td>950-4000cm⁻¹</td>
<td>Full IR range</td>
<td>910-1800cm⁻¹</td>
<td></td>
</tr>
<tr>
<td><strong>Speed (cells/hour)</strong></td>
<td>Few hundreds</td>
<td>30-50</td>
<td>50</td>
<td>Few hundreds</td>
</tr>
<tr>
<td><strong>Sample holder</strong></td>
<td>Compression cell</td>
<td>Liquid cell</td>
<td>Compression cell</td>
<td></td>
</tr>
<tr>
<td><strong>IR window</strong></td>
<td>0.5mm CaF₂</td>
<td>1mm CaF₂</td>
<td>0.5mm CaF₂</td>
<td></td>
</tr>
</tbody>
</table>
9.2 Comparison of Imaging and spectral quality

The quality of captured images and spectra from the three spectrometers will be discussed in this section. The comparison only focuses on the currently available results. The difference in the images and spectra recorded from three spectrometers will be summed up in Table.15.

Benchtop FTIR spectrometer

The image taken from the FTIR spectrometer with the low magnification objective has a large FOV which can capture up to a hundred cell. However, due to the low pixel resolution, the cell boundary was blurred. The spectra of cells were extracted with an automatic cell finder algorithm. Due to the unclear cell boundary and the instability in the water correction process when fitting the whole image with the same buffer reference, the variance between spectra extracted from the same image is relatively large. The experiments do not require complex calibration each time after changing the sample. Therefore, the repeatability of the spectra is guaranteed. Due to the saturation in the spectra and mathematic errors in the data processing, the final spectral region that was used to establish the classifier is 1200-1400cm⁻¹, 1700-1800cm⁻¹ and 2800-3000cm⁻¹. No obvious noise was noted in the spectra.

When measuring the cells with the high magnification objective, the FOV of the image is significantly reduced, such that it can only measure up to 8 cells in one image. The cell boundary was clearly resolved due to the high pixel resolution. Spectra of cells were manually extracted and fitted with the buffer spectrum taken from the region close to the cells. This leads to high consistency in the resultant spectra. In chapter.5, the spectral region was limited to the same range for consistency and cross-validation. However, the spectral region of 1000-1200cm⁻¹ in the spectra does not shift like in the spectra obtained with the low magnification objective. Therefore, the spectral region that can be used to establish the classifier is 1000-1400cm⁻¹, 1700-1800cm⁻¹ and 2800-3000cm⁻¹. Although the SNR in each pixel is reduced, the SNR in the resultant cell
spectra was still very good.

**SR-FTIR spectrometer**

Due to the FPA being unavailable at the time of experiments, the imaging of cells was not performed. Only one spectrum was recorded for each cell due to the large aperture size. The cells were manually located in the system and measured one by one. The spectra obtained from each sample loading are consistent. However, a significant variance was found between the spectra taken from different sample loadings, which results in low repeatability in the spectra. The spectra were also affected by the high cutoff of thicker CaF₂ windows which leads to the strong noise level in the low wavenumber region below 1300 cm⁻¹. The available spectral region used to build the classifier is 1300-1380 cm⁻¹, 1720-1800 cm⁻¹ and 2800 to 3000 cm⁻¹. Although the SNR in the spectra is expected to be large, the variance between spectra of the same cell line still affects the performance of the classifier.

**QCL microscope**

The image taken from the QCL microscope benefit from the large microbolometer FPA detector. The single image can capture more than a hundred cells and still clearly resolve the cell boundary. However, due to the coherent laser source, the correlated noise appeared in the recorded spectra which stop the implement of automatic cell extraction method. Therefore, the spectra of cells were manually selected and extracted. After noise reduction, the spectra exhibit excellent consistency. The repeatability of the spectra is limited by the correlated noise. Although the noise reduction method minimised the noise level in the spectra, the remaining noise still affects the spectra in the spectral region of 1200-1400 cm⁻¹. This issue was solved by applying a 9 point smoothing function. Therefore, the final spectral region used to build the classifier is the full spectra.
Table 15 Quality of image and spectra obtained from three spectrometers

<table>
<thead>
<tr>
<th></th>
<th>FTIR Low</th>
<th>FTIR High</th>
<th>SR-FTIR</th>
<th>QCL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell boundary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell boundary</td>
<td>blurred</td>
<td>clear</td>
<td>N/A</td>
<td>clear</td>
</tr>
<tr>
<td><strong>Spectra per cell</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectra per cell</td>
<td>10~20</td>
<td>150~200</td>
<td>1</td>
<td>120~150</td>
</tr>
<tr>
<td><strong>Extraction method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extraction method</td>
<td>Auto</td>
<td>Manual</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td>Low</td>
<td>High</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Repeatability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeatability</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td><strong>Appropriate spectral range (cm⁻¹)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate spectral range (cm⁻¹)</td>
<td>1200-1400</td>
<td>1000-1400</td>
<td>1300-1380</td>
<td>1000-1800</td>
</tr>
<tr>
<td></td>
<td>1700-1800</td>
<td>1700-1800</td>
<td>1700-1800</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2800-3000</td>
<td>2800-3000</td>
<td>2800-3000</td>
<td></td>
</tr>
</tbody>
</table>
9.3 Parallel study

The performance of classifier built with the spectra obtained from all three spectrometers are good. However, to achieve the expected performance, different spectral region and data process steps were employed. Therefore, the results from different spectrometer cannot be compared directly. In this section, the 100 water corrected spectra recorded from each spectrometer will undergo the same data processing and be limited to the same spectral regions of $1300-1380 \text{ cm}^{-1}$ and $1700-1800 \text{ cm}^{-1}$. The multivariate analysis and random forest classifier will be performed based on the 2nd derivate spectra.

*PCA*

The PCA results for the spectra obtained from the four sets of experiment are shown in Figure 9.1-9.4. The results clearly show the same pattern that the LNCAP is separated and the PC3 and PNT2 are overlapping in all four plots. The feature that separates the data point is consistently located in the lipid region. This proves that the water correction method is capable of removing the suitable amount of water and reveal the true chemical difference between those three cell lines.

*PC-LDA*

The PC-LDA results of the four groups of data are shown in Figure 9.5. It is evident that the data points of the three cell lines are well separated in the plots. Although the number of spectra and the spectral region in the spectra is limited, the remaining information was still able to separate those three different cell lines. The consistency of the spectra obtained from the three spectrometers can be evaluated by the degree of clustering of the data points from the same cell line.
Figure 9. 2 PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells from FTIR low mag setup. (right) 1st and 2nd loadings correspond to each PC.

Figure 9. 1 PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells from FTIR high mag setup. (right) 1st and 2nd loadings correspond to each PC.
Figure 9. 3 PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells obtained from SR-FTIR spectrometer. (right) 1st and 2nd loadings correspond to each PC.

Figure 9. 4 PCA plot for 1st PC against 2nd PC of raw spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells obtained from QCL microscope. (right) 1st and 2nd loadings correspond to each PC.
Random forest classifier

The spectra were also used to build a common random forest classifier. As the number of spectra was limited to 100 spectra per cell line, the data set only contains 300 spectra in total. To ensure there are enough spectra to cover the variance, the ratio between the training set and validation set was adjusted to 8:2. The accuracy of 100 classifiers built with the four group of data was shown in Table.16~19.

Figure 9. 5 PC-LDA results on 2nd derivate spectra of PC3 (red dot), LNCaP (blue dot) and PNT2 (green dot) cells obtained from FTIR low mag setup (top left), FTIR high mag setup (top right), SR-FTIR spectrometer (bottom left) and QCL microscope (bottom right).
### Table 16 Accuracy of classifier built with 2nd derivate spectra obtained from FTIR low mag setup

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>83.3%</td>
<td>93.1%</td>
<td>83.5%</td>
</tr>
<tr>
<td>AVG</td>
<td>96.6%</td>
<td>98.9%</td>
<td>94.2%</td>
</tr>
<tr>
<td>MAX</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 17 Accuracy of classifier built with 2nd derivate spectra obtained from FTIR high mag setup

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>93.5%</td>
<td>89.2%</td>
<td>81.2%</td>
</tr>
<tr>
<td>AVG</td>
<td>99.4%</td>
<td>98.6%</td>
<td>95.4%</td>
</tr>
<tr>
<td>MAX</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 18 Accuracy of classifier built with 2nd derivate spectra obtained from SR-FTIR spectrometer

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>80.0%</td>
<td>89.6%</td>
<td>88.2%</td>
</tr>
<tr>
<td>AVG</td>
<td>94.9%</td>
<td>97.4%</td>
<td>98.1%</td>
</tr>
<tr>
<td>MAX</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 19 Accuracy of classifier built with 2nd derivate spectra obtained from QCL microscope

<table>
<thead>
<tr>
<th></th>
<th>PC3</th>
<th>LNCAP</th>
<th>PNT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>68.3%</td>
<td>83.2%</td>
<td>72.4%</td>
</tr>
<tr>
<td>AVG</td>
<td>81.2%</td>
<td>95.7%</td>
<td>87.3%</td>
</tr>
<tr>
<td>MAX</td>
<td>96.3%</td>
<td>99.2%</td>
<td>98.9%</td>
</tr>
</tbody>
</table>
From the results above, the performance of the classifier built with the spectra obtained from the FTIR and SR-FTIR spectrometers are large and close to each other. This suggests that the FTIR spectrometer with Globar source is capable of performing live cell imaging to the same performance level with a thin water layer. The classifier built with the spectra recorded from QCL microscope has a lower performance. This may be a result of the lower number of the data points in the spectra. The spectra from the FTIR spectrometer benefit from the zero filling when performing the Fourier transform which doubles the spectral resolution mathematically. Due to the limitation in the demonstration software, the QCL microscope was not able to record spectra with 2 cm$^{-1}$ spectral resolution. However, this issue should be overcome in the near future.

To verify the feature that the classifier used to separate those three cell lines, the averaged importance (100 runs) of wavenumber for each classifier is shown in Figure 9.6. The most important features in all four classifiers are located in the lipid region. The value of importance at those wavenumber varies across the four classifiers, but the peak locations appear to be the same. Extra features in the spectral region of 1300-1380 cm$^{-1}$ were selected by the classifier for both FTIR spectrometers. This may result from two facts. First, the spectral region in the data set is too small. The feature in the lipid bands is not enough to cause separation, and therefore the classifier is looking for smaller chemical differences which increase the importance in this region. Second, the FTIR spectrometers were capable of capturing smaller chemical differences as the spectra were not affected by the correlated noise. Further experiments are required to explore this problem.
Figure 9. 6 averaged importance of each wavenumber in the classifier built from the spectra recorded by FTIR spectrometer with low mag setup (row 1), FTIR spectrometer with high mag setup (row 2), SR-FTIR spectrometer (row 3) and QCL microscope (row 4).
10. Conclusion

In this thesis, the imaging of cells in an aqueous environment using three types of IR spectrometer has been evaluated with a proposed novel water correction method. The fundamentals of the instrumentation of all three spectrometers were reviewed in detail at the beginning. The advantages and disadvantages of these spectrometers and the effects on the spectral quality were discussed. After studying how the water will affect the spectra of cells, the previous live cell imaging studies using different water correction algorithms and spectrometers were reviewed. The knowledge gap between these studies leads to the aims of this project. Before presenting the experiment, the experimental method was discussed. Due to the lack of a proper device, a microcompression cell used to measure solid samples, was modified to perform the experiments.

A novel water correction method based on fitting the protein region in the spectra of cell in water with a water reference spectrum and a cell reference spectrum has been proposed. The fitting parameter, including choice of reference, effects of fitting range and the reference dependency, was reported. In most of sample holder designed for live cell imaging, which normally create a thin water layer, the fringing effects can be easily triggered. Therefore, a fringe removal algorithm, originally from the EMSC method [1] was modified and attempted. Due to the uncertainty in the sinusoidal fluctuation, the algorithm failed. However, this can still potentially be improved in the future. A modified automatic cell extraction method [2] was employed to extract the spectra of cells from the image. Due to the extraction method being robust, the protein bands which used to perform the water correction is not suggested to be included in the further analysis. However, the spectra after correction contain enough information to classify cells from three different prostate cancer cell lines. The results evidently confirmed the feasibility of the new water correction algorithm. The image of the cell was taken in two resolutions. The high-resolution image was used to examine the
effects of the processing sequence in the water correction algorithm. The spectra obtained from high-resolution image exhibit greater consistency. The performance of the water correction on those spectra is excellent.

The water correction was then modified to remove the water contribution from the spectra of cell in water recorded by the SR-FTIR spectrometer at the Diamond synchrotron site. Due to the experimental limitations, the imaging of cells could not be performed, and the single beam spectra were obtained. The experiments with the SR-FTIR spectrometer utilised a properly sealed liquid cell with a well-designed spacer. This liquid cell requires large and thick CaF\textsubscript{2} windows which lead to worse cutoff [3] in the low wavenumber region of the spectra. The spectra have also been affected by the poor repeatability of experiments. However, the modified water correction method can still reveal enough chemical information in the spectra of cells which can be used to build a classifier to identify three different cell lines.

The QCL microscope, in theory, should benefit from the additional brightness of the QCLs which makes it ideal for the study of live cells in water. However, the correlated noise with fixed absorbance value was observed in the spectra recorded from QCL microscope. As the noise is correlated, normal noise removal methods, such as PCA noise reduction and Savitzky-Golay smoothing, were tried, and failed as expected. After analysing the property of noise, a new noise removal algorithm based on fitting the spectra from adjacent pixel was proposed. This method successfully reduced the noise level in the water corrected cell spectra to a similar level to the smoothed spectra and avoided the distortions from the smoothing method. The remaining noise in the spectra become problematic in the derivative spectra. To further eliminate the noise, a 9 point smoothing function was used. The different noise profile leads to the modification of the water correction method as well. The altered water correction method successfully revealed enough features in the cell spectra and the performance of the classifier built with the resulted spectra are excellent.

By comparing the results from different spectrometers, it can be concluded that the
novel water correction method is capable of removing a suitable amount of water contribution from the spectra of cell in water. The corrected spectra contain clear chemical information which can be used to distinguish different samples.

A parallel study was also carried out to compare the performance of the three spectrometers. The results suggest that all the three spectrometer is capable of live cell imaging, even the conventional FTIR spectrometer with a Globar source. The results of the QCL spectrometer were a bit worse, due to the instrumentation limit. However, as the QCL microscope is an emerging technology and under fast development, the existing issues may be resolved.

**Future Perspective**

Due to the limitation in the instrument, this project is focused on the simplified cell solution mode with fixed cells. The imaging of live cells in PBS solution should be easily achieved with a proper designed liquid cell device. The final goal of measuring live cells in culture medium or other bio-fluid will then come next.

The water correction method proposed in this project is a robust model. The correction is not aiming at obtaining an absolute correct cell spectrum, which is, in theory, not achievable at the moment [4]. However, the water correction successfully exhibits its diagnostic capability, which distinguishes the three different prostate cancer cell lines. The water band in Raman spectrum is weak but still presented [5]. Therefore, this method also can be applied to Raman spectrum of cell in water. When imaging live cells in complex bio-fluids, this method may need further adjustment as the bio-fluids contains similar features, such as protein bands, in the spectrum.

The synchrotron radiation is a powerful light source. However, the instability of the radiation requires a well-planned instrumental setup and experimental protocol. Currently, the radiation profile is manually controlled by tuning a set of mirrors. An automatic device which calibrates the radiation profile is expected in the future.

The QCL microscope as a newly developed technology has proved its capacity in
measuring cells in solution. The noise due to the coherent laser source should be solved
in the instrumental design in the future. The range of application needs to be broadened.
The Raman spectroscopy has been used in tumour margin detection [6]. Due to the
small illumination area of Raman probe, the surgery operation may take a long time to
identify the cancerous area. A similar application in IR spectroscopy is ATR imaging
[7]. However, due to the low intensity of the Globar source, the image suffered from
poor SNR. Therefore, coupling the QCL with the ATR should establish an imaging
probe system with a large field of view and excellent SNR.
Reference


